POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science degree in Engineering Physics

Analysis of the Down-Coupling Phenomenon in 3D NAND Flash Memories

Supervisor: Prof. Alessandro SOTTOCORNOLA SPINELLI
Co-Supervisor: Prof. Christian MONZIO COMPAGNONI

MSc Thesis of
Mattia GIULIANINI
Student ID number 920306

Academic year 2019-2020

Contents

List of Figures ii

Sommario iii

Abstract v

1 Introduction to NAND Flash memories 2

1.1 Flash memories success . . . 2
1.2 2D NAND Flash memories . . . 3
1.2.1 Functional principles . . . 3
1.2.2 Programming and erasing a memory cell . . . 4
1.2.3 Array structure . . . 6
1.3 Array operations . . . 7
1.4 Scaling issues in planar NAND memories . . . 10
1.4.1 Program noise . . . 11
1.4.2 Stress-Induced Leakage Current . . . 12
1.4.3 Cell-to-cell electrostatic interference . . . 13
1.4.4 Random Telegraph Noise . . . 14
1.5 3D NAND Flash memories . . . 15
1.5.1 Vertical channel structure . . . 15
1.6 Advantages and issues of 3D memories . . . 19
1.7 Thesis work goals . . . 20

2 Gate-induced drain leakage effect and down-coupling phenomenon 21

2.1 Gate-induced drain leakage . . . 21
2.1.1 Physical origin of GIDL . . . 22
2.1.2 GIDL in BiCS technologies . . . 23
2.2 Disturbs related to the program operation . . . 25
2.3 Self-Boosting effect . . . 27
2.3.1 Local Self-Boosting effect . . . 28

2.4 Down-Coupling Phenomenon . . . 29
2.5 Conclusions . . . 32

3 TCAD simulation about the DCP 33
3.1 Simulation environment . . . 33
3.2 Simulation definition . . . 34
3.2.1 Geometry of a single cell string . . . 34
3.2.2 Physical models included in the simulations . . . 36
3.3 Generation and recombination processes . . . 38
3.4 Threshold extrapolation . . . 39
3.5 Transient simulation definition . . . 41
3.6 DCP simulation results . . . 43
3.6.1 Analysis of the CB energy time evolution . . . 43
3.6.2 Time evolution of the carriers concentration . . . 45
3.6.3 Time evolution of the current . . . 47
3.7 Dependence on parameters . . . 50
3.8 Preliminary model of the return to equilibrium of the DCP . . . 53
3.8.1 Description of the compact model . . . 54
3.8.2 Circuital model . . . 56
3.8.3 Model results . . . 58
3.9 conclusions . . . 59

4 Simulations of BiCS and TCAT structures 61
4.1 DCP in BiCS memories . . . 61
4.1.1 Geometry and characteristics of a 16 WLs NAND string . . . 61
4.1.2 Physical differences between single WL structure and 16 WLs structure . . . 62
4.1.3 Time evolution of the conduction band energy in BiCS . . . 66
4.1.4 Carriers concentration in BiCS . . . 68
4.1.5 Time evolution of the current in BiCS . . . 71
4.2 Dependence on parameters for the BiCS structure . . . 75
4.3 DCP in TCAT memories . . . 76
4.3.1 TCAT geometry . . . 77
4.4 DCP results for a completely programmed TCAT structure . . . 78
4.5 DCP results for a not completely programmed TCAT structure . . . 81
4.5.1 Carriers concentration analysis in a not completely programmed TCAT structure . . . 83

