Etude d'architectures et d'empilements innovants de mémoires Split-Gate (grille séparée) à couche de piégeage discret

Lia Masoero

UNIVERSITÉ DE GRENOBLE

THÈSE

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité: Nano Électroniques et Nano Technologies

Arrêté ministériel : 7 août 2006

THÈSE DIRIGÉE PAR **GÉRARD/GHIBAUDO** ET

CO-ENCADRÉE PAR **GABRIEL/MOLAS**

PRÉSENTÉE PAR

LIA/MASOERO

PRÉPARÉE AU SEIN DU

CEA/LETI

DANS L'**ÉCOLE DOCTORALE:ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE & TRAITEMENT DU SIGNAL**

Etude d'architectures et d'empilements innovants de mémoires Split-Gate (grille séparée) à couche de piégeage discret

THÈSE SOUTENUE PUBLIQUEMENT LE **30 NOVEMBRE 2012**,

DEVANT LE JURY COMPOSÉ DE:

MONSIEUR, CHRISTOPHE MULLER

Prof. Université Aix-Marseille , Président

MONSIEUR, GIUSEPPE IANNACCONE

Prof. Université de Pise (Italie), Rapporteur

MONSIEUR, JORDI SUNE

Prof. Université autonome de Barcelone (Espagne), Rapporteur

MONSIEUR, JEAN-MICHEL MIRABEL

Ing. STMicroelectronics Rousset, Membre

MONSIEUR, GÉRARD GHIBAUDO

DR CNRS Alpes - IMEP/INPG, Membre

MONSIEUR, GABRIEL MOLAS

Ing. CEA-Grenoble, Membre

MONSIEUR, PHILIPPE BLAISE

Ing. CEA-Grenoble, Membre invité

Contents

General introduction

Context

In what is called "Digital Era" many applications in the day-to-day life utilise micro-controllers devices which contains non volatile memories [1].

It has been computed that in the last year the MCUs non volatile market size have just reached almost 16 Billion dollars [2]. Even more, this incredible number is going to grow faster and faster and nothing allows thinking that this growth will slow down.

The increasing in demand for non volatile embedded memories has thus driven research to provide always more capacitive, low power read/program/erase memories increasing in the same time velocity, reliability while obviously reducing the cost.

As this thesis is written, the prevalent non volatile solution for embedded applications, is the standard NOR flash cell. Some serious limitations to keep the required low consumption and endurance are driving research to new materials and architectures. In particular, the integration of a charge trapping layer in a 1.5T (split-gate) structure seems to be one of the most promising technology solution for replacing the standard flash memory [3].

Thesis presentation

The present thesis work focuses on the study of innovative stacks and architectures of Split-Gate charge trap memories.

In the first chapter, we will present the economic context, the evolution and the working of EEPROM-flash memories. Then, a detailed description of the technology, the functioning and their scaling limits will be provided. Finally we will expose the possible solutions to overcome these problems and the thesis framework.

The second chapter will present the multi-litho split-gate charge trap memories with electrical gate length down to 20nm. We will show the electrical results of Silicon nanocristals (Si-ncs), or silicon nitride (Si_3N_4) and hybrid Si-nc/SiN based split-gate memories, with SiO_2

or Al_2O_3 control dielectrics. Then we will present the study on the scalability of split-gate charge trap memories, investigating the impact of gate length reduction on the memory window, retention and consumption.

We thus present a possible solution to overcome the multi-litho approach issues: the spacer technology. First the integration scheme and the influence of process parameters on split-gate spacer memory will be presented. Then, electrical results on porgramming and erase operations on silicon nitride (Si_3N_4) based-memory will be shown.

In the third chapter we will study the role of defects in alumina we will first introduce the TANOS memory which employs Al_2O_3 as control dielectric with silicon nitride (SiN) charge trapping layer (CTL). Then, we will present the physical-chemical material analysis on alumina single layers needed for the atomistic simulations used in order to find the potential defects that could induce electronic levels inside the band gap of alumina. Finally the trap features estimated by quantum simulations are introduced in a TANOS device simulator. A very good agreement is obtained between model and device experimental data.

In the third chapter we will use atomistic simulation, consolidated by a detailed alumina (Al_2O_3) physico-chemical material analysis, to investigate the origin of traps in alumina. We will show that the leakage currents through Al_2O_3 layers, with different post-deposition anneals, are strictly correlated to the hydrogen content. Then the hydrogen-based trap features estimated by quantum simulations will be introduced in a TANOS device simulator allowing for a clear understanding of the role of alumina hydrogen content on the retention characteristics of charge-trap memories.

The manuscript finally ends on a general conclusion which summarizes the different results obtained in this work, before proposing some perspectives.

Chapter 1

Overview of the non volatile memories: needs and scaling limits

First of all in this chapter we will present the economic context, the evolution and the working of EEPROM-flash memories. Then, a detailed description of the technology, the functioning and their scaling limits will be provided. Finally we will expose the possible solutions to overcome these problems and the thesis framework.

1.1 Memories on the semiconductor industry

1.1.1 Economic context

The semiconductor industry born in 1947 when the first transistor at Bell Labs (US) was invented. Only 60 years later the worldwide Integrated Circuits (IC) market revenue totalled nearly 250 billion dollars (∼0.5% of world Gross Domestic Product).

Even though the single device cost has decreased by 50 per cent each year following the Moore's law, the IC industry revenues have grown at an average annual rate of 17% between 1970 and 2008. Over these years the demand in IC products has been driven by the new technology equipments introduced in the market (ICs, as PCs, cell phones, smart phone, tablet..) and it has thus been affected by constant booms and busts in demand for products (Fig 1.1).

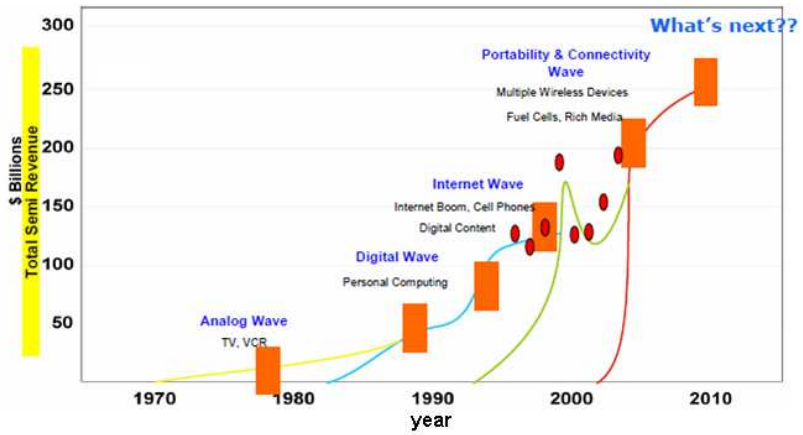


Figure 1.1: The IC industry cycle [4]

In this scenario, memories play an important role: they count for about a quarter of the total sold on IC market. The "ideal" memory should be a memory that retains the stored information even when it is not powered (non volatile), with a high integration density, that can be infinitely written/re-written (infinite endurance), with high speed program/erase/read operations, a low energy consumption and low price. Because the "ideal" device does not exist, different types of memories have been studied in order to develop one or more of these properties according to their final application (see fig. 1.2).
In the next section, the most important semiconductor memories will be summarized.

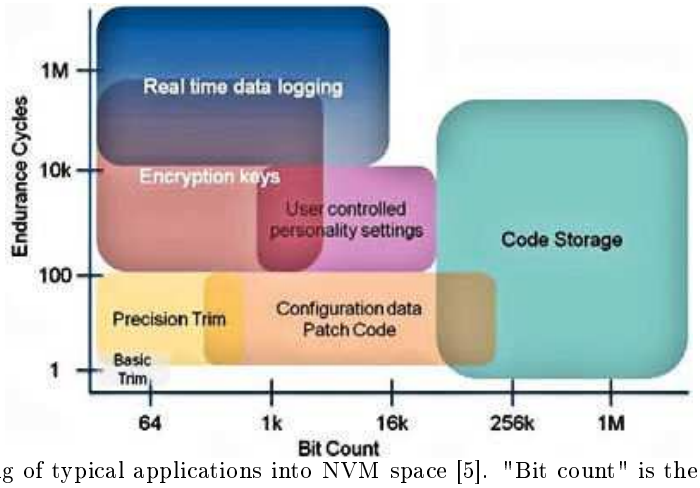


Figure 1.2: Mapping of typical applications into NVM space [5]. "Bit count" is the amount of data that can be stored in a given block.

1.1.2 Memory classification

There are various possibilities to classify semiconductor memories, one of these is to consider their electrical characteristics (Fig. 1.3).

Volatile Memories: are fast memories that are used for temporary storage data since they loose the information when the the power is turned off. We can divide them into two types:

Static Random access memory (SRAM): the information is maintained as long as they are powered. They are made up of flip-flop circuitry (six transistors in a particular configuration). Because of this large number of components SRAM is large in size and cannot compete with the density typical of other kind of memories.

Dinamic Random access memory (DRAM): these memories loose the informations in a short time. They are made up of a transistor and a capacity where the charge is stored. They are widely used in processors for the temporary storage of information. As the capacitor looses the charge, these memories need to be recharged or refreshed to maintain the achieve state.

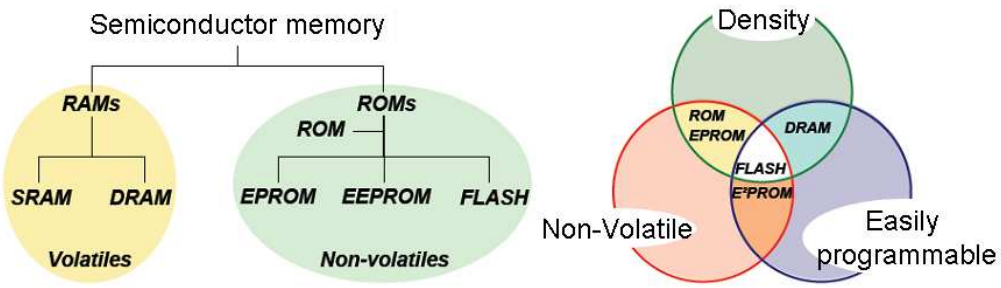


Figure 1.3: Left: Overview of the non volatile semiconductor memories; Right: Semiconductor memory classification by different performance criteria.

NON Volatile Memories they keep the information also when the power is down. They have been conceived in order to store the information without any power consumptions for long time. This thesis concerns charge storage non volatile memories that are a subgroup of the semiconductor memories. However it is important to remember that there are other media were the information used to be stocked. Historically the most used non volatile memory is the paper even if the density of information is very low. Others common storage media is the magnetic (we can find it in the hard disk, tape, credit card..) that has the drawback of a long access time and the sensitivity to magnetic fields. Another example of a common non volatile memory is the CD technology developed in the late 1970s, that use an optical media that can be read fast, but it needed a pre-recorded content.
Here we will detail only the memory based on semiconductor technology:

Read only memory (ROM) is the first non-volatile semiconductor memory. It consists in a simple metal/oxide/semiconductor (MOS) transistor thus its cell size is potentially the smallest of any type of memory device. The memory is programmed by channel implant during the fabrication process and can never be modified. It is mainly used to distribute firmware that are programs containing microcode that do not need frequent update (A typical example of firmware-controlled device is a television remote control).

Programmable read only memory (PROM) is similar to the ROM memory mentioned above, but the program step could be accomplished by the user. It was invented in 1956, it can be a cheaper alternative to the ROM memory because it does not need a new mask for each new programming.

Erasable Programmable read only memory (EPROM) This memory could be erased and programmed by the user, but the erase has to be done by extracting the circuit and putting it under ultraviolet (UV) radiations. The particularity of this device is the presence of a "floating gate" beetwen the control (top) and tunnel (bottom) oxides. This type of device has been described for the first time by [6] in 1971.

Electrically Erasable Programmable read only memory (EEPROM) In this memory both the write and erase steps can be electrically accomplished. The W/E operations could be made without removing the chip from the motherboard. The EEPROM cell features a select transistor in series to each floating gate cell. The select transistor increases the size of the memories and the complexity of array organization but it gives the possibility of erasing cell at a byte basis.

Flash memory is a synthesis between the compactness of EPROM and the enhanced functionality of a EEPROM. It looks like EEPROM memory but without the select transistor. This because the memory cell can have the role of the select transistor. Its name come from its fast erasing mechanism.
Because of these properties and the new applications (Fig. 1.4) the flash memory market is growing with a higher average annual rate than DRAM and SRAM, becoming nowadays the most produced memory (Fig. 1.5). Depending on their applications flash memories can be divided in two main families that we introduce here and we will detail in section §1.2.4.

NOR flash memory provides random memory access, fast reads useful for pulling data out of memory, but it writes data relatively slowly.

NAND makers. NAND, on the other hand, reads data slowly but has fast write speeds

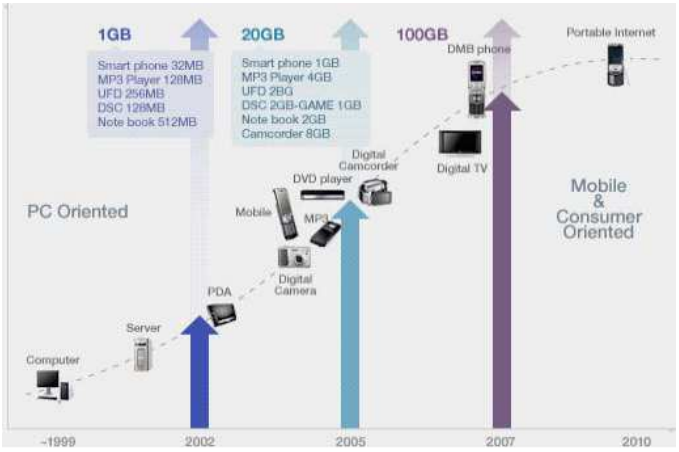


Figure 1.4: Time line of the NAND flash memory market and storage density needed for each product [7].

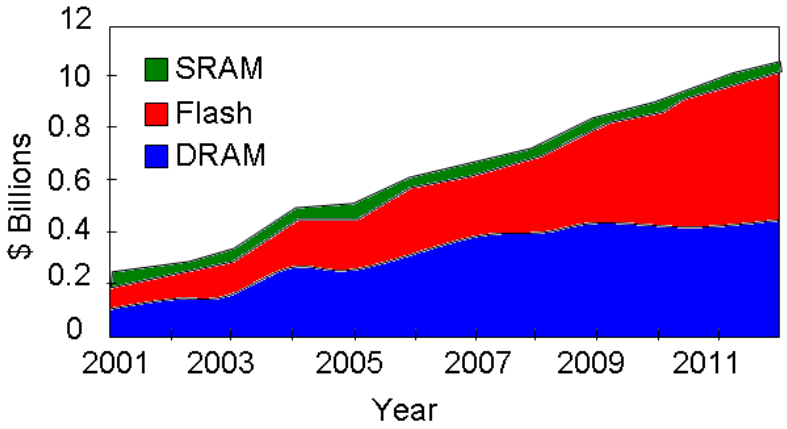


Figure 1.5: DRAM; SRAM ; Flash memory market evolution [7].

Split-Gate Memory is a mix between a flash and a EEPROM memory. it consists on a Flash memory followed or preceded by a select transistor. This allows to achieve fast program/erase operations, a low consumption and a better disturb immunity. On the other hand, the select transistor makes it too large for stand alone (high density) applications.

1.2 Floating gate cell overview

The floating gate cell is the basis of the charge trap split-gate memory studied in this thesis. In this part we will detail flash memory operations.
The operation principle is as follow (Fig. 1.6-a): when the cell is erased there are no charges in the floating gate and the threshold voltage (V_T) is low (V_{Terase}). On the contrary when the memory is written the injected charge is stored on the floating gate layer, and the threshold voltage value is high (V_{Twrite}). To know the state of the memory (e.g. the amount of trapped charge) just need to bias the gate with moderate read voltage (V_G) that is between (V_{Terase})

and (V_{Twrite}) and measure if the current flows through the channel (state 1) or not (state 0).

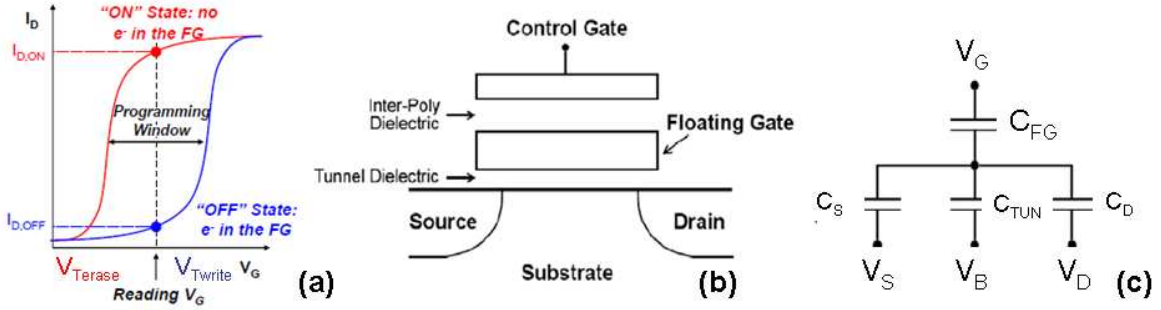


Figure 1.6: : (a) Written and erased $I_D(V_G)$ characteristics showing the programming window. (b) Schematic cross section of a floating gate memory. (c) Electrical model of a floating gate cell.

1.2.1 Structure and operation

The flash memory cell is based on a NMOS transistor were the gate stack is modified by adding a polysilicon gate between the tunnel oxide and the interpoly dielectric (Fig. 1.6b). Looking at the equivalent electrostatic scheme (Fig. 1.6c) it is easy to deduce the potential of the floating gate (V_{FG}):

$$V_{FG} = \frac{Q_{FG}}{C_T} + GRC \cdot V_G + \alpha_S V_S + \alpha_D V_D + \alpha_B V_B \qquad \text{with} \qquad C_T = C_{FG} + C_D + C_S + C_{TUN} \quad (1.1)$$

where $GRC = C_{FG}/C_T$, $\alpha_S = C_S/C_T$, $\alpha_D = C_D/C_T$, and $\alpha_B = C_{TUN}/C_T$ are the coupling factors; Q_{FG} is the trapped charge in the floating gate; C_{TUN}, C_S, and C_D are the capacitance between respectively the floating gate and the tunnel, the floating gate and the source; the floating gate and the drain.
If the drain potential is low, the source and the bulk are grounded and all the potentials referee to them; the expression (1.1) becomes:

$$V_{FG} = \frac{Q_{FG}}{C_T} + GRC \cdot V_G \quad (1.2)$$

this equation demonstrates that if we change the amount of trapped charge in the polysilicon floating gate (Q_{FG}) by injecting or removing electrons from it, the threshold voltage shift ΔV_{TH} which corresponds to the control gate voltage shift required to keep the potential of the floating gate V_{FG} constant, becomes:

$$\Delta V_{TH} = -\frac{Q_{FG}}{GRC \cdot C_T} = -\frac{Q_{FG}}{C_{FG}} \quad (1.3)$$

Equations (1.3) and (1.2) reveals the importance of the Gate Coupling Factor (GCR): (1.2) shows that high GCR induces a floating gate potential close to the applied control gate (V_G) bias and consequently the gate coupling ratio needs to be high to provide a good programming and erasing efficiency. On the other hand (eq. 1.3) indicates that high GCR reduces the impact of the storage charge to the programming window (ΔV_{TH}).
The international roadmap for semiconductor [8] indicates that the best trade-off is achieved with a GRC between $0 \div 6$ et $0 \div 7$.

1.2.2 Write mechanisms

The two main methods to program flash cell are the Fowler-Nordheim (FN) and the Channel Hot Electron (CHE) (Fig. 1.7).

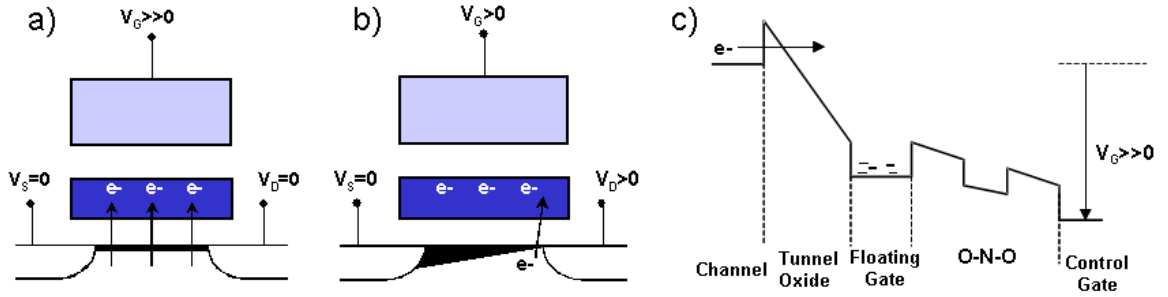


Figure 1.7: (a) Fowler-Nordhein and (b) Hot channel electron writing mechanism representation. (c) Band diagram of a floating gate during FN programming operation.

Fowler-Nordheim In this tunnel effect the electron flows from the conduction band of the silicon into the floating gate through the triangular energy barrier of the tunnel oxide (Fig. 1.7-c). This is possible keeping source drain and bulk grounded and applying a high positive voltage on the gate (about of 20V).
The programming done by FN is uniform. It is slower than CHE but is less degrading, moreover as no bias is applied at the source and the drain, the current consumption is negligible.

Channel Hot Electron (CHE) is done keeping bulk and source grounded and applying a positive high voltage on the gate (order of 10V) and on the drain (order of 5V). The electrons are first strongly accelerated in the pinchoff region by the high parallel electric field induced by the drain bias, then the electrons that have reached a sufficient high kinetic energy are injected into the polysilicon layer thanks to the vertical field induced by the positive voltage applied on the gate electrode.
Programming by hot channel electron is faster than FN, nevertheless due to the strong energy of the electrons the degradation of the oxide is higher. Another drawback of CHE is the poor efficiency (only few electrons are injected over the total amount of electrons that flow from source to drain), and consequently the high power consumption.

1.2.3 Erase mechanisms

The ways to erase the cell are mainly four (Fig. 1.8)

Fowler-Nordheim as for programming, the source, drain and bulk are generally kept grounded while a strong negative voltage (order of -15V) is applied to the gate. In this case, electrons are forced to flow from the floating gate to the semiconductor. This method is slow but the erasing is uniform.

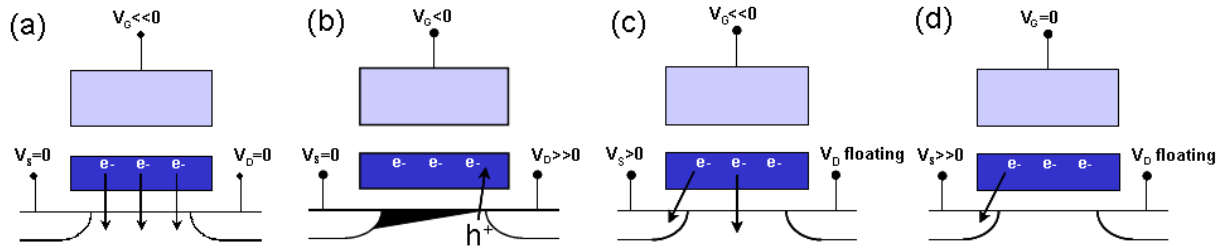


Figure 1.8: (a) Written and erased I_D(VG) characteristics showing the programming window. (b)Schematic cross section of a floating gate memory. (c) Electrical model of a floating gate cell

Source erasing force electrons to flow from the floating gate into the source junction by FN tunnelling. This erasing is done by applying a positive voltage of about 15V on the source and keeping bulk and gate grounded. In order to prevent current through the channel, the drain is kept floating. The drawbacks of this method are mainly three: the erasing is localised near the source, it needs a strong junction/gate overlap, it requires a really high voltage on the source.

Mix gate-source is a mix between the Source and the FN erasing. Electrons are erased both through the source and the channel. The interest is to share the high voltage needed in the source erasing between the gate and the source electrodes. As a result a negative bias of about -10V is applied to the gate and a positive bias of about 5V on the source. Again, the drain is kept floating in order to prevent source to drain current.

Hot hole injection (HHI) consists in accelerating the holes produced by forward biasing substrate-drain pn junction and inject these in the floating gate. This is done by keeping grounded the bulk and source and biasing positively the drain (order of 5V) and negatively the gate (abour -10V). HHI erasing method is fast, localised near the drain and, due to the presence of high energetic particles could induce, easier than the methods listed above, the SILC (Stress Induced Leakage Current) phenomenon.

1.2.4 Architecture

Flash memories are organised in arrays of rows (word lines or WL) and columns (bit lines or BL). The way they are connected determine the array architecture (Fig. 1.9).

NOR In nor architecture the cells are connected in parallel. The gates are connected together through the wordline, while the drain is shared along the bitline. The effect that the drain of each cell can be selectively selected, allows a random access of any cell in the array and the possibility to program by hot channel electrons. Programming is generally done by HCE and erasing by FN. NOR architectures provide fast reading and relatively slow programming mechanisms. The presence of a drain contact for each cell limits the scaling to $6F^2$ where F is the smallest lithographic feature. Fast Read, good reliability, relatively slow write/erase mechanism makes NOR architecture the most suitable technology for the embedded applications requiring the storage of codes and parameters, and more generally for execution-in-place.

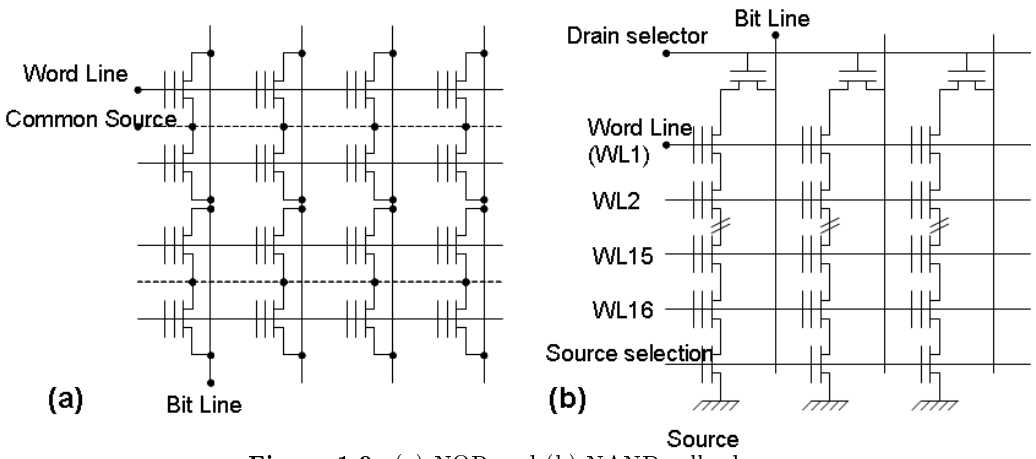


Figure 1.9: (a) NOR and (b) NAND cell scheme

NAND architecture is use for stand alone applications. The cells are connected in series. The gates are connected by a wordline and, differently from NOR technology, the drain and the source are not contacted. The absence of a drain contact induces that the cell cannot be selectively addressed and the programming can be done only by FN. On the other hand, the absence of shared contact allows reaching an optimal cell size of $4F^2$ thus a density 30% higher than in NOR cells.
In NAND architecture programming is relatively fast but the reading process is quite slow as the read of one cell is done forcing the cell in the same bit line to the ON state. The high density and the slow read but fast write speeds make NAND architecture suitable for USB key, storing digital photos, MP3 audio, GPS, and many other multimedia applications.

1.3 Floating gate cell evolution

The new applications have driven the semiconductor market and the research development. From 1999 when the flash cell was invented the progress on memory architectures and materials has been huge. The ideal memory should have:

- high density solution
- low power consumption
- non-volatility
- fast read/write/erase
- random read/write access
- endurance against write/erase cycles
- scalability with low cost
- compatibility with logic circuits and integration

that is the final objective of the semiconductor research. As the ideal device does not exist different types of memories have been invented in order to push some specific properties. For these reasons as shown in fig. 1.10, the memory technology development did not follow a single technology solution but the amount of branches have multiplied along the years [9]. In this section we will first introduce the device scaling and the related challenges, then we will present the developments of flash cell. It is worth to note that the solutions found for the flash memory cell can be use in embedded memory. In fact, even if in embedded memories there are less constraints on the cell dimensions, research is provided always smaller non-volatile memory for embedded applications that have to face-off with the same flash scaling limits.

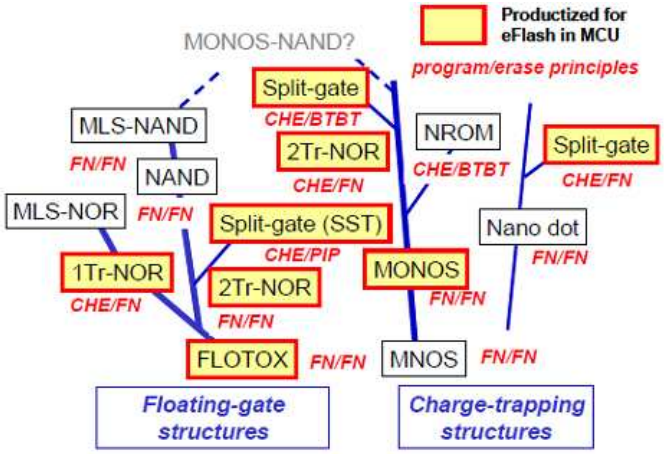


Figure 1.10: Evolution of flash technology [9]

1.3.1 Device scaling and challenges

Year of production	2012	2013	2014	2015	2016	2017	2018
Technology node F							
DRAM $^1/_2[nm]$	36	32	28	25	22.5	20.0	17.9
Flash NAND $^1/_2$ $[nm]$	28	25	22	20	18	16	14
Cell size - area factor in multiples of F^2SLC/MLC	4.0/1.3			4.0/1.0			
Gate coupling ratio [GCR]	0,6-0,7						
Non-volatile data retention [years]	10-20						
Endurance [erase/write cycles]	10^5			10^4			
Maximum number of bits per cell [MLC]	3	3	3	4			
Tunnel oxide thickness $[nm]$	6-7						
Interpoly dielectric material	*ONO*	*ONO/High-κ*		*High-κ*			
Interpoly dielectric thickness $[nm]$	9-11	4-6	3-5				

Table 1.1: Summary of the technological requirements for Flash NAND memories as stated in ITRS 2009 roadmap[8]. White cell color: manufacturable solutions exist and are being optimized; Yellow cell color: manufacturable solutions are known; Red cell color: unknown manufacturable solutions

In the last 30 years the Flash cell size shrunk from 1.5um to 25nm doubling the memory capacity every year. In table 1.1 we report the international technology roadmap for semiconductor 2009 that forecasts the future trends of the semiconductor technology. We can see that

if the trend is maintained, and the cell scaled down in the next years, some technological solutions are still not known. Moreover the scaling beyond the the 25nm will be very difficult if no revolutionary technology is adopted. The main issues that are limiting the device miniaturizing are:

SILC During each erase/write cycle the stresses degrade the tunnel oxide and the cell slowly loses its capacity of storing electric charges (see Fig. 1.11). This phenomena, that is called Stress Induced Leakage Current (SILC), increases as the tunnel oxide is thinned. This is due to the defects induced in the oxide by the electrons that pass through it during program/erase operations [10, 11, 12, 13, 14, 15, 16], and has as consequence that the retention depends on the number of cycling and the oxide scaling is limited to 6-7 nm.

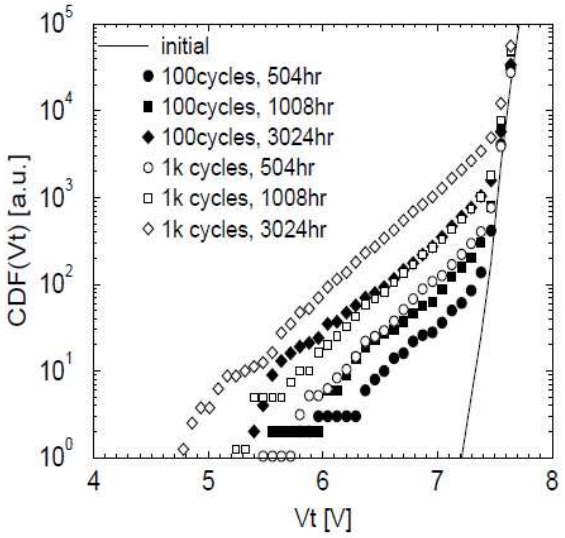


Figure 1.11: Experimental cumulative distribution functions of bits vs. threshold voltage, measured at different times after different P/E cycling conditions [12]

Short Channel Effects (SCE) SCE appear when the gate length dimensions are so shrunk that the gate control of the channel is lowered due to the influence of the source and drain potentials (Fig 1.12). This parasitic effect induces the Drain Induced Barrier Lowering (DIBL) phenomenon [17] which results in the threshold voltage decrease and the degradation of the subthreshold slope. Because of DIBL, the "OFF" current I_{OFF} increases and the power consumption reaches values incompatible with the advanced technology nodes requirements [8]. Moreover, elevated I_{OFF} currents result in some disturb of the memory cell, especially in the erased state.

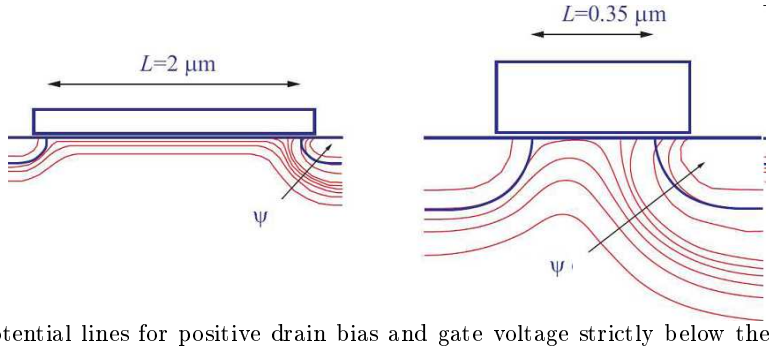


Figure 1.12: Potential lines for positive drain bias and gate voltage strictly below the threshold voltage [18].

Disturb The writing mechanism of a page, especially in NAND technology, requires high voltages to be applied to the cells. When the dimensions are scaled, the coupling between two cells becomes smaller and these high voltages might affect neighbouring cells and lead to parasitic charge leakage. This phenomenon can occur between bit line, word line or diagonally and is limited by modifying the architecture or the technology.

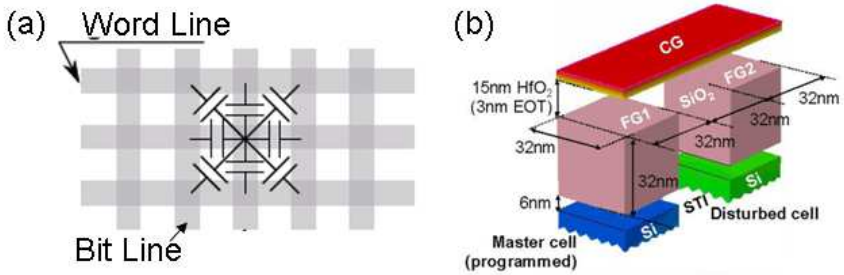


Figure 1.13: (a) Planar and (b) 3D view of the interaction between two neighbour cells [19]

Few electrons This phenomenon will be the ultimate intrinsic limit of NAND memories. As the dimension scaled down the possibility to store a charge decreases. If the trend shown in fig 1.14-a continues the number of electrons representing one bit will be reduced to some units. In this scenario the loss of a single electron would degrade dramatically the retention due to the stochastic nature of the discharging.

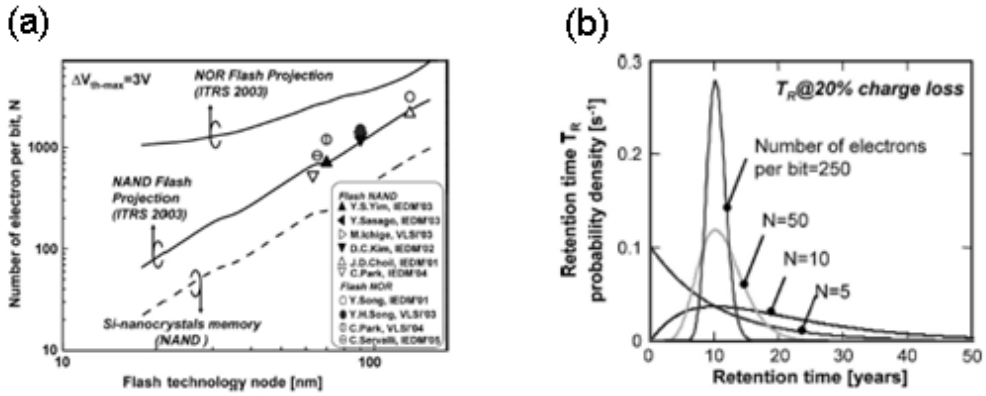
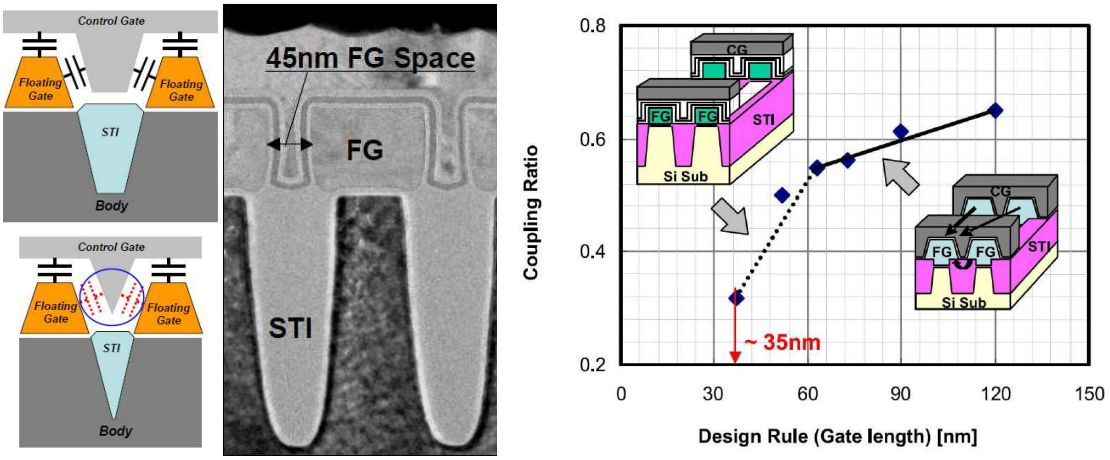


Figure 1.14: (a) Number of electrons representing 1 bit as a function of the Flash technology node. (b) Probability density of the retention time TR for memories with reduced number of electrons per bit N(5). The mean TR is fixed at ten years [20]

Coupling ratio As the cell dimensions are scaled down, the tunnel and interpoly dielectrics should be scaled accordingly. To maintain the GCR requirement, most NAND Flash structures have the word line (control gate) wrapped around the side-walls of the floating gate to increase the capacitance (i.e. figure 1.15-b).
However because the retention of the cell should remain higher than 10 years, the dielectrics have to scale at a slower pace. In 2009, the most advanced NAND technology (34nm half-pitch) uses an interpoly dielectric layer of 12nm, which makes wrap-around structures difficult to achieve when the spacing between cells becomes 20nm or less (i.e. figure 1.15-a). The loss of the extra coupling due to the sidewall extensions induces a strong decrease of the GCR [21] (i.e. figure 1.15-c). Therefore, maintaining a gate coupling ratio (GCR) above 0.6 appears to be a strong issue for the next NAND Flash generation [8].



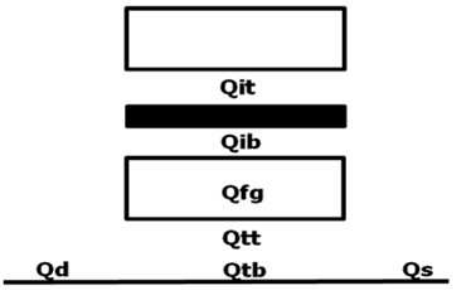
(a) Schematic diagram of the GCR loss.

(b) TEM picture of sub-50nm NAND Flash cell (Samsung [21]).

(c) Evolution of the coupling ratio with the gate length [21].

Figure 1.15: Impact of the inter-cells space reduction on the gate coupling ratio [21].

Parasitic Charge Trapping In scaled memories the reduction in the number of stored electrons leads to a higher influence of the parasitic electrons to the threshold voltage shift [22]. Fig. 1.16 shows various locations within a NAND cell where parasitic charge can be trapped and the results of a TCAD simulation showing that with the scaling of the memory dimensions, the number of electrons located outside the floating gate starts to dominate the cell V_{TH} shift.



	50nm	35nm	25nm
Qtb	4	2	1
Qtt	9	7	4
Qib	22	17	9
Qit	149	103	100
Qs	33	9	5
Qd	61	16	10
Qfg	18	12	10

Figure 1.16: Left: Locations of parasitic charge in a NAND cell. Right: number of electrons required in each location to shift the cell V_{TH} by 100mV [22].

Random Dopant Fluctuation The threshold voltage shift due to random variations in the number and position of dopant atoms is an increasingly problem as device dimensions shrink. In Fig. 1.17 are shown the mean and 3σ for the number of dopant atoms as a function of the feature size. As the device size scales down, the total number of channel dopants decreases, resulting in a larger variation of dopant numbers, and significantly impacting threshold voltage. It has been computed [23] that at 25nm node, the V_{TH} can be expected to vary by of about 30% purely due to the random dopant fluctuation.

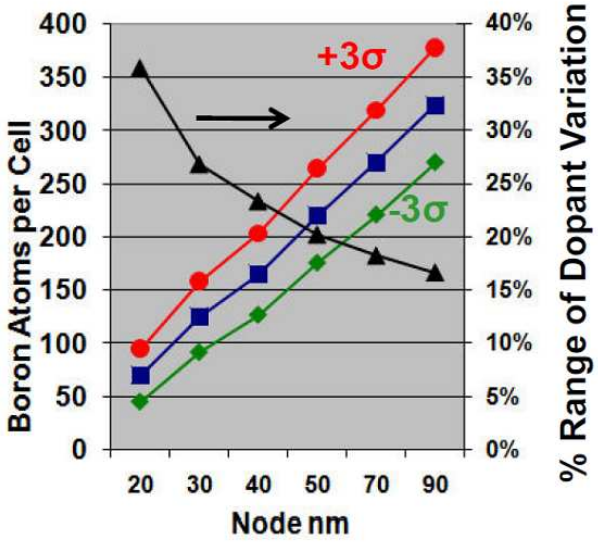


Figure 1.17: Number of Boron atoms per cell. (mean: square, -3σ: diamond, +3σ: circle vs. feature size. The triangle shows the $\pm 3\sigma$ percentage divided by the mean [22].

1.3.2 Proposed solutions

In this section we will describe some of the envisaged modification to the classical flash memory cell (see §1.2) in order to overcome the scaling limits presented in the previous section.

Tunnel dielectric In a flash memory the tunnel dielectric has the double role of tunnelling media during programming operations, and electrostatic barrier in order to preserve the stocked charge. Moreover we must avoid the creation of defects during the programming operations that can induce the SILC and degrade retention and cycling. This technological challenge can been solved by band engineering. As shown in Fig. 1.18 crested barriers can provide both sufficient programming and retention. Several crested barriers have been tested: the most common consists in an ONO layer [24], but other combinations have also been experimented ($SiO_2/Al_2O_3/SiO_2$ [25], SiO_2/AlN [26], etc.).

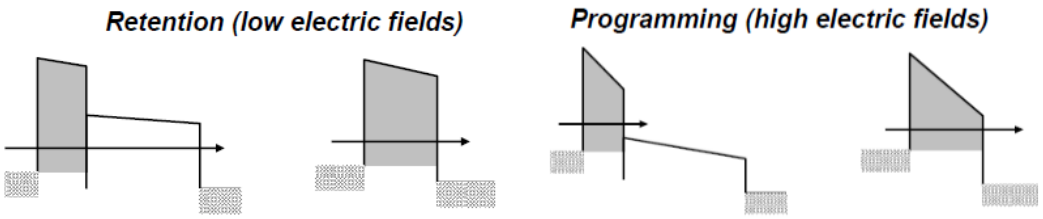


Figure 1.18: Principle of operation of crested barrier [27]

Trapping layer As schematically shown in fig 1.19-a since the charge is free to move along the conduction band, the polysilicon floating gate is very sensitive to SILC. The envisaged solution is to replace the polysilicon with a discrete charge trapping layer (fig. 1.19-b) where the charges, localised in the band gap of the medium, can not move freely reducing the impact of the SILC effect on data retention. The material most used as charge trapping layer is the silicon nitride (Si_3N_4) however others materials as HfSiON, AlN, Si, have also been studied (see table 1.2).

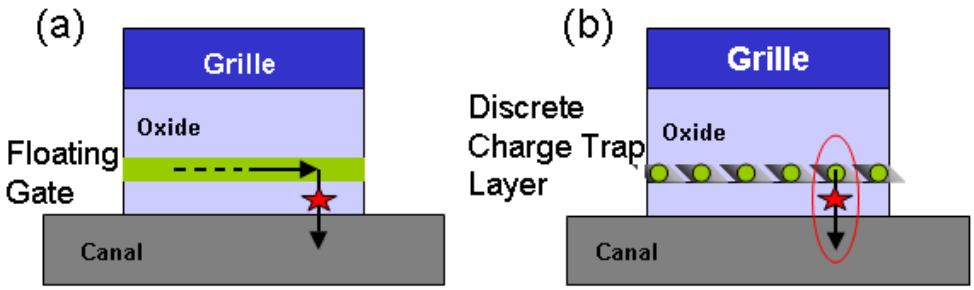


Figure 1.19: Schematic diagrams representing (SILC) phenomena for (a) continuous floating gate cell (b) discrete charge trapping layer.