4.5.2 Current time evolution in a not completely programmed TCAT structure 85
4.6 Conclusions 88

Conclusions **89**

Bibliografy **91**

List of Figures

1.1 F values for 2D NAND Flash memories manufactured by the leading companies in the semiconductor sector between 2001 and 2015 (from *International Solid-State Circuits Conference (ISSCC)*). 3
1.2 Schematic representation of a planar memory cell with *floating-gate*. . . . 4
1.3 Band diagrams of a NAND Flash memory where (a) positive charge and (b) negative charge is stored. 5
1.4 Band diagram for *2D NAND* memory during (a) the *programming* operation and (b) the *erasing* operation. 6
1.5 (a) Circuital scheme of a 2D NAND Flash array, highlighting the strings of floating-gate transistors driven by shared *WLs*. BL = *bit-line*; DSL = *drain-select line*; DUL = *dummy-line*; WL = *word-line*; SSL = *source-select line*; SL = *source-line*; (b)Planar layout of a 2D NAND Flash memory, (c) cross section of the array along the string direction (from [1]). 7
1.6 V_T distribution of cells in different states for (a) a *single level cell*, (b) a *multi level cell* and (c) *triple level cell* (from [1]). 8
1.7 Array representation of a NAND string with bias scheme for (a) *read*, (b) *program* and *erase* operation (from [1]). 9
1.8 Example of the ISPP process. In red is depicted the *control gate* bias while in black is shown the corresponding programme V_T (from [2]). 10
1.9 Pictorial representation of an electron undergoing the tunneling assisted by two traps states (from [3]). 12
1.10 Schematic representation of the capacitances between first neighbor FGs (from [4]). 13
1.11 (a) Example of simulated current density for a MOSFET where atomistic doping is considered, (b) comparison of the density current in the case of a trapped electron (from [5]). 14
1.12 Simple representation of a vertical channel memory (from [1]). 16

1.13 (a) Schematic view of a single cell for a GAA memory structure (b) Vertical section of a GAA 3D NAND memoriy(from [1]). 17
1.14 Manufacturing process flow (from [6]). 17
1.15 Vertical section of a *TCAT* flash memory (from [7]). 18
1.16 Schematic vertical section of a 3D NAND Flash memory highlighting the non-uniform pillar radius in the vertical direction (from [1]). 19

2.1 Subthreshold characteristic of planar *n-MOSFETs* showing a significant drain leakage current when V_D is high (from [8]). 22
2.2 (a) Schematic representation of the carrier movement during the B2BT and (b) representation of the band bending and tunneling of the electron (from [8]). 23
2.3 (a) Schematic representation of the band-to-band tunneling in BiCS memories, (b) Simulated result for a *erase* operation (from [9]). 24
2.4 Schematic representation of a NAND Flash array where the various connection are highlighted (from [10]). 25
2.5 Schematic representation of a program operation in a 2D NAND Flash array where in red are circled the cells subjected to pass disturbs and in green the cells subjected to program disturbs (from [10]). 26
2.6 (a) Bias scheme representation of the *local self-boosting effect* (from [12]) and (b) threshold voltage shift due program and pass disturb for a 2D NAND Flash memory (from [11]). 28
2.7 Array bias scheme for the program operation. 30
2.8 Simulation results of DCP occurring during a verify operation (from [13]). 31
2.9 Simulation results of DCP for different V_T on neighbors cell (from [13]). . 31

3.1 a) half geometry of the vertical-channel memory with the longitudinal lengths of the structure, b) radial lengths of the memory. 34
3.2 Doping profile used in the simulations 36
3.3 *BL* current comparison between a memory without trapped charge in the cell and a *P1* programmed memory. 41
3.4 Bias scheme of the simulation. 42
3.5 Band diagram at $t = 0$ s in (a) longitudinal direction at 0.25 nm from the silicon-oxide interface and (b) radial direction at the point of maximum Electric field located at the junction between the BL and the channel . . . 43
3.6 Time evolution of the conduction band energy at the middle of the string, since the end of the WL falling edge. The dashed black line represent the equilibrium value. 44

3.7 Longitudinal section of the electron density. The cut is taken at 0.25 nm from the silicon-oxide interface and at $t = 0$. 45

3.8 Time evolution of the hole density at the middle of the string since the end of the WL falling edge. 46

3.9 Comparison of the electron flow at the SL, blue curve, with respect the current due to the rate generation of the B2BT and SRH, red and yellow curves, respectively. 47

3.10 Longitudinal section of the hole density taken at 0.25 nm from the silicon-oxide interface for three different instants of the transitory. 48

3.11 Time evolution of the longitudinal component of the electric field that is present at the junction between the SL and the channel since the end of the WL falling edge. 49

3.12 Longitudinal section of the SRH rate generation taken at 0.25 nm from the silicon-oxide interface for three different instants of the transitory. 50

3.13 Time evolution of the conduction band energy at the middle of the string, since the end of the WL falling edge. 51

3.14 (a) Comparison of the SRH rate generation between the case of not implemented B2BT (blue curve) and implemented B2BT (dashed red curve). (b) Comparison of the SRH rate generation between the case where $\tau^n_{max} = 10^{-9}$ s (blue curve) and the case where $\tau^n_{max} = 10^{-8}$ s (red curve). The longitudinal cuts are taken at 0.25 nm from the silicon-oxide interface at the end of the verify phase. 52

3.15 Comparison of the electron flow, blue curve, with respect the current due to the rate generation of the B2BT and SRH, red and yellow curves, respectively. The plot is related to the simulation where $\tau^n_{max} = 10^{-9}$ s. 53

3.16 Schematic representation of the capacitive couplings in the string at the edge of the channel. 54

3.17 Schematic representation of the circuital model. 57

3.18 Comparison of the time evolution of the CB energy between the compact model result (blue curve) and the Sentaurus TCAD result (red curve). . . 58

3.19 Comparison of the time evolution of the currents between the compact model result and the Sentaurus TCAD result. 59

4.1 Half geometry of the BiCS structure. 62

4.2 Logarithmic comparison of the transcharacteristics belonging to a structure with 16 WLs (blue curve) and the structure with just one WL (red curve). 63

4.3 (a) Schematic representation of the contributions to the capacitance between the trapping layer and the channel and (b) pictorial view of the field lines in an elliptical system (from [23]). 64
4.4 Linear comparison of the transcharacteristics belonging to a structure with 16 WLs (blue curve) and the structure with just one WL (red curve). . . 66
4.5 Comparison of the CB energy between the structure with 16 WLs (blue curve) and the structure with one single WL (red curve). The longitudinal section is taken at 0.25 nm from the silicon-oxide and $t = 0$. 67
4.6 Time evolution of the conduction band energy underneath a trapping region in the middle of the string, since the end of the WL falling edge. The dashed black line represent the equilibrium value. 68
4.7 Comparison of the electron density between the structure with 16 WLs (blue curve) and the structure with one single WL (red curve). The longitudinal section is taken at 0.25 nm from the silicon-oxide interface taken at $t = 0$. 69
4.8 Comparison of the hole density between the structure with 16 WLs (blue curve) and the structure with one single WL (red curve). The longitudinal section is taken at 0.25 nm from the silicon-oxide interface taken at $t = 0$. 70
4.9 Longitudinal section of the string at 0.25 nm from the silicon-oxide. The figure reports: (a) the hole density comparison for $t = 0.005, 0.380, 2.1$ ms (b) the SRH rate at $t = 0.380$ ms. 71
4.10 Comparison of the electron flow(blue curve) with respect to the current due to the rate generation of the B2BT and SRH (red and yellow curves, respectively). For what concerns I_{SRH}, the dashed part comes from a recombination rate, instead the solid line comes from the generation rate. 72
4.11 Time evolution of the hole density underneath WL^7. 73
4.12 (a) Time evolution of the hole density underneath the spacer between the SSL and WL^0, (b) SRH generation rate for three different time instants. The longitudinal section is taken at 0.25 nm from the silicon-oxide. 74
4.13 (a) Longitudinal section of the string at 0.25 nm from the silicon-oxide. The figure reports the hole density comparison at $t = 5$ μs between the simulation where was set $\tau^n_{max} = 10^{-9}$ s (blue curve) and the simulation with $\tau^p_{max} = 3 \cdot 10^{-10}$ s (red curve). 76
4.14 Representation of the substrate region for a TCAT geometry. 77
4.15 Electrostatic potential profile of the substrate for (a) $N_a^{sub} = 5 \cdot 10^{18} [cm^{-3}]$ and (b) $N_a^{sub} = 1 \cdot 10^{18} [cm^{-3}]$. 78