Interpoly material Maintain a constant coupling ratio at a value of 0.6 is a great scaling challenge. The use of high-k dielectric in the interpoly dielectric is envisaged to reduce the total

EOT while maintaining or even increasing the gate coupling ratio. The choice of the high-k must be made taking into account that for most high-k materials higher dielectric constant comes at the expense of a narrower band gap (Fig. 1.20) which can itself result in leakage current during retention operation [28]. Among all the high-k materials the most studied are:

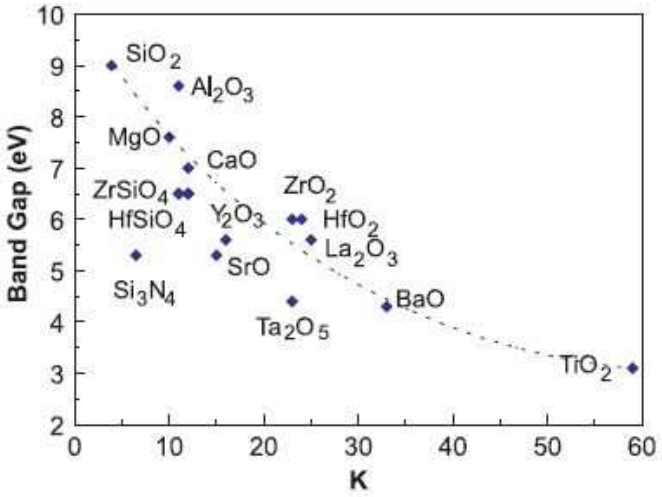


Figure 1.20: relationship between the dielectric constant and band gap [29].

- Alumina [30, 31].
- Hafnium Oxide [32].
- Aluminates hafnium [33].
- Silicates hafnium [34].

In particular Alumina dielectric is employed in the TANOS ($TaN/Al_2O_3/Si_3N_4/SiO_2/Si$) memory (see chapter 3), proposed for the first time by Samsung in 2005 [35], and studied in this thesis.

Despite the envisaged advantages, high-k materials are less known than the silicon oxide and they need further development before they can be integrated in the memory market. One of the main problem is that they inevitably introduce defects that can induce trap assisted conduction and degrade the memory operations [36, 37, 38].

Tunnel oxide	CT layer	Blocking gate	Control gate	Affiliation	Publication
SiO_2	Si_3N_4	Al_2O_3	TaN	Samsung	[35, 39]
SiO_2	$HfON$	$HfAlO$	TaN	Univ. Taiwan	[40]
SiO_2	AlN	$HfAlO$	$IrO2$	Univ. Taiwan	[41]
SiO_2	SiN, $HfAlO$, HfO_2, Al_2O_3	$HfAlO$, HfO_2, Al_2O_3, SiO_2	HfN	Univ. Singapore	[42] [43]
SiO_2	$AlGaN$	AlLaO3	TaN	Univ. Taiwan	[44]
$HfAlO$	$HfSiO$	$HfAlO$	IrO_2	Univ. Singapore	[45]
HfO_2	$Ta2O5$	Al_2O_3, HfO_2	TaN	Univ. Texas	[46, 47]
SiO_2	SiN	HfO_2, $HfSiON$	$Poly-Si$	NXP	[48]
SiO_2	SiN	Al_2O_3	$Poly-Si$	IMEC	[49]

Table 1.2: Example of memories employing different charge trapping layers and *High-κ*as inter-poly dielectric.

Control Gate During erasing operation flash memories employing SiN as charge trapping layer show a V_{TH} saturation phenomenon (fig. 1.21). This has been explained by the back-tunnelling effect. The back-tunnelling effect occurs during FN erasing operations when, due to the high negative bias applied on the gate electrode, the electrons flow from the poly-silicon gate to the charge trapping layer preventing the memory from a complete erasing. To overcome this problem and in order to suppress the depletion capacitance of poly-silicon floating gate, different metallic materials with a high work function have been successfully tried improving erasing dynamic [50].

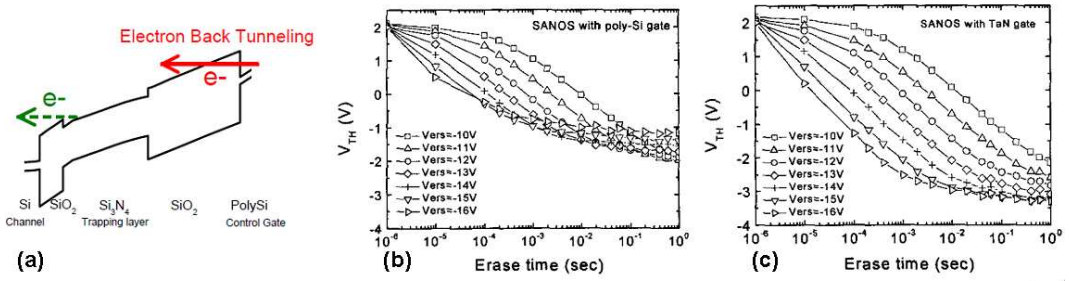
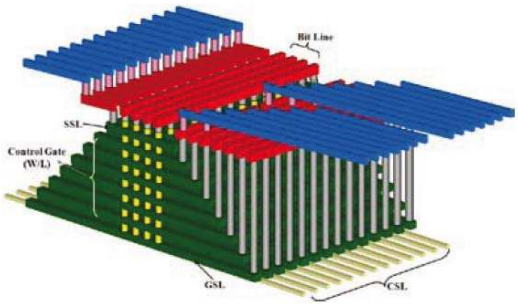
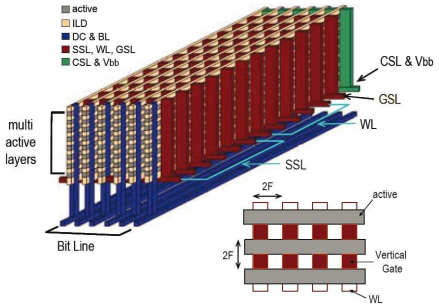


Figure 1.21: (a) Schematic explaining electron back tunneling phenomena. (b) Erase characteristics of SANOS device with n+ poly-Si gate and (c) TaN/n+ poly-Si gate [50]

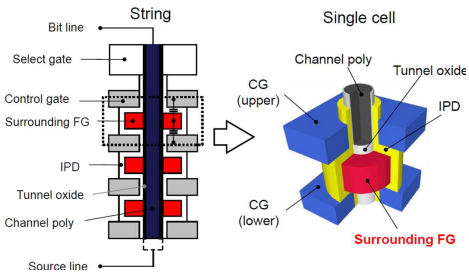
3D architectures The demand for reducing bit cost and increasing bit density have shrunk NAND Flash down to critical physical, electrical and reliability limits. A way to overcome the scaling limits is to use the third dimension in order to stack the cells inside the chip at arrays level. Many IC companies are currently developing this new generation of 3D memories. In fig. 1.22 are shown some of the last 3D solutions.



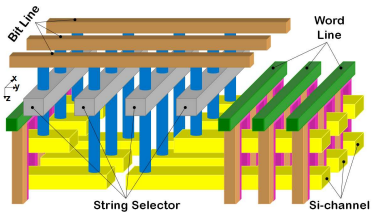
(a) TCAT: Vertical channel based on holes : GAA cell [51].



(b) VG-NAND: Horizontal channel based on trench : planar cell [52].



(c) DC-SF: Dual Control Gate Surrounding Floating Gate Cell [53].



(d) VG-NAND: Horizontal channel : FinFET cell [54].

Figure 1.22: Bird-eye views of others 3D array integration of charge trapped NAND Flash.

Split-gate memories Split gate memory has been invented for low power embedded applications in the 90's. This memory has been introduced especially for NOR applications in order to increase

- the injection efficiency
- the erase efficiency
- the disturb immunity

as reported in table 1.3 the split gate can have different geometries. The common idea is to add a separate access transistor to the flash memory. The access transistor, also called *select transistor*, controls the current that flows in the memory during program/erase operations. This implies a lowering of the current consumption that makes SG memories suitable for low power applications. Moreover it increases the disturb immunity due to the fact that when the access transistor is closed, no current flows through the cell.

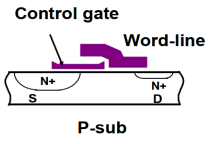
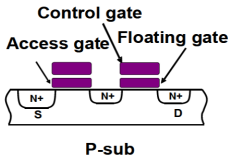
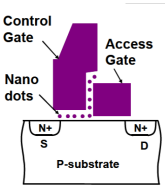
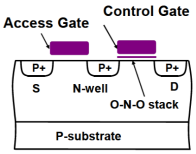
Type	1.5Tr cell (SuperFlashTM)	2Tr cell	1.5Tr	2Tr SONOS(PMOS)
Program	SSI	FN	SSI	CHE
Erase	FN (poly-poly)	FN (poly-sub)	FN	FN
Device structure				
Advantage	Fast program	Low power P/E	Fast, low-power program	Low power P/E
Publications	[55]	[56]	[56]	[57]

Table 1.3: Split gate memory technologies [9]

Programming Split-gate cells are programmed by channel hot electron. This is done keeping the bulk and the electrode on the select gate side (hereafter called *Drain*) grounded and applying a positive voltage on the select, memory and source electrodes (see fig. 1.23). The electrons are accelerated by the high parallel electric field induced by the high Source bias, and injected into the memory stack thanks to the vertical field due to the positive voltage applied on the memory gate. The injection comes at two points: at the pinch-off of the channel, corresponding to the region between the select gate and the memory gate; and at the Source junction where the lateral electric field is higher. In SG memory, differently from the classical flash cell, the channel current during CHE programming operation is efficiently controlled by the select gate allowing a reduction of the current consumed during CHE programming operations (see fig. 1.24).

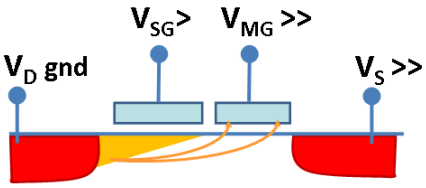


Figure 1.23: Split-Gate CHE schematic showing the two preferential injection points: at the limit between the two gates and at the Source junction.

Erasing Erasing efficiency is one of the main challenge in split gate memory. Depending on the memory geometry and memory gate stack composition, split-gate memory can be erased by:

HHI this method, in the follow also referred as Source Side Injection (SSI), is fast but it needs a high voltage on the Source electrode (fig.1.25-a) that can induce disturb phenomena and parasitic currents. Moreover the hot hole injection point is localized on the Source side causing a mismatch in spatial distribution between the trapped electrons and holes population that is a severe problem affecting the device endurance and retention [58, 59].

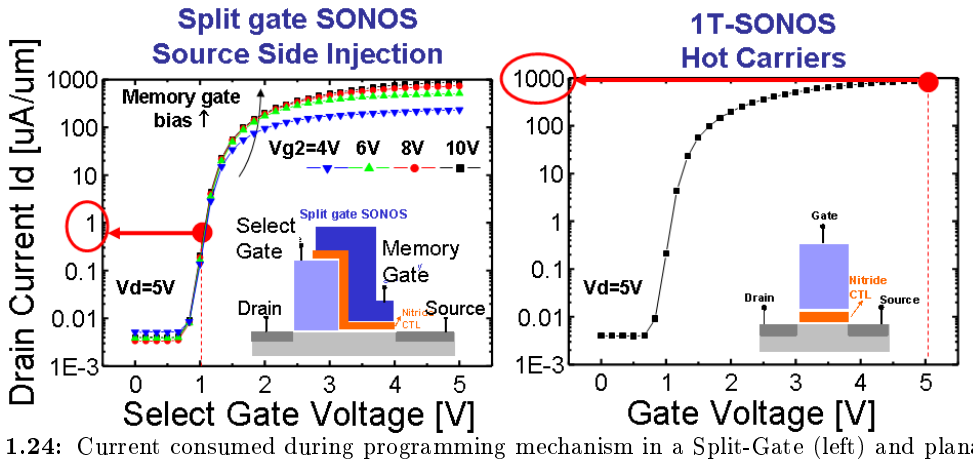


Figure 1.24: Current consumed during programming mechanism in a Split-Gate (left) and planar (right) memory

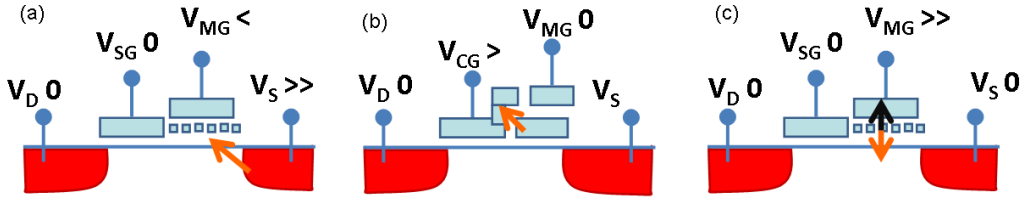


Figure 1.25: Split-Gate erase schematic (a) hot hole injection (b) poly-poly FN tunnelling (c) FN tunnelling through memory oxides.

FN allows an uniform erasing of the memory cell but it is slower than HHI. Split gate with polysilicon floating gate are erased using poly-poly FN tunnelling (fig. 1.25-b). In this case a floating-gate tip is used as a field enhanced tunnelling injector and lower voltages are needed. In charge trap-based split-gate memories this solution is not possible and the FN tunnelling is done through the interpoly dielectric or the tunnel oxide (fig.1.25-c).

Operation of split-gate charge trap memory cells made at CEA-LETI, and analysed in this thesis will be detailed in chapter 2. In the next section we will detail the state of the art of the split-gate (SG) charge trap memory (CTM) with the aim to describe the framework of this thesis.

1.4 Split-gate charge trap memory: state of the art

In the last ten years the SG technology dominating the split-gate cell market has been the *SuperFlash Split-gate* developed by the Taiwanese Silicon Storage technology (SST). The structure employing poly-silicon floating-gate memory, is licensed by the SST and it has been bought by the major IC companies (see fig. 1.26). A valid alternative to the super flash SST monopoly are the split gate charge trap memories studied in this thesis.

Split-gate charge trap memory combine the advantages of a discrete charge trapping layer

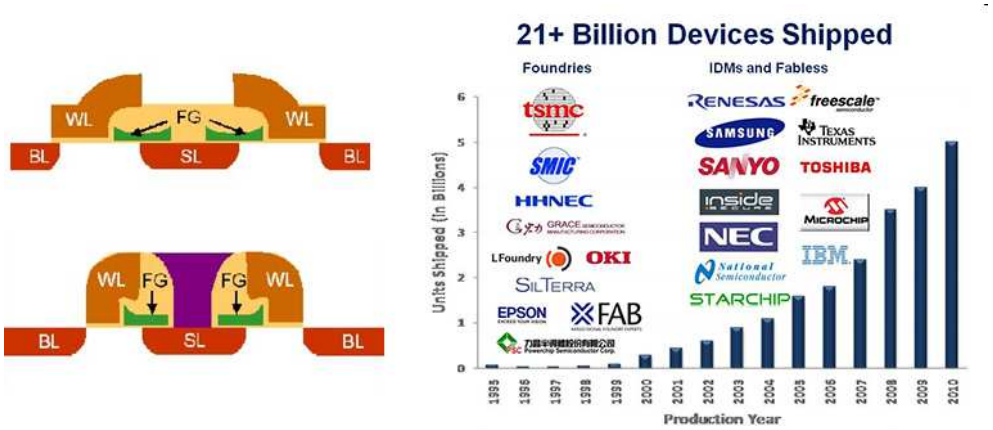


Figure 1.26: Left:Self-aligned SuperFlash cell (0.18-μm) for high density flash memory (top) and Non-self-aligned SuperFlash cell for low-density flash memory (bottom) [55]. Right: Number of produced SuperFlash cell licensed by SST [60].

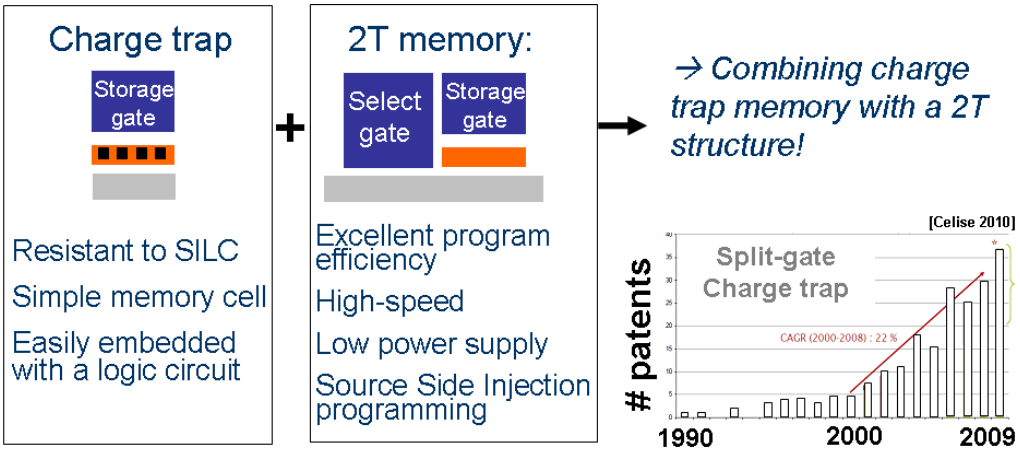


Figure 1.27: Left: Split-gate and charge trap memories advantages. Right: number of patents registered every year [61]

(scalability, resistance to oxide defects) and of the 2-Transistor (2T) configuration (high program efficiency, low consumption, fast access time...). Due to the increasing demand for consumer, industrial and automotive products, highly reliable, and low integration cost embedded memories are more and more required. In this context, split-gate CTM are a promising solution and the interest in this technology, evidenced by the number of registered patents every year (see Fig. 1.28), is increasing. Two main companies have reported development work on split-gate charge trapping memories (see fig. 1.28):

- **Freescale** integrating Si-ncs for microcontroller products in the 90nm technology node [62, 63].
- **Renesas** integrating Si_3N_4 as charge trapping layer. During the first half of 2012 Renesas announced the signature of an agreement with TSMC under which it would outsource manufacture of mcus at the 40nm node based on split gate MONOS (Metal/SiO_2/SiN_X/SiO_2/Si) architecture [64] .

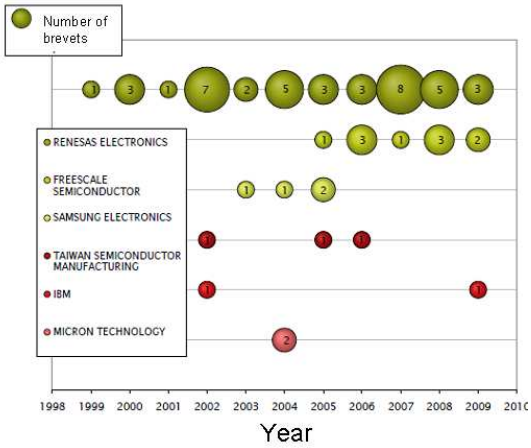


Figure 1.28: Number of patents registered by IC companies [61].

1.4.1 Architecture

A split gate memory can be done starting from the patterning of the select gate (SG "access first"), or the memory gate (SG "Memory First") as shown in fig.1.29 . Because of a better lateral isolation of the two control gates that allow a easier scalability, the select gate first is preferred. Moreover the access transistor first improves the coupling between the two gates avoiding the introduction of an additional insulator layer between them. As a result, in the following we study in details the "access first" Split-Gate. And by default Split-Gate in the following will refer to "access first" Split-Gate.

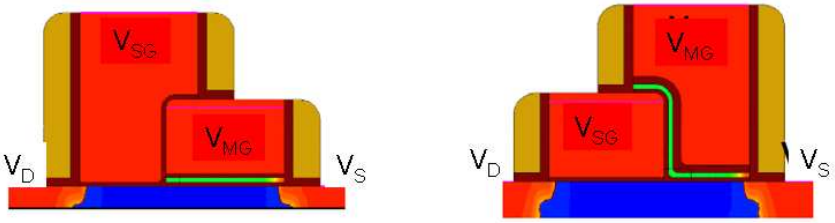


Figure 1.29: Schematic of SG with memory first Left) or Access first (Right) configuration.

As shown in fig.1.30 among all the possible architectures the two mainly used are the Spacer and the self-aligned multi-lithography. The multi-lithography approach, processed with select gate-first configuration, consists in a select gate transistor formed using standard logic processing and a memory gate stack that is deposited over it and etched to form the final split gate device. One of the main issues of the multi-lithography approach concerns the misalignment between the two control gates. To solve this problem, self aligned solutions have been proposed. In particular on a spacer approach the second control gate is defined by a spacer technology, on the edge of the access transistor. In this solution, one should notice that the second control gate does not require any lithographic step to define the spacer; only a non critical mask is used to remove the spacer on the adjacent sidewall of the select transistor. The structure and the principal advantages and drawbacks of these two architectures are reported in fig. 1.4 .

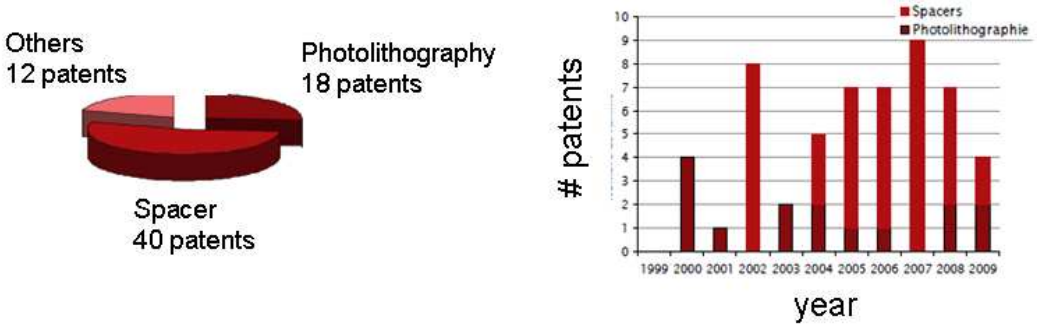


Figure 1.30: Study on patents: architecure [61]

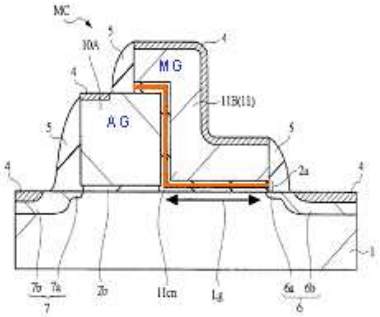
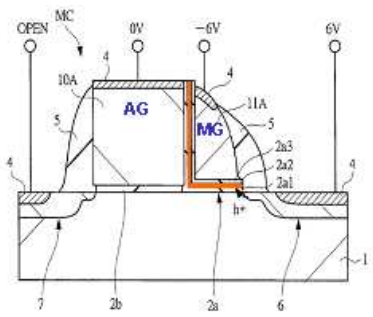
	MultiLitho	Spacer
		
Advantages	· Less critical masks (2nd gate length mainly controlled by poly deposition)	· Self-aligned, no alignment issue
Issues	· Misalignment directly leads to gate length variation (ΔVt variability)	· Memory Gate etching: technological challenge

Table 1.4: Comparison between Spacer and multi-litho split gate architecture.

1.4.2 Charge trap layers

At the beginning of this chapter at page 21 we described the advantage of integrating a discrete CT layer instead of the classical poly-silicon floating gate and the possible charge trap materials. A statistical study on registered patents employing CT split gate cell (Fig. 1.31), shows that among all the possible materials the most exploited are silicon nitride and silicon nanocrystal.

Silicon nanocrystal-based cell were introduced in 2006 by Freescale [65, 66, 67, 68, 69, 56, 70]. The technology were integrated in a array of 16 Mbit cell in 90 nm technology and the trapping medium consist of silicon nanocrystals (Si-Ncs). The select gate length is 150-200nm and the control gate length is 100-200nm. A TEM image of the cell is presented in Figure 1.35.

The memory is programmed and erased as described below. The w/e courbes are shown in fig. 1.33.

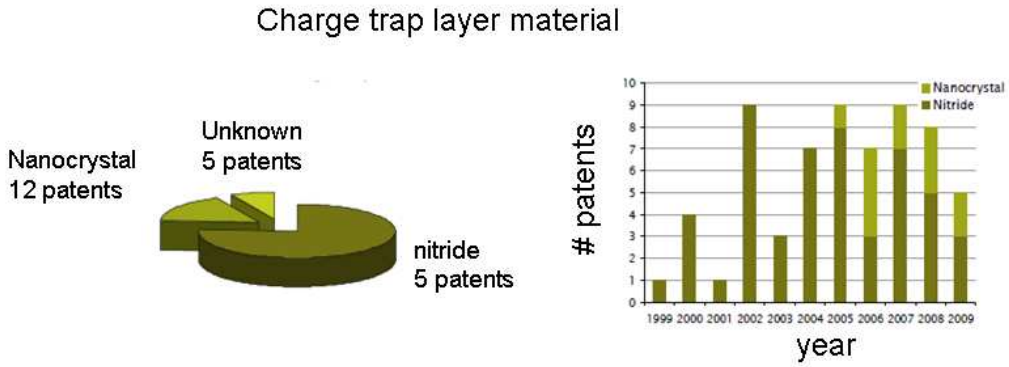


Figure 1.31: Study on patents: CT materials [61].

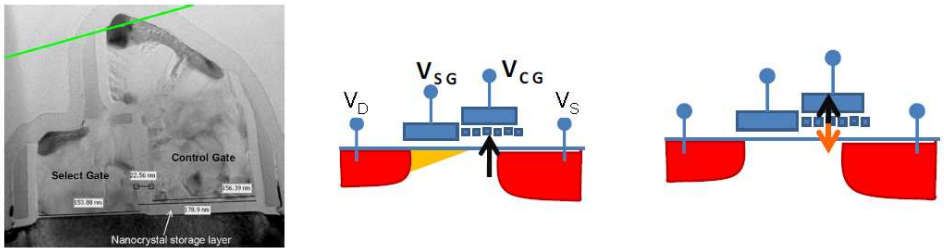


Figure 1.32: (left) TEM images of the cross section of nanocrystal Split-Gate Flash [66], (center) program schematic, (right) erase schematic.

Program The cells are programmed by hot electron tunneling (here called Source Side injection). A programming window of 3V is obtained in 1us, with a source voltage of 5V, a select gate voltage of 0.8V, a control gate voltage of 10V. Drain and bulk voltages are at 0V. The programming current is 2uA.

Erase Erasing is done by FN tunneling through the interpoly dielectric. The control Gate voltage is set to 14V, source, drain and bulk electrodes are keeping grounded. The erasing time is of the order of 10ms. There is no need for negative voltages.

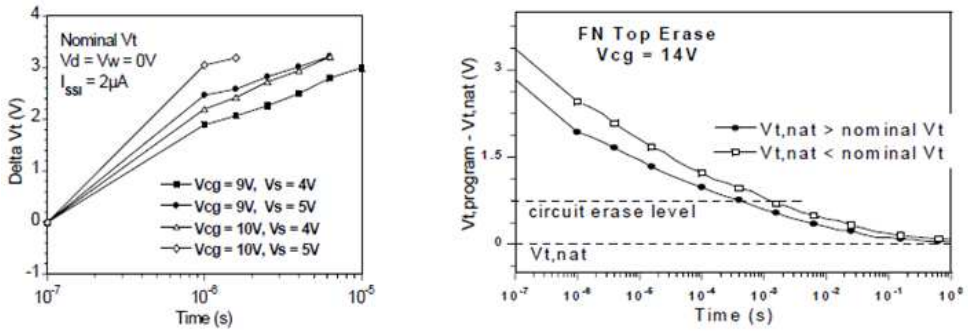


Figure 1.33: Left: Program vs. time for variations in control gate and source voltage. Right: FN erase of bitcells with Vcg = 14V. Each bitcell shows erase saturation near its natural Vt.

Silicon nitride is used as CTL in Renesas technology. A TEM image of their technology and its scaling is shown in fig 1.35. With this technology Renesas is now able to produce MCUs

consuming less than 10uW, aiming to produce nano watt MCUs (nW-MCUs) (see fig. 1.35).

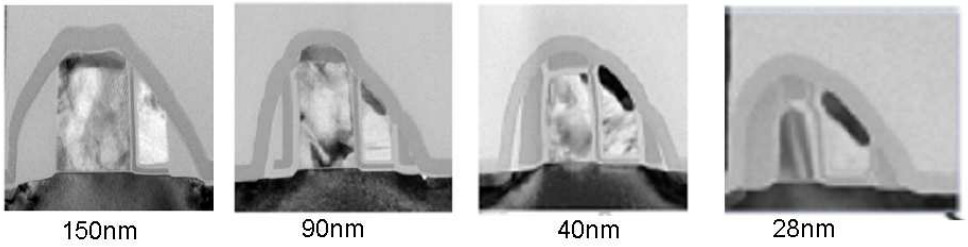


Figure 1.34: Tem images of scaled split-gate memories [3]

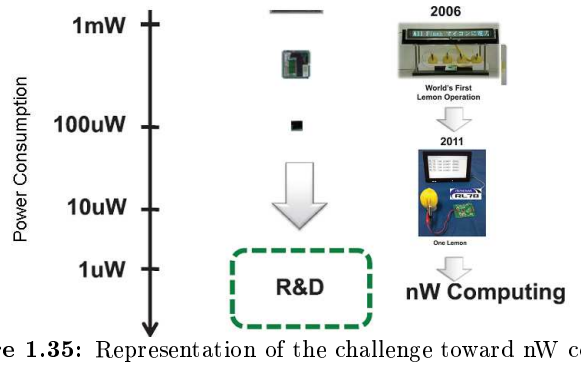


Figure 1.35: Representation of the challenge toward nW computing

1.5 Conclusion

In this chapter, we have presented the framework of this thesis.

In a first part the economic context, the evolution and the classification of semiconductors memory were presented. Then the Flash memory operations needed for understanding this thesis were reviewed. We thus presented the flash memory scaling limits and the proposed solutions: we explained the advantages of using a charge trapping layer instead of the continuous floating gate and a high-k control dielectric instead of the classical silicon oxide.

Finally, we introduced the split gate solution. In particular we reported the state of the art of charge trap Split-Gate cell, object of this thesis, that integrates in a split-gate structure the new materials presented as a solution to the flash memory scaling issues.

Chapter 2

Split gate charge trap memory

First of all in this chapter Multi-litho split-gate charge trap memories with electrical gate length down to 20nm are presented. Then Silicon nanocristals (Si-ncs), or silicon nitride (Si_3N_4) and hybrid Si-nc/SiN based split-gate memories, with SiO_2 or Al_2O_3 control dielectrics, are compared in terms of program erase and retention. The scalability of SiN based memories are thus studied, investigating the impact of gate length reduction on the memory window, retention, disturb, variability and consumption. Finally, the spacer-split gate memory is presented as a possible solution to overcame the multi-litho alignment scaling issue.

2.1 Basics of Split gate charge trap memories

Split-gate charge trap memories were processed with a 'memory last' configuration, meaning that the Memory Gate (MG) is deposited on the Select Gate (SG) electrode. Electron beam lithography was used to define select gates (L_{SG}) down to 40nm and the channel widths (W) down to 100nm. The electrical memory gate length is controlled by the poly-Si layer overlapping the memory channel (Fig.2.1-left), allowing to achieve gate length down to 20nm (Fig.2.1-right). In the following, L_{MG} will refer to the electrical length.

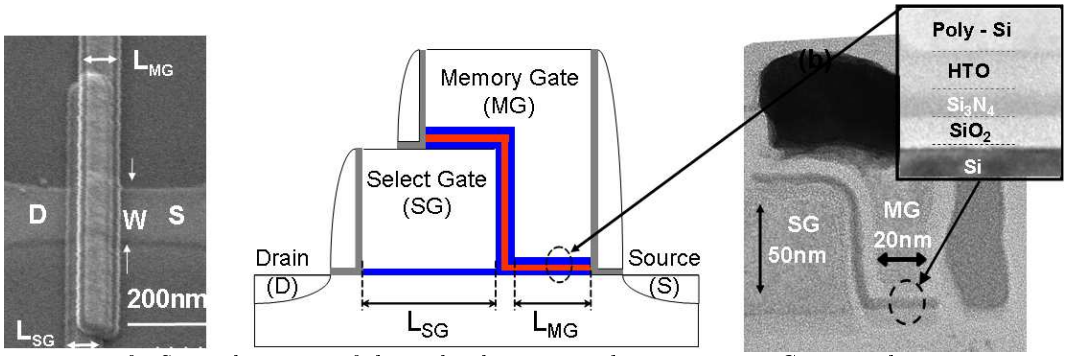


Figure 2.1: Left: SEM plane view of channel, select gate and memory gate. Center: schematic cross section of a split-gate memory. Right: TEM images of a 20nm SiN split-gate memory

In this structure, various gate stacks were integrated with Si-nc (Sample A), Si_3N_4 (B), and hybrid Si-nc/SiN (C) charge trapping layer CTL. Moreover we integrated HTO/Al_2O_3 control dielectrics to Si-nc/SiN CTL (D) to increase the gate coupling.

The silicon nanocrystals were grown on the tunnel oxide surface by Low Pressure Chemical Vapor Deposition (LPCVD) with a two-steps process. First, diffusion of Silane (SiH_4) on hydrophilic surface causes the germination phenomenon; then the Dichlorosilane (SiH_2Cl_2) is introduced in the process and the nanocrystals are selectively grow. Using this method the size and density of Si-nc were independently controlled. The first by the Dichlorosilane diffusion period the second by the Silane diffusion period.

In the case of hybrid Si-nc/SiN layer, the Si-nanocrystals were capped in-situ by a 2nm SiN layer, as described in [19], in order to boost the memory window.

In table 2.1 the technological details of the various stacks are reported, while TEM cross sections are shown in fig. 2.2. The samples with Si-ncs were analysed in Energy Filtered mode when Si is selected, putting in evidence the presence of Si clusters.

	Sample A	Sample B	Sample C	Sample D
Tunnel Oxide	SiO2 (5nm)	SiO2 (5nm)	SiO2 (5nm)	SiO2 (4nm)
Charge trapping layer	Si-ncs ($\Phi \sim$ 6nm)	SiN (6nm)	Si-ncs+SiN (3nm)	Si-ncs+SiN (3nm)
Control dielectrics	HTO (8nm)	HTO (10nm)	HTO (10nm)	HTO (3nm) Al2O3 (8nm)
Control Gate	Poly-Si	Poly-Si	Poly-Si	TiN

Table 2.1: Technological details of the studied split-gate charge trap memories

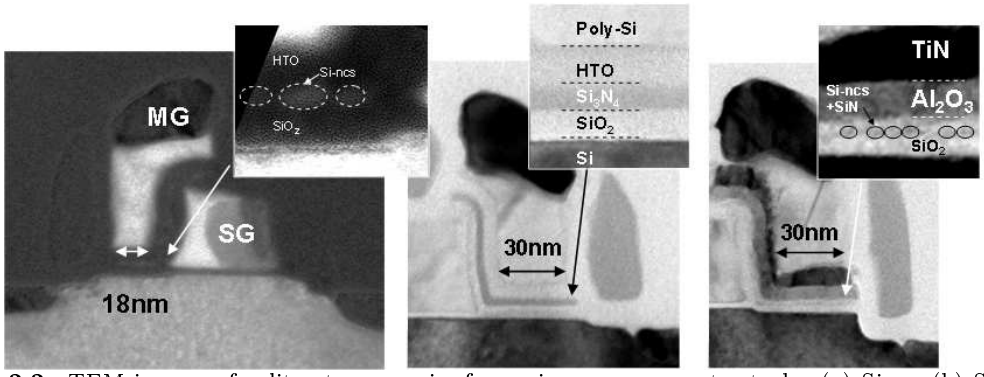


Figure 2.2: TEM images of split-gate memories for various memory gate stacks, (a) Si-nc, (b) Si_3N_4 (c) Si-nc/SiN charge trapping layers. (a) was obtained in Energy Filtered mode (selecting Si), (b) and (c) in High Resolution mode

$I_D(V_{SG},V_{MG})$ characteristic In split-gate memory the select gate allows to control the channel current. In fig 2.3 we can see a $I_D(V_{MG}, V_{SG})$ transfer characteristics measured for 20nm and 70nm memory gate length devices. On the 20nm memory, a slight V_{TH} reduction appears as V_{SG} increases, due to a parasitic control of the memory channel by the select transistor. Despite this issue, the channel current can efficiently be controlled from ON to OFF state by biasing the select gate.

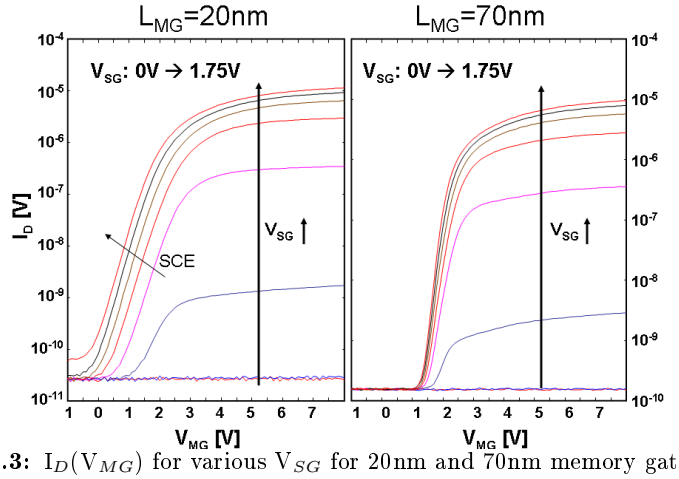


Figure 2.3: $I_D(V_{MG})$ for various V_{SG} for 20nm and 70nm memory gate lengths

Consumption measurement In order to quantify the programming consumption, the source current was measured during the programming operation using the dynamic technique proposed in [71]. The developed set-up (Fig 2.4) uses two waveform generators combined with two WGFMUs (Waveform Generator and Fast Measurement Units) integrated in an Agilent B1500A semiconductor device analyser.

The set-up was verified comparing the $I_S(V_{SG})$ transfer characteristics (at low V_S) with the average current consumed during a programming pulse for various select gate voltages. The good matching between the current measured in continuous and dynamic mode (Fig. 2.5) demonstrates the validity of our setup. Moreover it proves the capability of the select transistor to control the current even when high bias voltages are applied to the source and memory gate electrodes.

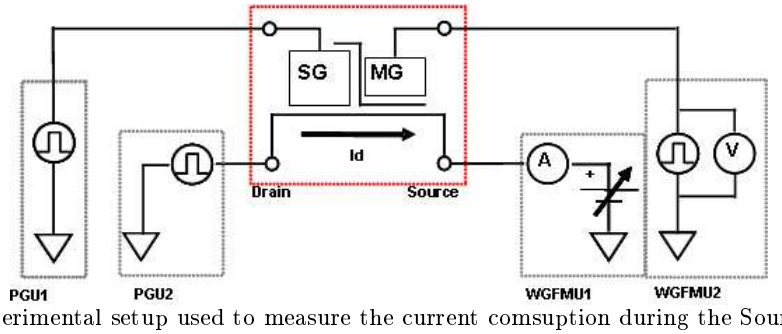


Figure 2.4: Experimental setup used to measure the current comsuption during the Source Side Injection (SSI) programming operation

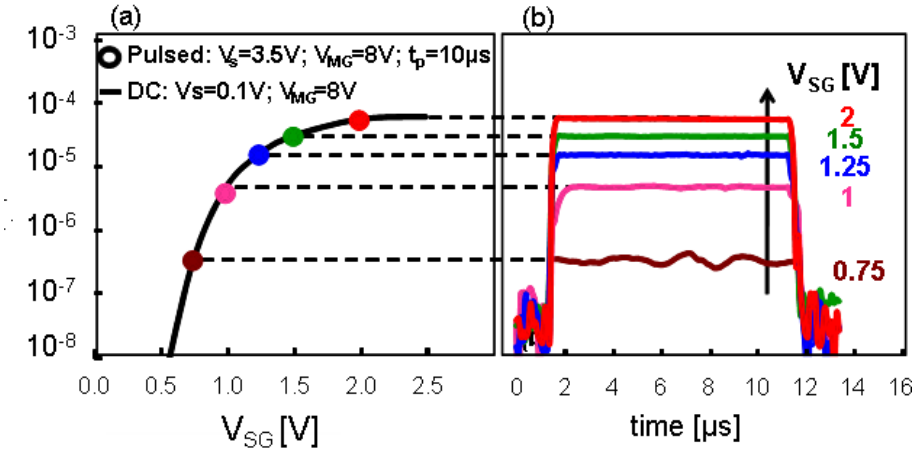


Figure 2.5: Comparison between (a) $I_S(V_{SG})$ transfer characteristic and (b) channel consumption current as a function of the select gate voltage (V_{SG}) during a 10μ s programming pulse. In dynamic mode, each point corresponds to the average current measured during the pulse

2.2 Impact of the memory gate stack on the memory performances.

For the samples described above (see table 2.1) we investigated the impact of the memory gate stack on the memory performances.

2.2.1 Programming

The split-gate memories are programmed using Source Side Injection (see §1.3.2), biasing both the memory gate and the source electrode at high voltages. The select gate potential is set to 1V in order to operate close to the threshold regime (I_S ~10μA).

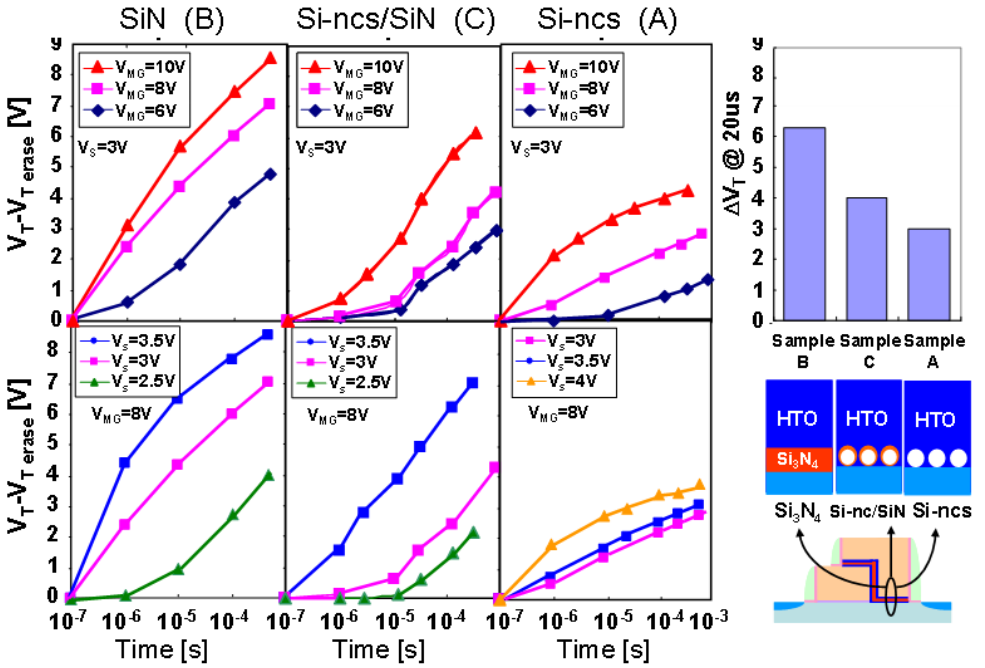


Figure 2.6: Left: Program characteristics in Source Side Injection mode of sample B (Si_3N_4), sample C (Si-nc/SiN), sample A (Si-nc) with 30nm memory electrical gate length, for various programming V_{MG} and V_S. Right: Memory windows of Si-nc, Si-nc/SiN and Si_3N_4 split-gate charge trap memories for same programming conditions (VGM=10V, V_S=3V, tw=20μs)

Figure 2.6 shows the normalized program characteristics of Si_3N_4, Si-ncs, and hybrid SiN/ncs split-gate memories for various programming V_{MG} and V_S. The memory gate length is of 40nm. Due to a higher density of trapping sites, nitride memories exhibit a higher memory window. In particular, for a 10μs programming pulse with V_{MG}=10V and V_S=3.5V, the threshold voltage shift is about 3V, compared with 1.25V of Si-nc samples. Thus, higher voltages or longer programming times are required for Si-nc memories to achieve the same memory window. Interestingly, hybrid Si-ncs/SiN layers offer a good memory window improvement and allow to partly compensating the reduced ΔV_{TH} of Si-nc memories.

These results are summarized on the right side of fig. 2.6 . In this bar diagram we show

the memory window for the different stack after a programming time of 20μs and a given programming condition. It appears that Si_3N_4 memory has the highest memory window. On the contrary, Si-nc memory exhibit the lowest ΔV_{TH}. Whereas Hybrid Si-ncs/SiN shows an intermediate behaviour.

2.2.2 Erasing

Various erasing methods were used depending on the nature of the charge trapping layer and the control dielectric:

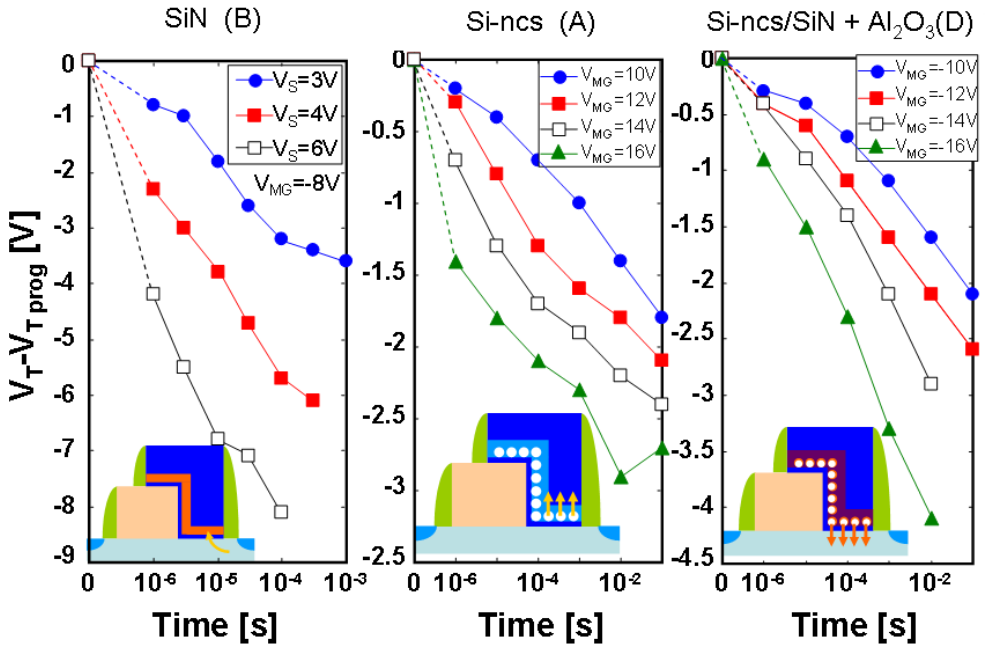


Figure 2.7: Erase characteristics of 30nm memories with various erasing mechanisms: Hot Hole Injection (sample B: Si_3N_4), FN through top dielectric (sample A: Si-ncs) and FN through bottom dielectric (sample D: Si-ncs with high-k top oxide).

Nitride based memories are erased using Hot Hole Injection (HHI), this method allows faster erasing speed but suffer from a higher current consumption. Moreover the high voltage needed on the source electrode could induce disturb phenomena especially when the memory dimensions shrink (see §2.3.3.2).

Si-nc memories with **HTO control dielectrics** are erased by Fowler-Nordheim (FN) injection from the top oxide: the electrons are removed from the ncs by tunnelling through the HTO to the memory gate electrode. This erasing method is slower than HCI but no current flows in the channel as both drain and source electrodes are kept grounded. The high voltage applied on the gate electrode creates a high potential difference between the select gate and the memory gate that can cause irreversible damages. Keeping floating the select gate electrode we were able to erasing the memory voltages up to 20V without breaking the device.

Si-nc memories with **high-k control dielectrics** are erased by FN injection through the bottom oxide. In this case the electrons are removed from the ncs through the SiO_2 tunnel oxide. Thanks to the high-k dielectric, for a given programming condition, the potential that drops on the tunnel oxide is higher comparing with the samples integrating silicon oxide as control dielectric. This electric field enhancement leads to faster erasing speed.

Fig.2.7 presents the corresponding erasing characteristics; HHI allows faster erasing speed but suffer from a higher current consumption. In FN mode, +16V and -16V gate voltages are respectively used to erase memory samples with HTO and Al_2O_3 control dielectrics.

2.2.3 Retention

For the four samples we measured the retention characteristics after a initial pulse of 10 μs with V_D=4V, V_{MG}=10V ans V_{SG}=1. We studied devices with a memory gate length of 90nm.

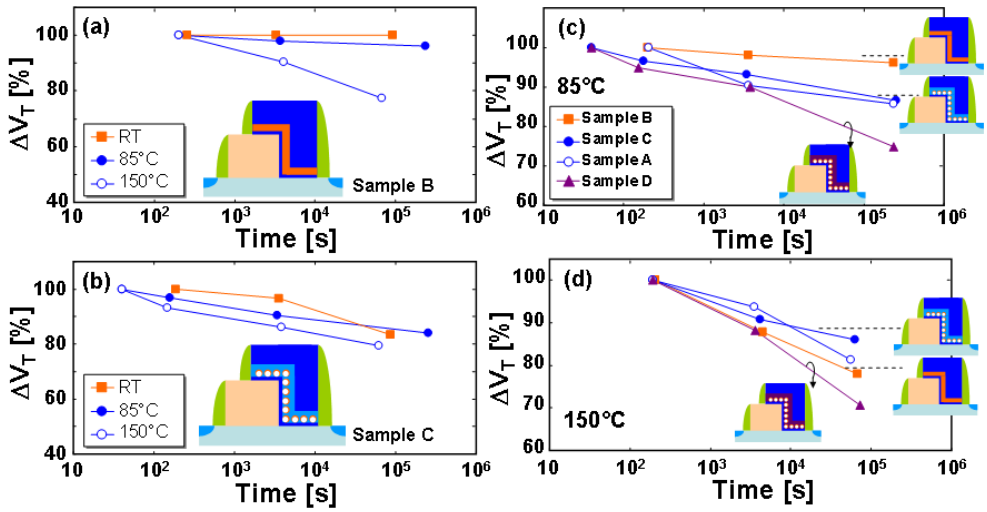


Figure 2.8: (a, b): Retention characteristics for various temperature for Si_3N_4 (sample B) and Si-nc / SiN (sample C) charge trapping layers. (c, d): Comparison of retention characteristics at 85°C and 150°C of split-gate charge trap memories with various gate stacks.

The retention characteristics are presented in fig.2.8. A high temperature activation is measured with a nitride CTL with a strong charge loss at 150°C (Fig. 2.8-a), while Si-nc/SiN CTL exhibits a more stable behaviour as the temperature is increased (Fig.2.8-b). Fig.2.8-c and Fig.2.8-d show the comparison of the memory samples at 85°C and 150°C. Up to 85°C, Si_3N_4 CTL offers the best retention performances, while for higher temperatures, an inversion of trend is observed as Si-nc memories present the smallest charge loss in agreement with [72].

For all the investigated temperatures, memories with high-k control dielectrics show faster charge decay, due to the thinner tunnel oxide and the lower barrier height of Al_2O_3 compared to SiO_2. These results are summarized in fig. 2.9 where we reported the charge retention after about three days from the initial writing.

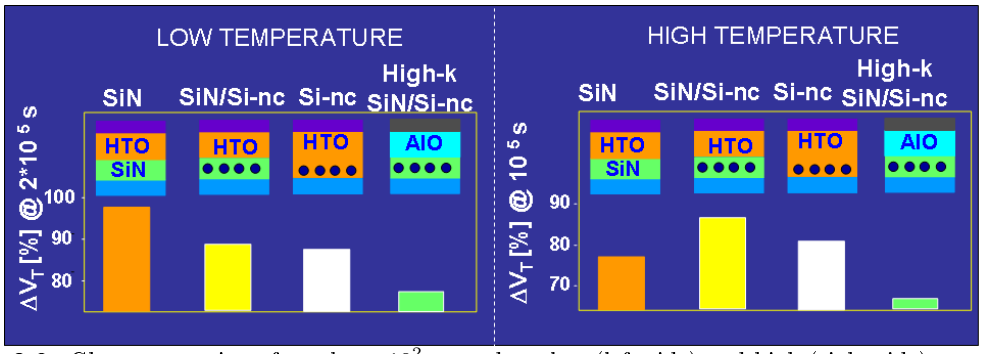


Figure 2.9: Charge retention after about 10^2 seconds at low (left side) and high (right side) temperature

The choice between SiN or Si-ncs CTL should be made considering the final application of the memory device: at low-temperature applications, the charges trapped deeply in SiN (1.2 to 1.6eV [73]) are hardly de-trapped and SiN is preferable to Si-ncs. On the contrary, at high-temperature applications (e.g. embedded applications for automotive products), silicon nanocrystals show a better retention than SiN. Indeed, in nano crystal memories the electrons are stored in the Si conduction band, and do not have temperature-activated de-trapping processes.

2.3 Scaling the memory dimensions

In this section we investigate the impact of the select gate and the memory gate scaling on the memory operation.

Thanks to multi-lithography approach and e-beam lithography, we processed devices with select gate lengths from 500nm down to 40nm. For each select gate length we processed memory gate with electrical gate lengths from 350nm down to 20nm. In the time this thesis is written this is the smallest device presented in the world. Investigating the consequence of this aggressive scaling is crucial for the development of the split-gate technology.

The study on scaling of the dimensions (memory and select gate) were done on SiN memories. This choice was motivated by the higher programming window of nitride memories (see §2.2.1) that allows a better understanding on the physical mechanism involved in the memory operations. The experimental results were explained by means of TCAD simulations.

2.3.1 TCAD simulation

Simulations were performed with the Synopsis TCAD Sentaurus software. The interest in this program resides in the fact that we can reproduce the structure and the physics of a 2D system. In this section we will show the steps and the tools that allowed us to calibrate the software on our devices before using it for the physical understanding of split-gate behaviour.

Synopsis tools The synopsis program is made of many tools that are linked with each other. The facilities we used are:

Sprocess is an advanced 1D, 2D and 3D process simulator for developing and optimizing silicon semiconductor process technologies. The output is a device structure which can be used for device electrical simulations. We used it mainly for the simulation of the Spacer architecture (see §2.5.2) .

SDE Sentaurus Structure Editor is a 2D and 3D device structure editor. With this tool we can draw the device structure or, starting from the output of the process simulation, it is possible to add the contact and generating the mesh.

Sdevice is a device simulation tool that simulates the electrical characteristics of semiconductor devices, as a response to external electrical, thermal or optical boundary conditions imposed on the structure. The input device structure comes from process simulation with the aid of tools like Sentaurus Structure Editor.

Tecplot Specialized plotting software. Dedicated for scientific visualization of the simulation results, for example, energy band diagrams and cross-sectional 2D or 3D data.

Inspect Inspect Curve display program. Simulation output can be plotted using Inspect, such as current-voltage characteristics. We used it also for the threshold voltage extraction.

SWB Sentaurus Workbench is the primary graphical front end that integrates TCAD Sentaurus simulation tools into one environment. We used it to organize, and run simulations.

Structure creation The structure and its mesh fig. 2.10 has been implemented under Structure Device Editor software. We can notice that the mesh size is smaller in the region where the electric field is higher. The choice of the mesh has to be done considering that the Poisson's equations are solved for each point of the mesh. If the mesh size is too large, the results obtained from the simulation are imprecise on the contrary with a too small mesh the computation time increases too much.

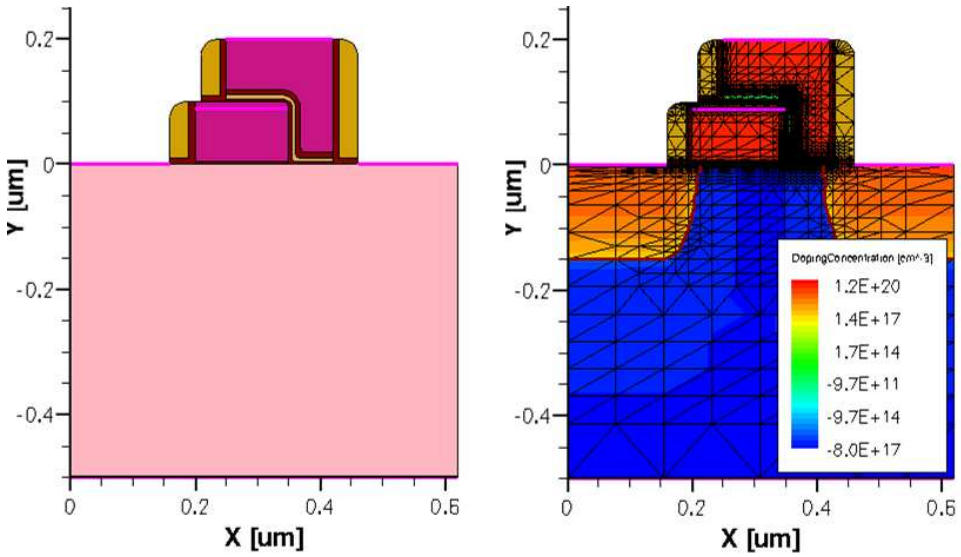


Figure 2.10: Boundary (left) and mesh (right) of the simulated SiN split-gate memory (L_{MG}=50nm L_{SG}=100nm).

Parameters calibration The junction size, the mobility parameters, the dopant concentrations were found reproducing the $I_D(V_{MG},V_{SG})$ characteristic of various split gate devices with different dimensions. The initial simulation parameters were taken from previous simulations [74] made on planar memory fabricated at CEA/LETI and processed similarly to our split-gate memory. In fig. 2.11 we reported fitting between the measured and simulated threshold voltage for device with various memory gate lengths. We can see that the simulation fits the data both when the reading is done in reveres mode (V_S=1.5, V_D=0) or in forward mode (V_D=1.5, V_S=0). Differently from a planar SONOS memory we can notice a difference between forward and reverse in the fresh cell due to the asymmetric structure of the split gate cell.

The simulation parameters were calibrated by fitting the programming characteristics over different V_S/V_{MG} for two memories with respectively 20nm and 40nm memory gate lengths (Fig. 2.12). To be sure of the dimensions, we made a TEM image on a tested 20nm device. In our structures, 3V of programming V_S is sufficient to generate hot carriers due to the short L_{MG}. Note that the numerical simulations correctly reproduce our experimental results.

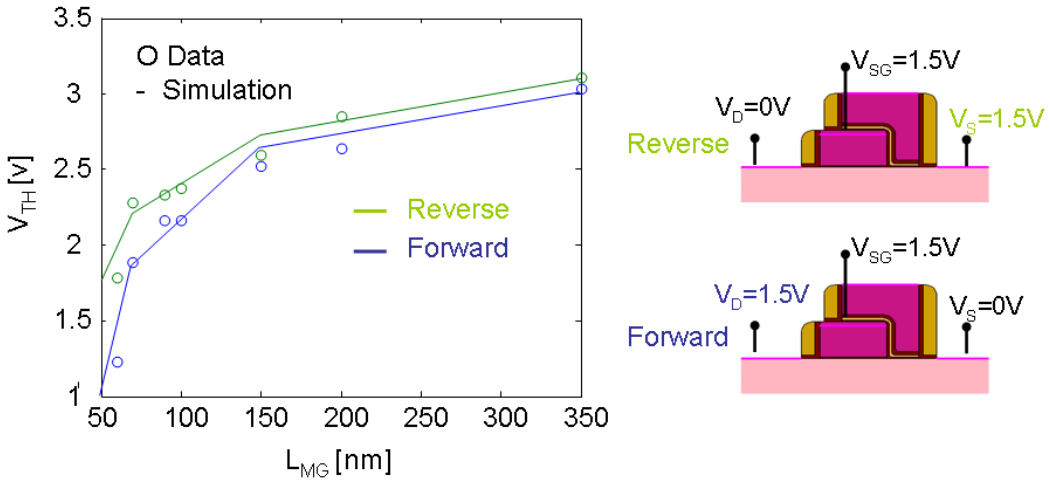


Figure 2.11: Left: Simulated and measured threshold voltage in reverse and forward mode. Right: schema representing the forward and reverse reading conditions.

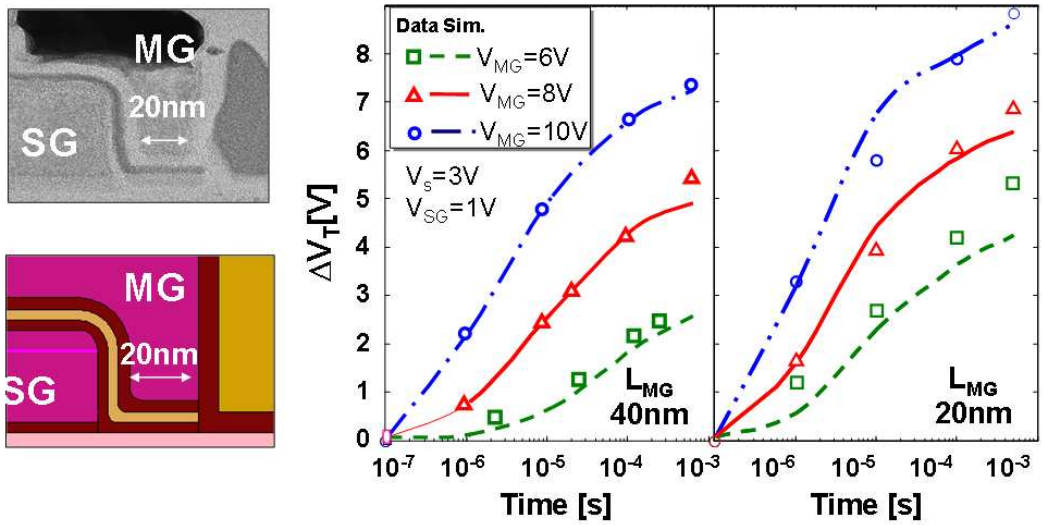


Figure 2.12: Left: (up) TEM image of the tested device with a 20nm memory gate length; (down) corresponding TCAD simulated structure. Right: Measured and simulated programming characteristics using the Fiegna model.

Fiegna model The simulation of the SSI mechanism is crucial for the understanding of the SG physical behaviour. In order to simulate the hot channel injection programming we activated Fiegna model [75] in the Synopsis tool suit.

In Fiegna's model, the hot carrier injection current I_G (see fig. 2.13) is calculated as an integral along the semiconductor-insulator interface (s) over the product of energy dependent normal to interface carrier velocity (v), carrier distribution energy (f), and carrier density of states (g), so that:

$$I_G = \int P_{ins} \left[\int_{E_{B0}}^{\inf} v_{\perp}(\epsilon) f(\epsilon) g(\epsilon) d(\epsilon) \right] ds \tag{2.1}$$

where E_B is the Si-SiO_2 barrier energy and P_{ins} the probability that an electron does not scatter in the image potential well.

The carrier distribution energy is approximated to the case of a parabolic and an isotropic band structure, and equilibrium between lattice and electrons leading to a simplified expression of the gate current:

$$f(\epsilon) = Aexp\left(-\chi\frac{\epsilon^3}{F_{eff}^{1.5}}\right) \tag{2.2}$$

where A is a fitting parameter; χ is a constant of the high-energy distribution function; n the electron density; and F is the effective electric field that replaces the local electric field to capture at first order the effects of the non-locality of hot electron injection [76].
Putting 2.2 in 2.1 the gate current can be rewritten as:

$$I_G = q\frac{A}{3\chi}\int P_{ins} n \frac{F_{eff}^{3/2}}{\sqrt{E_B}} e^{-\chi\frac{\epsilon^3}{F_{eff}^{1.5}}} ds \tag{2.3}$$

2.3.2 Understanding of SSI operation

The understanding of the physical mechanisms beside the SSI operation in split-gate memories is crucial for the interpretation of different aspects of the experimental results [77, 78, 79, 80, 81]. In this section we use simulations to understand the impact of the memory and select gate scaling on the measured programming efficiency and current consumption.

In split-gate charge trap memories, the electrons flow through the channel below the select gate to be subsequently injected toward the charge trapping layer. This happens because the source voltage induces a high electric field parallel to the interface that gives to the electrons the needed energy to pass over the Si/SiO_2 barrier. Then, the electrons are driven to the charge trapping layer by the attractive transversal electric field induced by a strong positive voltage applied on the memory gate (Fig. 2.13).

Dummy gate experiment We started with measuring a reference split-gate structure with a dummy memory transistor, composed by a 5nm oxide instead of the Oxide/Nitride/Oxide memory stack, in order to directly monitor the injected current through the tunnel oxide. We measured the source current (I_S); the memory gate current I_{MG}; and we computed the injection

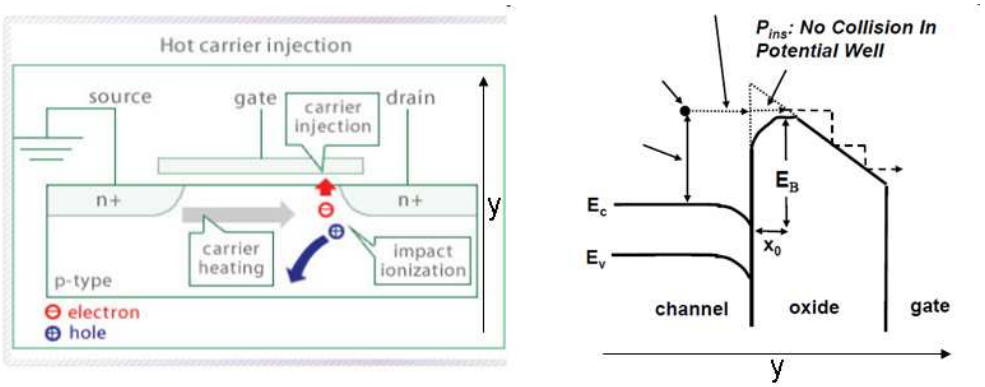


Figure 2.13: Schematic representation of HCI mechanism.

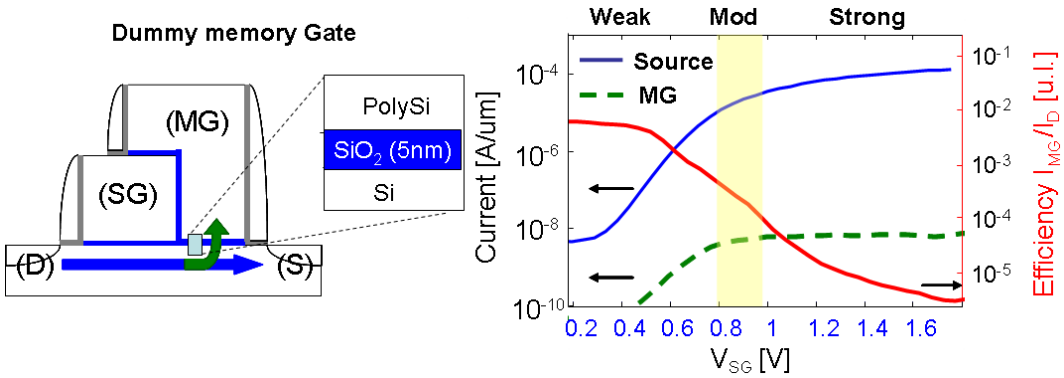


Figure 2.14: Left: Schematic representation for a split-gate with a dummy memory stack composed by a 5nm SiO_2. Right: Measured source current (I_S); memory gate current IMG; and computed injection efficiency (I_{MG}/I_S) for a dummy memory gate with L_{SG}=200nm and L_{MG}=100nm.

efficiency (I_{MG}/I_S) as a function of the select gate voltage for V_S=3V and V_{MG}=3V. Fig 2.14 shows that the best choice for the select gate voltage during the programming operation is close to the threshold voltage, giving the best compromise between a low current consumption and a high current injection. Indeed, when the select transistor is in weak inversion, the gate and source currents increase nearly exponentially, on the contrary when the select transistor is in strong inversion, the injection current saturates and using a higher V_{SG} is inefficient.

The experimental results were figured out by the device simulations. The simulated channel potential during programming operation when the select transistor is in weak (V_{SG}=0.3V), moderate (V_{SG}=0.9V), and strong inversion (V_{SG}=2V), is reported in Fig.2.15-a. First it should be noted that most of the hot electrons are generated by the strong electric field created across the weak-controlled gap between select and memory gates. Indeed, in this thin region occurs the major voltage drop between the select and memory gates. To understand the main parameters that influence the injection efficiency, we rewritten the equation of the injection current (2.6) as:

$$I_G = q\frac{A}{3\chi}\int P_{ins}\cdot p2\cdot ds \tag{2.4}$$

where

$$p2 = n\cdot\frac{F_{eff}^{3/2}}{\sqrt{E_B}}e^{-\chi\frac{\epsilon^3}{F_{eff}^{1.5}}} \tag{2.5}$$

depends on:

- the product, defined as $p2$ (2.5), between a monotonic function of the local electric field F and the number of channel electrons n;
- the probability P_{ins} that an electron does not scatter with the image potential well.

Because P_{ins} depends on the height of the Si-SiO_2 barrier and this is a function of the insulator field, we will neglect it in this analysis as we can consider it as constant for a given V_{MG}.

The product p2 (Fig.2.15-b), in agreement with the experimental results, is low at V_{SG}=0.3V and remains in the same order of magnitude between V_{SG}=0.9V and V_{SG}=2V. This can be explained by the fact that F and n have opposite behavior as the select gate bias increases.

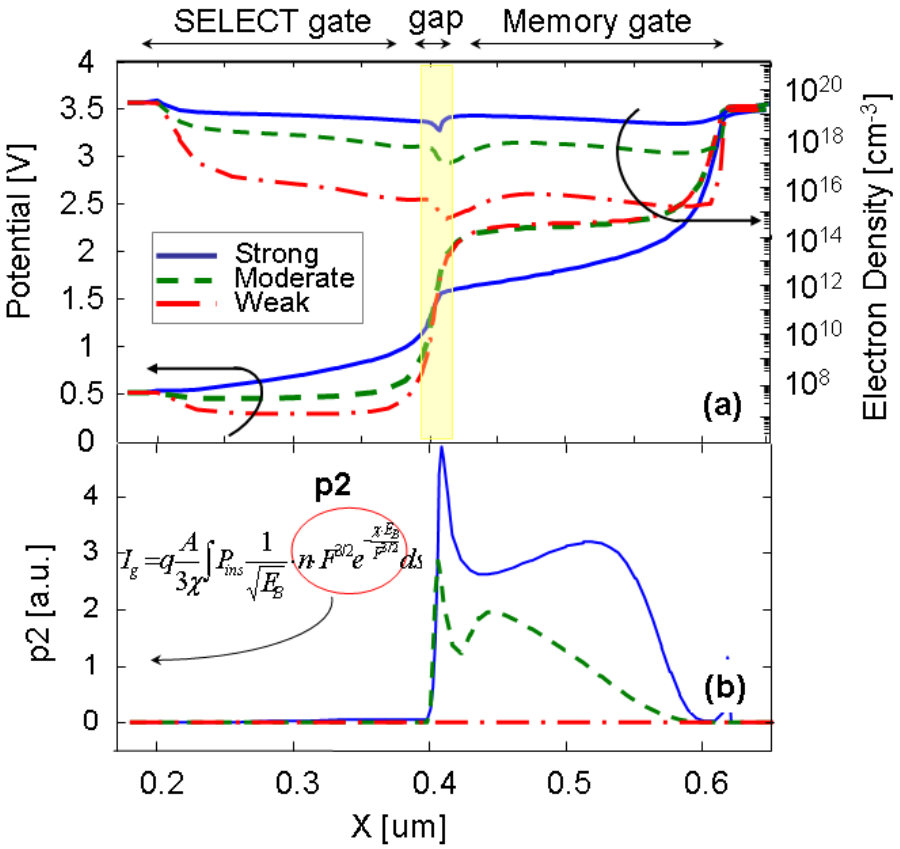


Figure 2.15: (a) Simulated potential profile and electron concentration at the Si/SiO2 interface during programming operation. The bias conditions are: V_{MG}=3V V_S=3V V_{SG}=0.3V (weak) V_{SG}=0.9V (moderate) V_{SG}=2V (strong inversion). The device dimensions are L_{SG}=200nm and L_{MG}=200nm. (b) Corresponding simulated injected current (in arbitrary units), plotted along the memory channel.

When the select transistor operates in weak inversion, the injected current increases with V_{SG}, due to the increasing of the amount of electrons provided by the select transistor. On the other hand, in strong inversion, the injected current is limited by the reduction of the electric field, as V_{SG} increases. Indeed, in strong inversion, the select gate potential is disturbed by the memory gate, causing a lowering of the potential difference at the gap side that results in a lower electric field.

Analogue experiments have been done for the memory devices described in section §2.1 where we measured the consumed current at the source electrode and the memory gate threshold voltage shift during a programming pulse (V_{MG}=8V, V_D=3V τ=100μs). Fig.2.16 shows the memory window (ΔV_{TH}) and the current consumption I_S when the select gate is in weak, moderate and strong inversion. The previous behaviour (Fig.2.14) was found again, confirming that the optimal choice for V_{SG} is in a region strictly above the select gate threshold voltage, insuring the best compromise between a high programming window and a limited current consumption.

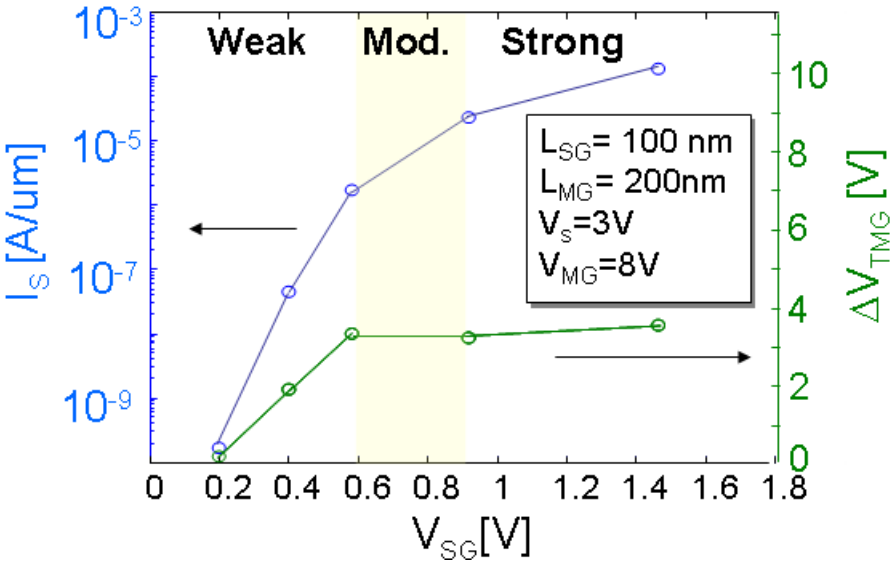


Figure 2.16: Measured channel current and programming window versus the normalised select gate voltage during a 10μs programming pulse with V_S=3V and V_{MG}=8V. L_{MG}=200nm

2.3.3 Select Gate scaling

The scaling of the select gate is limited by the lithography challenges associated with printing sub-wavelength features and its sensitivity to variation. Thanks to e-beam lithography we achieved select gate length down to 40nm. At this scaled dimension the functionality of the select gate must be proved. In this part we will investigate the impact of the memory gate scaling on programming/consumption and disturb.

2.3.3.1 Programming and consumption

The effect of the select gate scaling on the programming current consumption was investigated by measuring the select gate threshold voltage lowering and the programming window for devices with a select gate length from 350nm down to 40nm.

Fig.2.17-a shows that as the select gate dimensions scale, the memory window remains unchanged but the select gate threshold voltage decreases due to DIBL (drain induced barrier lowering). This parasitic effect causes, for a given V_{SG}, an increase of the consumed current during program operation. For instance, as the select gate scales from 90nm to 40nm we measured during a pulse of 10μs with V_S=3V; V_{MG}=10V; V_{SG}=1V (corresponding to the same ΔV_{TH} in the two devices: see Fig.2.17-b), a current consumption increase of about one decade. Indeed, the DIBL in devices with scaled L_{SG} is a consequence of the insufficient control of the channel potential by the select gate. At high applied source voltages this induces, similarly to the case of the strong inversion described above, a lowering of the electric field that results in an increasing of the consumed current and consequently a decreasing of injection efficiency ($\Delta V_{TH}/I_S$). Therefore in ultra-scaled devices reducing DIBL phenomenon, optimizing for example the junction implantations, is of great importance to control the consumption.

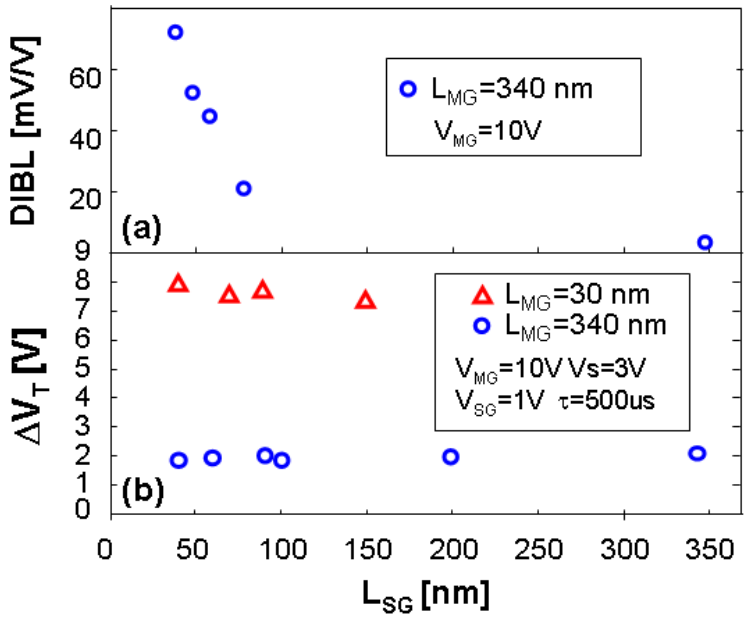


Figure 2.17: (a) DIBL due to select gate scaling. (b) Programming window as a function of the select gate length for devices with a long (L_{MG}=340nm) and short (L_{MG}=40nm) memory gate length.

2.3.3.2 Disturb

The way split gate memories are linked together is shown in fig 2.18 . While writing a cell in a memory array, the neighbour cells can be effected by source disturb.

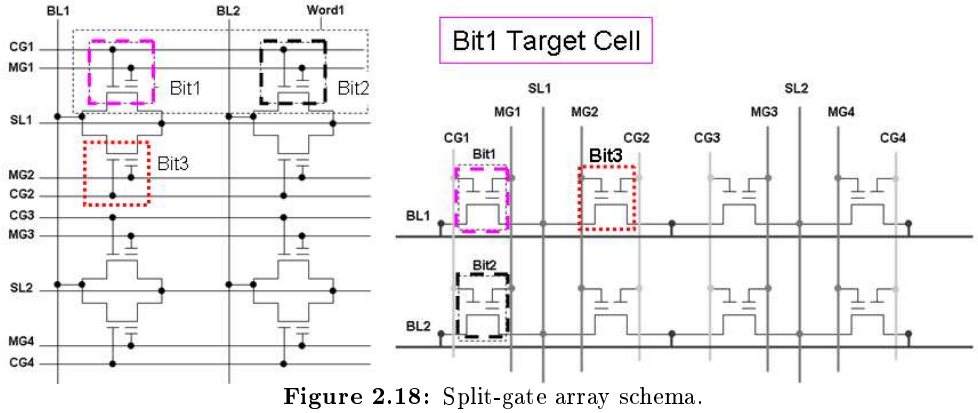


Figure 2.18: Split-gate array schema.

We analysed the sensitivity to source disturb effect for two types of neighbour cells:

Bit3 The cells connected through the bit line to the target cell (Bit1) have the same source voltage. The sensitivity to this source disturb effect is examined by measuring the threshold voltage shift of an isolated split gate cell when a high voltage (V_S=6V) is applied on the source electrode while all the other contacts are kept grounded. The results are illustrated

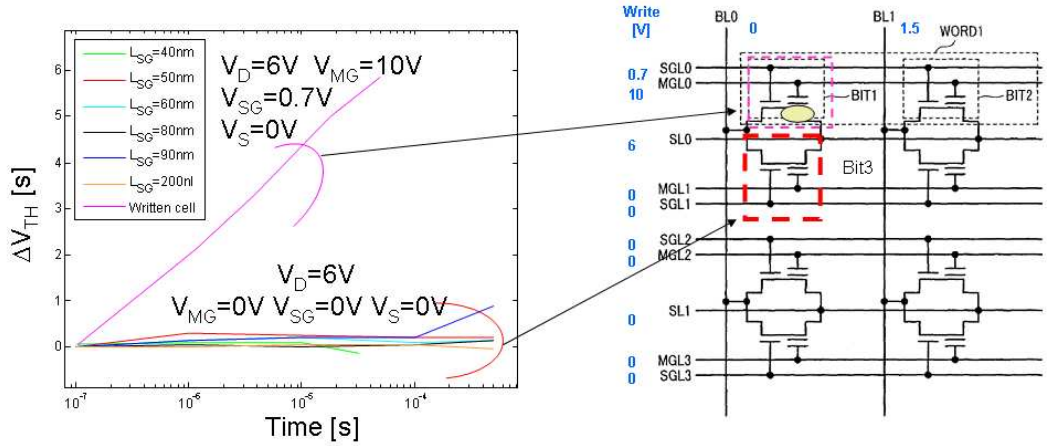


Figure 2.19: Left: programming characteristic of the target cell and threshold voltage shift induce in the neighbour cell (with common source line) for memories with different select gate lengths and memory gate length of 200nm. Right: schematic of the cell array.

in fig. 2.19 where we compared the programming window of the target cell with that of the neighbour memory cell. We can state that disturb phenomenon induces a very small V_{TH} shift in fresh cell also in memories with scaled dimensions.

Bit2 This cell is connected to the target cell through the memory gate, select gate and drain lines. The sensitivity to the disturb effect is examined by measuring the threshold voltage shift of an isolated split gate cell when the two gates and the drain are biased with the same programming condition of the target cell, while the drain electrode is biased with a positive bias in order to prevent inversion charges in correspondence with the select gate.

In a first experiment a bias of 6V was applied on the Source electrode, with V_{MG}=10V, V_D=1.5V, V_{SG}=1V. The high potential induces DIBL phenomenon that affects the capability of the select gate to control the current increasing the disturb (see fig. 2.20).

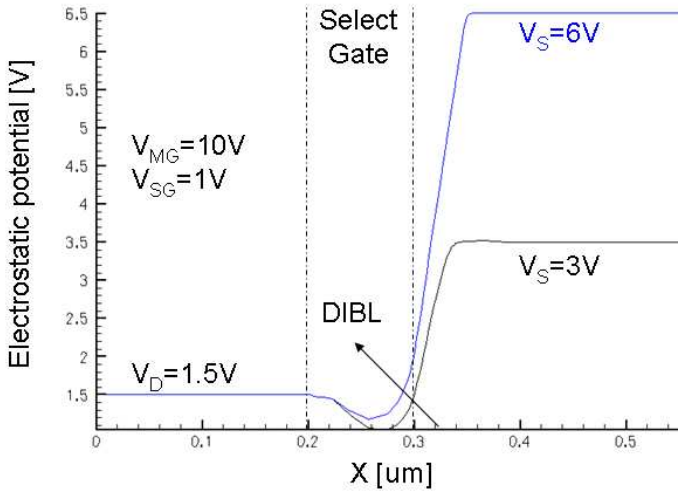


Figure 2.20: Electrostatic potential of a split-gate memory with L_{MG}=20nm, L_{SG}=100nm for V_{MG}=10V, V_{SG}=1V and two source bias conditions: V_S=3V (black line) and V_S=6V (blue line). DIBL effect is shown.

To overcome this issue two ways were exploited:

- increasing the Drain bias,
- decreasing the Source bias.

The first solution was discarded to avoid oxide breakdown due to leakage current through the thin select gate oxide (2.5nm). We thus opted for the second solution performing programming measurement using a Source voltage of 3V that, in scaled memory devices, leads to a programming window up to 9V (see §2.3.4).

Programming measurement with V_{MG}=10V, V_D=1.5V, V_{SG}=1V and V_S=3V are reported in fig.2.21 for a memory gate length of 200nm and various select gate lengths. As summed up in fig. 2.22 the disturb increases with the decreasing of the select gate length. When the select gate length is smaller than 50nm the disturb becomes an important effect that may prevent the select gate scaling.

In order to reduce the disturb, the control of the select gate over the channel must be improved. An easy way is to optimize the drain junction.

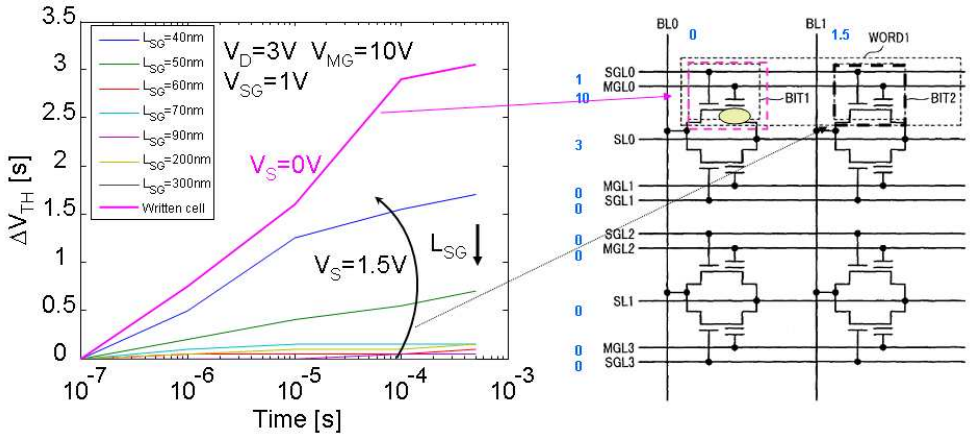


Figure 2.21: Left: programming characteristic of the target cell and threshold voltage shift induced in the neighbour cell (with different Drain line) for memories with different select gate lengths and memory gate length of 200nm. Right: schematic of the cell array.

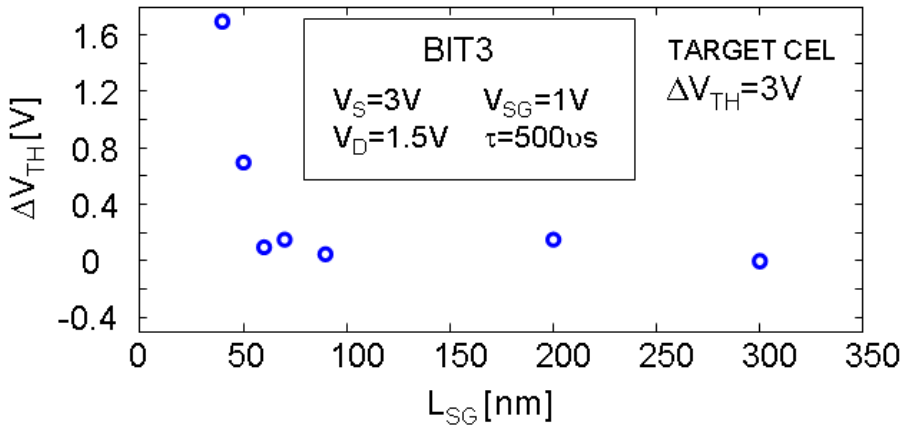


Figure 2.22: Threshold voltage shift due to disturb phenomenon after 500us as a function of the select gate length. The memory gate length is 200nm.

2.3.4 Memory gate scaling

In the previous section we described the effect of scaling the select gate dimension. Here we focus on the memory gate scaling. We will see that the memory gate scaling improves the programming and erasing efficiency without increasing the current consumption. On the other hand the memory gate scaling can introduce a higher ΔV_{TH} variability due to a gate length variation directly linked to a bad control of the lithography alignment.

2.3.4.1 Programming

The impact of the memory gate scaling on the injection efficiency has been investigated by studying the programming characteristics of devices with a 100nm select gate length and a memory gate length from 180nm down to 20nm. Fig.2.24-a shows the programming window after a 500μs program pulse (V_{MG}=10V; V_S=3V; V_{SG}=1V) as a function of the memory gate length. With the shrinking of the memory dimensions the programming window strongly increases from 3V to 9V. This result has been explained by the means of TCAD simulations.

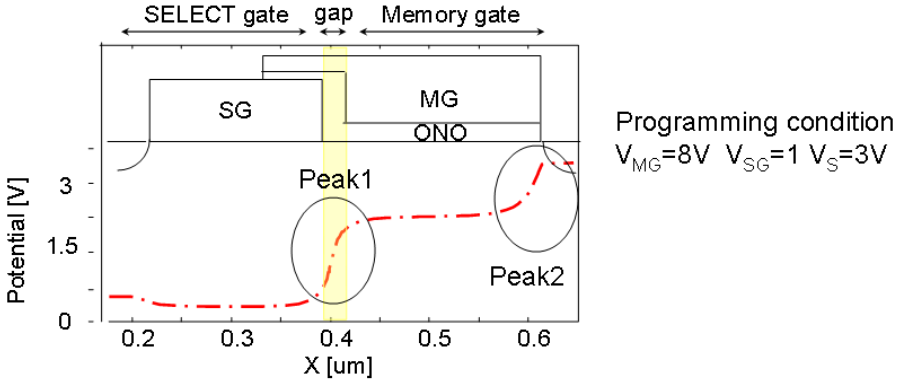


Figure 2.23: Channel potential during SSI operation

In long devices the electric field in the memory channel shows two peaks (Fig.2.24-c), the first one is located in the gap, due to the difference between the memory gate and the select gate potentials; the second peak is created at the channel source junction (see fig. 2.23). As the gate length is further reduced, the two peaks merge and the maximum of the electric field increases, leading to an enhanced injected charge in the nitride layer (Fig.2.24-b).