4.16 Longitudinal section of the CB energy at different time instants. The cut is of taken at 0.25 nm from the silicon-oxide interface. The dashed line shows the equilibrium value. 79
4.17 holes current density profile at (a) $t = 1.5[\mu s]$ before the end of the falling edge and (b) $t = 10[\mu s]$ after the falling edge of WLs. 80
4.18 Time evolution of the conduction band energy of the channel underneath WL^{15}. The time origin corresponds to the end of the falling edge of the verify operation. 81
4.19 Comparison of the CB energy as a function of the position for different time instants. The longitudinal section is taken at 0.25 nm from the silicon-oxide interface in a string where WL^8 is erased. The dashed line shows the equilibrium value. 82
4.20 Time evolution of the conduction band energy in a point of the channel underneath WL^{12} that is the middle cell in the BLS region. 83
4.21 Representation of the hole density as a function of the longitudinal coordinate. The cut is taken at 0.25 nm from the silicon-oxide interface and at $t = 0$ s. The blue curve it is related to a string where just WL^7 is ERASED, the red curve represents the hole density section for a string with WL^7 and WL^8 ERASED. 84
4.22 Time evolution of the hole density underneath the spacer between the DSL and WL^{15}. 85
4.23 Comparison of the electron flow at the BL (blue curve) with respect to the currents calculate by the SRH and B2BT generation rate (red and yellow curves, respectively) and the hole current coming from the BODY contact (green curve). 86
4.24 Time evolution of the barrier height present in the channel position correspondent to WL^8. 87

Sommario

La tecnologia NAND flash rappresenta una delle principali soluzioni nel mercato delle memorie non volatili. La costante richiesta di performance migliori e un maggiore risparmio di spazio ha spinto le industrie di semiconduttori a svolgere continuamente diversi sforzi nel processo di *scaling* di queste memorie. La miniaturizzazione delle memorie continua da oltre trenta anni e negli ultimi tempi la caratteristica *feature size (F)* che descrive la dimensione delle celle è giunta a valori di circa 15 nm. Grazie a ciò, tali memorie possiedono una elevata densità oltre a mantenere un basso costo di produzione, il che le rende un ottima evoluzione che possa sostituire gli *hard-disk* magnetici.

Nonostante ciò, negli ultimi anni l'estrema miniaturizzazione di queste memorie ha portato al sorgere di nuovi problemi tecnologici e di affidabilità, cosi che la riduzione di F è diventata sempre più difficile. Questi problemi sono correlati alla natura quantizzata della carica, quindi sono intrinsici e inevitabili. Per questo motivo è stato necessario trovare un nuovo design di queste memorie. La soluzione è stata quella di passare da una intrisica struttura bidimensionale dei MOSFET ad una struttura tridimensionale. Questo nuovo design ha portato alla creazione delle memorie 3D NAND Flash, che sfruttano la dimensione verticale per l'implementazione di ulteriori celle. Questo tipo di tecnologia è in grado di aumentare la densità di dati salvati senza ridurre eccessivamente la F. Anzi, la capacità di memoria è elevata già ad F sufficientemente grande da evitare i problemi dell'estrema miniaturizzazione e questo la rende una tecnologia veramente affascinante. La variazione principale di questa nuova tecnologia risiede in una innovativa struttura del canale. Infatti esso cessa di essere essenzialmente uno strato planare di silicio sotto i transistor e diventa una struttura a *pillar* dove l'inversione di carica è comandata da *control gates (CGs)* che circondano il canale. Questa tecnologia però è ancora lontana dall'essere senza difetti, infatti in tale struttura sono comparsi disturbi durante la scrittura delle celle che non erano mai stati visualizzati nelle memorie 2D. Questi disturbi possono produrre errori nel salvataggio dei dati ed uno di questi disturbi, il Down-Coupling Phenomenon (DCP), è analizzato in questo lavoro.

Il DCP è un disturbo che avviene durante la fase di verifica delle celle appena prima dell'inizio della programmazione di queste. Nella parte finale della verifica, operazione

molto simile ad una lettura delle celle, la tensione ai *control gates (CGs)* scende da un valore superiore alla massima *tesione di soglia (V_T)* fine alla tensione di *ground*. Durante questo transitorio, i transistor della stringa, nel caso siano programmati, si spengono nel momento in cui la tensione ai *gate* scende sotto la tensione di soglia delle celle e il canale entra in stato flottante. A questo punto si instaura tra i gates ed il canale un accoppiamento capacitivo che fa in modo che la tensione del canale segua la discesa della tensione ai gates e che l'operazione si concluda con un voltaggio negativo nel canale. Il disturbo sorge dal momento che dopo tale operazione inizi la programmazione delle celle. Questa, affinchè celle non selezionate non vengano scritte indesideratamente, sfrutta il *Self-Boosting Effect (LSBE)* dove il canale viene portato ad alte tensioni. Nel momento in cui al canale venga eseguita la procedura del boosting partendo da tensioni negative, il voltaggio finale risulterà abbassato e ciò può produrre errori di programmazione.

Al momento si è a conoscenza della presenza del DCP ma non è stato ancora analizzato il modo in cui la stringa di memoria riesca a tornare alla situazione e in quanto tempo essa riesca a farlo. Questo lavoro è incentrato proprio sullo studio della fisica che giace sotto questo fenomeno e particolare attenzione verrà data alla fase di ritorno all'equilibrio da parte della memoria. Abbiamo utilizzato il software *Sentaurus TCAD* per simulare il transitorio che descrive la fine della fase di verifica. Siamo partiti dall'analisi l'analisi di una struttra tridimensionale in cui è stato implementato un solo gate che controlla l'intero canale. Abbiamo utilizzato tale struttura per ottenere dei risultati inizialmente più semplici da analizzare. Infatti tale design permette di trascurare tutte quelle interazioni elettrostatiche presenti tra i vari gate che complicano l'analisi del fenomeno. Da ciò che abbiamo analizzato su tale struttura, abbiamo creato un preliminare modellino compatto che possa simulare il fenomeno. In seguito siamo passati allo studio delle simulazioni eseguite sulle due principali strutture della tecnologia 3D: la Bit-Cost Scalable (BiCS) e la Terabit Cell Array Transistor technology. Nello specifico abbiamo analizzato come i portatori di carica riescono ad entrare ed uscire dal canale una volta che i transistor sono nella stato di spenti. Abbiamo comparato le modalità di ripristino dell'equilibrio tra il caso di memoria BiCS e di memoria TCAT. Alla fine abbiamo concluso il lavoro con un analisi quantitativa del transitorio dovuto al DCP, chiarendo quale sia la fisica di base che caratterizza il ritorno all'equilibrio di questo fenomeno. In questo modo aperto le porte per una dettagialta valutazione del DCP sull'affidabilità delle memorie 3D NAND Flash e per lo studio di nuovi designs della stringa che portino ad un miglioramento delle prestazioni.

Abstract

The NAND Flash technology represents a main solution in the market of non-volatile memories. The constant requests of higher performances and a higher space saving have brought the semiconductor industries to continuously carry out several efforts in the scaling process of these memories. The miniaturization of these memories began more than thirty years and today it has achieved a so advanced level that the *characteristic size (F)* has reached the value of about 15 *nm*. Thank to this, NAND Flash memories have a high storage density, other than a low production cost, and so they represent a good evolution of the magnetic *hard-disk*.

Even though, in the last years, the extreme miniaturization of the memories has brought to the outcome of new stability problems such that the scaling of F has become more difficult. These problems are related to the quantized nature of the charge, so they are intrinsic and unavoidable. For this reason a new design of the memories was necessary. The solution was to change from an intrinsic bi-dimensional structure of the MOSFETs to a three dimensional structure. This new design has brought to the creation of the 3D NAND Flash memories that exploit the vertical dimension to create further cells. This kind of technology is able to increase the bit storage density without reducing too much the characteristic size F. The storage capacity is so much high already for sufficiently large F that this technology avoid the problems of the extreme scaling and for this reason this solution is so fascinating. The main variation of this new technology lies in the innovative channel structure. In fact, essentially the channel is no longer a planar layer of silicon underneath the transistor but is a *pillar* where the charge inversion is controlled by *control gates (CGs)* that surround the channel. This technology is still far away from being flawless, indeed during the writing of the cells unprecedented disturbs arise in the 3D structure while in the 2D design they were never seen. This disturbs can produce errors of the data and one of them, the *Down-Coupling Phenomenon (DCP)* is analyzed in the current work.