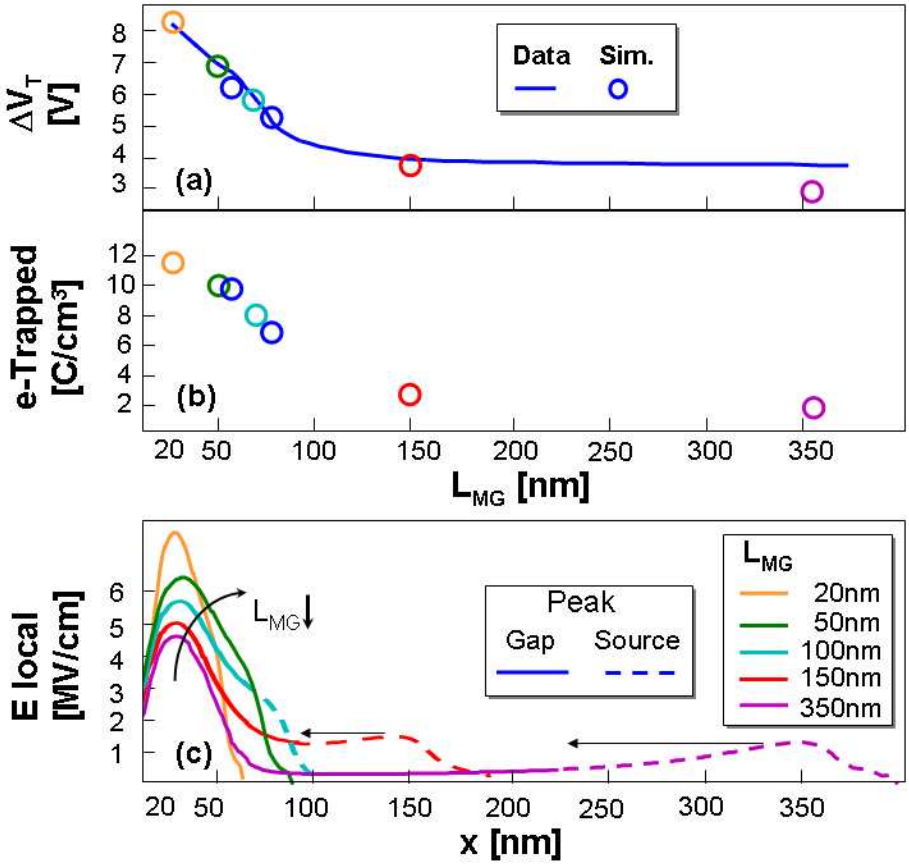


Figure 2.24: a) Measured and simulated programming windows of split-gate SiN memories for various gate lengths (V_{MG}=10V; V_S=3V; V_{SG}=1V, t=500μs). b) Corresponding simulated trapped charge concentration in the SiN layer. c) Simulated local electric field in the channel during Source Side Injection programming. Two peaks appear, one near the gap, the other close to the source electrode. In short devices, the two peaks merge

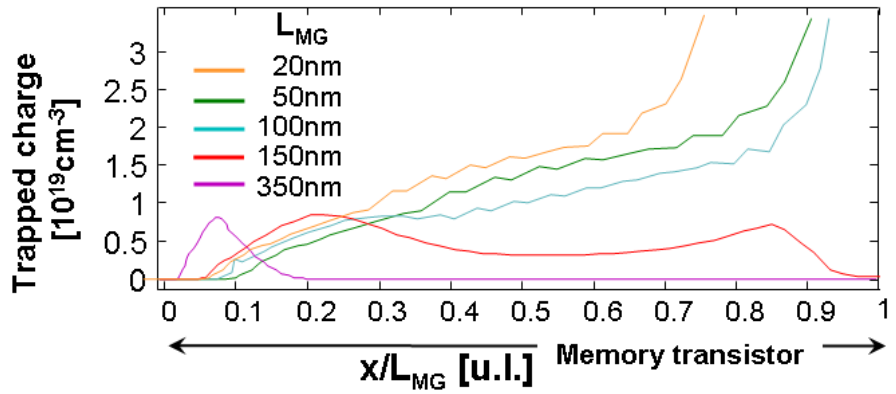


Figure 2.25: Trapped charge location after a programming pulse of 500μs (V_S=3 V_{MG}=10)

2.3.4.2 Erasing

In fig. 2.25 we reported the simulated trapped charge in Si_3N_4 during Source Side Injection programming operation for various gate lengths. Accordingly to what described in the previous section (see Fig. 2.24), in long device the charge is injected near the gap region and the maximum concentration of trapped charge gradually moves to the source electrode in short devices.

During HHI erasing operation, the hot holes are generated at the source junction and then injected in the SiN region near it. This implies that in long devices, as schematically shown in fig. 2.26-Right, occurs a mismatch between the injected holes and the trapped electrons population that prevents a complete erasing of the cell [82, 83, 84, 80]. On the contrary in devices with a small memory gate length, due to the short distance between the gap and the source, the holes during HHI erasing and the electrons during SSI programming are injected in the same location, and thus the cell can be completely erased.

In order to verify this assumption we computed, for memories with different L_{MG}, the percentage of the erased charge after the same series of a programming (V_D=3 V_{MG}=10 V_{SG}=1V t=500μs) and an erasing pulses (V_D=5 V_{MG}=-10 V_{SG}=0V t=500μs). Accordingly to what is described above, the same signal that erase the 100% of the charge in a 30nm memory is insufficient to completely erase a 50nm memory and can not erase a memory with a gate length larger than 150nm. In order to achieve a complete erase of the cell the HHI erasing time and voltages have thus to increase with the increasing of memory gate dimension. When the memory gate length is too large ($\gtrsim$100nm) the HHI becomes useless and the FN erasing method is the only method to erase the cell.

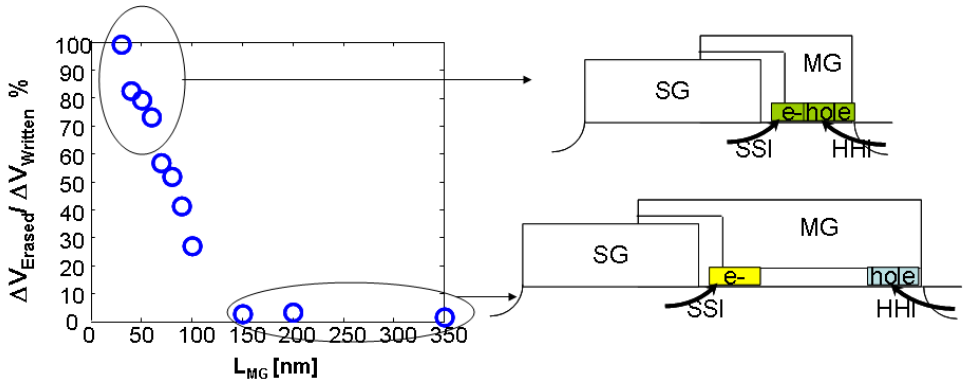


Figure 2.26: Left: Erase efficiency versus memory gate length for given program/erase conditions. Right: Representation of the HHI/SSI mismatch issue

2.3.4.3 Consumption

To analyse the effect of the memory dimensions scaling on programming consumption, we measured for memory gate lengths from 350nm down to 20nm, the programming characteristics and the current consumption as a function of the programming time (Fig. 2.27). The consumed energy is calculated as the integral along the programming time of the channel current times the applied source voltage:

$$ENERGY = \int_0^{time} V_S I_S(t) dt \tag{2.6}$$

In split-gate memories, during programming, the memory transistor is ON, and the channel current is controlled by the access gate voltage. Consequently the programming current I_S for a given V_{SG} is constant and for a given program V_S, the energy only depends on programming time (see Eq. 2.6 and Fig.2.27-b). In scaled memory gate devices, as the programming efficiency is higher, shorter programming times are sufficient to reach a given ΔV_{TH} (Fig.2.27-a), and a lower programming energy is achieved (Fig.2.27-b).

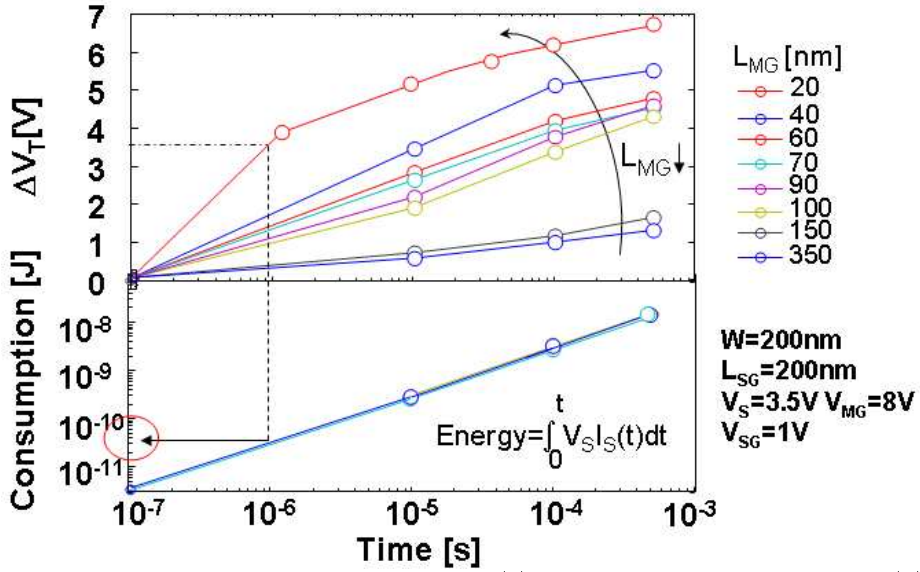


Figure 2.27: Measured programming characteristics (a) and measured consumed current (b) as a function of the programming time for various devices with different memory gate lengths

In Fig.2.28 we plotted the channel current consumed to reach a given programming window of 3.5V as a function of the memory gate length. The required programming time is extrapolated from Fig.2.27-a. The result shows an improvement of over 10 times of energy consumption when the memory gate length scales from 100nm to 40nm. In particular, for sub-40nm gate length devices, $<$0.1nJ of programming energy is reached, suitable for low power applications.

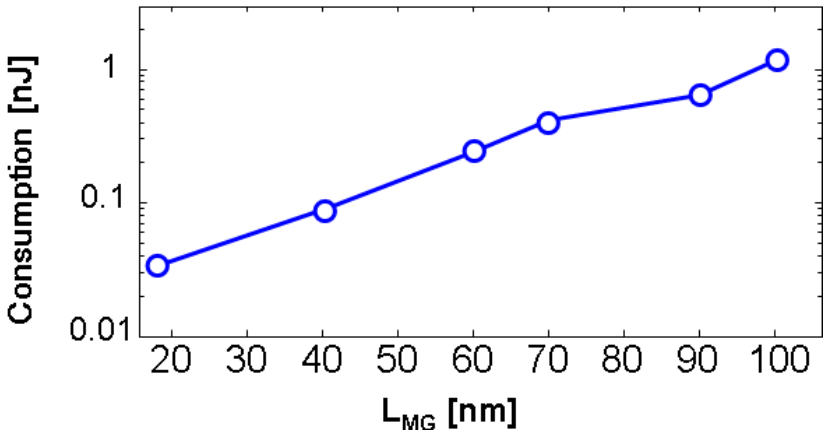


Figure 2.28: Measured current consumption during a programming pulse with V_S=3.5V; V_{MG}=8V to reach a programming window of 3.5V as a function of the memory gate length.

2.3.4.4 Variability

In fig 2.29-left we show the programming window as a function of the nominal memory gate length. In the last years the scaling of the device dimensions has been much more aggressive than the improvements in the resolution of the lithography-process. The placement of the image with respect to underlying layers needs to be accurate in all locations on each integrated circuit to achieve adequate precision. In particular in split-gate memories with multi-litho architecture, the misalignment between the select gate and the memory gate causes the variation of the memory dimensions that induces a non-negligible programming window shift. The variation of the programming window ($\sigma_{\Delta V_{TH}}$) due to a variation of the memory gate length ($\sigma_{L_{MG}}$) can be written, in a first approximation, as:

$$\Delta V_{TH} = f(L_{MG}) - - > \sigma_{\Delta V_{TH}} = \frac{\partial f(L_{MG})}{\partial L_{MG}} \sigma_{L_{MG}} \tag{2.7}$$

To quantify this variation, we first found the coefficients a, b, c, d, e of a fourth-order polynomial that reproduces the experimental data $\Delta V_{TH} = aL_{MG}^4 + bL_{MG}^3 + cL_{MG}^2 + dL_{MG} + e$, then we derived it in order to compute the threshold voltage shift due to a memory gate length variation:

$$\frac{\sigma_{\Delta V_{TH}}}{\sigma_{L_{MG}}} = 4aL_{MG}^3 + 3bL_{MG}^2 + 2cL_{MG}^1 + d \tag{2.8}$$

Fig. 2.29 shows that depending of the lithography process and the chip variability constrains, the choice of the memory gate dimension has to be done carefully: for instance to guarantee a ΔV_{TH} variability lower than 5%, the memory gate length must be

- larger than 110nm with lithography overlay of ±10nm typical of Dry deep ultraviolet (DUV 193nm) lithography.
- larger than 80nm with lithography overlay of ±7nm typical of Immersion DUV 193nm.
- larger than 40nm with ±5nm (state of the art DUV 193nm).

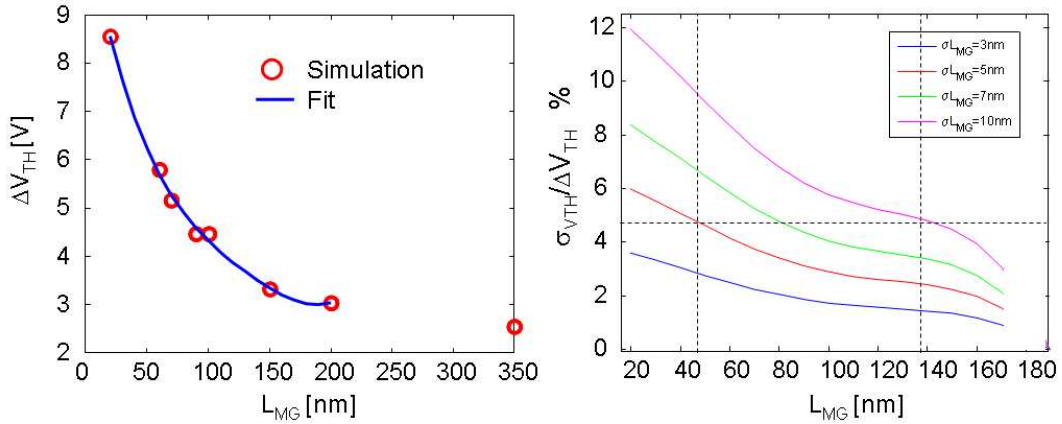


Figure 2.29: Left: programming windows versus memory gate lenght simulations and its fitting. Right: Variability due to L_{MG} variation for different lithography process

- with a overlay of ±3nm (Extreme UV) the variation of the threshold voltage is always lower than 5%.

The results presented above put in evidence the misalignment issue: the overlay variability of the memory gate over the select gate, limits the memory scaling. To overcome this problem a self-aligned solution must be introduced as it is described in section §2.5.

2.4 Study of the trapped charge location

In order to analyse the trapped charge location, V_{TH} shift after a 500μs program pulse (V_{MG}=10V V_S=3V) was measured in forward (V_{DS}>0) and reverse (V_{DS}<0) modes. Fig.2.30-up shows the memory window and its TCAD fitting in both reverse and forward mode. Moreover, Fig.2.30-down shows that the difference between forward and reverse V_{TH} is larger for scaled devices.

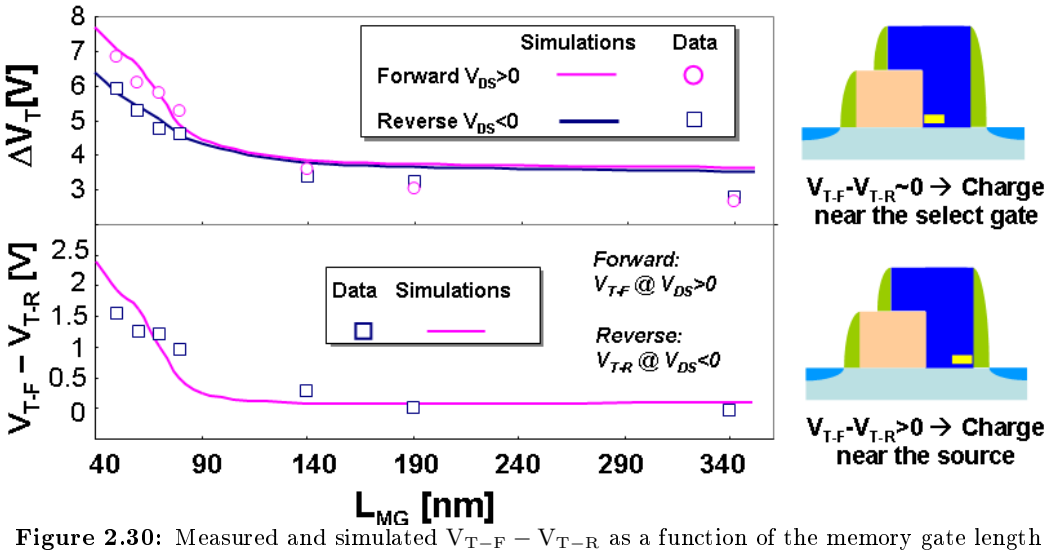


Figure 2.30: Measured and simulated $V_{T-F} - V_{T-R}$ as a function of the memory gate length

This behaviour can be explained by means of TCAD simulation.

Fig.2.31 shows the simulated parallel electric field in the channel during SSI operation. We can see that two peaks of electric field appear: one close to the select gate, the other near the source electrode. For large gate length devices (down to 150nm), the injection point is located on the side of the select gate. As the gate length is further reduced, the two peaks merge (see pag. 52) and the maximum electric field peak is shifted toward the source electrode, in agreement with [82].

Fig. 2.32 shows the channel potential during reading operation. When the electrons are injected close to the source, the trapped charge is partly screened leading to positive $V_{T-F} - V_{T-R}$ value [85]. On the contrary, $V_{T-F} - V_{T-R} \backsim 0$ when the charge is injected near the select transistor. Based on these considerations, we analyzed the retention characteristics.

Fig. 2.33-left demonstrates that even after 10^5s, short devices still exhibit a larger ΔV_T and in forward mode the threshold voltage is nearly constant during retention indicating a small charge loss. On the contrary fig.2.33-right shows that the $V_{T-F} - V_{T-R}$ evolves during time. Indeed, the measured time evolution of $V_{T-F} - V_{T-R}$ shows two behaviours: in long devices, due to presence of pocket of charge close to the select transistor, V_{T-F} remains similar to V_{T-R} during time. For short devices the charge initially close to the source diffuses toward the select gate and $V_{T-F} - V_{T-R}$ decreases.

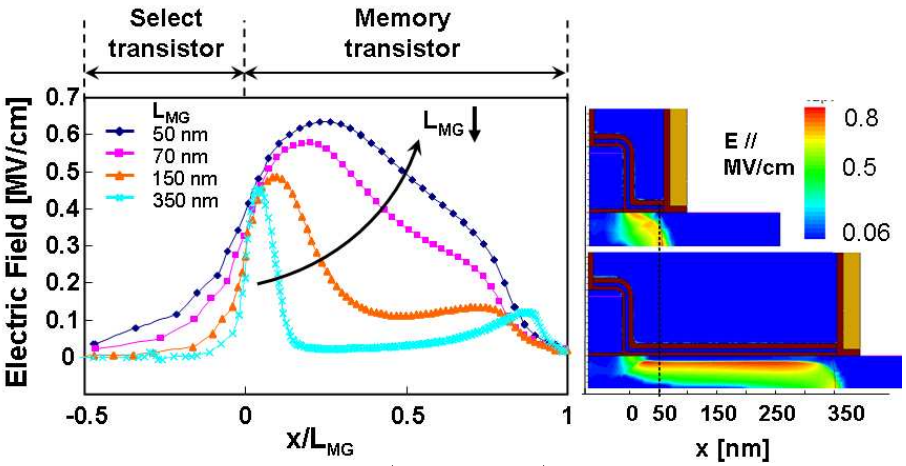


Figure 2.31: Simulated parallel electric field (Fiegna model) during Source Side Injection programming operation for various gate lengths

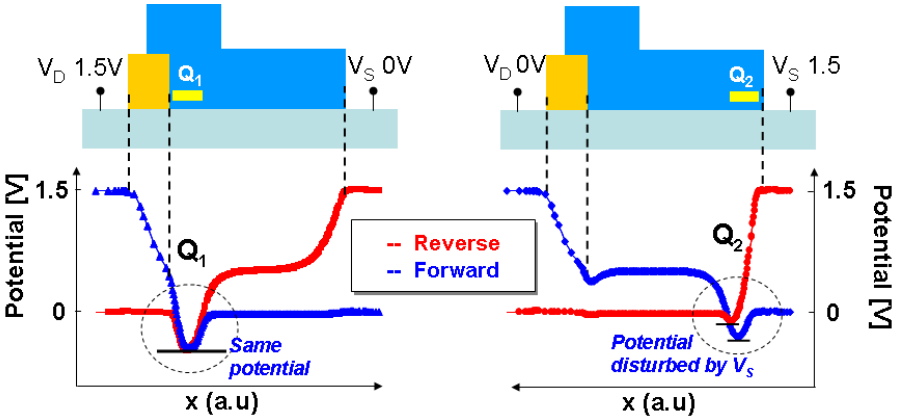


Figure 2.32: Simulated $V_{T-F} - V_{T-R}$ assuming the spread of a pocket of charge located close to the select gate (Q1) for a long device or to the source electrode (Q2) for a shorter device.

TCAD simulations were performed to confirm the experimental results. The idea was to compute the threshold voltage shift when a pocket of charge initially located at the source or select gate side, spreads along the channel.
In a first attempt, a charge density of 10^{13}cm^{-2} closed to the select gate and covering 35nm was added to a 350nm L_{MG} memory. Similar simulations were repeated considering the same charge spread over different intervals from 35nm up to 350nm. With the same method, a charge with density of 10^{13}cm^{-2} covering 7nm and located close to the source was added in a 70nm memory and spread over the channel with a step of 7nm.

The simulated threshold voltages are shown in (Fig. 2.34). In agreement to what described above, in short devices the difference between forward and reverse V_T decreases as the charge moves toward the select gate, on the contrary in longer devices the difference between forward-reverse remains $\backsim 0$ even if the charge, initially located close to the select gate, moves toward the source.

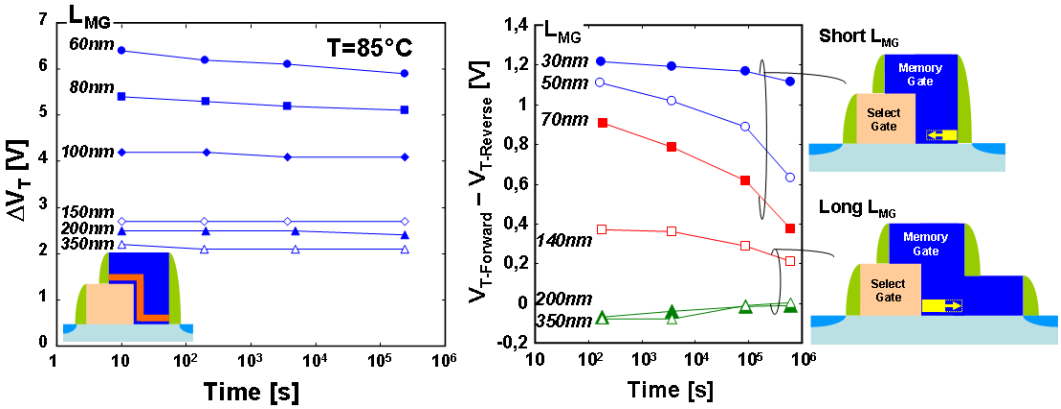


Figure 2.33: Left: Impact of the memory gate length on the retention characteristics at 85°C Right:Time evolution at RT of $V_{T-F} - V_{T-R}$ during retention for various memory gate lengths. $V_{T-F} - V_{T-R}$ gives indications on the trapped charge location. V_{TH} is measured at V_DS=1.5V

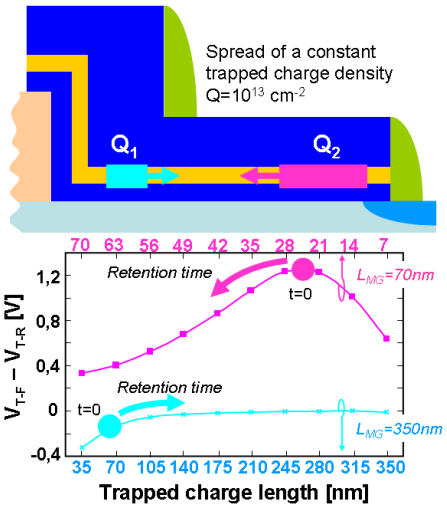


Figure 2.34: Simulated $V_{T-F} - V_{T-R}$ assuming the spread of a pocket of charge located close to the select gate for a long device or to the source electrode for a shorter device

2.5 Multi-litho SG-CTM evolution: The Spacer technology.

The multi-litho solution presented in the first part of this chapter appears very promising for high-speed and low consumption embedded memories. However, the misalignment of the two gates can hinder the scaling of this architecture (see pag. 55). Indeed, an insufficient control of the position of the memory gate leads to a variation of the electrical gate length of the memory transistor, inducing variability of the electrical performances of the device. To solve this problem, self aligned, spacer, solutions can be proposed.

In the spacer approach the memory gate is defined by a spacer technology, on the edge of the access transistor (Fig. 2.35-Left). Therefore the memory gate patterning requires only a non-critical lithographic step, needed to remove the spacer on the select gate side, making this solution particularly adapted to scaled technologies.

First of all in this section spacer split-gate charge trap memories are introduced. Then, the integration scheme and the influence of process parameters on memory characteristic are presented. Finally electrical results on silicon nitride (Si_3N_4) based-memory are shown and compared with the multi-litho technology.

2.5.1 Spacer presentation

A silicon nitride layer (CTL) was integrated in a spacer structure (fig. 2.35).

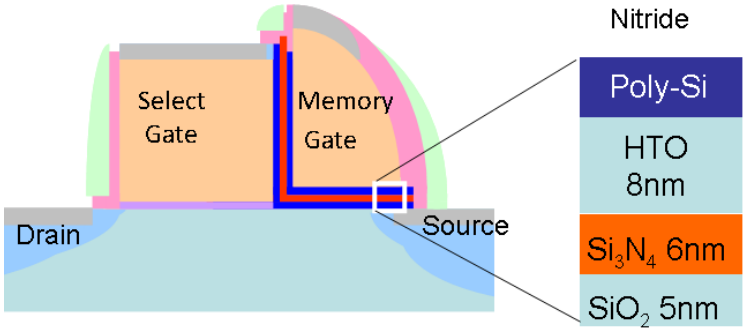


Figure 2.35: Left: Schematic cross section of the spacer split-gate memory. Right: representation of the memory gate stack.

2.5.2 Process simulation

Process simulations were realized with the software Sentaurus Process (release 2010.12), module of Synopsys in order to optimize the process parameters. In this section, after presenting the spacer split-gate memory fabrication process, we will analyse the influence of the implant conditions and the spacer shape on the memory performances.
The the main front-end steps of split gate spacer process are described below:

1. Shallow Trench isolation (STI) on silicon p-doped substrate
2. Boron well implantation and annealing
3. Select Gate stack deposition (SiO_2/polysilicon/HTO)
4. Nitride hard mask deposition
5. Select Gate lithography, etching and stripping
    - See fig. 2.36

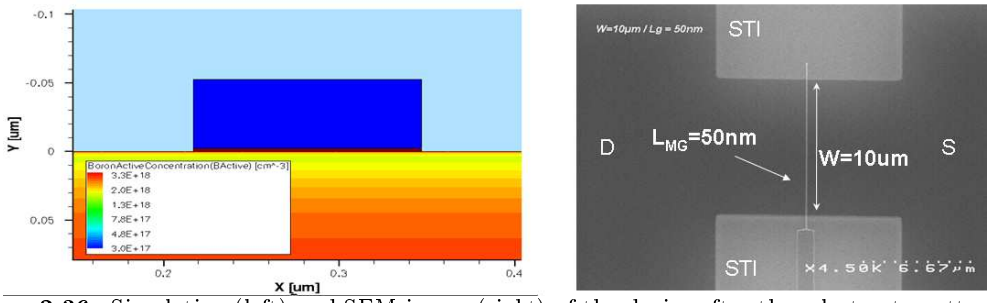


Figure 2.36: Simulation (left) and SEM image (right) of the device after the select gate patterning.

6. Memory gate stack depositions ONO/polysilicon
7. Memory spacers etching and stripping
    - See fig. 2.37

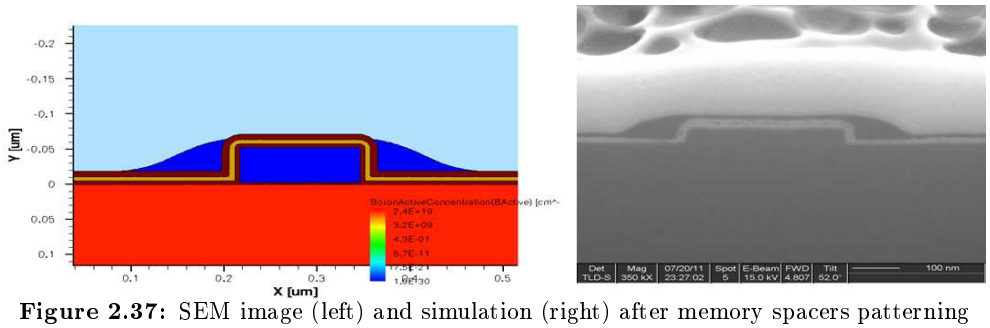


Figure 2.37: SEM image (left) and simulation (right) after memory spacers patterning

8. Hard mask nitride deposition
9. Select gate spacer removal: Lithography etching and stripping
10. Lightly Doped Drain (LDD) implantation and annealing
    - See fig. 2.38
11. Nitride Spacer deposition and etching
12. Highly Doped Drain (HDD) source/drain implantation and annealing

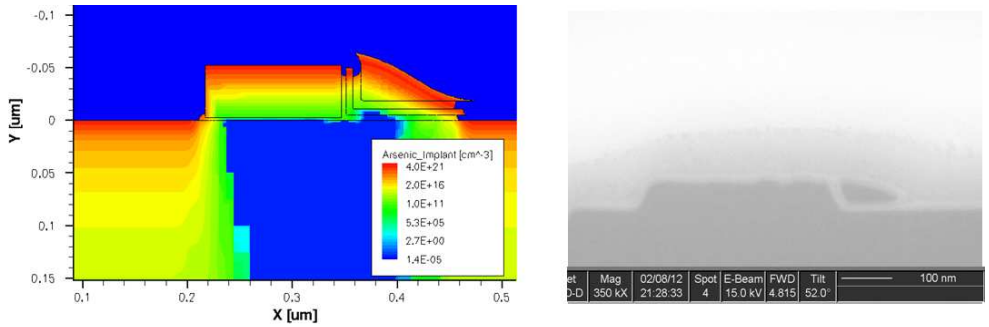


Figure 2.38: Left: Simulation after MG stripping step showing LDD doping concentration. Right: SEM image of the device after select gate spacer removal

13. silicidation, deposition of the passivation oxide

    - See fig. 2.39

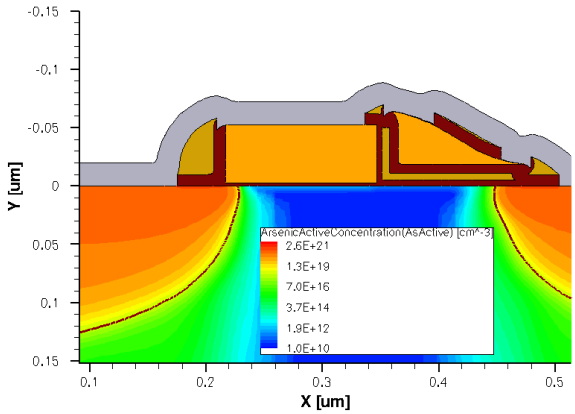


Figure 2.39: Results of the spacer SG process simulation.

14. Back end of line (BEOL)

The process steps listed above will be use to study the optimization of the source drain implantation energy.

2.5.2.1 Optimization of the Source/Drain implantation energy

In order to optimize the cell characteristic the doping impact on memory performances has been evaluated. The simulated process flow corresponds to the Spacer architecture process with a memory gate length of 60nm. The select gate thickness is set to 50nm what is the worst case in terms of channel counter-doping.

In order to find an energy that prevents to have counter-doping on the channel, the process simulations were done introducing a split on the energy HDD implantation with the following values: 30, 25, 20, 15 and 10 keV and keeping constant the others parameters of the HDD SD

implantation like the dose set at $2x10^{15}cm^{-2}$. The implantations are simulated by the analytical model of SProcess following the steps described above.

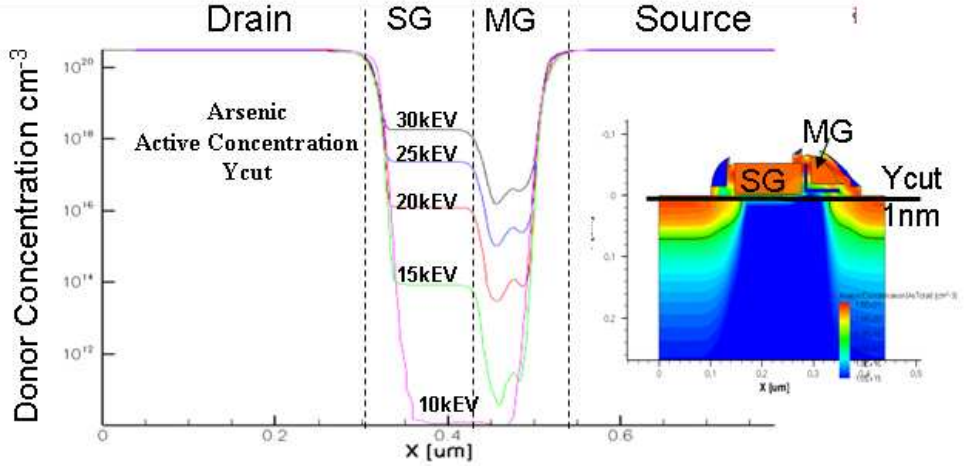


Figure 2.40: Profile of active arsenic concentration along an horizontal line situated 1nm under the surface of the channel.

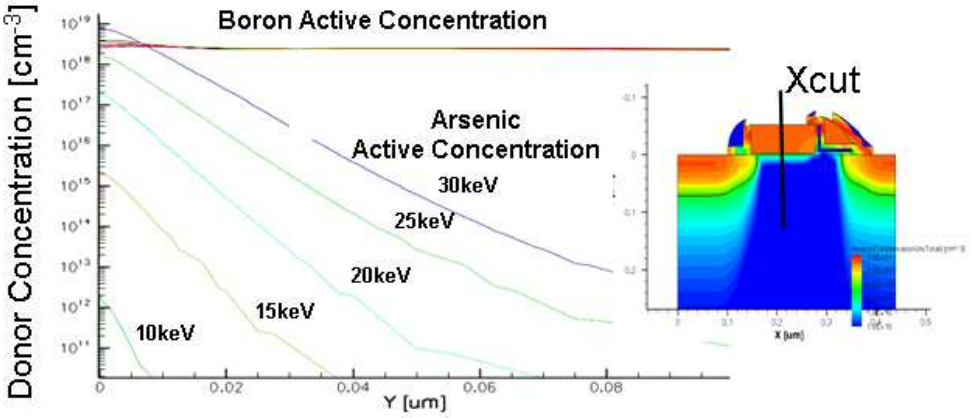


Figure 2.41: Profile of active arsenic and active boron concentration along a line situated at the middle of the select gate.

The results presented here correspond to the profiles at the end of the whole simulated process (LDD implantation, HDD S/D implantation and a spike annealing at 1050 ℃ for the S/D activation). The results indicate that in the case of high energies the polysilicon gate of 50 nm does not completely screen the implantation and the resulting concentration of Arsenic in the channel corresponds to counter-doping.

Figure 2.42 illustrates that the counter-doping in the channel corresponds to barrier lowering and modifies the electrical characteristics of the device (see 2.42-inset). This study has allowed us to choose the best HDD implantation energy: the energy of 10keV for the HDD SD implantation has been retained.

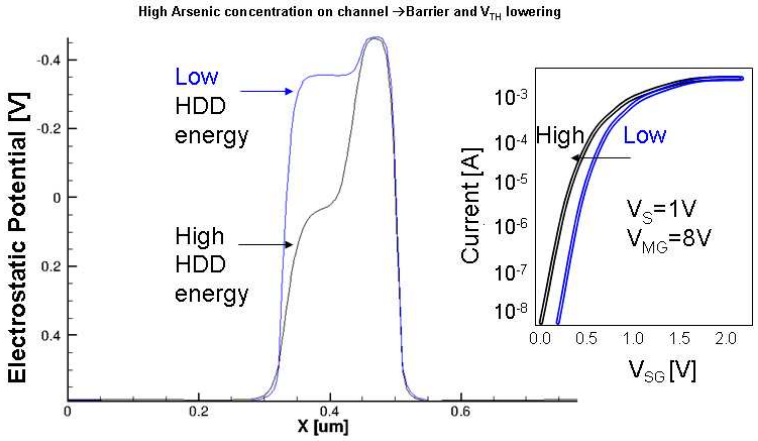


Figure 2.42: Electrostatic potential along an horizontal line situated 1nm under the surface of the channel in the case of a high (black line) and low (blue line) HDD implantation energy. Inset: the corresponding electrostatic characteristics.

2.5.2.2 Spacer shape

Because of the process complexity the final memory spacer geometry remains uncertain. Nevertheless a final geometry intermediate between the three schematic geometries, triangular, square and natural, is expected. In consequence, the influence of the memory gate geometry on arsenic profiles obtained after LDD and HDD implantation has been studied.

Three different devices have been simulated with Sprocess; they only differ by the geometry of the memory gate. The following parameters are used for the implantations :

- **LDD** implant: Arsenic dose = $5{\cdot}10^{14}$; energy = 5 keV
- **HHD** implant: Arsenic dose = $2{\cdot}10^{15}$; energy = 10 keV

The profile of arsenic active concentration 1nm under the surface of the channel plotted in fig. 2.44 confirms that the arsenic profile after implantation and diffusion depends on the gate geometry. The triangular geometry corresponds to the case where the arsenic doping penetrates more in the channel. This case is the less favourable as the hardly controlled arsenic concentration in the channel may effects electrical behaviour of the memory. The desired configuration results to be the square geometry that is technologically difficult to achieve without any additional lithography step.

2.5.3 Electrical results

In this section we will analyze the first results on the spacer technology processed at CEA/LETI. We first showed the functionality of the SiN spacer memory. Then we compared the programming characteristic with what found for the analogue SiN memory built with the multi-litho solution.

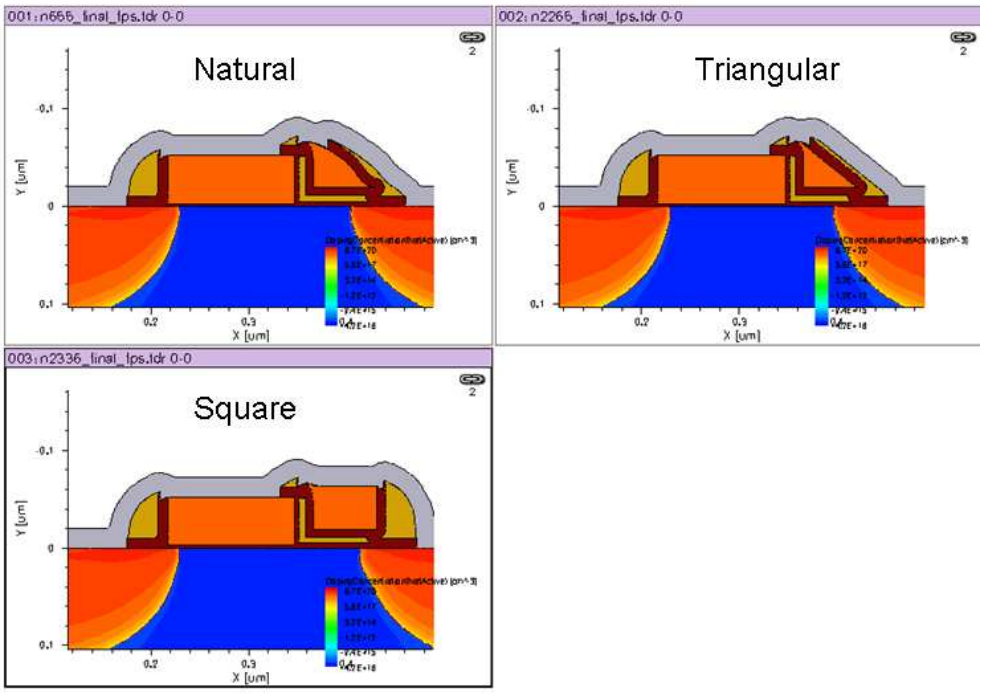


Figure 2.43: Spacer architecture with triangular-natural and square shape

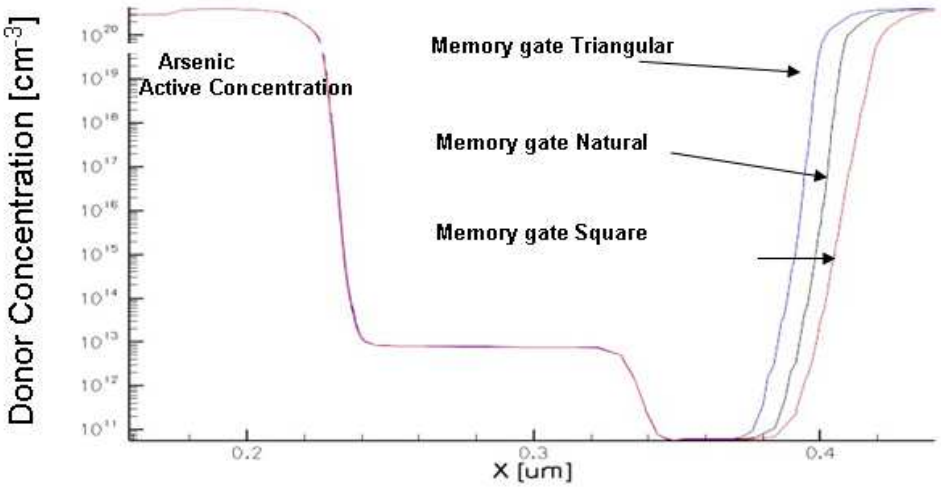


Figure 2.44: Arsenic active concentration potential along a horizontal line situated 1nm under the surface of the channel

Differently from the self aligned architecture where the memory gate length (L_{MG}) is defined by a lithography step, in spacer architecture, the L_{MG} depends on the thickness of the deposited memory gate poly-silicon. Consequently the available memory gate dimensions are limited to one for each stack. The tested memory is processed with select gate lengths (L_{SG}) from 5um down to 40nm (limit of e-beam lithography) and memory gate length (L_{MG}) of 100nm.

The transfer characteristic of the device are shown in fig. 2.45 where one can see that the

current can be efficiently controlled biasing the select gate. During programming operation the select gate bias has been set to 1V. We considered that, similarly to what described in section 2.1, biasing the select gate just above its threshold voltage is the best compromise between a low consumption and a high programming window.

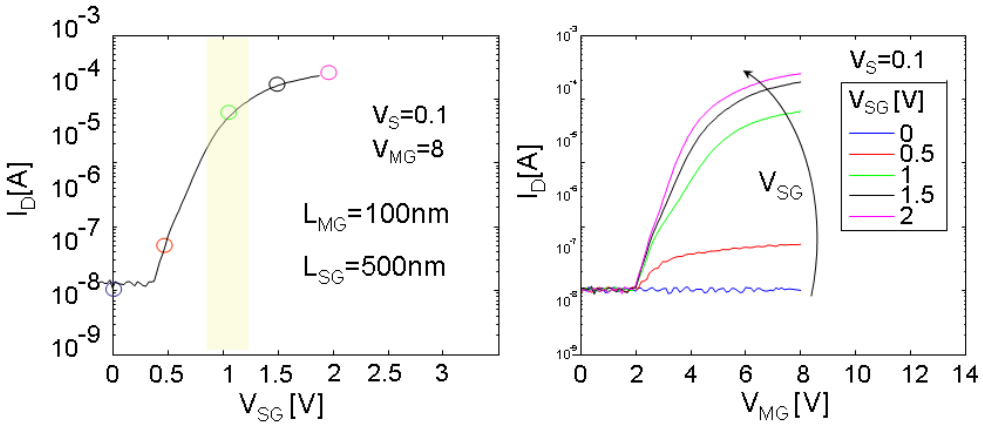


Figure 2.45: $I_D(V_{SG}, V_{MG})$ and $I_D(V_{MG}, V_{SG})$ of a SIN split-gate spacer memory with L_{MG}=100nm and L_{SG}=500nm.

Programming The split-gate spacer memories are programmed using Source Side Injection, biasing both the memory gate and the source electrode at high voltages. Fig.2.46 shows the program characteristics of Si_3N_4, for V_{MG}=10V V_{SG}=1V and various programming V_S. The memory gate length is 100nm and the L_{SG} is 750nm. The memory shows a promising behaviour: after some microseconds, a programming window of 4V is achieved applying on the source electrode a bias of 4.5V.

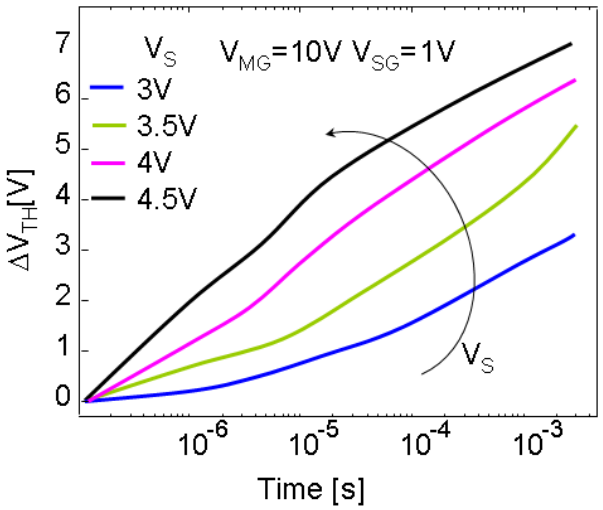


Figure 2.46: Program characteristics in Source Side Injection mode of Si_3N_4 memory for various programming V_S. The ΔV_{TH} has been averaged over 35 devices.

To analyse the influence of the select gate length on the programming performances we performed programming measurements on more than 30 devices with two select gate lengths: 750nm and 5μm. The programming windows for a pulse of 500μs (V_{MG}=10V, V_S=3V) are

reported in Fig. 2.47-Left. The results show that there is no substantial difference in the programming window distribution between the two memories under test, confirming that the injection efficiency, at these relaxed dimensions, is not influenced by the select gate length in agreement to what described for the multi-litho approach (see pag. 47).

Finally the spacer programming window induced by a 500us pulse with V_S=3V and V_{MG}=10V has been compared with what obtained for the multi-litho architecture with the same programming condition. Figure 2.47 shows the simulated multi-litho programming window as a function of the memory gate length together with the average programming window measured for a spacer memory with memory gate length 100 nm. We can see that in spacer memory the programming window is always lower than that of the multi-litho solution. Indeed, the spacer technology the S/D-channel junctions are less abrupt than in multi-litho memory leading a lowering of the electric field and thus of the programming efficiency [71].

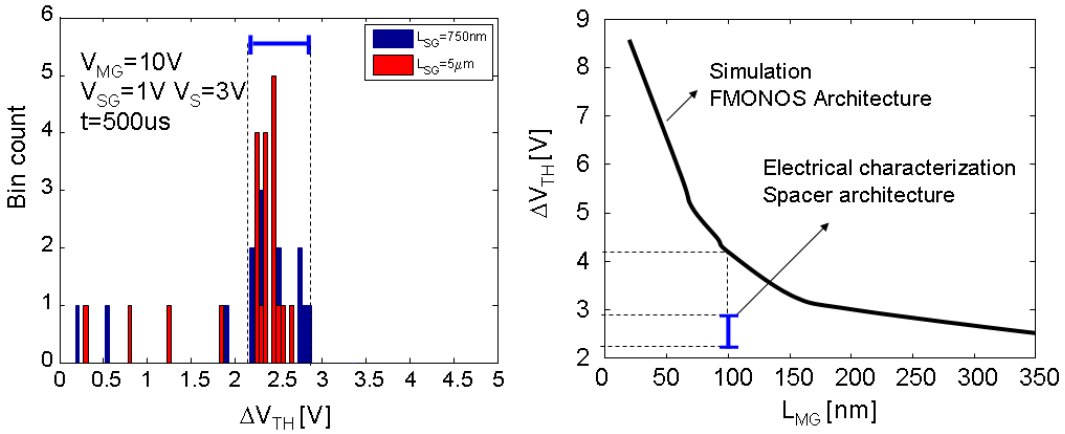


Figure 2.47: Program characteristics in Source Side Injection mode of Si_3N_4 spcar memory with ∼100nm memory electrical gate length, 750nm select gate length, for various programming V_S.

Erasing SiN memory was erased by hot hole injection biasing the source electrode at 8V and the memory gate with -8V, -10V and -12V. Fig.2.48 presents the corresponding erasing characteristics. The erase efficiency at -8V is very low and after 1ms only a slight reduction of the initial threshold voltage appears. Increasing the memory gate bias the erasing is improved but, due to the HTO used as blocking layer, when the memory gate bias is higher then -12V, the back tunneling effect becomes stronger than the SiN de-trapping, preventing the erasing of the cell. To overcome this issue a possibility is to maintain a low memory gate voltage and increasing the Source bias (see fig.2.49).
As a matter of fact, with a source bias of 9V (V_{MG}=-8) the memory can be efficiently erased but a high source voltage results in a higher programming consumption and a higher disturb effect.

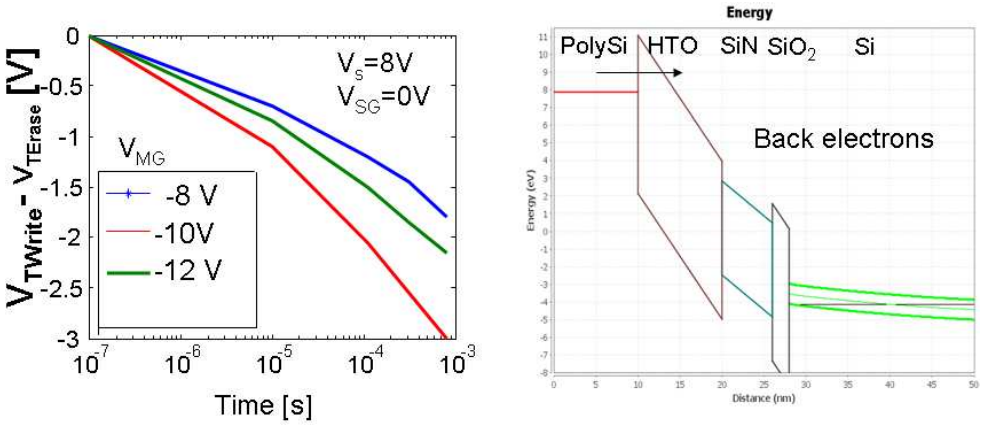


Figure 2.48: Left:SiN spacer memory erase characteristics for V_S=8V, V_{SG}=0V and various V_{MG}. Right: Band diagram of the spacer memory gate stack ($Si/SiO_2/SiN/HTO/Poly\text{-}Si$) illustrating the back tunnelling effect during erasing mechanism (V_{MG}=-12V).

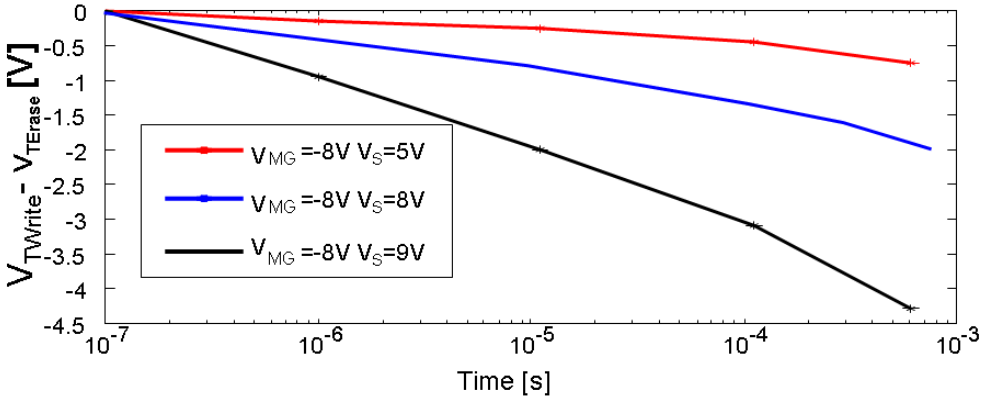


Figure 2.49: Erasing characteristics for V_{MG}=-8V, V_{SG}=0V, and various V_S

2.6 Conclusion

Charge trap memories with Si-ncs or Si_3N_4 charge trapping layers and multi-litho architecture were processed for the first time with electrical memory gate lengths down to 20nm.

In the first section of this chapter we studied the impact of the charge trapping layer on memory performances:

- Nitride memories show a higher ΔV_{TH} during HCI programming than Si-nc. Hybrid Si-nc/SiN memories, allow to improve Si-nc windows
- Si-nc memories can be erased by FN allowing 0 current consumption
- Si_3N_4 exhibits higher memory window and better retention up to 85°C, Si-nc presents better retention at 150°C.

In the next section we analysed the effect of scaling the memory dimensions. Experiments on ultra-scaled memories coupled to static and dynamic TCAD simulations allowed to study the physical mechanisms during Source Side Injection programming. In particular, we showed that scaling the dimension of the **select gate** can induce:

- lower injection efficiency (higher current in smaller devices)
- lower disturb immunity (L_{SG}<50nm bad control of the channel current)

while the scaling of the **memory gate** is favourable in terms of :

- programming window (up to 9V);
- programming energy (down to 0.1nJ);
- erasing efficiency (see §2.3.4.2),

but it can introduce

- ΔV_{TH} variability (up to 6%) due to the gates misalignment.

The study on multi-litho architecture was concluded with the analysis of the trap charge location after a SSI programming pulse and its time evolution during retention. We demonstrated that in short devices the charge is injected close to the source electrode while in larger device the charge is injected close to the select gate. Moreover we showed that in short devices the charge that is initially injected in the Source side, during retention, diffuses toward the select gate.

In the last section we introduced spacer technology as a possible evolution to multi-litho split gate memory. We presented the device, the process steps and we compared the first electrical

results with the that of the multi-litho approach. We concluded that spacer technology is a promising candidate to replace the multi-litho approach if a particular attention is done in controlling the fabrication process and the implantations.

Chapter 3

Role of alumina on TANOS memory

In the previous chapter we presented nitride CT split-gate devices integrating alumina in the memory gate stack. Split-gate memories employing Al_2O_3 as control dielectric showed a better erasing and programming efficiency but a faster charge decay during retention probably related to some trap assisted tunnelling. In this chapter, after introducing TANOS ($TaN/Al_2O_3/Si_3N_4/SiO_2/Si$) memory, we will use atomistic calculations to find potential defects that could induce electronic levels inside the band gap of alumina. We will link them with the electrical characterization of TANOS memory and physical-chemical material analysis on alumina single layers.

3.1 TANOS memory

Maintaining a high coupling ratio while reducing parasitic coupling between adjacent cells poses a great scaling challenge. The most promising alternative to floating gate NAND Flash memory, is the integration of a discrete charge trapping layer instead of poly-silicon floating gate (see tab. 3.1).

Among all the possible stacks employing a CT layer, the TANOS ($TaN/Al_2O_3/Si_3N_4/SiO_2/Si$) structure, proposed for the first time by Samsung in 2005 [35], seems to be the most suitable solution for high density NAND memories for stand-alone applications. It is worth to notice that CT memories, developed for the conventional linear technology, are integrable in split-gate charge memories (see pag. 37) and in vertically stacked 3D Flash memory architectures [86, 87, 88]. In particular in [89] we can find a successful example of three dimensionally stacked NAND Flash memory cell strings with 63nm dimension and a TANOS structure.

In TANOS memory the **nitride** charge trapping layer, allows a reduction of the tunnel oxide and a higher reliability to oxide defects. Moreover the employment of an **high-k** as control dielectric permits the integration of a thicker tunnel oxide to increase the retention (∼2.5nm in a classical SONOS memory versus 3∼4nm in a TANOS). Finally, the **TaN** control gate, with a work function of 4.8eV, reduces the back-tunnelling current and thus the erase saturation phenomenon.

Year of production	2011	2012	2013	2014	2015	2016	2017
technology node F							
Planar (2D) Flash NAND $[nm]$	22	20	18	17	15	14	13
Flash NOR (SONOS/NROM)$[nm]$	65	65	45	45	38	38	38
1 — Flash NAND – charge trapping							
Endurance [cycles]	10^4						$5 \cdot 10^3$
Max number of bits per cell	3	4					
Tunnel oxide thickness $[nm]$	$6-7$					$5-6$	
Interpoly dielectric material	ONO						*hk*
Interpoly dielectric thickness $[nm]$	10-13	11		10		9	
Control gate material	n-Poly				n-Poly/metal		metal
2 — Flash NOR – SONOS/NROM							
Endurance [cycles]	10^5		10^5				
Max number of bits per cell $[nm]$	2						
Tunnel oxide thickness $[nm]$	$4-5$						
Charge trapping layer thickness $[nm]$	$5-7$						
Blocking (top) dielectric thickness $[nm]$	$7-9$						
Highest W/E voltage $[V]$	$7-9$						

Table 3.1: 2011's Summary of Non-Volatile Memory Technology Requirements [90].

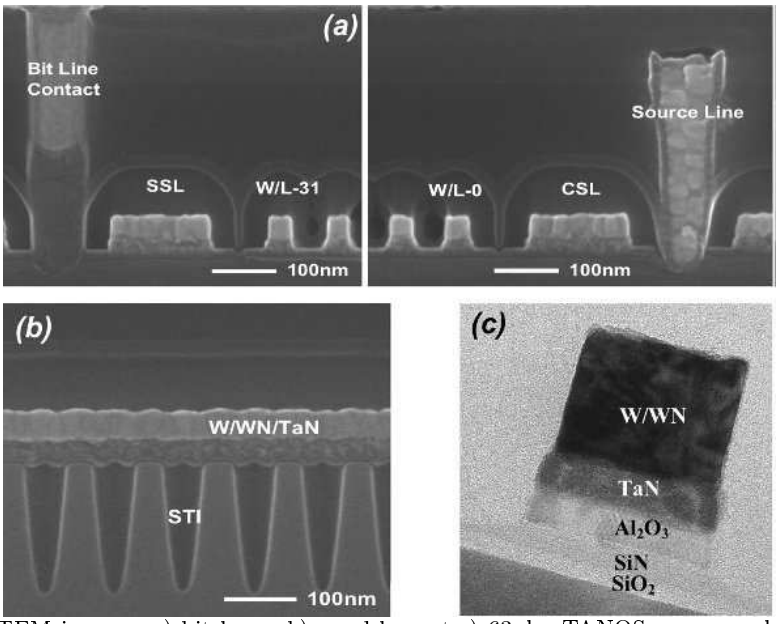


Figure 3.1: TEM images: a) bit lane , b) word lane et c) 63nlm TANOS memory planar view [35].

Figure 3.1 shows the TEM images of a Samsung TANOS memory matrix for a technology node of 63nm. We can see the TANOS stack, together with the 32 memory cells on the bit lane in series with the select transistor that is typical of the NAND matrix.

TANOS memories have shown interesting electrical characteristic. The graph 3.2 presents a ΔV_T of 6V with programming and erasing condition of respectively 17V/100μs and -19V/10ms [91]. Moreover, for the 63nm technology, no interference between neighbour cells has been observed.

Table 3.2 summarizes the differences between the continuous floating gate memories and the discrete charge trapping memories. We can see that the TANOS device seems more favourable for the next generation of scaled memories. This because TANOS memories are more resistant to cell to cell coupling and SILC effect. Moreover, the low dispersion of their V_T, makes them

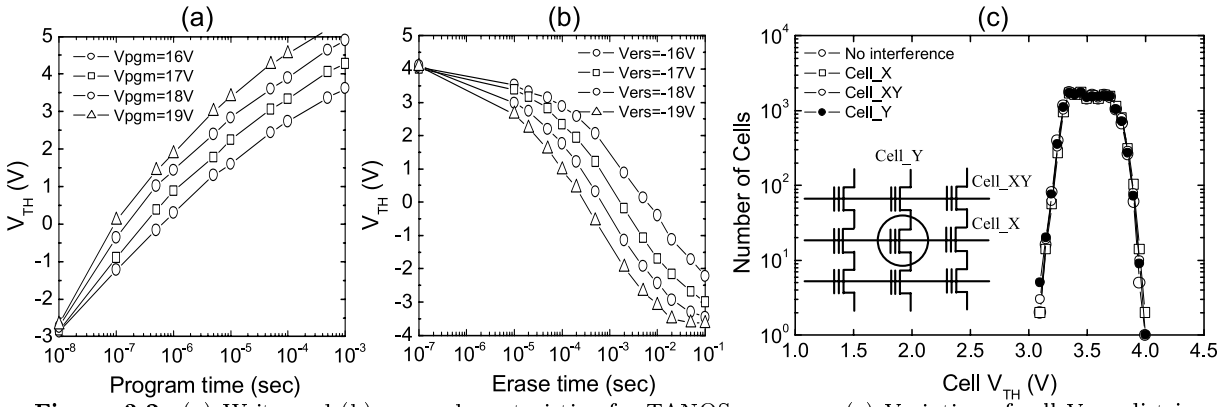


Figure 3.2: (a) Write and (b) erase characteristic of a TANOS memory. (c) Variation of cell V_{TH} distribution as adjacent cells are programmed [91].

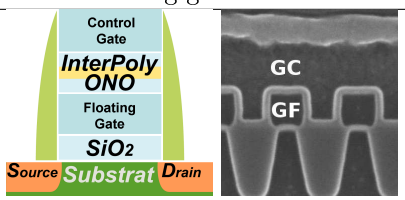
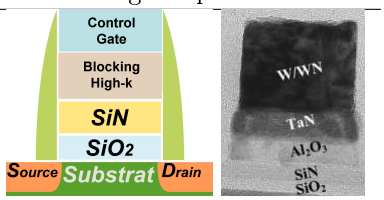
	Floating gate NAND	Charge trap NAND
Structure		
Programming : Erasing :	FN electrons injection FN electrons injection	FN electrons injection FN holes injection
Advantages	Improved retention	Low V_T dispersion promising technology for scaled devices.
Drawbacks	Abnormal V_T dispersion Less scalable	lower retention

Table 3.2: Properties of floating gate and charge trapping memories [92].

more attractive to be used as multi-level memories.

Even if the TANOS memories offer several improvements (tunnel oxide thicker, high-k control dielectric, metal gate), they have a retention that degrade especially at high temperature (see §2.2.3).
Moreover the knowledge of classical Flash memory materials and technology facilitates the resolution of manufacturing problems related to miniaturization process. This implies that even if the TANOS solution was previewed for the next sub $20 \sim 30nm$ generation [21], at the time of this writing the classical approach, based on poly-silicon floating gate, resists.

In order to improve erasing and retention, further studies on TANOS stack have to be done. In fact the high biases applied during programming strongly degrade the endurance and retention of these devices [35]. For example [93] shows the interest in optimising the alumina stack. In this chapter, we will show the interest in increasing the alumina post deposition annealing temperature to decrease the amount of interstitial hydrogen defects on it.

3.1.1 Alumina deposition

Aluminium oxide is an amphoteric oxide of aluminium with the chemical formula Al_2O_3. It is also commonly referred to as alumina.

In the fabrication of the devices used in this thesis the Al_2O_3 layers were deposited by atomic layer chemical vapor deposition (ALCVD) at 350°C using trimethyaluminuiom and H_2O precursors.

Atomic layer chemical vapor deposition Atomic layer chemical vapor deposition (ALCVD), which is based on the controlled growth of thin films, has attracted much attention as an

advanced materials processing for nano-thickness thin film deposition. Several unique advantages of ALCVD compared with other conventional deposition techniques make it a powerful deposition method in nano-fabrication. These advantages are accurate control of film thickness, uniformity over large areas, excellent conformality over complex-shaped substrates, low temperature mildly oxidizing process, multilayer processing capability, and layer by layer control.

Since ALCVD is based on a saturated, self-limiting surface reaction of precursors, a surface chemistry is important to understand the growth mechanism and optimize the ALCVD process. The film growth rate, film structure, composition, and surface morphology are affected by the surface chemical reactions of precursor.

The main surface reactions involved in ALCVD are exchange reactions between functional groups of precursors. For example, MO_2 oxide thin films can be deposited through the following exchange reactions [94]:

$$M - OH(s) + ML_n(g) \rightarrow MOM - L_{n-k}(s) + kLH(g)$$
$$MOM - L_{n-k}(s) + H_2O(g) \rightarrow MOM - OH_k(s) + (n-k)LH(g)$$

Where M and L are a metal and a ligand, respectively. Fig. 3.3 shows a schematic diagram of direct exchange reactions occurring during ALCVD of oxide thin films. A saturation of chemisorption takes place when the available surface bonding sites (OH) are all occupied by precursor (ML_n) and the adsorbed species do not create new bonding sites for the dosed precursors. OH functional groups are replaced by new functional groups L. New functional group L does not act as bonding sites for ML. Physisorbed multilayers over the layer chemisorbed on the substrate surface are removed during purge by inert gases. In the next saturated chemisorption, the surface is exposed to a different precursor (H_2O) and new surface bonding sites (OH) are formed by exchange reactions between L and H_2O on the surface. A series of repetitions of these exchange reactions gives a layer by layer growth of metal oxide thin film.

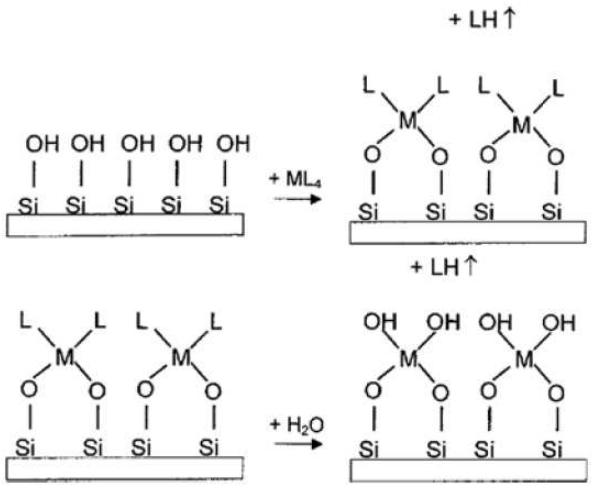


Figure 3.3: Growth mechanism of metal oxide by ALCVD, showing exchange reactions between functional groups of precursors.

Alumina layer fabrication In this work TMA(trimethyaluminum=$Al(Ch_3)_3$) has been used for as precursor in Al_2O_3 ALCVD. The growth mechanism of Al_2O_3 ALCVD using TMA and water is the follow:

$$Al(CH_3)* + H_2O(g) \rightarrow AlOH* + CH_4(g)$$
$$AlOH* + Al(CH_3)_3(g) \rightarrow Al(CH_3)* + CH_4(g)$$

TMA reacts readily with surface -OH groups by depositing -$Al(CH_3)_2$ and liberating CH [95].

Although there is no chlorine residue in films grown by using TMA, there are still H contaminations in grown film.

3.2 Material analysis

In order to study the behaviour of alumina in TANOS memory, we first characterize the physical characteristic of various 15nm alumina single layers. To do this we have exploited different recipes to deposit the alumina layer. In particular we performed different Rapid Thermal Anneals (RTA) from 700 to 1050℃ under N_2 or O_2. The results, here reported, are described by Colonna et Al. in [96].

The samples, summarized in Table 3.3, have been studied by means of different experimental techniques

- **Stress measurements** were done by the curvature method using the Stoney formula [97].
- **Thickness, and density measurements** were done by x-ray reflectometry (XRR) along with infrared spectroscopy in attenuated total reflection mode (ATR-FTIR).
- **Species distributions and stoichiometry** were analyzed with Secondary Ion Mass Spectroscopy (SIMS).
- **Hydrogen content and the H bond** were indentify with Multi Internal Reection (MIR) Spectrometry .
- **Band gap and refractive Index** extractions were done by optical measurement.

The informations extracted from these techniques are fundamental to create the chemical and physical models for the Al_2O_3 structure and to investigate the trapping properties of the different Al_2O_3 layers.

Anneal Ambiance		N_2				O_2	
Anneal Temperature °C	As dep	700	850	950	1050	700	1050
Stress (Mpa tensile)	300	300	1980	1700	1540	1880	1400
XRR Thickness A	150	142	122	122	125	150	124
XRR density (g/cm^3)	3	3	3.5	3.5	3.5	3	3.5

Table 3.3: Thickness, density, and stress.

3.2.1 Experimental and physical characterization

Stress, thickness and density measurements are summarized in Table 3.3.

Two populations can be distinguished from these measurements: a low thermal budget population with as deposited and 700 °C annealed samples, and a high thermal budget population (samples annealed at 850, 950, and 1050 °C):

The low thermal budget population has stress values below a few hundred of MPa, a thickness of 15 nm and a density of 3 g/cm^3.

The high thermal budget population has higher stress values (between 1400 and 2000 MPa), a lower thickness (12-12.5 nm) indicating a shrink and thus a higher density (3.5 g/cm^3).

It is noticeable that the sample annealed at 700 °C under O_2 ambiance shows a thickness and a density coherent with the low thermal budget population, but a high value of stress.

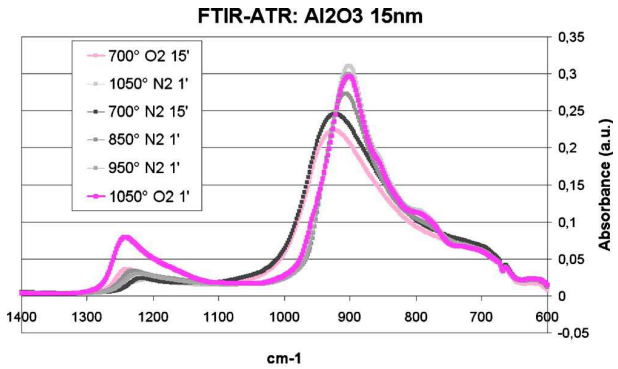


Figure 3.4: FTIR-ATR Spectra of alumina with different annealing conditions

These high and low thermal budget populations (see Tab. 3.3) could be distinguished analysing the Infrared spectroscopy results that are presented in Fig. 3.4.

Fig. 3.4 shows two peaks: the main peak is located at 900-950 cm^{-1} and corresponds to the LO mode (longitudinal optic) of the Al-O bond. The second peak located around 1250 cm^{-1} corresponds to the interfacial Si-O bond (LO mode). If we consider the second peak, located at around 1250 cm^{-1}, we can notice that after the oxygen anneal at 1050 °C the absorbance increases, what indicates an interfacial oxide regrowth.

Analysing the peak position and the full width half maximum (FWHM) of the main Al-O peak,we can distinguish the high and low thermal budget populations of table 3.3.

Because the FWHM corresponds to the order of the Al-O bonds in the alumina layer, we can state that the high thermal budget population with a narrower peak is more ordered than the low thermal budget population suggesting an amorphous and a crystalline population. One can be more precise: according to [98] the crystalline γ-phase can be identified from the shoulder at 780 cm^{-1}.

High resolution transmission scanning microscopy (HRTEM) pictures confirm the FTIR-ATR results. Figure 3.5 shows HRTEM pictures of the memory stack of SiO_2/ Si_3N_4/Al_2O_3 annealed at 700 and 1050 ℃. We can notice the crystalline structure of the 1050 ℃ annealed alumina and its subsequent thickness shrink noticed in Table 3.3 an in Fig. 3.5. This result confirms that the low thermal budget annealed samples correspond to amorphous state and the high thermal budget samples correspond to crystalline state.

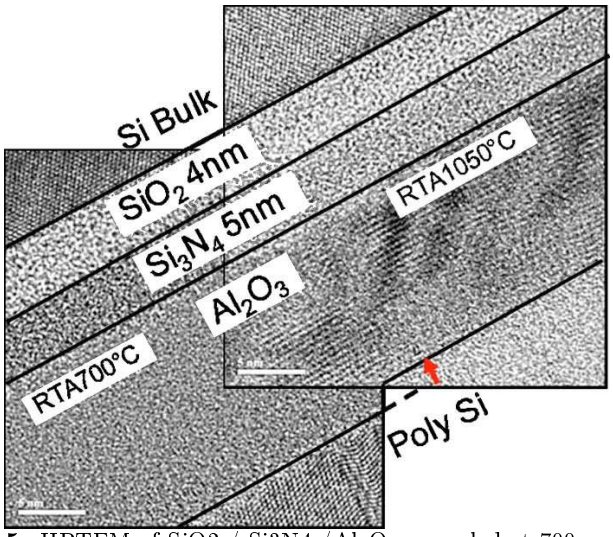


Figure 3.5: HRTEM of SiO2 / Si3N4 /Al_2O_3 annealed at 700 and 1050 ℃.

Electron energy-loss spectroscopy (EELS) was also performed on these two samples as shown in fig.3.6. The peak at 85 eV on the 1050 ℃ annealed sample is typical of the γ-phase according to [99], which confirms the FTIR-ATR results.

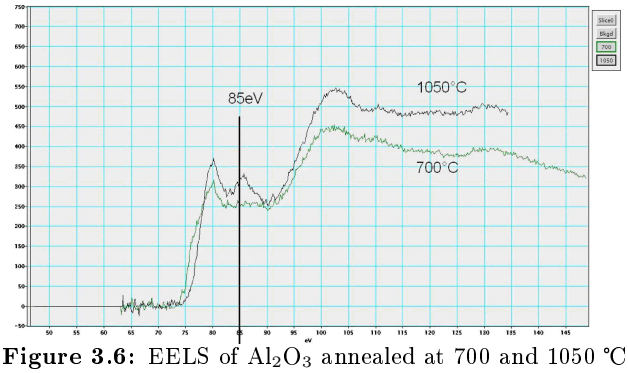


Figure 3.6: EELS of Al_2O_3 annealed at 700 and 1050 ℃.

Finally vacuum ultraviolet spectroscopic ellipsometry was used to determine the optical properties of alumina samples. Refractive index n and extinction coefficient k are plotted for both 700 and 1050 ℃ annealed samples as shown in Fig. 3.7. Refractive index is similar for both crystalline and amorphous sample: n(1.5 eV)=1.73 for crystalline alumina and 1.68 for amorphous alumina. The extinction coefficient k is rather different for the two samples. The energy band gap was extracted from the extinction coefficient measurement as explained in [100]. The amorphous sample show an energy band gap of 6.5 eV while the crystalline sample has a band gap value of 7 eV.

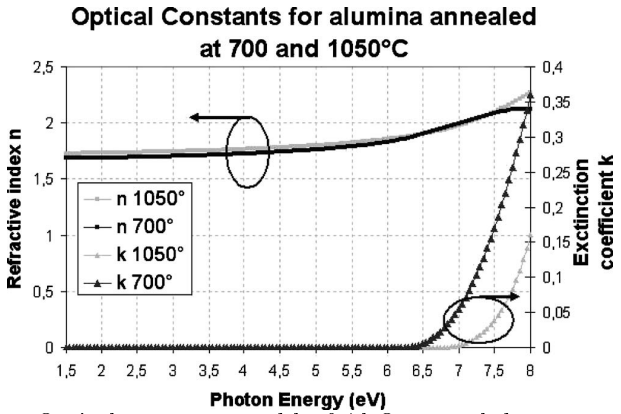


Figure 3.7: Optical constant n and k of Al_2O_3 annealed at 700 and 1050 ℃.

3.2.2 summary

In this section we have reported the characterization of Alumina layers of 15 nm deposited using Atomic Layer Deposition from water and TMA followed by Rapid Thermal Anneal. After a low temperature anneal (up to 700 ℃) alumina remains amorphous with low stress and density values. At higher temperatures (above 850 ℃) alumina is crystalline (γ-phase) with higher stress, density and optical band gap.

We thus choose the gamma phase for the atomistic simulations used to explore the role of defects on alumina. Indeed in a TANOS memory the alumina experiences a high thermal annealing step, typical of the source drain activation in the memory integration process.

3.3 Atomistic simulation

The study of the physical phenomena at atomic scale is crucial to understand the electrical behaviour of microscopic devices. In order to model the electronic structure of the matter different approaches exist. In this part we will detail the theory and the computational procedure of the Density Functional approach that has been used to compute defects electronic effects on a bulk γ alumina.

3.3.1 Density functional theory

Among various ab-initio methods the Density Functional Theory (DFT) has become the standard tool to study the electronic properties of many electrons system. Instead of being based on the resolution of a complicated many-body Hamiltonian, in DFT the problem of the iteration between many particles is substituted with considering a system of fictitious non interacting particles having the same density as the interacting many electron system.

The central object of DFT is thus the electron density $\rho(\mathbf{r})$ that represents the the probability of finding an electron in a specific location **(r)**. The DFT is an exact theory for calculating ground state properties of the system because it completely determines all the ground state properties of the interacting many-electron system.

In this paragraph we will introduce the formal basis of DFT theory.

Hohenberg-Kohn theorems In 1964 Hohenberg and Kohn [101] demonstrated in the *First Hohenberg-Kohn theorem* that the ground state of a system of N interacting electrons under an external potential v (i.e. the Coulomb potential of the nuclei on the electrons (v_{ne})) is completely described by its charge density ρ(r), which only depends on the three spatial coordinates, r.

$$E = E[\rho] \tag{3.1}$$

This means that with the knowledge of ρ(r), one also knows the particle number N and the external potential v(r), and hence all the ground state properties of the system such as the total energy $E\,[\rho]$, the potential energy $V\,[\rho]$,the kinetic energy $T\,[\rho]$, etc.

The first theorem would be useless without the second HK theorem demonstrating that the ground state electronic density is the one that minimises the total energy functional $E\,[\rho]$. In this way, DFT reduces the N-body problem to the determination of a 3-dimensional function ρ(r) which minimizes a functional $E\,[\rho(\mathbf{r})]$.

The Kohn-Sham equations [102] The practical way to apply the HK theory was theorized by Kohn and Sham (KS) in 1965.

The central idea in density functional energy is to replace the real system with an equivalent system of *non-interacting electrons* moving in an effective potential having the same ground state density of the real system:

$$E[\rho] = F\left[\tilde{\rho}\right] + \int v_{ext}(\mathbf{r})\tilde{\rho(\mathbf{r})} \tag{3.2}$$

where v_{ext} is the external potential, $\tilde{\rho}$ is the density of the *noninteracting* electron system, and $F\left[\rho\right]$ is the density functional:

$$F\left[\rho\right] = T_s\left[\rho\right] + E_H\left[\rho\right] + E_{xc}\left[\rho\right]. \tag{3.3}$$

The first term $T_s\left[\rho\right]$ is the *noninteracting* part of the kinetic energy of the electron system:

$$T_s\left[\rho\right] = \frac{1}{2}\sum_{i=1}^{N}|\nabla\psi_i(r)|^2 \tag{3.4}$$

the second term (E_H) in equation (3.3) describes the classical electrostatic (Hartree) interaction between the fictitious particles,

$$E_H\left[\rho\right] = \frac{1}{2}\int\int\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' \tag{3.5}$$

the third term $E_{xc}\left[\rho\right]$ is the so-called exchange-correlation energy, covering all electron-electron interaction effects beyond the Hartree term:

- the exchange energy due to the Pauli exclusion principle,
- the correlation energy
- the difference in kinetic energy between the interacting and non-interacting systems.

The second and third terms can be grouped with the external potential to define the so-called effective Kohn-Sham potential:

$$v_{eff}(\mathbf{r},\rho(\mathbf{r})) = v_{ext}(\mathbf{r}) + \int\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + v_{xc}. \tag{3.6}$$

where the v_{xc} is the exchange-correlation potential:

$$v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}. \tag{3.7}$$

Equation (3.2) is thus equivalent to

$$E[\rho] = T_s\left[\rho\right] + \int v_{eff}(\mathbf{r})\rho(\mathbf{r}), \tag{3.8}$$

and the so-called Kohn-Sham equations finally have the form

$$\left[-\frac{1}{2}\nabla^2 + v_{eff}(\mathbf{r})\right]\varphi_i = \epsilon_i\varphi_i\ . \tag{3.9}$$

In practice the system is calculated with an iterative method (see fig. 3.8): the density is calculated with 3.10, then the obtained density is inserted in the exchange-correlation potential 3.7, then the new eigenstates and new density are calculated, until convergence is reached.

$$\rho(\mathbf{r}) = \sum_{i=1}^{N}|\varphi_i(\mathbf{r})|^2 \tag{3.10}$$

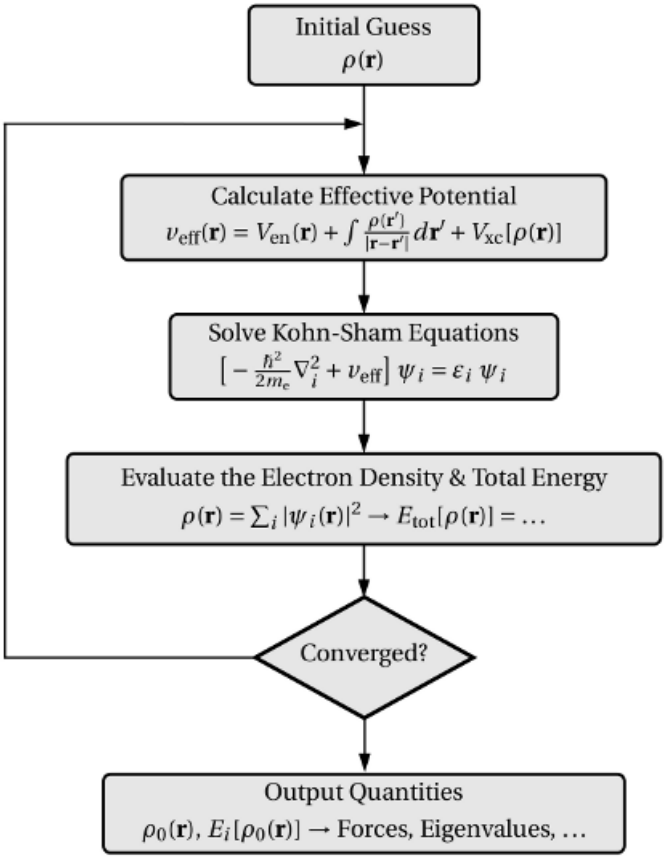


Figure 3.8: DFT iterative resolution.

Approximations to the exchange-correlation potential In equation (3.7) we defined the exchange-correlation potential as

$$v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}\,[\rho]}{\delta \rho}\,. \tag{3.11}$$

Unfortunately, the exact E_{xc} is unknown and has to be approximated. In this part we will present the two major approximations: the Local Density Approximation (LDA) and the General Gradient Approximation (GGA).

The Local density approximation The Local Density Approximation (LDA) [103] replaces the true exchange-correlation density at each point $\mathbf{r}$ in space by the xc-energy density of a homogeneous electron gas of the same (global) density.

Let ϵ_{xc} be the exchange-correlation energy per electron of a homogeneous electron gas of density ρ, then the exchange-correlation energy functional is approximated by:

$$E_{xc}[\rho] \approx E_{xc}^{LDA}[\rho] \equiv \int \rho(\mathbf{r})\epsilon_{xc}(\rho(\mathbf{r}))\,d\mathbf{r}\,. \tag{3.12}$$

Taking the functional derivative of E_{xc} with respect to ρ, one obtains the following expression for v_{xc}:

$$v_{xc}^{LDA} = \frac{\delta E_{xc}}{\delta \rho} = \epsilon_{xc}(\rho) + \rho\frac{\partial \epsilon_{xc}}{\partial \rho} \tag{3.13}$$

and, since the system is not homogeneous, $\rho = \rho(\mathbf{r})$. In order to calculate ϵ_{xc} it can be divided into exchange and correlation parts $\epsilon_{xc} = \epsilon_x + \epsilon_c$. The exchange part is known analytically and given by:

$$\epsilon_x = \frac{3}{4}\left(\frac{3\rho(\mathbf{r})}{\pi}\right)^{1/3} \tag{3.14}$$

and the LDA exchange potential is:

$$v_x(\mathbf{r}) = -(\frac{3}{\pi}\rho(\mathbf{r}))^{1/3}\,, \tag{3.15}$$

while only limiting expressions for the correlation density are known exactly, leading to numerous different approximations for ϵ_c.

For a homogenuous electron gas, the LDA exchange-correlation functional is exact. Despite the fact that for most applications, especially for isolated systems, the electron density is by far different to the one of a homogenuous electron gas, the approximation can yield good results.

Generalized gradient approximations One can think about Generalized Gradient corrections (GGA's) as next order corrections to the LDA, since a functional dependence on the gradient of the density is added to ϵ_{xc}, i.e.,

$$E_{xc}^{GGA}[\rho] = \int d^3\mathbf{r}\,\rho(\mathbf{r})\epsilon_{xc}\left(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})\right) d\mathbf{r}\,. \tag{3.16}$$

Compared to the LDA approximation, the error for ionization energies is diminuished by a factor 3-5 for the generalized gradient approximation [102]. Still, accuracy achieved with wavefunction-methods remains higher.

Pseudopotentials In order to solve the the Kohn-Sham equations, besides the exchange-correlation functional, also an influence of the latter is needed.

As we can see in figure 3.9, the contribution of the valence states to the total electron density is negligible within the core region and dominant beyond it. This allows to separate the set of states in core sets $\left|\psi^{(c)}\right\rangle$ and valence set $\left|\psi^{(v)}\right\rangle$. For describing a solid, the core states do not need to be described by Kohn-Sham equations. Together with the atom cores they can be approximattely described by a softer ion potential experienced by the valence electrons, the so-called *pseudopotential*.

Choice of initial wavefunctions To obtain the energy of the system with DFT calculations, the initial wave-functions (ϕ) have to been set at the beginning of simulation. DFT calculations use either plane waves, as in the ABINIT code, or spatially localized functions, as in the SIESTA code. The main advantage of the plane waves is that an absolute convergence criterion exists unfortunately this comes at the expense of a high computational time.

In the SIESTA code [105], the basis set is written as linear combination of atomic orbitals (LCAO). This technique was introduced to calculate molecular orbitals in quantum chemistry

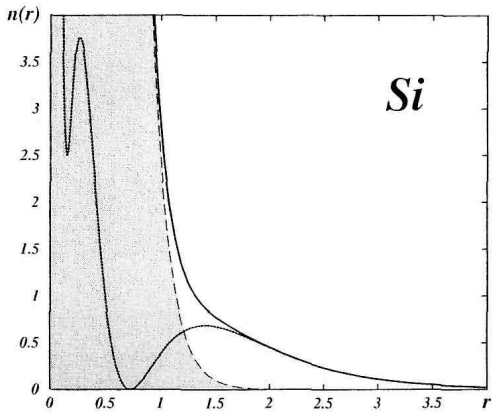


Figure 3.9: Electron density n(r) as a function of the radius in Angström. The dashed line with the shaded area underneath represents the density of the core electrons, the solid line (with wiggles) represents the density of the valence electrons, the other solid line the total electron density.[104]

and it is used in the DFT calculations in order to reduce the number of basis functions and consequently the computation time. The disadvantage of this method there is no theoretical guarantee of an absolute convergence and the accuracy of the basis adopted for the LCAO calculations should be controlled with what obtained for a planar basis set.

3.3.2 Structure of Al_2O_3

In page 78 we proved that among the many polymorphs of alumina, after a high thermal annealing step, typical of the source drain activation, alumina layer results in γ-Al_2O_3 crystalline form in agreement with [106].

The structure of γ-alumina is usually described as a defective spinel, denoted as $\square_{22/3}Al_{21\ 1/3}O_{32}$, ($\square$ = vacancy). This formula is deduced from the fact that a spinel cubic cell (typified by $MgAl_2O_4$) has 32 O atoms on a face-centered cubic (fcc) lattice and 24 cation sites in tetrahedral (Mg) and octahedral (Al) positions (fig.3.10).

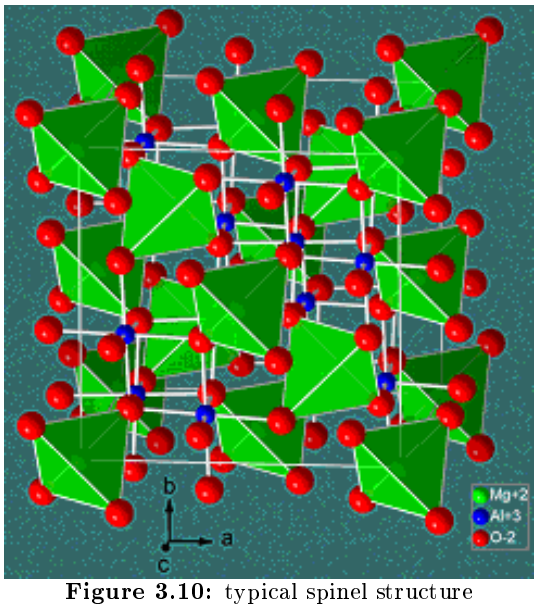


Figure 3.10: typical spinel structure

In order to satisfy the Al_2O_3 stoichiometry, one must introduce to this structure an average of $2\frac{2}{3}$ cation vacancies per cell . The question of the vacancy distribution between the tetrahedral and octahedral sites has led to several conflicting reports. For example, calculations based on classical molecular dynamics (MD) and Monte Carlo simulations in conjunction with *ab initio* calculations show that more than 50% of the vacancies are on tetrahedral sites. However, the calculations of Mo et all [107] based on empirical pair potentials in conjunction with the *ab initio* LCAO method, as well as the classical MD simulations by Streitz and Mintmire [108], suggest that these vacancies occupy octahedral sites. This last theory is confirmed by Gutierrez et all [109]: using ab-initio simulations, they showed that when considering all possible configurations, the one which has both vacancies on octahedral sites has the lowest total energy.

In this thesis we use the structure (Fig. 3.11) found by Mendez et all [110] that use [109] conclusions as starting point.

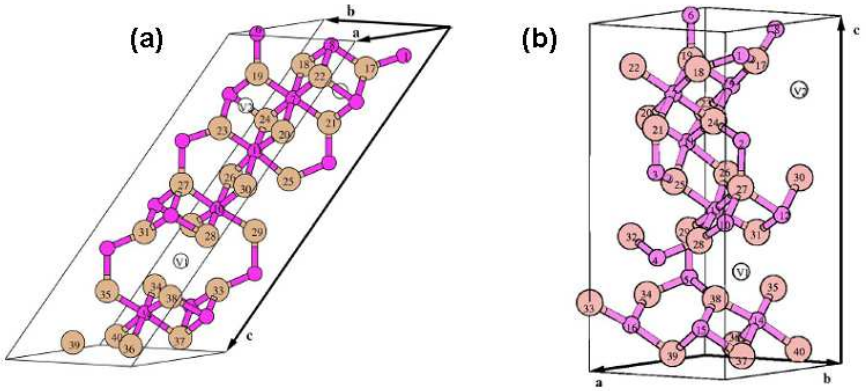


Figure 3.11: (a) The unit cell of the optimized structure for γ-Al_2O_3. (b) Compact unit cell obtained by a change of basis $a' = -b$, $b' = b - a$, $c' = a + b - c$, followed by a relaxation of both the cell and the ionic positions. Big circles represent O atoms, small circles represent Al atoms, and the white circles show the positions of the two Al vacancies in 16d positions.