The DCP is a disturbs that occurs during the verify phase of the cells, just before the writing operation. In the last part of the verify phase, an operation alike the read one, the voltage of the CGs decreases from a value higher than the maximum *threshold voltage*

(V_T) to the ground voltage. During this transient, the string transistors, in case they have a positive V_T, are turned off when the gates voltage gets lower than the threshold and the channel enter in a floating state. At this point a capacitive coupling arises between the gates and the channel and this makes the potential voltage of the channel negative once the verify operation ends. The disturb take place due to the fact that then the program operation begins. This operation, in order to not write unselected cells, exploits the *Local Self-Boosting Effect (LSBE)* where the channel is brought to high voltages. When the boosting effect takes place starting from a negative potential, the final voltage is lower than the wanted one and this produces programming errors.

At the moment the DCP is known just for this behavior but the way in which the system could return to the equilibrium and in how ch time it has not been analyzed yet. Indeed, this work is focused on the study of the physics that lies underneath this phenomenon with particular attention to the return to the equilibrium of the memory. We have used the software *Sentaurus TCAD* in order to simulate the transitory that describe the end of the verify operation. We started from the analysis of a simplified structure in which there was implemented just one control gate that controls the channel. In this way we have obtained results easier to analyze because this structure allows to avoid those electrostatic interactions that are present between the various gates. With what we have obtained by this analysis, we have implemented a preliminary compact model that could replicate the simulate the phenomenon. Next, we moved to the analysis of the effect in the two main designs for the 3D NAND Flash technology: the Bit-Cost Scalable (BiCS) and the Terabit Cell Array Transistor technology. In particular we have analyzed how the carriers can enter and exit from the channel once the transistors are switched off. We have compared the ways in which the BiCS and the TCAT structures can return to the stationary condition. In the end we have concluded our work with a quantitative analysis of the DCP transient, pointing out the physical events that characterize the return to the equilibrium of this phenomenon. In this way, we have opened new perspectives for a detailed evaluation of the impact of the DCP on the reliability of 3D NAND Flash memories and for the study of new string designs that could bring to an improvement of the performances of these memories.

Chapter 1

Introduction to NAND Flash memories

The NAND Flash technology has drawn many interests due to the several advantages that this technology can offer. For this reason, many efforts were done by the semiconductor industries in oder to make the Flash technology an even better resource. In the last years the mechanical stress immunity and the huge capabilities of the NAND memories made this technology a good replacement to the hard disk drives. In this chapter, an introduction on the 2D NAND technology will be given, the main array operations will be presented and the principal issues stated. Next, the motivations that pushed the investors to move to the 3D NAND Flash technology, in order to keep going with the miniaturization of cells, will be explained. In the end a brief presentation of this new technology will be given.

1.1 Flash memories success

The huge growth of the electronic market during the last years has been feasible thanks to the development of nonvolatile solid-state memories which are able to store a huge amount of data by using a small quantity of silicon and so keeping the cost-per-cell relatively small. With the passing of the years, the introduction of better technology processes and the development of better designs led the Flash technologies to become a very reliable solution for non-volatile memories. The success was achieved thanks to the constant miniaturization of the memories that brought to a memory capacity, defined in terms of *gross bit storage capacity (GBSD)*, exceeding the 1 $Gbit/mm^2$. Fig. 1.1 shows the trend of the *Feature size (F)* along the years, highlighting a reduction of a factor $\sqrt{2}$ every two years, in agreement with what the *Moore's law* states. Despite the cost savings, the miniaturization drove the technologies to its physical limits and brought out reliability problems and new phenomena that both worsened the device functionalities and increased the system complexity. In particular, at very small dimensions, the discrete nature of the