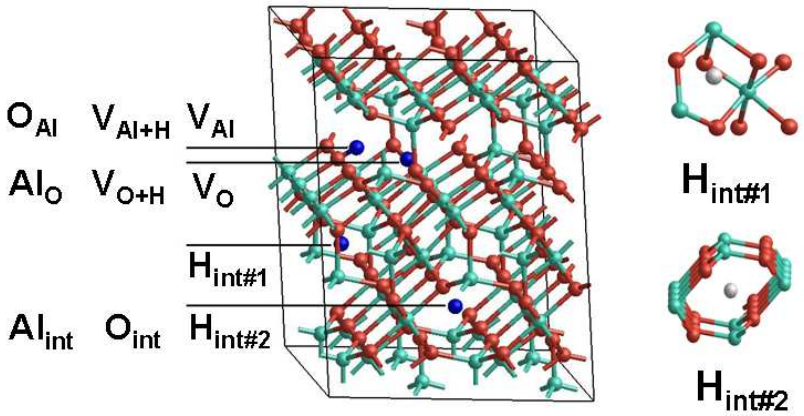


Figure 3.12: Left: Simulated γ-Al_2O_3 structure with a 160 atoms supercell periodically repeated: relaxed defective spinel, 96 Al + 64 O. Al in green, O in red, points of interests in blue. Right: structural position of $H_{int\#1}$ and $H_{int\#2}$.

3.3.3 Computational method

In order to determine the most relevant defects that can play a role in the memory device we have focused our attention on the position of their electronic levels with respect to the band gap of bulk γ-Al_2O_3 and also on their Gibbs formation energy that governs primarily the concentration of defects inside the deposited film.

Two kinds of defects were considered (see fig. 3.12): intrinsic oxide point defects and hydrogenated defects. The latter are motivated by the fact that hydrogen is an ubiquitous impurity in all fabrication processes. Moreover as we will see in the next section, both the leakage current and the H amount in the Al_2O_3 layer are reduced when a higher thermal budget is used, what is an experimental indication that the H concentration and the electrical features of Al_2O_3 are correlated.

DFT parameters All the results reported in this work were taken allowing the system to relax with a lowering of the residual forces below 0.04 eV/Angström. For the more stable neutral defects we also performed GW calculations at first order G0W0, correcting the underestimated γ-Al_2O_3 bandgap in DFT/LDA from 4.1 eV to 6.5 eV, consistent with our experimental measurements (see fig. 3.7). We used ABINIT [111] for the G0W0 calculations, starting from a LSDA calculation with a reduced model made of 40 atoms of γ-Al_2O_3, by employing a sufficient number of bands and accurate cut-offs for the evaluation of the dielectric matrix and the self-energy with a Plasmon-pole approximation (800 bands, cut-offs between 8 and 12 Hartree)

Formation energy The concentration of defects in the deposited film is governed primarily by their formation energies that depend on the chemical potential of atomic constituents and the electron chemical potential.

The formation energies of Al_2O_3 defects were calculated from the total energy of the supercell as a function of the oxygen chemical potential (μ_O). We based our calculations on the formalism by [112] used to find the chemical Potential dependence of defect formation energies in the case of GaAs. In alumina, the formation energies of defects depend on the chemical potential of the atomic species Al (μ_{Al}) , O (μ_O) and, in the case of hydrogenated defects, H (μ_H). If the defects are charged, they also vary with the electronic chemical potential (μ_e) of the electron (that is the Fermi level). If we consider a charged defect, its formation energy G_F is given by:

$$G_F = E_D - n_e\mu_e - n_{Al}\mu_{Al} - n_O\mu_O - n_H\mu_H \tag{3.17}$$

where E_D is the total energy of the supercell with the defect, n_{Al} and n_O are the number of Al and O atoms and n_e is the number of electrons that are added (with negative sign) or removed (with positive sign) from the system. The chemical potential of the species present in the equation above is unknown and depends on the fabrication conditions (e.g. temperature, pression, precursors..). Despite this fact it is possible to determine the chemical potential range in values which may or may not be physically allowed. To do this we have applied some restrictions on the specie chemical potentials:

Al The chemical potential of Al may not exceed the chemical potential of bulk Al $\mu_{Al} < \mu_{Al(bulk)}$. In the contrary, the excess Al may form a bulk Al precipitate.

H The chemical potential of H in the bulk is in equilibrium with the reservoir of H_2 molecule, and thus, μ_H must be $= \frac{1}{2}\mu_{H_2}(T, P)$. In this case the dependence of gaseous molecules H_2 to pressure and temperature is stronger than bulk-aluminium so further correction are needed.

O The restriction on the oxygen chemical potential (μ_O) is calculated from the Al_2O_3 stoichiometry: the sum of the chemical potential of the single species forming the ideal Al_2O_3 must be equal to the total energy of the perfect supercell (E_0), therefore, $E_0 = 2\mu_{Al} + 3\mu_O$.

Hence the formation energy has been calculated solving for each defect the system:

$$\begin{cases} G_F = E_D - n_e\mu_e - n_{Al}\mu_{Al} - n_O\mu_O - n_H\mu_H \\ \mu_H = \frac{1}{2}\mu_{H_2}(T,P) \\ \mu_{Al} < \mu_{Al(bulk)} \\ E_0 = 2\mu_{Al} + 3\mu_O \end{cases} \tag{3.18}$$

The total energy of the perfect supercell E_0, as well as that of the supercell with the defect E_D, is an output of the Al_2O_3 atomistic simulations. The chemical potential $\mu_{Al(bulk)}$ is found dividing the total energy of the unit cell used to simulate the bulk Aluminum (see fig. 3.13) with the number of its atoms. The chemical potential of the gas hydrogen μ_{H_2} (dihydrogen or molecular hydrogen) as was computed drawing up two hydrogen atoms and starting an ab-initio simulation allowing the system to relax.

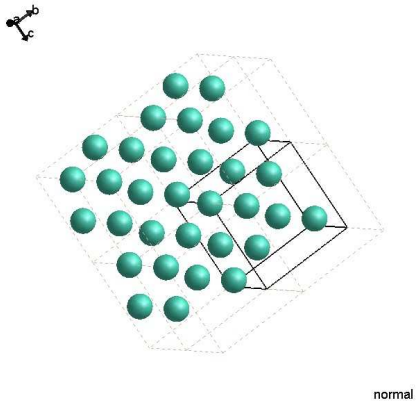


Figure 3.13: Simulated Aluminium crystal structure.

In the follow, the Gibb's free energy will be reported for a range of allowed values of μ_O. For each μ_O the μ_{Al} can be computed from 3.18.
For the sake of simplicity in the next section we will only report the formation energy of the defects for their neutral charge state: considering the formation energy of charged defects would multiply the number of possible cases without clarifying the physical behaviour of alumina, as the electron chemical potential during trap assisted conduction mechanisms is hardly known.

3.3.4 Results

The results of many calculations are presented in the next sections: the aluminum, oxygen and hydrogen inerstitials (Al_{int} O_{int} H_{int}) in different sites of the defective spinnel structure; the oxygen and the aluminum vacancies (V_O and V_{Al}) possibly passivated with H (V_{O+H} and V_{Al+H}); the substitutional defects Al replaced with O and vice versa O_{Al} and Al_O.

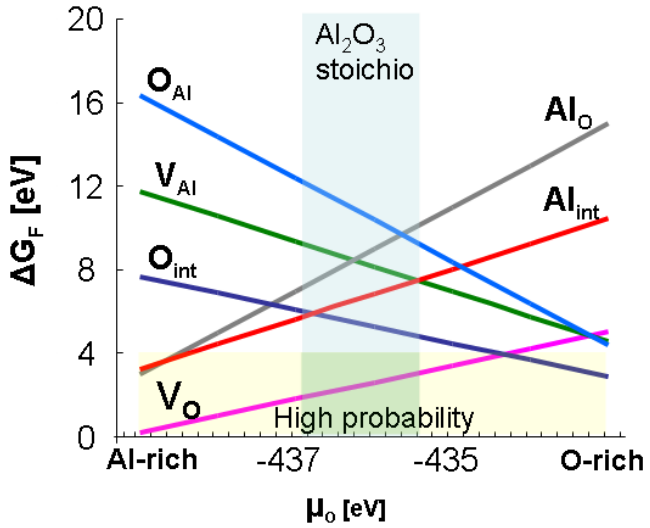


Figure 3.14: Defect's Gibbs free energy of formation versus the O chemical potential (μ_O). Lowest G_F corresponds to most probable defects.

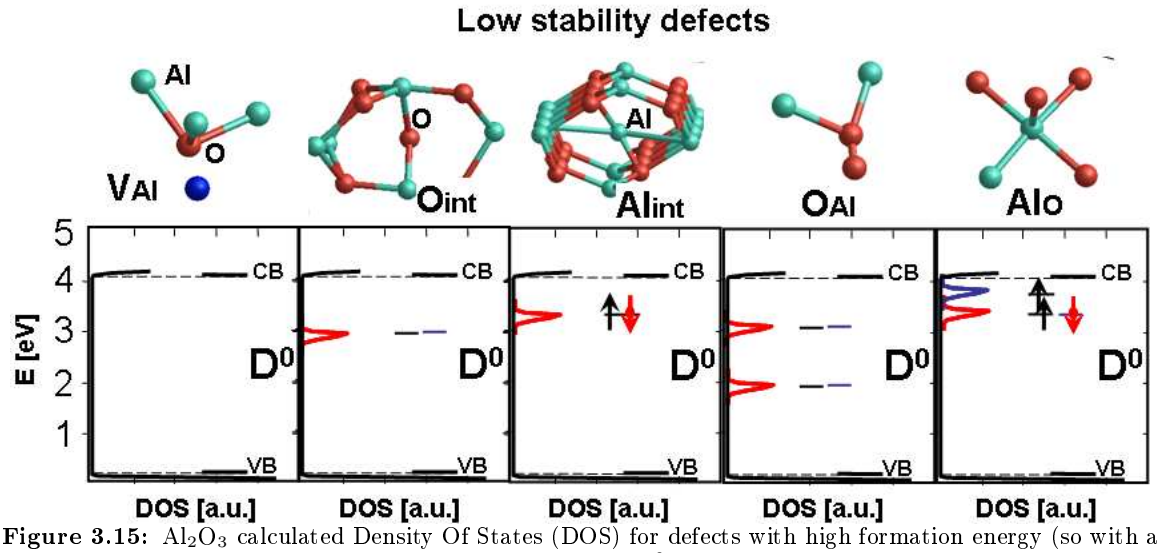


Figure 3.15: Al_2O_3 calculated Density Of States (DOS) for defects with high formation energy (so with a low occurrence probability: see Fig. 3.14). Neutral states (D^0) are represented.

3.3.4.1 Intrinsic defect

In this first section, we present the calculations corresponding to intrinsic defects present in an Al_2O_3 layer in the γ-phase. For V_{Al} ,V_O, Al_{int}, O_{int}, O_{Al}, Al_O, we computed the formation energy (Fig. 3.14) and the Density Of States (DOS) of the neutral charged state (Fig. 3.15 + Fig. 3.16 for V_O).

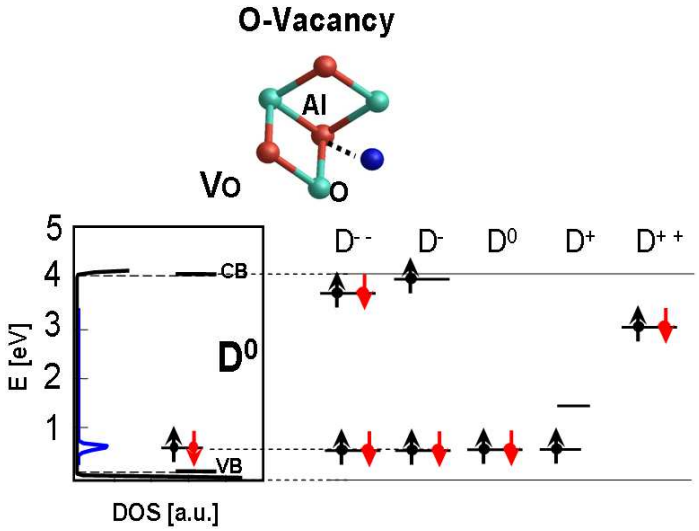


Figure 3.16: Al_2O_3 calculated DOS for various charge states of oxygen vacancy. In the negatively charged states (D^0 and D^{--}) a peak slightly below the conduction band appears.

For a quasi-stoichiometric γ-alumina, the V_O is the most stable among the different considered point defects. For an oxygen-rich alumina O_{Al} and O_{int} could also be considered as they provide energy levels close to the conduction band. Moreover O_{int} for a stoichiometric alumina is less stable than V_O but remains one of the most stable γ-Al_2O_3 active defect with a Gibbs free energy comparable to that of H_{int} (see next section and fig.3.17) at high pressure and temperature conditions. The oxygen vacancy can usually exist in all five charged states from -2 to +2. The computed DOS of V_O in these charged states is reported in Fig. 3.16. In the D^0 state, V_O introduces a vacant level slightly above the conduction band, unlikely to participate to electrical conduction. However, even by opening correctly the band gap, GW calculations confirmed the behaviour reported in [106]: the energy level decreases below the CB only when the trap is progressively charged by electrons. It can thus be imagined that if V_O plays a role in trap assisted conduction, it would be with D^- and D^{--} states.

3.3.4.2 H-related defects

This section describes the study of H-related defects in γ-Al_2O_3, which is motivated by the presence of H-atoms in the layer as put in evidence by SIMS measurements (see fig. 3.22). Among all the possible H-related defects, the interstitial H in position # 1 and # 2 are the most stable one at standard pressure and temperature conditions (Fig. 3.17, lower panel).

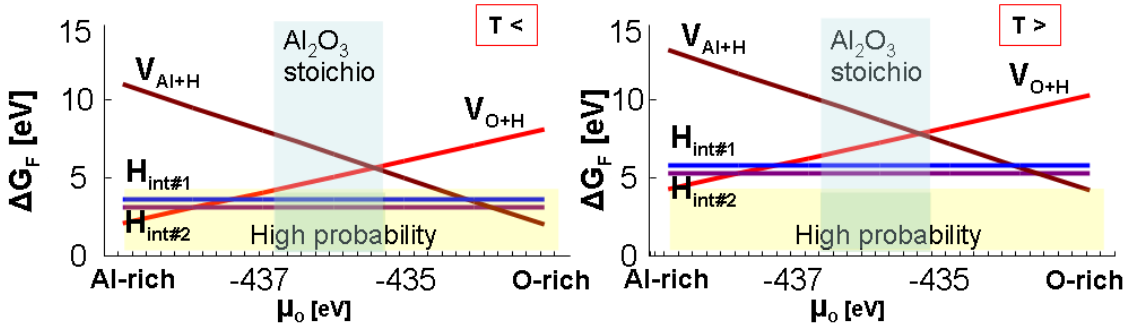


Figure 3.17: Defect's Gibbs free energy of formation versus the O chemical potential (μ_O). Lowest Gf corresponds to most probable defects. High T °C PDA reduces the H defect stability, as $G_F$increases.

In order to study the impact of the Al_2O_3 post deposition anneal, we have also calculated the Gibbs free energy for a chemical potential of hydrogen corresponding to a temperature of 1000K and a lower partial pressure of 1 mTorr. In this case the relative stability of these defects is decreased of more than 1eV (Fig. 3.17 right panel), what is directly related to the lower H amount measured on our SIMS experiments (Fig. 3.22). Moreover Fig. 3.17 indicates that the introduction of an H atom in an oxygen or aluminium vacancy is less stable than in the interstitial positions.

The Density Of States of the less probable defects (V_{O+H} V_{Al+H}) in their neutral charged state are reported in Fig. 3.18. It's interesting to notice that H in substitution of an oxygen or aluminium atom does not change the electrical nature of the defect: V_{Al+H} does not give any energy level in the band gap while V_{O+H} introduces a filled level at about 0.6eV from conduction band.

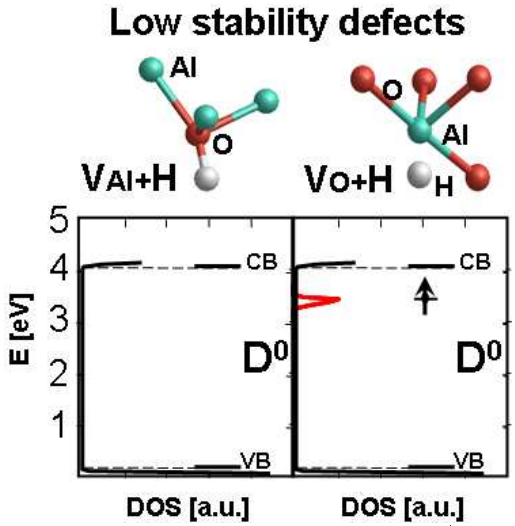


Figure 3.18: DOS for H-related defects with high formation energy (see Fig. 3.17). Neutral states (D^0) are represented.

For H_{int} in position #1 we computed the DOS of its three charge states. As shown in Fig. 3.19 the $H_{int\#1}$ in the neutral charge state (D^0) provides two energy levels inside the band gap:

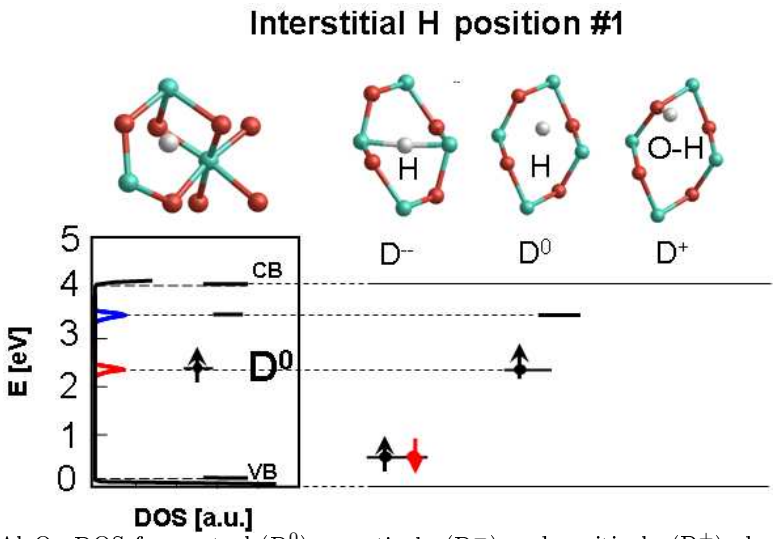


Figure 3.19: γ-Al_2O_3 DOS for neutral (D^0), negatively (D^-) and positively (D^+) charged interstitial H defect in position # 1. Only D^+ does not give energy state in the band gap, due to H migration in the network and O-H bond formation.

one is empty and the other one occupied by one electron, confirmed by our G0W0 calculations (table 3.4). From this configuration when an electron is introduced in the system (D^-) the trap depth is strongly increased and H remains in interstitial position. When the system loses an electron (D^+), $H_{int\#1}$ is attracted by the negative charge of a nearest-neighbour oxygen and binds to it forming a hydroxyl bond that does not give any level in the bandgap. The behaviour of interstitial $H_{int\#2}$ is different. In this case H is always bond to its nearest oxygen neighbour with almost the same global configurations for its three charge states. H acts in this case as a shallow donor as the D^0 level lies above the conduction band [113]. As general conclusions concerning the role of hydrogen, we can say that (i) H does not passivate O vacancies, (ii) is able to generate stable energy levels inside the γ-Al_2O_3 band gap.

DEFECTS	V_{Al}	V_O	O_{int}	Al_{int}	Al_O	O_{Al}	V_{Al+H}	V_{O+H}	$H_{int\#1}$	$H_{int\#2}$
Stability	Low	High	Med.	Low	Low	Low	Low	Med.	High	High
Active	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
En. from CB (eV)										
DFT/LSDA	None	~0	1.1	None	None	0.5	None	None	0.8	None
G0W0	~0								2.1	None

Table 3.4: Synthesis of defects inside γ-Al_2O_3. Defect stability is given by the Gibbs free energy of formation. Activity concerns the potential presence of an energy level in the band gap. Energy level precises the trap depth of the 1st free level based on DFT/LSDA calculations for D^0 state. G_0W_0 corrections are given for some of them.

3.3.4.3 Summary

Atomistic simulations were performed to study electronic structures and formation energies of point defects in γ-alumina. As resumed in Tab. 3.4 it is found that among all the possible defects, the H interstitial, the oxygen vacancy and in a less extent the O in interstitial position

are energetically favourable. These defects give free energy levels inside the band gap.

Trap assisted conduction involving V_O could only be envisaged between D^- and D^{--} states, as the neutral charged state gives energy levels too close to the Conduction Band. On the other hand H_{int} could be a good candidate to explain both the trap assisted conduction and the correlation between leakage current and PDA temperature. Eventually, O_{int} could also explain some trap assisted conduction when the H concentration becomes negligible

3.4 Electrical characterization

Based on the atomistic simulation considerations (see §3.3.4), it appears that the H_{int} defects could be the main responsible of the trap assisted leakage current through the Al_2O_3 layers, as they offer an extended distribution of energy state and a good stability with respect to other defects. To validate this theory the electrical characterization on alumina based devices were confronted with some analytic simulations as described below.

3.4.1 Electrical characterization of alumina single layer

To point out the relation between trapping and PDA temperature we have done stress-CV measurements on alumina single layers. The CV measurements were performed on transistors with ALCVD Al_2O_3 dielectric, with different PDAs, and AVD TaN control gate (deposited in a Tricent reactor). In the fabrication process of a memory device a rapid thermal anneal of 1050°C is performed in order to activate the Source/Drain implants. We have thus performed a rapid anneal of 1 min at 1050 °C also on the alumina single layers samples. The results reported in fig. 3.20 are in agreement with the atomistic study: trapping in Al_2O_3 is reduced as the PDA temperature is increased.

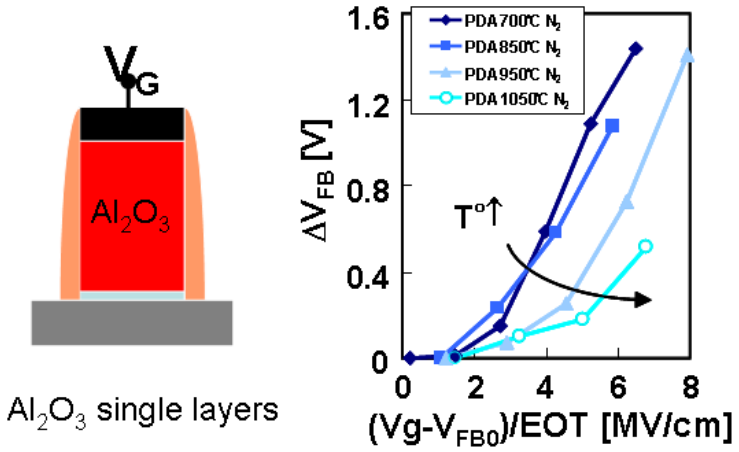


Figure 3.20: Flat band voltage shift as a function of the applied stress voltage (V_G), measured on 16nm Al_2O_3 single layers with various PDA temperatures. Trapping is reduced as PDA is increased.

Another experimental validation of the atomistic simulation results comes from the analysis of the leakage currents of the Al_2O_3 single layers. Currents are strongly activated in temperature; moreover a higher temperature PDA gives rise to lower leakage currents (Fig.3.21).

The experimental results were compared with the analytic simulations performed using a model which considers a multi-phonon trap-assisted tunneling conduction mechanism, including random defect generation [114]. For both PDAs, a similar trap energy distribution (between 1.5eV and 2eV) was extracted from the fitting of the experimental curves (Fig.3.21). Note that the values are in good agreement with the H_{int} defect in the D^0 state simulated by atomistic calculations. On the other side, the extracted trap density decreases of about a factor 5 as the PDA temperature is increased from 700°C to 900°C .

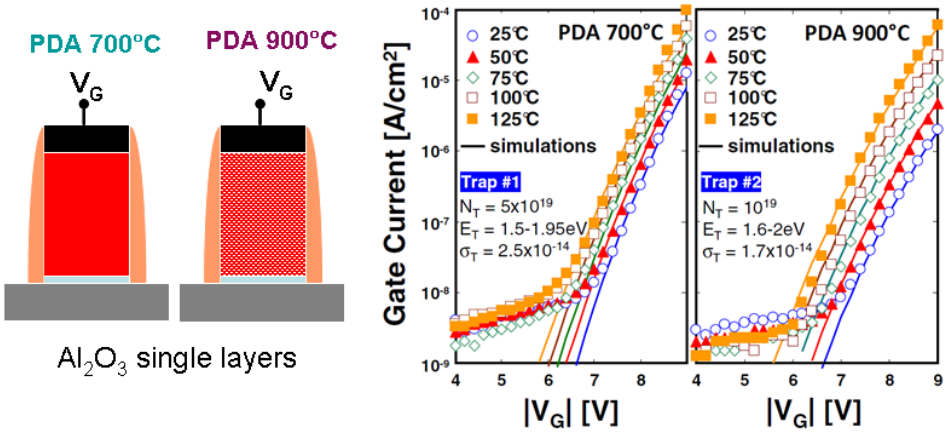


Figure 3.21: Left: schematic representation of Alumina single layers. Right: Experimental (symbols) and simulated (lines) $I_G(V_G)$ characteristics at various temperatures of 16nm Al_2O_3 layers processed with various PDA (TaN control gates are used). Simulations are based on the model presented in [114], traps parameters are indicated.

This reduction can be explained by the decreasing of the H content at higher thermal budgets, in agreement with the SIMS measurements reported in fig. 3.22 and the simulation results (fig. 3.17).

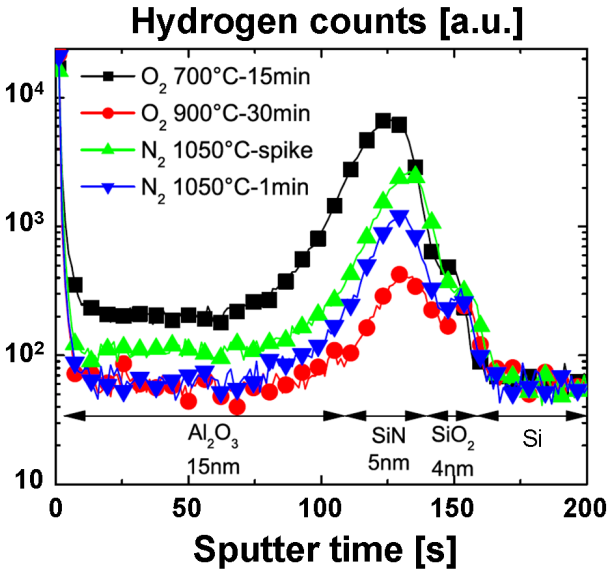


Figure 3.22: H profile measured by SIMS in $Al_2O_3/Si_3N_4/SiO_2/Si$ stacks. The H content (and thus the H-related defects) is reduced as the thermal budget of the Post Deposition Anneal is increased.

3.4.2 Electrical characterization and physical modelling of TANOS memories

To evaluate the role of Al_2O_3 H-related defects on the retention characteristics of charge-trap memories, the trap parameters coming from atomistic simulations and validated through the fitting of Al_2O_3 leakage currents were introduced in a complete device physical simulator of TANOS memory [114, 115] (Fig.3.23). In parallel, a detailed experimental study of retention

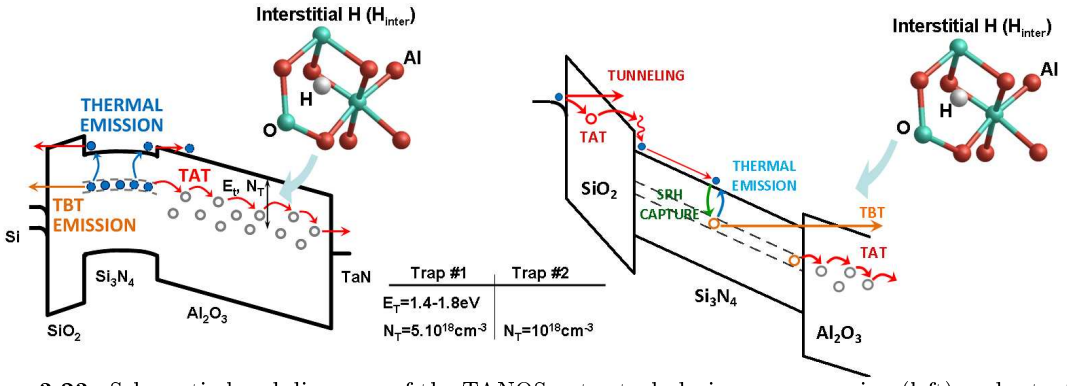


Figure 3.23: Schematic band diagrams of the TANOS gate stack during programming (left) and retention (right), describing the physical mechanisms taken into account in the physical modelling reported in [115]. Traps parameters (from Fig.3.21) used for device simulations are reported.

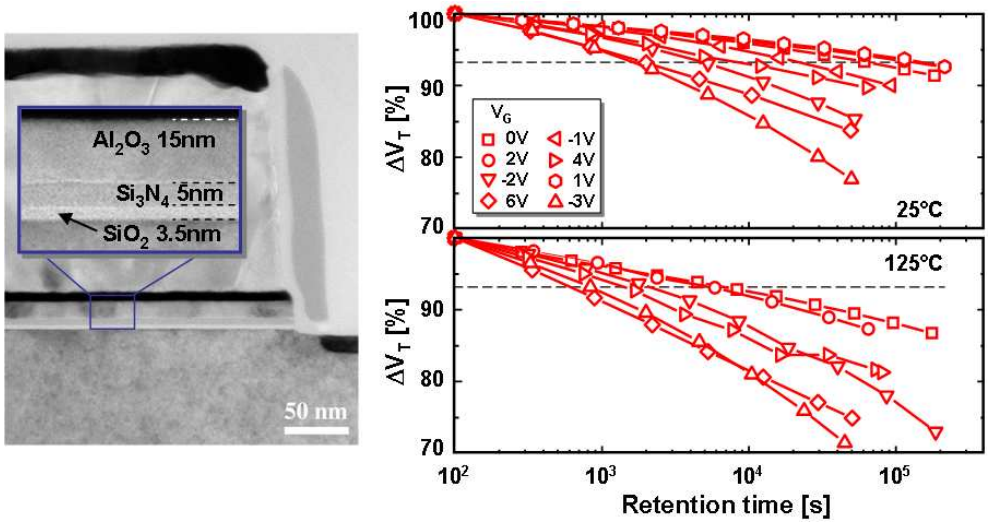


Figure 3.24: Left: TEM cross section of the fabricated TANOS memory devices. Right: Retention measurements performed on TANOS memories (with 700°C PDA for Al_2O_3) at various applied V_G and temperatures. Dashed lines indicate the 7% charge loss criterium used to extract the retention life time reported in Fig. 3.25 .

behaviour was performed on TANOS memory devices, with a 3.5nm tunnel oxide, 6nm LPCVD Si_3N_4 charge trapping layer, 16nm ALCVD Al_2O_3 layer with two different PDAs - and AVD TaN control gate (see TEM in Fig.3.24).

To enhance the leakage current through the alumina stack during retention we have applied a positive voltage on the gate electrode. Charge-loss measurements with different applied V_G and different temperatures are illustrated in Fig.3.24. For negative V_G, retention is governed by the leakage through the tunnel oxide (being the electric field in the tunnel oxide increased while reduced in Al_2O_3). Similarly, for positive V_G, retention is governed by the leakage through Al_2O_3.

Fig.3.25 shows the memory lifetime (retention time corresponding to 7% of charge loss) as a function of the applied V_G. A bell shaped curve appears, with a maximum of life time

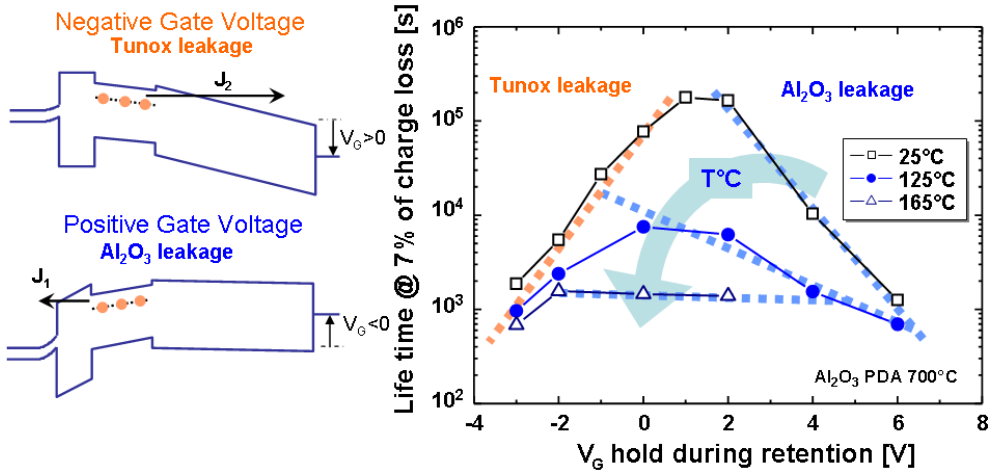


Figure 3.25: Life time (corresponding to 7% of charge loss) extracted from retention measurements reported in fig.3.24. Dashed lines, illustrating tunnel and Al_2O_3 leakage, serve as guides to the eyes.

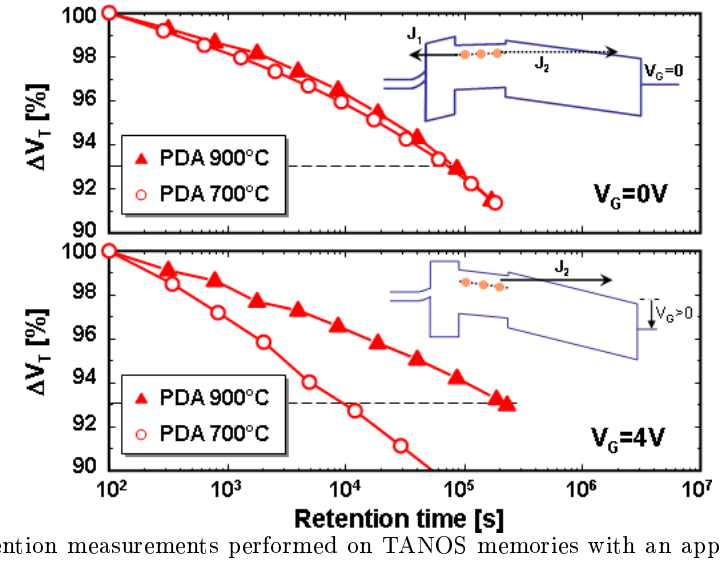


Figure 3.26: Retention measurements performed on TANOS memories with an applied V_G of 0V or 4V during measurements. PDAs of 700°C and 900°C are compared.

corresponding to the equality of the leakage currents through Al_2O_3 and SiO_2 (J_1=J_2). As the measurement temperature is increased, the bell curve is flattened, the retention being more and more limited by the electron thermal emission in the conduction band of nitride.

On figs.3.26-3.27 we analyse the impact of the alumina PDA on the retention characteristics. The PDA does not impact retention when the charge loss through the tunnel oxide is dominant (V_G=0V). On the other hand, when the charge loss through Al_2O_3 is dominant (V_G=4V), the 900°C alumina PDA offers an improved retention behaviour, due to the lowering of the Al_2O_3 leakage current. Based on this experimental understanding, the device physical modelling was used to simulate experimental data.

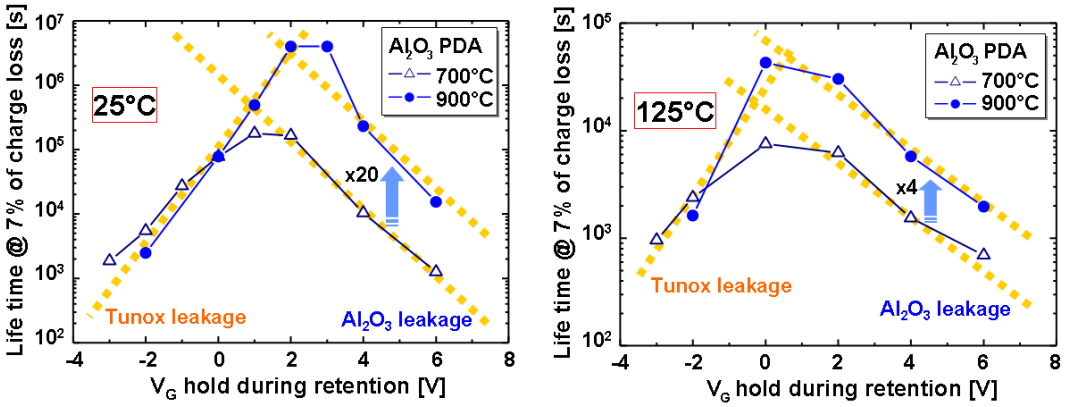


Figure 3.27: Life times (corresponding to 7% of charge loss) extracted from retention measurements performed at 25°C and 125°C and various applied V_G on TANOS memories with different Al_2O_3 PDAs. Dashed lines, illustrating tunnel and Al_2O_3 leakage, serve as guides to the eyes.

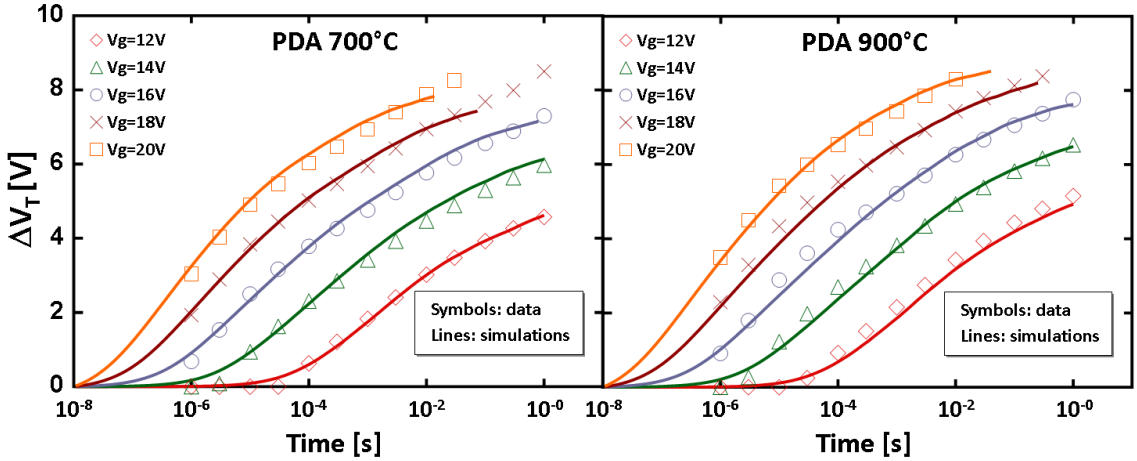


Figure 3.28: Comparison between measured (symbols) and simulated (lines) program transients for TANOS memories, processed with 700°C (trap# 1 parameters are considered) or 900°C (trap#2 are considered) Al_2O_3 PDAs.

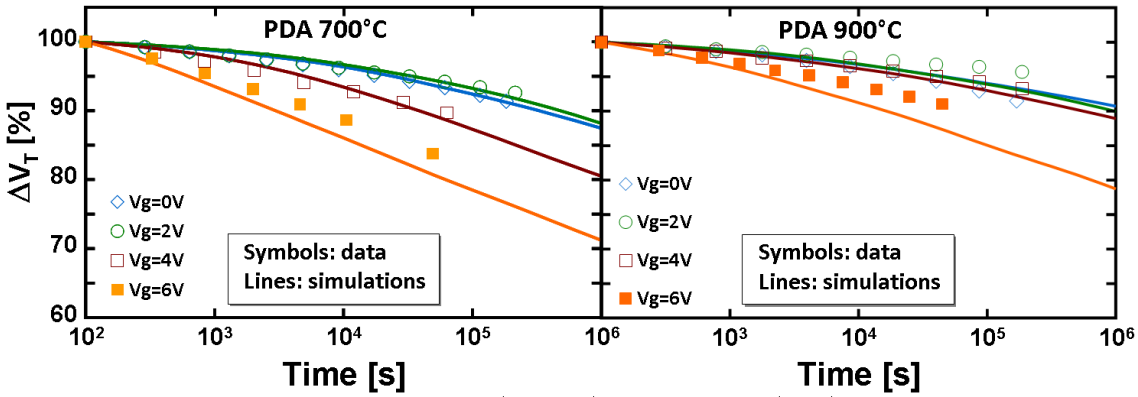


Figure 3.29: Comparison between measured (symbols) and simulated (lines) retention curves of TANOS memories, with 700°C (trap#1 parameters are considered) or 900°C (trap#2 parameters are considered) Al_2O_3 PDAs.

We started by accurately reproducing the program characteristics of TANOS memories with the two different PDAs, allowing to extract the electron distribution in the nitride layer immediately after charge injection (before retention), (Fig.3.28). Then, the retention characteristics were simulated for various applied V_G, assuming trap assisted currents through Al_2O_3 with 700℃ and 900℃ PDAs. The trap parameters, extracted in fig.3.21, are almost the same for the two PDAs temperature, but the trap concentration is lower in 900℃ in agreement with the ab-initio simulations of hydrogen defects. The simulated retention characteristics show a very good agreement with the experimental data (fig.3.29).

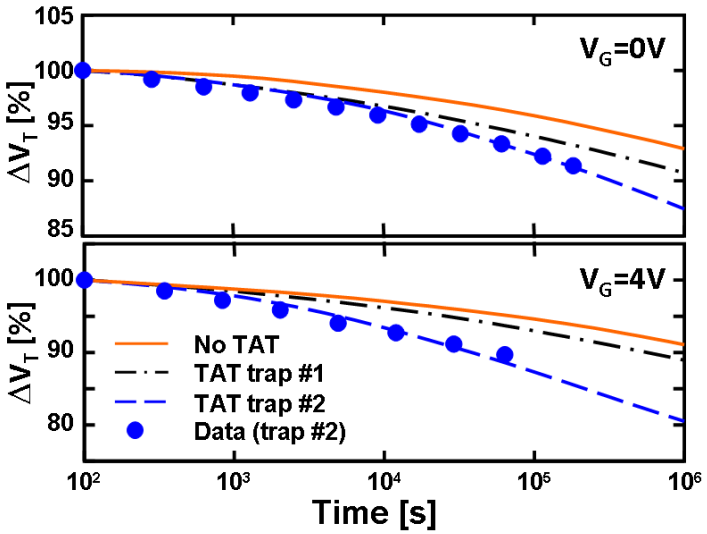


Figure 3.30: Simulated retention curves for TANOS memories with with 700℃ (trap#1 parameters are considered), 900℃ (trap#2 parameters are considered) or without traps in Al_2O_3.

Finally, fig.3.30 presents the simulated retention curves for TANOS memories, assuming no traps in Al_2O_3, or a trap assisted conduction using trap#1 parameters (corresponding to 700℃ Al_2O_3 PDA) or trap#2 parameters (900 ℃ Al_2O_3 PDA). Curves are simulated for two applied V_G. These simulations made with and without traps in Al_2O_3, clearly put in evidence the role of Al_2O_3 defects on the memory charge loss. The improved retention with the 900℃ Al_2O_3 PDA is due to the significant H-related defect density decrease and thus decreased alumina leakage.

3.4.3 Summary

In this section, we used results from quantum calculations as parameters for higher-scale device physical modelling. A very good agreement has been achieved between simulations and data of charge-trap memories with Al_2O_3 with different PDAs. This analysis clearly elucidates the critical role played by the alumina H-related defects (in particular interstitial H, H_{int}) on the charge trap memory retention.

3.5 Conclusion

With the decreasing of the memory dimensions, the integration of SiO_2 interpoly dielectric becomes more and more difficult (see 1.3.1). The envisaged solution, adopted in charge trapping memories, is to use alumina instead of silicon oxide. One of the main issue about the integration of alumina is that, especially at high temperatures, the memory degrades in retention. In particular it has been shown that the leakage current through Al_2O_3, that affects the memory retention characteristic, is related to some trap assisted tunnelling.

In this chapter we used atomistic calculations to find potential defects that could induce electronic levels inside the band gap of alumina and we will linked them with the electrical characterization of TANOS memory and physical-chemical material analysis on alumina single layers.

First we presented the material analysis performed on alumina single layers, annealed with different temperatures. The results show that after a high temperature spike, typical of the source drain activation, alumina crystallized in the (γ-phase).

In the second part atomistic simulations were performed to study electronic structures and formation energies of point defects in γ-alumina. We found that among all the possible defects, the Hydrogen interstitial could be a good candidate to explain both the trap assisted conduction and the correlation between leakage current and PDA temperature.

In the third part we used the results of the two previous sections to simulate the electrical behaviour of TANOS memories proving the critical role played by interstitial Hydrogen defect on the charge trap memory retention.

Chapter 4

Conclusion and perspective

This work focused on the study of innovative stacks and architectures of Split-Gate charge trap memories.

In the **first chapter** the economic context, the evolution and the classification of semiconductors memory was presented. Then the Flash memory operations needed for understanding this thesis were reviewed. We thus presented the flash memory scaling limits and the proposed solutions: we explained the advantages of using a charge trapping layer instead of the continuous floating gate and a high-k control dielectric instead of the classical silicon oxide.

Finally, we introduced the split gate solution. In particular we reported the state of the art of charge trap Split-Gate cell, object of this thesis. This technology integrates in a split-gate structure the new materials that have been presented in the first part of the chapter as a possible solution to the flash memory scaling issues.

In chapter two we studied split-gate charge trap memories with multi-litho architecture. Devices with memory gate length down to 20nm were presented for the first time integrating Si-nc or SiN charge trap layers and alumina or silicon-oxide as control dielectrics.

We have concluded that Si-nc charge trapping layers could be more adapted to high temperatures (>150°C) applications and enable Fowler Nordheim erasing, while SiN offers a wider memory window. Moreover we showed that integrating alumina instead the classical silicon oxide as control dielectric can improve the erasing efficiency but it leads to a faster charge loss during retention operation. These results, although promising, highlight the need for further investigations on the gate stack to satisfy embedded memory requirements.

The study on multi-litho split-gate memory was concluded with the analysis of the cell scaling effects. We showed that the programming consumption decreases with the decreasing of the memory gate length, leading to a program energy <0.1nJ for sub 50nm devices. Moreover the study on the trapped charge location for various memory gate lengths, based on experimental results and TCAD simulations, allowed us to understand the higher erasing efficiency of shorter devices. Finally we pointed out the misalignment issue between the two control gates that leads to a higher programming variability, becoming critical in advanced technological nodes (<40nm).

We thus considered the select gate scaling. Consumption and disturb measurements in scaled devices, indicate that if any junction optimizations is done, in devices with select gate length <50nm, the select transistor loses the ability to control the channel current during program operations.

In last part of the **second chapter** we presented a possible evolution for the split gate architecture: the spacer approach, which is a self align solution. This solution avoid the multi-litho control gate misalignment issue. The first results on spacer technology were presented: SiN memory show good programming/erasing behaviour even if we remarked a lowering of the programming window with respect to multi-litho split gate memory probably related to a bad control of the source/drain implants during the device fabrication process. Indeed, we show how the spacer shape could cause a notable variation on the dopant concentrations and consequently on the memory electrostatic.

In the **third chapter** we present the problem of integrating Al_2O_3 as high-k control dielectric in a planar SONOS memory. In particular, we evidenced that electronic conduction through alumina that affects the memory retention characteristic is probably related to some trap assisted tunnelling. We used atomistic calculations to study the potential Al_2O_3 point defects and hydrogen-related defects that could induce electronic levels inside the band gap of alumina. Results from quantum calculations have been used as parameters for higher-scale device physical modelling. A very good agreement has been achieved between simulations and data of charge-trap memories with Al_2O_3 with different post deposition annealing temperature. Our analysis clearly elucidates the critical role played by the alumina interstitial hydrogen on the charge trap memory retention: with the increasing of the post deposition annealing temperature, the amount of hydrogen decreases and the TANOS retention improves leading to a charge loss of 15% after 10 years of data retention at room temperature.

Perspective

Split gate architecture

Split gate charge trap memories show a promising behaviour: scaling the dimensions is possible and leads to a lowering of the current consumption suitable for low power applications. The program window variability induced by the misalignment of the select and memory gate in the multi-litho approach, especially for sub 40nm node, could be overcome by the spacer architecture. Nevertheless the spacer architecture process is more difficult to control and a particular attention on the doping implantation must be done. In this thesis, we used process simulations to avoid the channel counter-doping and we showed how the spacer shape could influence the memory performances. To improve the reliability of the split-gate memory without adding any critical lithographic step, an interesting solution would be to use a sacrificial gate to self align the second gate.

Split gate materials

Memory gate

We showed the functionality of SiN and Silicon Nano-Crystal charge trapping layers. An improvement of the memory performances can be achieved integrating alumina as control dielectric in the memory gate stack. Unfortunately memories with alumina layer show a fast charge loss. In this thesis we show a possible way to improve the alumina behaviour by avoiding interstitial hydrogen defects through the increasing of the post deposition annealing temperature. However, due to the number of possible defects in alumina that can induce trap assisted conduction, and despite the improvements described above, alumina is not ready to be integrate on memories for embedded applications that require a good data retention at high temperature (>150 °C). The integration of alumina on embedded memories, at least for smart cards, where requirements are less aggressive in terms of temperature, requires the study of further improvements (new materials, doping, interface optimisation..).

Select gate

Split gate memories are built in "select first" configuration, meaning that the select gate is processed first. In order to be compatible with the CMOS transistor of the logic, the integration of a high-k metal-gate based select gate transistor must be studied for the sub-40nm CMOS technology node, evaluating the advantages (SCE,DIBL ..) and drawbacks (gate leakage current), of this configuration.

Chapitre 5

Résumé du travail de thèse en français

5.1 Présentation de la Thèse

Ce travail de thèse concerne l'étude d'architectures et d'empilements innovants de mémoires Split-Gate (grille séparée) à couche de piégeage discret.

Dans le premier chapitre, nous présenterons le contexte économique, l'évolution et le fonctionnement des mémoires flash EEPROM. Ensuite, une description détaillée de la technologie, son fonctionnement et ses limites d'échelle seront fournis. Enfin, nous allons exposer les solutions possibles pour éviter ces problèmes et le cadre de la thèse.

Le deuxième chapitre présentera les mémoires Split-Gate avec *multi-litho* architecture et une longueur de grille allant jusqu'à 20 nm. Nous montrerons les résultats de mémoires split-gate à base nanocristaux de Silicium (Si-ncs), nitrure de silicium (Si_3N_4) et hybrides Si-nc/SiN, avec SiO_2 ou Al_2O_3 comme diélectrique de contrôle. Ensuite, nous présenterons l'étude de l'impact de la réduction des dimensions des mémoires split-gate à piégeage de charge, sur la fenêtre de programmation, la rétention et de la consommation.

Nous présenterons donc une solution possible pour surmonter les problèmes de l'approche *multi-litho* : la technologie *Spacer*. Son schéma d'intégration et l'influence des paramètres du procédé sur la fenêtre de programmation seront présenté. Ensuite, les résultats électriques sur des mémoires à base nanocristaux de silicium (Si-ncs) et nitrure de silicium (Si_3N_4) seront montrés.

Dans le troisième chapitre, après l'introduction de la mémoire TANOS qui emploie l'Al_2O_3 comme diélectrique de contrôle et le nitrure de silicium (SiN) comme couche de piégeage de charge (CTL), nous utiliserons des calculs atomistiques pour trouver des défauts potentiels qui pourraient induire des niveaux électroniques à l'intérieur du *Band Gap* de l'alumine. Nous les relierons avec le comportement électrique de la mémoire TANOS pendant la rétention et l'analyse physico-chimique des matériaux sur des simples couches d'alumine.

Le manuscrit termine avec une conclusion générale qui résume les différents résultats obtenus dans ce travail de thèse, avant de proposer quelques perspectives.

5.2 Introduction

Contexte économique

L'industrie des semiconducteurs naît en 1947 lorsque le premier transistor est inventé aux *Bell labs* (Etats-Unis). Seulement 60 ans plus tard, l'industrie des circuits intégrés (IC) totalise revenus pour plus de 250 milliards de dollars ($\sim$ 0.5 % du PIB mondial).

Même si le coût d'un seul dispositif a diminué de 50 % chaque année en suivant la loi de Moore, les revenus de l'industrie d'IC ont augmenté à un taux annuel moyen de 17% entre le 1970 et le 2008. Au cours de ces années, la demande en produits IC a été alimentée par des les nouvelles technologies introduite sur le marché (IC, les PC, les téléphones cellulaires, les smartphone, tablets ..) et il a donc été influencée par fluctuations dans la demande des produits (figure 5.1).

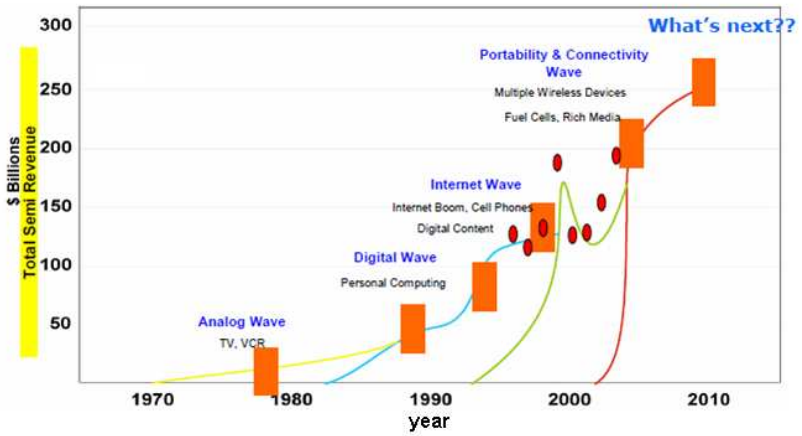


Figure 5.1 – Le cycle de l'industrie IC [4]

Dans ce scénario, les mémoires jouent un rôle important : ils comptent pour environ un quart du total vendu sur le marché IC.

Le mémoire «idéal» doit être une mémoire qui conserve les informations stockées même quand elle n'est pas alimentée (non volatile), avec une haute densité d'intégration, qui peut être infiniment écrite/ ré-écrite (endurance illimitée), avec une grand vitesse de programmation/effacement /opération de lecture, une faible consommation d'énergie et à bas prix. Puisque le dispositif idéal n'existe pas, différents types de mémoires ont été étudiées afin de développer une ou plusieurs de ces propriétés en fonction de l'application finale.

Présentation des mémoires split-gate a stockage discrète de charge

La mémoire à grille separée (split-gate) a été inventé pour les applications embarquées de faible puissance dans les années 90. Cette mémoire a été introduit afin d'augmenter

- l'efficacité d'injection
- l'efficacité effacement
- l'immunité au phénomène de *disturb*

comme indiqué dans le tableau 5.1 les mémoires split-gate peuvent avoir des géométries différentes. L'idée commune est d'ajouter à la mémoire flash un transistor d'accès séparé. Le transistor d'accès, appelé aussi transistor de sélection, commande le courant qui circule dans la mémoire au cours des opérations de programmation/effacement. Ceci implique un abaissement de la consommation de courant qui rend les mémoires SG appropriées pour des applications de faible puissance. En outre, elle augmente l'immunité au disturb dû au fait que, lorsque le transistor d'accès est fermée, aucun courant ne circule à travers la cellule.

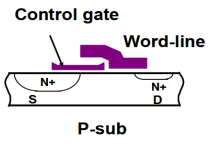
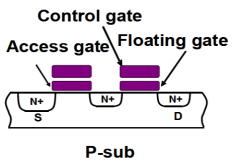
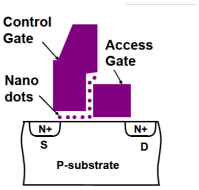
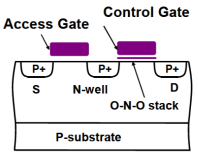
Type	1.5Tr cell (SuperFlashTM)	2Tr cell	1.5Tr	2Tr SONOS(PMOS)
Programme	SSI	FN	SSI	CHE
Effacement	FN (poly-poly)	FN (poly-sub)	FN	FN
Dispositif structure				
Avantage	Fast program	Low power P/E	Fast, low-power program	Low power P/E
Publications	[55]	[56]	[56]	[57]

TABLE 5.1 – Memoires split-gate [9]

Programmation Les cellules Split-gate sont programmées par électrons chauds (CHE). Ceci est fait en gardant le substrat et l'électrode à côté de la grille de sélection (ci-après appelé *Source*) à la masse et en appliquant une tension positive sur la grille de sélection, la mémoire et la Source (voir fig. 5.2). Les électrons sont accélérés par le fort champ électrique horizontal induite par un fort voltage appliqué sur la source, et ils sont injecté dans l'empilement de la mémoire grâce au champ vertical dû à la tension positive appliquée sur la grille de la mémoire. L'injection vient à deux points : au *pinch-off* du canal, correspondant à la région entre la grille de sélection et la grille de mémoire, et à la jonction de source où le champ électrique latéral est le plus élevé. Dans la mémoire SG, à la différence de la cellule flash classique, le courant pendant la programmation CHE est efficacement contrôlé par la grille de sélection permettant une réduction de la consommation du courant pendant la programmation CHE (voir fig. 5.3.).

effacement L'efficacité d'effacement est l'un des principaux défis dans la mémoire split-gate. En fonction de la géométrie et de l'empilement de la mémoire, elle peut être effacée par :

HHI cette méthode, dans la suite également appelé injection côté source (SSI), est rapide mais a besoin d'une haute tension sur l'électrode de source (fig. 5.4-a) qui peuvent induire

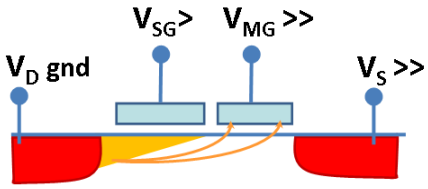


Figure 5.2 – Split-gate CHE schématique montrant les deux points d'injection préférentiels : à la limite entre les deux grilles et à la jonction de la Source.

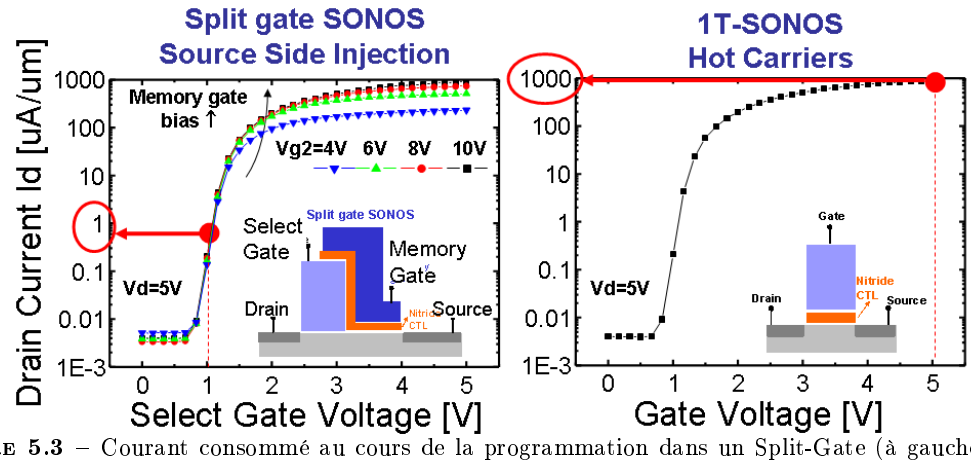


Figure 5.3 – Courant consommé au cours de la programmation dans un Split-Gate (à gauche) et dans une memoire plane (à droite)

phénomènes de "disturb" et des courants parasites. Par ailleurs le point d'injection de trous chauds est localisé du côté Source provoquant un déséquilibre dans la répartition spatiale entre les électrons piégés et la population de trous qui est un problème grave affectant la résistance périphérique et la rétention [58, 59].

FN permet d'effacer un uniforme de la cellule de mémoire, mais il est plus lent que l'HHI. Les mémoires split-gate avec grille flottant en polysilicium sont effacés à l'aide de FN tunnel (fig. 5.4-b). Dans ce cas, une pointe à grille flottante est utilisée comme un injecteur accrue et des tensions plus faibles sont nécessaires. Dans les memoires split-gate à piège des charges discret cette solution n'est pas possible et le tunnel FN se fait à travers le diélectrique de contrôle ou l'oxyde tunnel (fig. 5.4-c).

Architecture La figure 5.5 montre que parmi toutes les architectures possibles les deux principalement utilisées sont l'auto-alignée SPACER et le multi-lithographie. L'un des problèmes majeurs de l'approche multi-lithographie concerne le décalage entre les deux grilles. Pour résoudre ce problème, cdes solutions auto-alignée adaptées ont été proposées. En particulier l'approche SPACER semble être la mieux adaptée pour les technologies avec des petites dimensions. En effet, dans cette solution, il faut remarquer que la seconde grille de contrôle ne nécessite aucune étape de lithographie pour définir le SPACER, seulement un masque non critique est utilisé pour enlever le spacer adjacente au transistor de sélection. La structure et les principaux avantages et inconvénients de ces deux architectures sont présentées dans la tableau : 5.2.

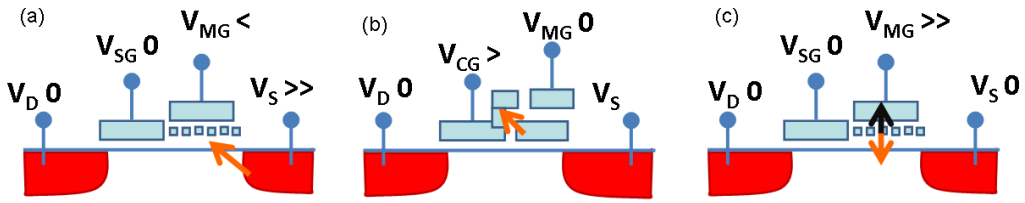


FIGURE 5.4 – Schéma de l'effacement Split-Gate (a) d'injection de trous chauds (b) poly-poly FN à effet tunnel (c) FN tunnel à travers les oxydes de mémoire.

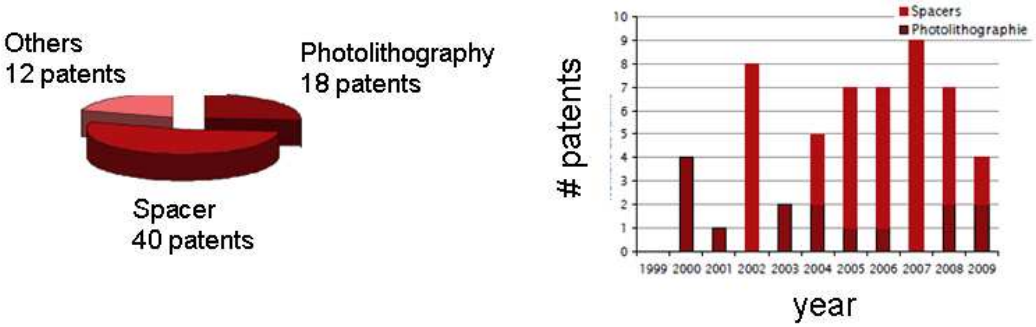


FIGURE 5.5 – Étude sur les brevets : l'architecture [61]

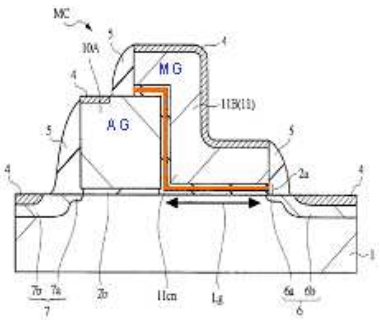
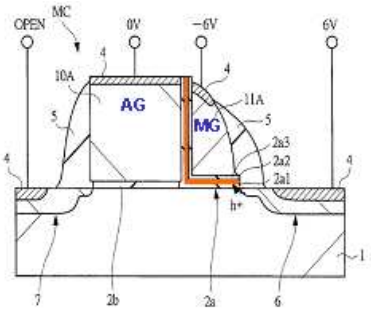
	MultiLitho	**Spacer**
		
Advantages	· Less critical masks (2nd gate length mainly controlled by poly deposition)	· Self-aligned, no alignment issue
Issues	· Misalignment directly leads to gate length variation (ΔVt variability)	· Memory Gate etching : technological challenge

TABLE 5.2 – Comparaison entre Spacer et multi-litho architecture.

5.3 Mémoires split-gate

Fonctionnement des mémoires Split-Gate à piégeage de charge

Les memoires split-gate a piégeage de charge ont été traitées avec une configuration "mémoire dernière", qui signifie que la grille de la mémoire (MG) est déposée sur la grille de sélection (SG). La lithographie E-beam a été utilisé pour définir la grille de sélection (L_{SG}) jusqu'à 40nm. La longueur électrique de la grille mémoire est déterminée par la couche de poly-silicium recouvrant le canal de la mémoire (Fig. 5.6-gauche) permettant d'obtenir une longueur de grille mémoire de 20nm (Fig. 5.6-droite). Dans ce qui suit, L_{MG} se réfère à la longueur électrique.

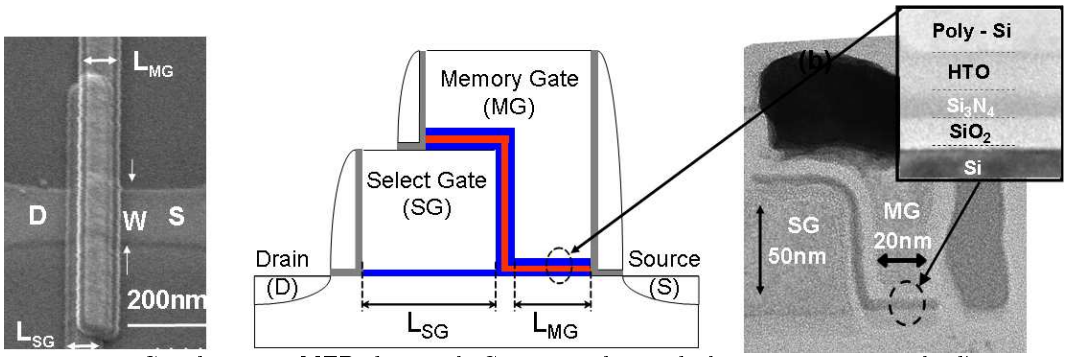


Figure 5.6 – Gauche : vue MEB du canal. Centre : schéma de la coupe transversale d'une mémoire split-gate. Droite : images TEM d'un mémoire de 20nm avec SiN comme couche de piégeage

$I_D(V_{SG},V_{MG})$ caractéristiques Dans la mémoire split-gate la grille de sélection permet de contrôler le courant dans le canal. Dans la figure 5.7 nous pouvons voir les courbes caractéristiques $I_D(V_{SG},V_{MG})$ mesurées pour des dispositifs avec une longueur de grille mémoire de 20nm et 70nm. Dans la mémoire de 20nm, on peut noter une diminution de la tension de seuil V_{TH} avec l'augmentation de V_{SG} en raison d'un contrôle parasite du transistor de sélection sur le canal de la mémoire. Malgré ce problème, le courant du canal peut être efficacement contrôlé par la grille de sélection.

Impact de la couche de piégeage sur les performances de la mémoire.

Dans cette section, nous allons étudier l'impact de la couche de piégeage sur les performances de la mémoire. Cela a été fait en intégrant dans l'empilement de grille mémoire différentes couche de piégeage (CTL)[19] avec Si-nc (échantillon A), SiN (B), et hybride Si-nc/SiN (C). De plus, nous avons intégré comme diélectriques de contrôle HTO ou AlO (échantillon D).

Dans le tableau 5.3 les détails techniques des différent empilements sont montré, tandis que l'image TEM de leur section transversale (figure 5.8). Les échantillons avec des Si-ncs ont été analysés en mode de "énergie filtrée" lorsque Si est sélectionné.

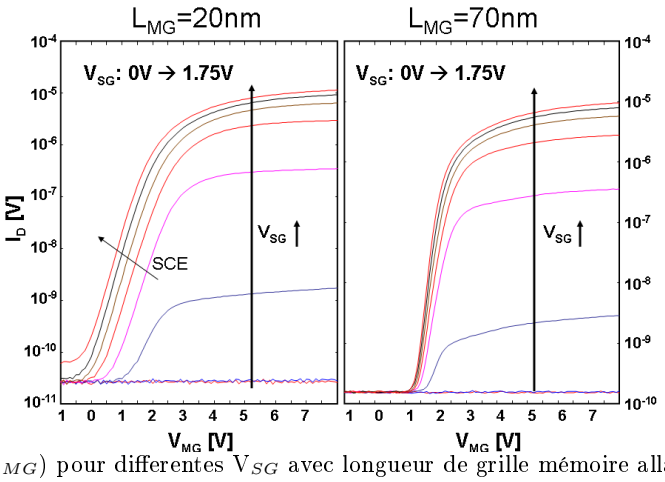


FIGURE 5.7 – $I_D(V_{MG})$ pour differentes V_{SG} avec longueur de grille mémoire allant de 20nm à 70nm

	Sample A	Sample B	Sample C	Sample D
Tunnel Oxide	SiO_2 (5nm)	SiO_2 (5nm)	SiO_2 (5nm)	SiO_2 (4nm)
Charge trapping layer	Si-ncs ($\Phi \sim$ 6nm)	SiN (6nm)	Si-ncs+SiN (3nm)	Si-ncs+SiN (3nm)
Control dielectrics	HTO (8nm)	HTO (10nm)	HTO (10nm)	HTO (3nm) Al_2O_3(8nm)
Control Gate	Poly-Si	Poly-Si	Poly-Si	TiN

TABLE 5.3 – Détails technologiques des memoires splitgate

Programmation

Les mémoires split-gate sont programmées avec CHE. Le potentiel de la grille de sélection est réglé à 1V afin d'être proches du régime de seuil ($I_S \sim 10\ \mu$ A).

La figure 5.9 montre les caractéristiques de programmation des mémoires ayant comme CTL Si_3N_4, Si-ncs, et hybride SiN/ncs pour divers programmes V_{MG} et V_S. La longueur de grille de la mémoire est 40nm. En raison d'une plus grande densité de sites de piégeage, les mémoires à base de nitrure présentent une plus grand fenêtre de programmation. En particulier, pour une impulsion de programmation de 10 mus avec V_{MG}=10V et V_S=3.5V, le décalage de la tension de seuil est d'environ 3V, alors que les mémoires avec Si-nc CTL présentent une fenêtre de programmation de seulement 1.25V. De plus la figure 5.9 montre que les mémoires ayant comme CTL hybride Si-ncs/SiN offrent une amélioration de la fenêtre de programmation en permettant de compenser la réduction de $\Delta\ V_{TH}$ typique des mémoires avec un CTL en Si-nc.

Ces résultats sont résumés sur le côté droit de la figure 5.9. Dans le diagramme à barres, nous montrons la fenêtre de programmation pour des émpilements différentes après un temps de programmation de 20 μs et une condition de programmation donné. On peut voir que la mémoire avec un CTL en SiN a la fenêtre de programmations la plus élevée. Au contraire, les mémoires avec Si-nc présentent la plus faible ΔV_{TH}. Alors que les mémoires hybrides Si-ncs/SiN présente un comportement intermédiaire.

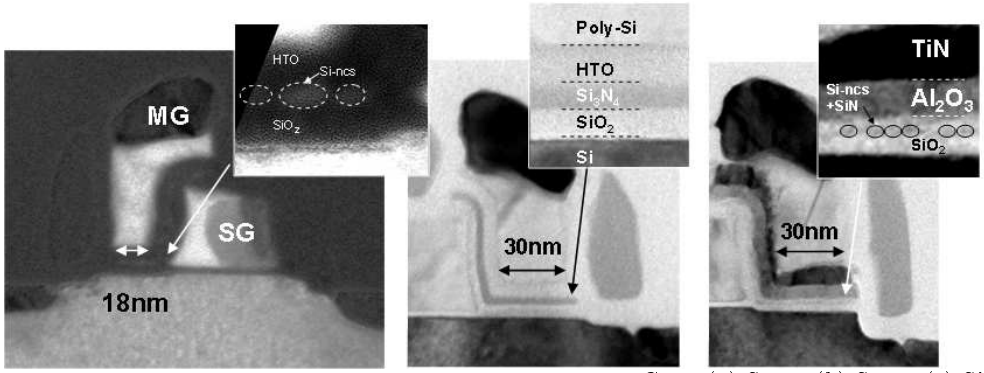


Figure 5.8 – Image TEM d'une memoire split-gate avec comme CTL, (a) Si-nc, (b) Si_3N_4 (c) Si-nc/SiN.

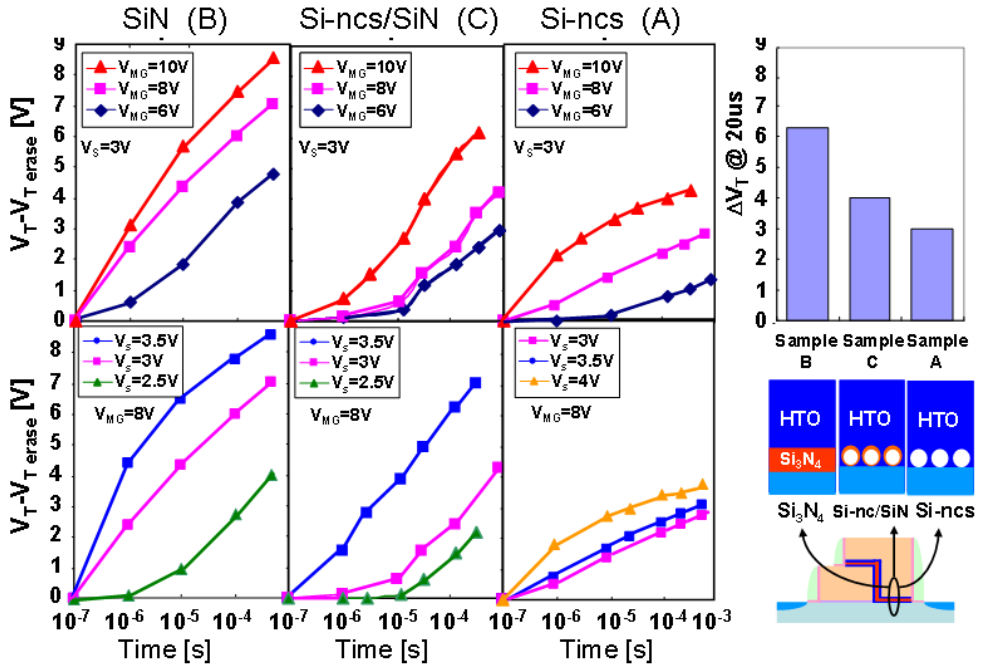


Figure 5.9 – Gauche : fenêtre de programmation des trois échantillons (Si_3N_4), (Si-nc/SiN), (Si-nc) avec une longueur de grille de 30nm. Droite : fenêtre de programmation pour les mémoires avec Si-nc, Si-nc/SiN et Si_3N_4 avec une condition de programmation (V_{MG}=10V, V_S=3V, t=20μs)

Effacement

Diverses méthodes d'effacement ont été utilisées en fonction de la nature de la couche de piégeage du diélectrique de contrôle (figure 5.10) :

Nitrure Les mémoires à base de SiN sont effacées à l'aide d'injection de trous chauds (HHI), cette méthode permet d'effacer rapidement, mais souffre d'une consommation élevée de courant. En outre, la haute tension nécessaire sur l'électrode de source pourrait induire des phénomènes de *disturb* en particulier lorsque les dimensions de la mémoire diminues.

Si-ncs Les mémoires intégrant **Si-ncs** comme CTL et **HTO** comme diélectrique de contrôle sont

effacées par Fowler-Nordheim (FN) depuis l'oxyde supérieure. Ce procédé d'effacement est plus lent que HCI, mais aucun courant ne circule dans le canal vu que les électrodes de drain et source sont maintenues à la masse. La haute tension appliquée sur l'électrode de grille crée une différence de potentiel élevée entre la grille de sélection et la grille de mémoire qui peut causer des dommages irréversibles.

Si-nc Les mémoires avec **high-k diélectriques** sont effacées par FN à travers l'oxyde de tunnel. Le high-k, a comme effet une amélioration du champ électrique dans l'oxyde tunnel et donc une vitesse d'effacement de la mémoire plus rapide.

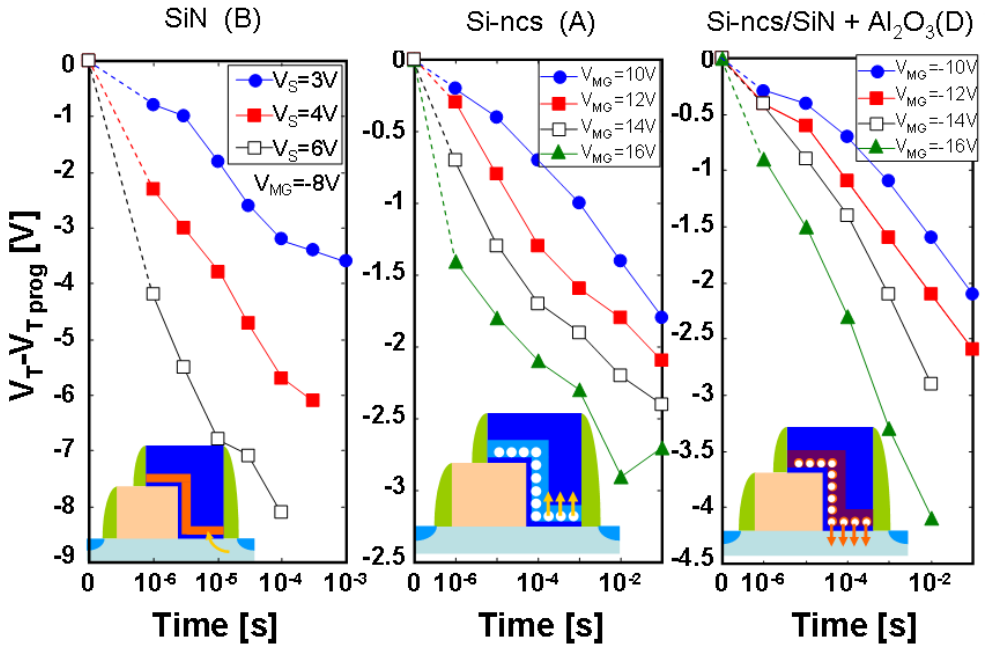


Figure 5.10 – Caractéristiques d'effacement des mémoires avec différents mécanismes d'effacement : injection de trous chaud (échantillon B : Si_3N_4), FN à travers le diélectrique supérieur (échantillon A : Si-ncs) et FN à travers le diélectrique inférieur (échantillon D :Si-ncs avec l'oxyde high-k)

Diminution de la longueur de la grille de sélection

Dans cette section, nous étudions l'impact de la réduction de la longueur des grilles de sélection et mémoire sur le fonctionnement de la mémoire.

La structure multi-lithographie a permis de fabriquer des dispositifs avec une longueur de grille de sélection allant de 350nm jusqu'à 40nm et de grille mémoire allant de 350 jusqu'à 20nm qui est en 2012 le plus petit dispositif à grille séparée présent dans le monde. Étudier les conséquences de la diminution des dimensions est donc cruciale pour le développement de la technologie à grille séparée.

Cet étude a été effectuée sur des mémoires SiN. Ce choix a été motivé par la grande fenêtre de programmation qui permet une meilleure compréhension du fonctionnement de la mémoire. Les résultats expérimentaux ont été expliqués par la simulation TCAD.

Programmation et consommation

L'effet de diminution de la grille de sélection sur la consommation et la programmation a été étudiée mesurant l'abaissement de la tension de seuil de la grille de sélection ainsi que la fenêtre de programmation.

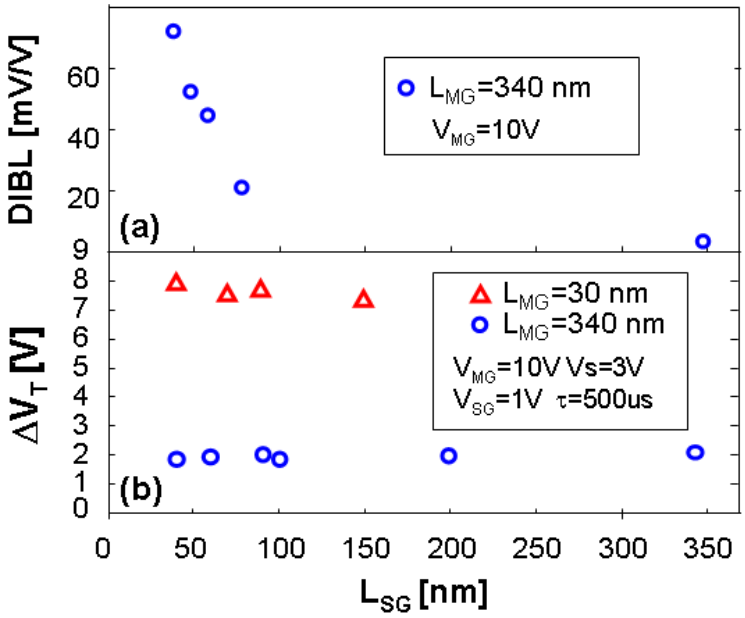


Figure 5.11 – DIBL effect (a) et fenêtre de programmation (b) en function de L_{SG}

Fig. 5.11-a montre que la fenêtre de programmation reste inchangée avec la diminution de L_{SG}, mais la tension de seuil de la grille de sélection diminue en raison du DIBL. Cela provoque l'effet parasite d'une augmentation de la consommation de courant pendant la programmation.

Diminution de la dimension de la grille de la mémoire

Dans la section précédente, nous avons décrit l'effet de la diminution de la dimension de la grille de sélection. Ici, nous nous concentrons sur la diminution de la grille de la mémoire. Nous verrons que la diminution de la grille de la mémoire améliore l'efficacité de la programmation et l'effacement sans augmenter la consommation de courant. D'autre part, la diminution de L_{MG} peut augmenter la variabilité du ΔV_{TH} due à une variation de longueur de grille directement lié à un mauvais contrôle de l'alignement pendant le processus de la lithographie.

Programmation

La figure 5.12-a montre la fenêtre de programmation après une impulsion de programmation de 500μs (V_{MG} =10V ; V_S=3V ; V_{SG}=1V) en fonction de la longueur de grille de mémoire. Avec la diminution des dimensions de la mémoire, la fenêtre de programmation augmente fortement.

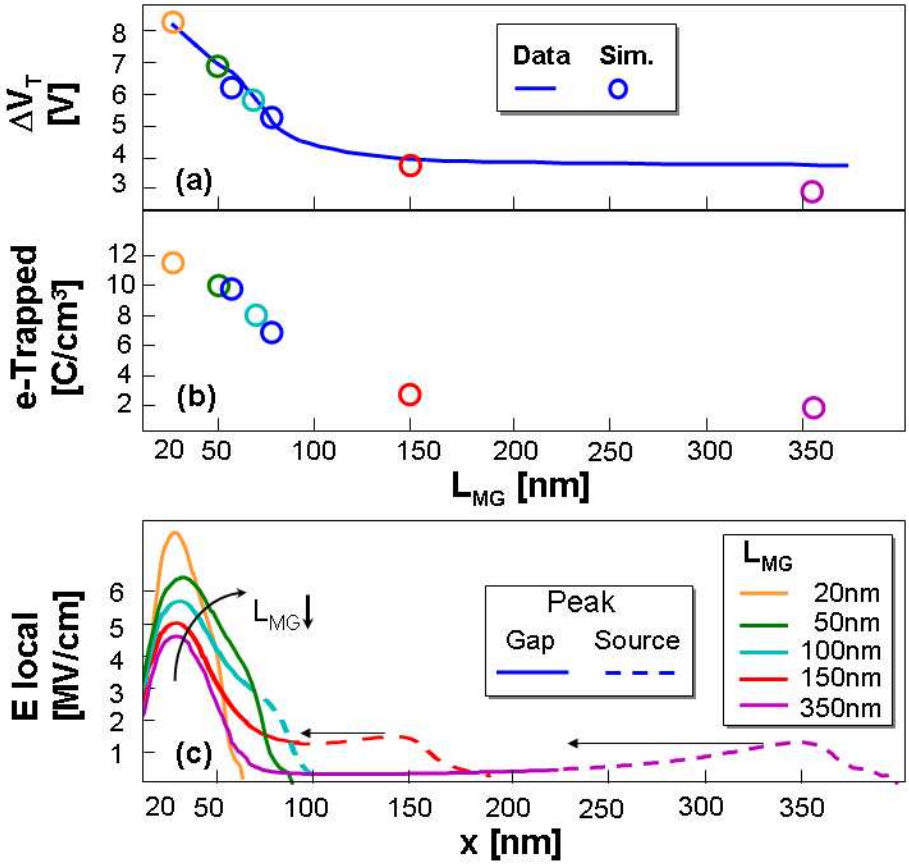


FIGURE 5.12 – a) Fenêtres de programmation mesurées et simulées de mémoires split-gate pour différentes longueurs de grille (V_{MG}=10V ; V_S=3V ; V_{SG}=1V, t=500μs). b) Charge piégée dans la couche de SiN. c) Champ électrique local dans le canal lors de la programmation par injection du côté source. Deux pics apparaissent, l'un près du gap, l'autre près de l'électrode de source. Dans les dispositifs courts, les deux pics se superposent

Effacement

Pendant l'écriture, dans les dispositifs longs la charge est injectée à proximité du *gap* et la concentration maximale de charge piégée se déplace progressivement vers l'électrode de source dans les dispositifs plus courts.

Pendant l'effacement HHI, les trous chauds sont générés à la jonction de la source puis injecté dans la région de SiN près de lui. Cela implique que dans les dispositifs longues, comme schématisé sur la figure 5.13-droite, se produit un décalage entre les trous injectés et la population électrons piégés qui empêche une effacement complète de la cellule [82, 83, 84, 80]. Au contraire, dans les dispositifs avec une longueur de grille de mémoire petite, en raison de la courte distance entre le *gap* et la source, les trous lors de l'effacement HHI et les électrons sont injectés dans le même lieu, et donc la cellule peut être complètement effacée .

Afin de vérifier cette hypothèse, nous avons calculé, pour des mémoires avec différents L_{MG}, le pourcentage de la charge effacée après une même série de programmation (V_D = 3V L_{MG}=10V, V_{SG} =1V t = 500μs) et un effacement de (V_D = 5 V_{MG}=-10V V_{SG}=0V, t=500μs).

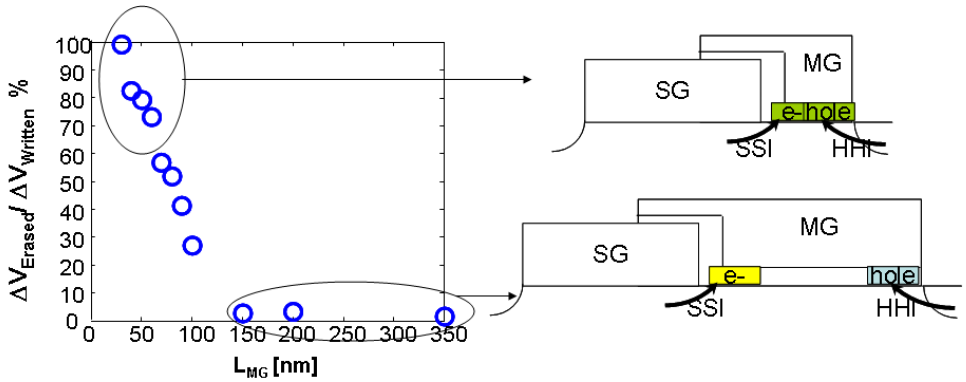


Figure 5.13 – Efficacité d'effacement par rapport à la longueur de grille.

Consommation

Pour analyser de l'effet de la réduction de dimension de la mémoire de la mise à l'échelle sur la consommation de la programmation, nous avons mesuré pour des longueurs de grille de mémoire allant de 350nm jusqu'à 20 nm, les caractéristiques de programmation et la consommation de courant en fonction de la durée de programmation (Fig.5.14). L'énergie consommée est calculée comme l'intégrale sur le temps de programmation du canal courant multiplié par la tension de source appliquée :

$$ENERGY = \int_0^{time} V_S I_S(t) dt \tag{5.1}$$

Dans les mémoires split-gate, lors de la programmation, le transistor de mémoire est activé, et le canal e donc la courant est contrôlé par la tension de la grille d'accès. Par conséquent, le courant de programmation I_S pour un V_{SG} et V_S donné est constante ! l'énergie ne dépend que du temps de programmation (voir l'équation 5.1 et la figure 5.14-b). Dans les dispositifs à courte grille de mémoire, l'efficacité de la programmation est plus élevée : temps de programmation plus courts sont suffisants pour atteindre un Δ V_{TH} donné (Fig. 5.14-a), et une énergie donc une faible energie de programmation (Fig. 5.14-b).

Dans la figure 5.15 nous avons tracé le courant consommé pour atteindre une fenêtre de programmation proposée de 3.5V en fonction de la longueur de la grille mémoire. Le temps de programmation nécessaire est extrapolée à partir de la figure 5.14-a. Les résultats montrent une amélioration de plus de 10 fois de l'énergie de consommation lorsque la longueur de grille mémoire passe de 100nm à 240nm. En particulier, pour les dispositifs de longueur de grille sous-40nm , une énergie de programmation $<$0.1nJ est atteinte, et convient pour des applications de faible puissance.

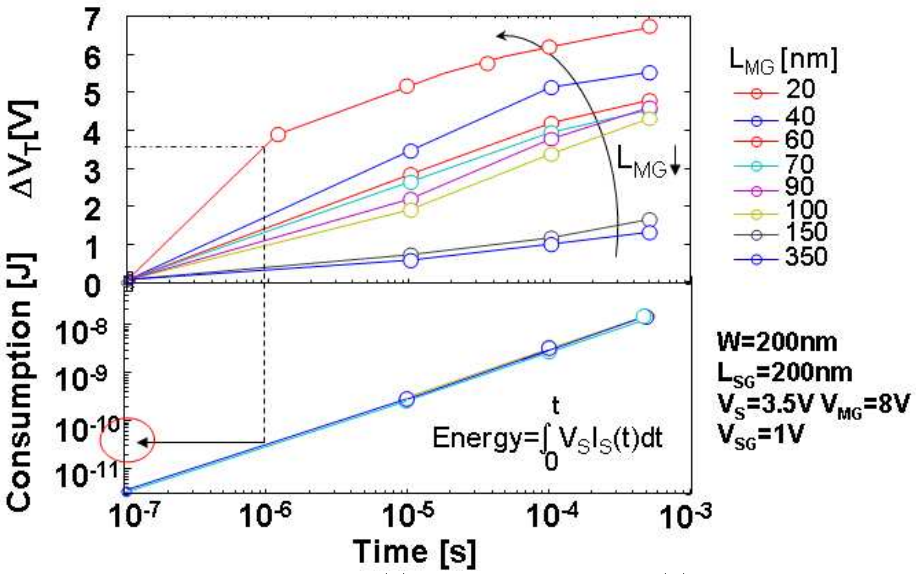


Figure 5.14 – Fênetre de programmation (a) et courant consommée (b) pour des dispositifs avec différentes longueurs de grilles de mémoire

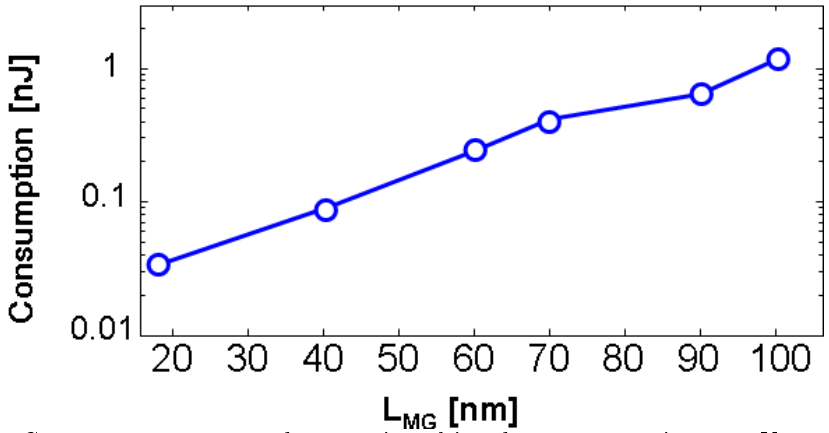


Figure 5.15 – Courant consommé pendant une impulsion de programmation avec V_S=3.5V ; V_{MG}=8V pour obtenir une fenêtre de programmation de 3.5V, en fonction de la longueur de la grille mémoire

Variabilité

Dans la figure 5.16 à gauche, nous montrons la fenêtre de programmation en fonction de la longueur de grille mémoire. Dans les dernières années, la diminution des dimensions des dispositifs a été beaucoup plus agressives que les améliorations de résolution du processus de lithographie. En particulier pour les mémoires split-gate avec architecture muli-lithographie, le défaut d'alignement entre la grille de sélection et la grille mémoire provoque unr variation des dimensions de la mémoire qui induit un décalage non négligeable de la fenêtre de programmation. La variation de la fenêtre de programmation ($\sigma_{\Delta V_{TH}}$) en raison d'une variation de la longueur de la grille de mémoire ($\sigma_{L_{MG}}$) peut s'écrire, en première approximation, en tant que :

$$\Delta V_{TH} = f(L_{MG}) - - > \sigma_{\Delta V_{TH}} = \frac{\partial f(L_{MG})}{\partial L_{MG}} \sigma_{L_{MG}} \tag{5.2}$$

Pour quantifier cette variation, nous avons d'abord calculé les coefficients a, b, c, d, e d'un polynôme de quatrième ordre qui reproduit les données expérimentales $\Delta V_{TH} = aL_{MG}^4 + bL_{MG}^3 + cL_{MG}^2 + dL_{MG}^1 + e$, puis nous l'avons dérivé pour calculer le décalage de la tension de seuil dû à une variation de la longueur de mémoire :

$$\frac{\sigma_{\Delta V_{TH}}}{\sigma_{L_{MG}}} = 4aL_{MG}^3 + 3bL_{MG}^2 + 2cL_{MG} + d \tag{5.3}$$

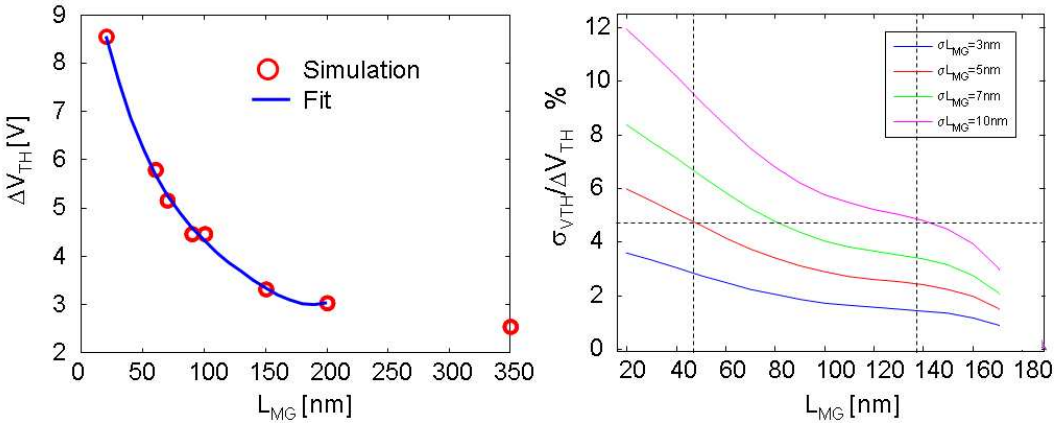


Figure 5.16 – Left : programming windows versus memory gate lenght simulations and its fitting. Right : Variability due to L_{MG} variation for different lithography process

La figure 5.16 montre la variabilité en fonction du procédé de lithographie. Le choix de la dimension de grille de mémoire doit être fait avec soin : par exemple pour garantir un variabilité de Δ V_{TH} inférieure à 5 %, la longueur de grille de mémoire doit être

- supérieure à 110nm avec une erreur dans l'overlay de la lithographie de $\pm$ 10nm typique de lithographie DUV 193 nm .
- supérieure à 80nm avec une superposition de la lithographie de pm 7 nm typique d'immersion DUV 193 nm.
- supérieure à 40nm avec $\pm$ 5 nm (193 nm état DUV).
- avec un recouvrement de $\pm$ 3nm (extrême UV) la variation de la tension de seuil est toujours inférieure à 5%.

5.4 Conclusions générales

Du fait de l'augmentation de la demande de produits pour les applications grand public, industrielles et automobiles, des mémoires embarquées fiables et à faible coût de fabrication sont de plus en plus demandées.

Dans ce contexte, les mémoires split-gate à piégeage discret sont proposées pour des microcontrôleurs. Elles combinent l'avantage d'une couche de stockage discrète et de la configuration split-gate.

Durant ce travail de recherche, des mémoires split-gate à couche de piégeage discret ayant des longueurs de grille de 20nm sont présentées pour la première fois. Celles-ci on été réalisées avec des nanocristaux de silicium (Si-nc), du nitrure de silicium (SiN) ou un hybride Si-nc/SiN avec diélectrique de control de type SiO2 ou AlO et sont comparées en termes de performances lors des procédures d'effacement et de rétention.

Le rôle des défauts dans le diélectrique de contrôle (alumine) a enfin été étudié. Nous avons montré que la concentration de pièges dans AlO pouvait être réduite par ajustement des conditions de procédé de fabrication, débouchant ainsi sur l'amélioration de la rétention dans les mémoires à piégeage de charge.

dans les mémoires à piégeage de charge.

Bibliography

[1] R. Strenz. Embedded flash technologies and their applications: Status amp; outlook. In *Electron Devices Meeting (IEDM), 2011 IEEE International*, pages 9.4.1 –9.4.4, dec. 2011.

[2] K. Baker. Embedded nonvolatile memories: A key enabler for distributed intelligence. In *Memory Workshop (IMW), 2012 4th IEEE International*, pages 1 –4, may 2012.

[3] Y. Yano. Take the expressway to go greener. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 24 –30, feb. 2012.

[4] IC Insigths, editor. *The McClean Report 2007 Edition: An In-Depth Analysis and forecast of the Integrated Circuit Industry.* http://www.icinsights.com/prodsrvs/mcclean/mccr2007.pdf, 2007.

[5] Craig Zajac. Choose the right non-volatile memory ip, 2010.

[6] D. Frohman-Bentchkowsky. A fully-decoded 2048-bit electrically-programmable mos rom. In *Solid-State Circuits Conference. Digest of Technical Papers. 1971 IEEE International*, volume XIV, pages 80 – 81, feb 1971.

[7] B. De Salvo. Paths of innovation in silicon non-volatile memories. Technical report, 2007.

[8] Process integration, devices, and structures. Technical report, INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS, 2009.

[9] Hideto Hidaka. Evolution of embedded flash memory technology for mcu. *IC Design and Technology (ICICDT), 2011 IEEE International Conference on*, 2011.

[10] P. Pavan, R. Bez, P. Olivo, and E. Zanoni. Flash memory cells-an overview. *Proceedings of the IEEE*, 85(8):1248 –1271, aug 1997.

[11] Ling-Chang Hu, An-Chi Kang, J.R. Shih, Yao-Feng Lin, K. Wu, and Ya-Chin King. Statistical modeling for postcycling data retention of split-gate flash memories. *Device and Materials Reliability, IEEE Transactions on*, 6(1):60 – 66, march 2006.

[12] A. Hoefler, J.M. Higman, T. Harp, and P.J. Kuhn. Statistical modeling of the program/erase cycling acceleration of low temperature data retention in floating gate nonvolatile memories. In *Reliability Physics Symposium Proceedings, 2002. 40th Annual*, pages 21 – 25, 2002.

[13] H.P. Belgal, N. Righos, I. Kalastirsky, J.J. Peterson, R. Shiner, and N. Mielke. A new reliability model for post-cycling charge retention of flash memories. In *Reliability Physics Symposium Proceedings, 2002. 40th Annual*, pages 7 – 20, 2002.

[14] M. Kato, N. Miyamoto, H. Kume, A. Satoh, T. Adachi, M. Ushiyama, and K. Kimura. Read-disturb degradation mechanism due to electron trapping in the tunnel oxide for low-voltage flash memories. In *Electron Devices Meeting, 1994. IEDM '94. Technical Digest., International*, pages 45 –48, dec 1994.

[15] R. Yamada, Y. Mori, Y. Okuyama, J. Yugami, T. Nishimoto, and H. Kume. Analysis of detrap current due to oxide traps to improve flash memory retention. In *Reliability Physics Symposium, 2000. Proceedings. 38th Annual 2000 IEEE International*, pages 200 –204, 2000.

[16] A. Chimenton, P. Pellati, and P. Olivo. Analysis of erratic bits in flash memories. In *Reliability Physics Symposium, 2001. Proceedings. 39th Annual. 2001 IEEE International*, pages 17 –22, 2001.

[17] L.D. Yau. Simple i/v model for short-channel i.g.f.e.t.s in the triode region. *Electronics Letters*, 11(2):44 –45, 23 1975.

[18] Yannis Tsividis. *Operation and Modeling of the MOS Transistor*. Oxford University Press, USA, 2 edition, June 2003.

[19] G. Molas, M. Bocquet, J. Buckley, J.P. Colonna, L. Masarotto, H. Grampeix, F. Martin, V. Vidal, A. Toffoli, P. Brianceau, L. Vermande, P. Scheiblin, M. Gely, A.M. Papon, G. Auvert, L. Perniola, C. Licitra, T. Veyron, N. Rochat, C. Bongiorno, S. Lombardo, B. De Salvo, and S. Deleonibus. Thorough investigation of si-nanocrystal memories with high-k interpoly dielectrics for sub-45nm node flash nand applications. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pages 453 –456, dec. 2007.

[20] G. Molas, D. Deleruyelle, B. De Salvo, G. Ghibaudo, M. GelyGely, L. Perniola, D. Lafond, and S. Deleonibus. Degradation of floating-gate memory reliability by few electron phenomena. *Electron Devices, IEEE Transactions on*, 53(10):2610 –2619, oct. 2006.

[21] Kinam Kim. Technology for sub-50nm dram and nand flash manufacturing. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 323 –326, dec. 2005.

[22] K. Prall and K. Parat. 25nm 64gb mlc nand technology and scaling challenges invited paper. In *Electron Devices Meeting (IEDM), 2010 IEEE International*, pages 5.2.1 –5.2.4, dec. 2010.

[23] D.J. Frank, Y. Taur, M. Ieong, and H.-S.P. Wong. Monte carlo modeling of threshold variation due to dopant fluctuations. In *VLSI Technology, 1999. Digest of Technical Papers. 1999 Symposium on*, pages 169 –170, 1999.

[24] Hang-Ting Lue, Szu-Yu Wang, Erh-Kun Lai, Yen-Hao Shih, Sheng-Chih Lai, Ling-Wu Yang, Kuang-Chao Chen, J. Ku, Kuang-Yeu Hsieh, Rich Liu, and Chih-Yuan Lu. Besonos: A bandgap engineered sonos with excellent performance and reliability. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 547 –550, dec. 2005.

[25] P. Blomme and J. Van Houdt. Scalability of fully planar nand flash memory arrays below 45nm. In *Memory Workshop, 2009. IMW '09. IEEE International*, pages 1 –2, may 2009.

[26] G. Molas, J.P. Colonna, R. Kies, D. Belhachemi, M. Bocquet, M. Gee andly, V. Vidal, P. Brianceau, L. Vandroux, G. Ghibaudo, and B. De Salvo. Investigation of charge-trap memories with aln based band engineered storage layers. In *Memory Workshop (IMW), 2010 IEEE International*, pages 1 –4, may 2010.

[27] Julien Buckley. *Etude de memoires Flash integrant des dielectriaues high-k en tant qu'oxyde tunnel ou couche de stockage*. PhD thesis, Ecole Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal, December 2006.

[28] J. D. Casperson, L. D. Bell, and H. A. Atwater. Materials issues for layered tunnel barrier structures. *Journal of Applied Physics*, 92:261–267, July 2002.

[29] J. Robertson. Interfaces and defects of high-k oxides on silicon. *Solid-State Electronics*, 49(3):283 – 293, 2005.

[30] Wellekens D., Blomme P. Govoreanu B., De Vos J. Haspeslagh L., van Houdt J., Brunco D.P. et van der Zanden K. Al_2O_3 Based Flach Interpoly Dielectrics: a Comparative Retention Study. 2006.

[31] Power J.R., Shum D., Gong Y., Bogacz S., Haeupel J., Estel H., Strenz R., Kakoschke R., van der Zanden K., et Allinger R. Improved Retention for a Al2O3 IPD Embedded Flash Cell without Top-Oxide. pages 93–96, 2008.

[32] Van Duuren M., van Schaijk R., Slotboom M., Tello P.G., Akil N., Miranda H.A. et Golubović D.S. Pushing the scaling limits of embedded non-volatile memories with *High-*κ materials. 2006.

[33] Molas G., Grampeix H., Buckley J., Bocquet M., Garros X., Martin F., Colonna J.P., Brianceau P., Vidal V., Gély M., De Salvo B., Bongiorno C., Lombardo S. et Deleonibus S. In-depth Investigation of HfAlO Layers as Interpoly Dielectrics of Future Flash Memories. 2006.

[34] Miranda A.H., van Schaijk R., van Duuren M., Akil N. et Golubović D.S. Reliability Comparison of Al2O3 and HfSiON for use as Interpoly Dielectric in Flash Arrays. 2006.

[35] Shin Y., Choi J., Kang C., Lee C., Park K.-T., Lee J.-S., Sel J., Kim V., Choi B., Sim J., Kim D., Cho H.-J.et Kim K. A novel NAND-type MONOS nemory using 63nm process technology for Multi-Gigabit Flash EEPROMs. 2005.

[36] Leroux C., Mitard J., Ghibaudo G., Garros X., Reimbold G., Guillaumot B., Martin F. Characterization and modeling of hysteresis phenomena in *High-*κ dielectrics. 2004.

[37] Wilk G.D., Wallace R.M. et Anthony J.M. *High-*κ gate dielectrics: Current status and materials properties considerations. 89:5243–5275, 2001.

[38] Tsai P.-H., Chang-Liao K.-S., Liu C.-Y., Wang T.-K., Tzeng P. J., Lin C.H., Lee L.S., et Tsai M.-J. Novel SONOS-Type nonvolatile memory device with optimal Al doping in HfAlO charge-trapping layer. 29:265–268, 2008.

[39] Lee C.H., Choi K.I., Cho M.K., Song Y.H., Park K.C. et Kim K. A Novel SONOS Structure of $Si0_2/SiN/A1_20_3$ with TaN metal gate for multi-giga bit flash memerie. pages 613–616, 2003.

[40] Lai C.H., Chin A., Kao H.L., Chen K.M., Hong M., Kwo J. et Chi C.C. Very Low Voltage SiO_2/HfON/HfAlO/TaN Memory with Fast Speed and Good Retention. pages 210–211, 2006.

[41] Lai C.H., Chin A., Chiang K.C., Yoo W.J., Cheng C. F., McAlister S.P., Chi C.C. et Wu P. Novel $SiO_2/AlN/HfAlO/IrO_2$ Memory with Fast Erase, Large ΔVth and Good Retention. pages 210–211, 2005.

[42] Tan Y.N., Chim W.K., Choi W.K., Joo M.S., Ng T.H. et Cho B.J. *High-κ* HfAlO Charge Trapping Layer in SONOS-type Nonvolatile Memory Device for High Speed Operation. 2004.

[43] Tan Y.N., Chim W.K., Choi W.K., Joo M.S. et Cho B.J. Hafnium Aluminum Oxide as Charge Storage and Blocking-Oxide Layers in SONOS-Type Nonvolatile Memory for High-Speed Operation. 53:654–662, 2006.

[44] Chin A., Laio C.C., Chen C., Chiang K.C., Yu D.S., Yoo W.J., Samudra G.S., Wang T., Hsieh I.J., McAlister S.P. et Chi C.C. Low Voltage High Speed $SiO_2/AlGaN/AlLaO_3/TaN$ Memory with Good Retention. 2005.

[45] Wang Y.Q., Singh P.K., Yoo W.J., Yeo Y.C., Samudra G., Chin A., Hwang W.S., Chen J.H., Wang S.J. et Kwong D.-L. Long Retention and Low Voltahe Operation Using IrO2/HfAlO/HfSiO/HfAlO Gate Stack for Memory Application. pages 169–172, 2005.

[46] Wang X. et Kwong D.-L. A Novel High-k SONOS Memory Using TaN/Al2O3/Ta2O5/HfO2/Si Structure for Fast Speed and Long Retention Operation. 53:78–82, 2006.

[47] Wang X., Liu J., Bai W., Kwong D.-L. A Novel MONOS-Type Nonvolatile Memory Using High-k Dielectrics for Improved Data Retention and Programming Speed. 51:597–602, 2004.

[48] Van Schaijk R., van Duuren M., Neuilly F., Baks W., Miranda A.H., Slotboom M., Akil N. et Tello P.G. SONOS flash memories with HfO_2 or $HfSiON$. pages 219–221. Philips, 2005.

[49] Cacciato A., Furnemont A., Breuil L., De Vos J., Haspelagh L., van Houdt J. Effect of Al_2O_3 morphology on the erase saturation performance in SANOS-type memory cells. pages 217–220, 2007.

[50] Chang Hyun Lee, Kyung In Choi, Myoung Kwan Cho, Yun Heub Song, Kyu Charn Park, and Kinam Kim. A novel sonos structure of sio2/sin/al2o3 with tan metal gate for multi-giga bit flash memories. In *Electron Devices Meeting, 2003. IEDM '03 Technical Digest. IEEE International*, pages 26.5.1 – 26.5.4, dec. 2003.

[51] Tae-Su Jang, Joong-Sik Kim, Sang-Min Hwang, Young-Hoon Oh, Kwang-Myung Rho, Seoung-Ju Chung, Su-Ock Chung, Jae-Geun Oh, S. Bhardwaj, Jungtae Kwon, D. Kim, M. Nagoga, Yong-Taik Kim, Seon-Yong Cha, Seung-Chan Moon, Sung-Woong Chung,

Sung-Joo Hong, and Sung-Wook Park. Highly scalable z-ram with remarkably long data retention for dram application. In *VLSI Technology, 2009 Symposium on*, pages 234 –235, june 2009.

[52] Jiyoung Kim, A.J. Hong, Sung Min Kim, E.B. Song, Jeung Hun Park, Jeonghee Han, Siyoung Choi, Deahyun Jang, Joo Tae Moon, and K.L. Wang. Novel vertical-stacked-array-transistor (vsat) for ultra-high-density and cost-effective nand flash memory devices and ssd (solid state drive). In *VLSI Technology, 2009 Symposium on*, pages 186 –187, june 2009.

[53] SungJin Whang, KiHong Lee, DaeGyu Shin, BeomYong Kim, MinSoo Kim, JinHo Bin, JiHye Han, SungJun Kim, BoMi Lee, YoungKyun Jung, SungYoon Cho, ChangHee Shin, HyunSeung Yoo, SangMoo Choi, Kwon Hong, S. Aritome, SungKi Park, and SungJoo Hong. Novel 3-dimensional dual control-gate with surrounding floating-gate (dc-sf) nand flash cell for 1tb file storage application. In *Electron Devices Meeting (IEDM), 2010 IEEE International*, pages 29.7.1 –29.7.4, dec. 2010.

[54] Eun-Seok Choi, Hyun-Seung Yoo, Han-Soo Joo, Gyu-Seog Cho, Sung-Kye Park, and Seok-Kiu Lee. A novel 3d cell array architecture for terra-bit nand flash memory. In *Memory Workshop (IMW), 2011 3rd IEEE International*, pages 1 –4, may 2011.

[55] Xian Liu, V. Markov, A. Kotov, Tho Ngoc Dang, A. Levi, I. Yue, A. Wang, and R. Qian. Endurance characteristics of superflashÂ® memory. In *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, pages 763 –765, oct. 2006.

[56] J. Yater, C. Hong, S.-T. Kang, D. Kolar, B. Min, J. Shen, G. Chindalore, K. Loiko, B. Winstead, S. Williams, H. Gasquet, M. Suhail, K. Broeker, E. Lepore, A. Hardell, W. Malloch, R. Syzdek, Y. Chen, Y. Ju, S. Kumarasamy, H. Liu, L. Lei, and B. Indajang. Highly optimized nanocrystal-based split gate flash for high performance and low power microcontroller applications. In *Memory Workshop (IMW), 2011 3rd IEEE International*, pages 1 –4, may 2011.

[57] H.M. Lee, S.T. Woo, H.M. Chen, R. Shen, C.D. Wang, L.C. Hsia, and C.C.-H. Hsu. Neoflash - true logic single poly flash memory technology. In *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st*, pages 15 –16, feb. 2006.

[58] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan. Electrons retention model for localized charge in oxide-nitride-oxide (ono) dielectric. *Electron Device Letters, IEEE*, 23(9):556 –558, sept. 2002.

[59] K. Harber C. M. Hong C. B. Li C. T. Swift E. J. Prinz, G. L. Chindalore. An embedded 90nm sonos flash eeprom utilizing hot electron injection programming and 2-sided hot hole injection erase. In *NVM WORKSHOP*, 2003.

[60] SST. Superflash technology, 2012. http://www.sst.com/.

[61] Avenium Consulting. Panorama brevet - split gate. Technical report, 2011.

[62] Ko-Min Chang. Sg-tfs: A versatile embedded flash with silicon nanocrystals as the storage medium. In *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, pages 943 –946, oct. 2008.

[63] Paulo Knirsch and Donnie Garcia. Shifting the balance of power, 2011. Available on line.

[64] Simon Fogg. Renesas to outsource 40nm mcu manufacture to tsmc, 2012. Available on line.

[65] J.A. Yater, T. Kirichenko, E.J. Prinz, M. Sadd, R. Steimle, C.T. Swift, and K.-M. Chang. 90nm split-gate nanocrystal non-volatile memory with reduced threshold voltage. In *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st*, pages 60 –61, 2006.

[66] J.A. Yater, S.T. Kang, R. Steimle, C. Hong, B. Winstead, M. Herrick, and G. Chindalore. Optimization of 90nm split gate nanocrystal non-volatile memory. In *Non-Volatile Semiconductor Memory Workshop, 2007 22nd IEEE*, pages 77 –78, aug. 2007.

[67] Cheong Min Hong, J. Yater, Sung-Taeg Kang, H. Gasquet, and G. Chindalore. Reliability study of split gate silicon nanocrystal flash eeprom. In *Non-Volatile Semiconductor Memory Workshop, 2007 22nd IEEE*, pages 75 –76, aug. 2007.

[68] G. Chindalore, J. Yater, H. Gasquet, M. Suhail, Sung-Taeg Kang, Cheong Min Hong, N. Ellis, G. Rinkenberger, J. Shen, M. Herrick, W. Malloch, R. Syzdek, K. Baker, and Ko-Min Chang. Embedded split-gate flash memory with silicon nanocrystals for 90nm and beyond. In *VLSI Technology, 2008 Symposium on*, pages 136 –137, june 2008.

[69] Sung-Taeg Kang, J. Yater, CheongMin Hong, J. Shen, N. Ellis, M. Herrick, H. Gasquet, W. Malloch, and G. Chindalore. Si nanocrystal split gate technology optimization for high performance and reliable embedded microcontroller applications. In *Non-Volatile Semiconductor Memory Workshop, 2008 and 2008 International Conference on Memory Technology and Design. NVSMW/ICMTD 2008. Joint*, pages 59 –60, may 2008.

[70] Sung-Taeg Kang, B. Winstead, J. Yater, M. Suhail, G. Zhang, C.-M. Hong, H. Gasquet, D. Kolar, J. Shen, B. Min, K. Loiko, A. Hardell, E. LePore, R. Parks, R. Syzdek, S. Williams, W. Malloch, G. Chindalore, Y. Chen, Y. Shao, L. Huajun, L. Louis, and S. Chwa. High performance nanocrystal based embedded flash microcontrollers with exceptional endurance and nanocrystal scaling capability. In *Memory Workshop (IMW), 2012 4th IEEE International*, pages 1 –4, may 2012.

[71] V.D. Marca, A. Regnier, J. Ogier, R. Simola, S. Niel, J. Postel-Pellerin, F. Lalande, and G. Molas. Experimental study to push the flash floating gate memories toward low energy applications. In *Semiconductor Device Research Symposium (ISDRS), 2011 International*, pages 1 –2, dec. 2011.

[72] Yu-Chung Lien, Jia-Min Shieh, Wen-Hsien Huang, Cheng-Hui Tu, Chieh Wang, Chang-Hong Shen, Bau-Tong Dai, Ci-Ling Pan, Chenming Hu, and Fu-Liang Yang. Fast programming metal-gate si quantum dot nonvolatile memory using green nanosecond laser spike annealing. *Applied Physics Letters*, 100(14):143501, 2012.

[73] E. Vianello, F. Driussi, P. Palestri, A. Arreghini, D. Esseni, L. Selmi, N. Akil, M. van Duuren, and D.S. Golubovic. Impact of the charge transport in the conduction band on the retention of si-nitride based memories. In *Solid-State Device Research Conference, 2008. ESSDERC 2008. 38th European*, pages 107 –110, sept. 2008.

[74] Etienne Nowak. *Impact of geometry on charge trap non volatile memories.* PhD thesis, Ecole Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal, 2010.

[75] C. Fiegna and E. Sangiorgi. Modeling of high-energy electrons in mos devices at the microscopic level. *Electron Devices, IEEE Transactions on*, 40(3):619 –627, mar 1993.

[76] A. Zaka, Q. Rafhay, P. Palestri, R. Clerc, D. Rideau, L. Selmi, C. Tavernier, and H. Jaouen. On the accuracy of current tcad hot carrier injection models for the simulation of degradation phenomena in nanoscale devices. In *Semiconductor Device Research Symposium, 2009. ISDRS '09. International*, pages 1 –2, dec. 2009.

[77] P. Palestri, N. Akil, W. Stefanutti, M. Slotboom, and L. Selmi. Effect of the gap size on the ssi efficiency of split-gate memory cells. *Electron Devices, IEEE Transactions on*, 53(3):488 – 493, march 2006.

[78] K. Sridhar, P. Bharath Kumar, S. Mahapatra, E. Murakami, and S. Kamohara. Controlling injected electron and hole profiles for better reliability of split gate sonos. In *Physical and Failure Analysis of Integrated Circuits, 2005. IPFA 2005. Proceedings of the 12th International Symposium on the*, pages 190 – 194, june-1 july 2005.

[79] Y.-H. Wang, M.-C. Wu, W.-T. Chu, C.-J. Lin, Y.-T. Lin, and C.S. Wang. On the dynamic coupling ratio of drain-coupling split gate flash using quasi-two-dimensional analysis. *Circuits, Devices and Systems, IEE Proceedings* -, 153(2):115 – 123, april 2006.

[80] L. Breuil, L. Haspeslagh, P. Blomme, D. Wellekens, J. De Vos, M. Lorenzini, and J. Van Houdt. A new scalable self-aligned dual-bit split-gate charge-trapping memory device. *Electron Devices, IEEE Transactions on*, 52(10):2250 – 2257, oct. 2005.

[81] W. Stefanutti, P. Palestri, N. Akil, and L. Selmi. Monte carlo simulation of substrate enhanced electron injection in split-gate memory cells. *Electron Devices, IEEE Transactions on*, 53(1):89 – 96, jan. 2006.

[82] Y. Tsuji, M. Terai, S. Fujieda, T. Syo, T. Saito, and K. Ando. Lateral profile of trapped charges in split-gate sonos memory. *Electron Devices, IEEE Transactions on*, 57(2):466 –473, feb. 2010.

[83] Y. Tsuji, M. Terai, S. Fujieda, T. Syo, T. Saito, and K. Ando. A novel method for evaluating electron/hole mismatch in scaled split-gate sonos memories. In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pages 1 –4, dec. 2008.

[84] Y. Okuyama, T. Furukawa, T. Saito, Y. Nonaka, T. Ishimaru, K. Yasui, D. Hisamoto, Y. Shimamoto, S. Kimura, M. Mizuno, K. Toba, D. Okada, T. Hashimoto, and K. Okuyama. Determination of lateral charge distributions of split-gate sonos memories using experimental devices with nanometer-size nitride piece. In *Non-Volatile Semiconductor Memory Workshop, 2007 22nd IEEE*, pages 85 –87, aug. 2007.

[85] L. Perniola, G. Iannaccone, B. De Salvo, G. Ghibaudo, G. Molas, C. Gerardi, and S. Deleonibust. Experimental and theoretical analysis of scaling issues in dual-bit discrete trap non-volatile memories. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 857 –860, dec. 2005.

[86] Erh-Kun Lai, Hang-Ting Lue, Yi-Hsuan Hsiao, Jung-Yu Hsieh, Chi-Pin Lu, Szu-Yu Wang, Ling-Wu Yang, Tahone Yang, Kuang-Chao Chen, Jeng Gong, Kuang-Yeu Hsieh, Rich Liu, and Chih-Yuan Lu. A multi-layer stackable thin-film transistor (tft) nand-type flash memory. In *Electron Devices Meeting, 2006. IEDM '06. International*, pages 1 –4, dec. 2006.

[87] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama. Bit cost scalable technology with punch and plug process for ultra high density flash memory. In *VLSI Technology, 2007 IEEE Symposium on*, pages 14 –15, june 2007.

[88] Jiyoung Kim, A.J. Hong, M. Ogawa, Siguang Ma, E.B. Song, You-Sheng Lin, Jeonghee Han, U-In Chung, and K.L. Wang. Novel 3-d structure for ultra high density flash memory with vrat (vertical-recess-array-transistor) and pipe (planarized integration on the same plane). In *VLSI Technology, 2008 Symposium on*, pages 122 –123, june 2008.

[89] Soon-Moon Jung, Jaehoon Jang, Wonseok Cho, Hoosung Cho, Jaehun Jeong, Youngchul Chang, Jonghyuk Kim, Youngseop Rah, Yangsoo Son, Junbeom Park, Min-Sung Song, Kyoung-Hon Kim, Jin-Soo Lim, and Kinam Kim. Three dimensionally stacked nand flash memory technology using stacking single crystal si layers on ild and tanos structure for beyond 30nm node. In *Electron Devices Meeting, 2006. IEDM '06. International*, pages 1 –4, dec. 2006.

[90] Process integration, devices, and structures. Technical report, INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS, 2011.

[91] Lee C.-H., Choi J., Kang C., Shin Y., Lee J.-S., Sel J., Sim J., Jeon S., Choe B.-I., Bae1 D., Park K. et Kim K. Multi-Level NAND Flash Memory with 63 nm-node TANOS (Si-$Oxide$-SiN-Al_2O_3-TaN) Cell Structure. 2006.

[92] Choi S., Baik S.J. et Moon J.-T. Band Engineered Charge Trap NAND Flash with sub-40nm Process Technologies. pages 925–928, 2009.

[93] Park Y., Choi J., Kang C., Lee C., Shin Y., Choi B., Kim J, Jeon S., Sel J., Park J., Choi K., Yoo T., Sim J. et Kim K. Highly Manufacturable 32Gb Multi - Level NAND Flash Memory with 0.0098 μm^2 Cell Size using TANOS(Si - Oxide - Al_2O_3 - TaN) Cell Technology. 2006.

[94] Y. Kijung and J. Joonhee. Applications of atomic layer chemical vapor deposition for the processing of nanolaminate structures. *Korean Journal of Chemical Engineering*, 19(3):451–456, May 2002.

[95] Manik Kumer Ghosh and Cheol Ho Choi. The initial mechanisms of al_2o_3 atomic layer deposition on oh/si(1 0 0)−2 *yuy* 1 surface by tri−methylaluminum and water. *Chemical Physics Letters*, 426(4-6):365–369, 2006.

[96] J. P. Colonna, M. Bocquet, G. Molas, N. Rochat, P. Blaise, H. Grampeix, C. Licitra, D. Lafond, L. Masoero, V. Vidal, J. P. Barnes, M. Veillerot, and K. Yckache. Study of parasitic trapping in alumina used as blocking oxide for nonvolatile memories. *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, 29(1):01AE02 –01AE02–5, jan 2011.

[97] Milton Ohring. Chapter 12 - mechanical properties of thin films. In *Materials Science of Thin Films (Second Edition)*, pages 711 – 781. Academic Press, San Diego, second edition edition, 2002.

[98] Chung-Kwei Lin, Hsien-Ta Hsu, Chin-Te Chen, and Tsong-Jen Yang. The effect on the microstructures of electroless nickel coatings initiated by pulsating electric current. *Thin Solid Films*, 516(2-4):355–359, 2007.

[99] Dani**E**le Bouchet and Christian Colliex. Experimental study of elnes at grain boundaries in alumina: intergranular radiation damage effects on al-l23 and o-k edges. *Ultramicroscopy*, 96(2):139 – 152, 2003.

[100] C. Licitra, E. Martinez, N. Rochat, T. Veyron, H. Grampeix, M. Gely, J. P. Colonna, and G. Molas. Coupling of advanced optical and chemical characterization techniques for optimization of high-kappa dielectrics with nanometer range thickness. *AIP Conference Proceedings*, 931(1):292–296, 2007.

[101] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[102] W. Kohn. An essay on condensed matter physics in the twentieth century. *Rev. Mod. Phys.*, 71:S59–S77, Mar 1999.

[103] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[104] E. Kaxiras. *Atomic and Electronic Structure of Solids,*. Cambridge University Press, 2003.

[105] Ordejen P. Artacho E. Sanchez-Portal, D. and J. M. Soler. Density-functional method for very large systems with lcao basis sets. *International Journal of Quantum Chemistry*, 65:453–461, Sep 2000.

[106] T.V. Perevalov, A.V. Shaposhnikov, and V.A. Gritsenko. Electronic structure of bulk and defect α- and γ-al$_2$o$_3$. *Microelectronic Engineering*, 86(7-9):1915 – 1917, 2009. INFOS 2009.

[107] Shang-Di Mo and W. Y. Ching. Ab initio calculation of the core-hole effect in the electron energy-loss near-edge structure. *Phys. Rev. B*, 62(12):7901–7907, Sep 2000.

[108] F. H. Streitz and J. W. Mintmire. Energetics of aluminum vacancies in gamma alumina. *Phys. Rev. B*, 60(2):773–777, Jul 1999.

[109] G. Gutiérrez, A. Taga, and B. Johansson. Theoretical structure determination of γ-Al_2O_3. *prb*, 65(1):012101, January 2002.

[110] E. Menéndez-Proupin and G. Gutiérrez. Electronic properties of bulk $\gamma - al2o3$. *Phys. Rev. B*, 72(3):035116, Jul 2005.

[111] R. Caracas F. Detraux M. Fuchs G.-M. Rignanese L. Sindic M. Verstraete G. Zerah F. Jollet M. Torrent A. Roy M. Mikami Ph. Ghosez J.-Y. Raty D.C. Allan X. Gonze, J.-M. Beuken. First-principles computation of material properties : the abinit software project. *Computational Materials Science*, 25:478 – 492, 2002.

[112] S. B. Zhang and John E. Northrup. Chemical potential dependence of defect formation energies in gaas: Application to ga self-diffusion. *Phys. Rev. Lett.*, 67:2339–2342, Oct 1991.

[113] J. Robertson and P. W. Peacock. Doping and hydrogen in wide gap oxides. *Thin Solid Films*, 445(2):155 – 160, 2003. Proceedings of the 3rd International Symposium on Transparent Oxide Thin films for Electronics and Optics.

[114] A. Padovani, L. Larcher, S. Verma, P. Pavan, P. Majhi, P. Kapur, K. Parat, G. Bersuker, and K. Saraswat. Statistical modeling of leakage currents through sio2/high-k dielectrics stacks for non-volatile memory applications. In *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pages 616 –620, 27 2008-may 1 2008.

[115] A. Padovani, L. Larcher, D. Heh, and G. Bersuker. Modeling tanos memory program transients to investigate charge-trapping dynamics. *Electron Device Letters, IEEE*, 30(8):882 –884, aug. 2009.

[116] D. Liu and J. Robertson. Oxygen vacancy levels and interfaces of al_2o_3. *Microelectronic Engineering*, 86(7-9):1668 – 1671, 2009. INFOS 2009.

Author's publication list

International conferences – Conférences internationales

[1] **Masoero, L.**, Molas, G., Della Marca, V., Gély, M., Cueto, O., Colonna, J.P., De Luca, A., Brianceau, P., Charpin, C., Lafond, D., Delaye, V., Aussenac, F., Carabasse, C., Pauliac, S., Comboroure, C., Boivin, P., Ghibaudo, G., Deleonibus, S., De Salvo, B. *Physical understanding of program injection and consumption in ultra-scaled SiN split-gate memories* (2012) 2012 4th IEEE International Memory Workshop, IMW 2012, art. no. 6213686.

[2] **Masoero, L.**, Molas, G., Brun, F., Gély, M., Colonna, J.P., Della Marca, V., Cueto, O., Nowak, E., De Luca, A., Brianceau, P., Charpin, C., Kies, R., Toffoli, A., Lafond, D., Delaye, V., Aussenac, F., Carabasse, C., Pauliac, S., Comboroure, C., Ghibaudo, G., Deleonibus, S., De Salvo, B. *Scalability of split-gate charge trap memories down to 20nm for low-power embedded memories* (2011) Technical Digest - International Electron Devices Meeting, IEDM, art. no. 6131522, pp. 9.5.1-9.5.4.

[3] **Masoero, L.**, Molas, G., Blaise, P., Colonna, J.P., Vianello, E., Selmi, L., Papon, A.M., Lafond, D., Martin, F., Gely, M., Licitra, C., Barnes, J.P., Ghibaudo, G., De Salvo, B. *Study of defects in Al_2O_3 blocking layers of TANOS memories by atomistic simulation, electrical characterization and physico-chemical material analyses* (2011) International Symposium on VLSI Technology, Systems, and Applications, Proceedings, art. no. 5872268, pp. 148-149.

[4] **Masoero, L.**, Blaise, P., Molas, G., Colonna, J.P., Gély, M., Barnes, J.P., Ghibaudo, G., De Salvo, B. *Defects-induced gap states in hydrogenated γ-alumina used as blocking layer for non-volatile memories* (2011) INFOS Conference on "Insulating Films on Semiconductors" 2011 .

[5] Molas, G., **Masoero, L.**, Blaise, P., Padovani, A., Colonna, J.P., Vianello, E., Bocquet, M., Nowak, E., Gasulla, M., Cueto, O., Grampeix, H., Martin, F., Kies, R., Brianceau, P., Gély, M., Papon, A.M., Lafond, D., Barnes, J.P., Licitra, C., Ghibaudo, G., Larcher, L., Deleonibus, S., De Salvo, B. *Investigation of the role of H-related defects in Al2O3 blocking layer on charge-trap memory retention by atomistic simulations and device physical modelling* (2010) Technical Digest - International Electron Devices Meeting, IEDM, art. no. 5703414, pp. 22.5.1-22.5.4.

[6] Della Marca, V., **Masoero, L.**, Molas, G., Amouroux J., Petit-Faivre, E., Postel-Pellerin, J., Lalande, F., Jalaguier, E., Deleonibus, S., De Salvo, B., Boivin,P., Ogier,J-L. *Optimization of Programming Consumption of Silicon Nanocrystal Memories for Low Power Applications* (2012) International Semiconductor Conference Dresden-Grenoble, ISCDG.

[7] Colonna, J.P., Bocquet, M., Molas, G., Rochat, N., Blaise, P., Grampeix, H., Licitra, C., Lafond, D., **Masoero, L.**, Vidal, V., Barnes, J.P., Veillerot, M., Yckache, K. *Study of parasitic trapping in alumina used as blocking oxide for nonvolatile memories* (2010) Workshop on Dielectrics in Microelectronics, WODIM.

Journal articles – Articles de revues

[8] **Masoero, L.**, Blaise, P., Molas, G., Colonna, J.P., Gély, M., Barnes, J.P., Ghibaudo, G., De Salvo, B. *Defects-induced gap states in hydrogenated γ-alumina used as blocking layer for non-volatile memories* (2011) Microelectronic Engineering, 88 (7), pp. 1448-1451.

[9] Colonna, J.P., Bocquet, M., Molas, G., Rochat, N., Blaise, P., Grampeix, H., Licitra, C., Lafond, D., **Masoero, L.**, Vidal, V., Barnes, J.P., Veillerot, M., Yckache, K. *Study of parasitic trapping in alumina used as blocking oxide for nonvolatile memories* (2011) Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures, 29 (1), pp. 01AE021-01AE025.

[10] Della Marca, V., Just, G., Regnier,A., Ogier,J-L., Simola R., Niell, S., Postel-Pellerin, J.,Lalande F., **Masoero L.**, Molas, G. *Push the Flash Floating Gate Memories Toward the Future Low Energy Application* (2012) Solide State Electronic 2012.

Other communications – Autres communications

[11] **Masoero, L.**, Molas, G., De Salvo, B.,Brillouet, M. *Study of defects in Al_2O_3 used as blocking layer of TANOS memories* (2012), 8th International Nanotechnology Conference (INC8), POSTER

[12] **Masoero, L.**, *Split-gate charge trap memories: impact of scaling on performances and consumption for low-power embedded applications* (2012), Workshop Memories 2012, June 21 - MINATEC Grenoble - France, ORAL

Book chapter – Chapitre d'ouvrage

[13] **Masoero, L.**, Molas, G., De Salvo, B., *Silicon nanocrystal (Si-nc), split-gate, charge trap memories for embedded applications* (2012), Advances in nonvolatile memory and storage technology: Woodhead Publishing, BOOK CHAPTER, to be published.

TITRE:

Etude d'architectures et d'empilements innovants de mémoires Split-Gate (grille séparée) à couche de piégeage discret

Résumé:

Du fait de l'augmentation de la demande de produits pour les applications grand public, industrielles et automobiles, des mémoires embarquées fiables et à faible coût de fabrication sont de plus en plus demandées. Dans ce contexte, les mémoires split-gate à piégeage discret sont proposées pour des microcontrôleurs. Elles combinent l'avantage d'une couche de stockage discrète et de la configuration split-gate. Durant ce travail de recherche, des mémoires split-gate à couche de piégeage discret ayant des longueurs de grille de 20nm sont présentées pour la première fois. Celles-ci on été réalisées avec des nanocristaux de silicium (Si-nc), du nitrure de silicium (SiN) ou un hybride Si-nc/SiN avec diélectrique de control de type SiO2 ou AlO et sont comparées en termes de performances lors des procédures d'effacement et de rétention. Le rôle des défauts dans le diélectrique de contrôle (alumine) a enfin été étudié. Nous avons montré que la concentration de pièges dans AlO pouvait être réduite par ajustement des conditions de procédé de fabrication, débouchant ainsi sur l'amélioration de la rétention dans les mémoires à piégeage de charge.

Spécialité :

Micro- et Nano-électronique

Mots-clês:

Mémoires, Split-gate, SONOS, Alumine, Simulation atomistique, Défauts, Hydrogène.

TITLE:

Study of innovative stacks and architectures of Split-Gate charge trap memories

Abstract:

Due to the increasing demand for consumer, industrial and automotive products, highly reliable, and low integration cost embedded memories are more and more required. In this context, split-gate charge trap memories were proposed for microcontroller products, combining the advantage of a discrete storage layer and of the split-gate configuration. In this thesis, split-gate charge trap memories with electrical gate length down to 20nm are presented for the 1st time. Silicon nanocristals (Si-nc), or silicon nitride (SiN) and hybrid Si-nc/SiN based split-gate memories, with SiO_2 or AlO control dielectrics, are compared in terms of program erase and retention. Then, the scalability of split-gate charge trap memories is studied, investigating the impact of gate length reduction on the memory window, retention and consumption.
We thus studied the role of defects on alumina control dielectric employed in TANOS-like memory. We used atomistic simulation, consolidated by a detailed alumina physico-chemical material analysis, to investigate the origin of traps in alumina. We showed that the trap concentration in AlO can be decreased by adjusting the process conditions leading to improved retention behaviour in charge trap memory, suitable for embedded applications.

Speciality:

Micro- et Nano-electronics

Key words:

Non-volatile memory, Split-gate, SONOS, Alumina, Atomistic simulation, Defects, Hydrogen.

Thèse préparée au sein: Du Laboratoire d'Électronique et de Technologie de l'Information (LETI), CEA-Grenoble, Minatec, 17 av. des Martyrs, 38054 Grenoble Cedex 9, France.