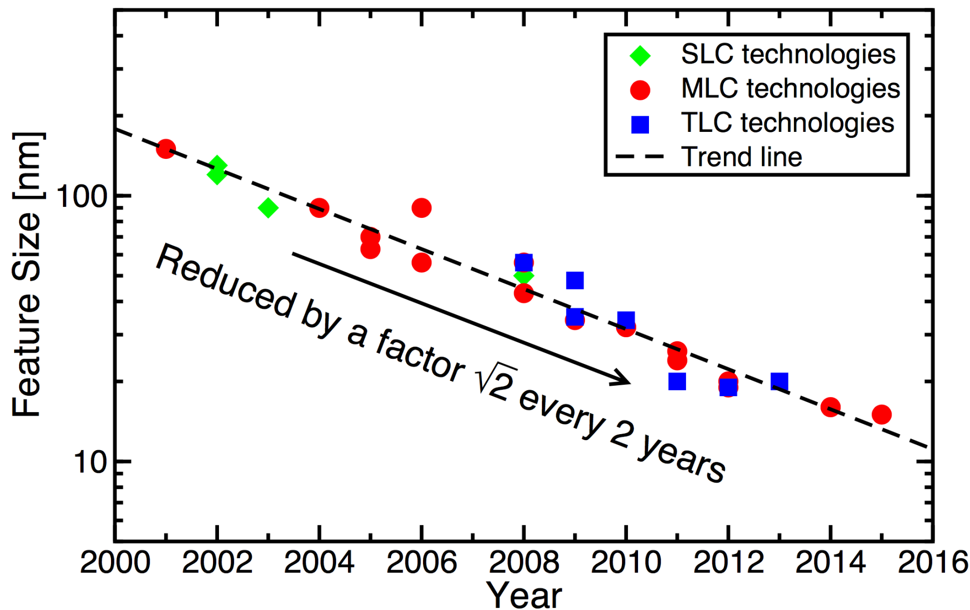


Figure 1.1: F values for 2D NAND Flash memories manufactured by the leading companies in the semiconductor sector between 2001 and 2015 (from *International Solid-State Circuits Conference (ISSCC)*).

charge plays a huge limitations because it induces an increase of the variability of the device, giving some problems or reliability. For this reason, in the last years, the interest of moving to the 3D Flash technology grew a lot, due to the benefits that it can produce, in particular the ability of increasing the density of stored data without decreasing the feature size.

1.2 2D NAND Flash memories

1.2.1 Functional principles

In this section, we want to introduce more in detail the technology which this work is about. The Fig. 1.2 reports the schematic structure of a planar Flash memory cell. In the image we can recognize the basic structure of a MOS n-type transistor with: two $\mathbf{n}^+$ wells that constitute the *source* and the *drain*, a **p** substrate and a *control gate (CG)*. In addiction, there is an electrode between the CG and the active area in the substrate. This electrode is not electrically connected and for this reason it is called *floating gate (FG)*. The oxide layer in between the active region and the FG is named *tunnel oxide* while that one between the FG and the CG is called *blocking oxide*.

In order to program the cell, a certain amount of electrons has to move inside the floating gate where the charge must be stored until a new operation on the cell is performed.

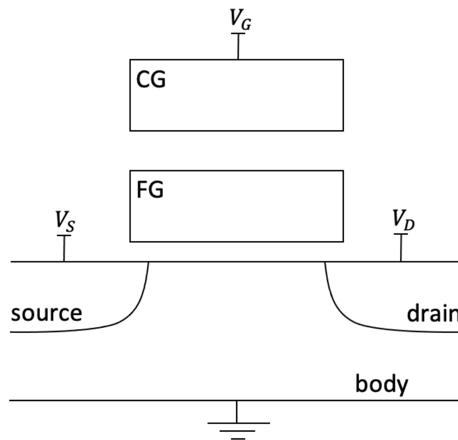


Figure 1.2: Schematic representation of a planar memory cell with *floating-gate*.

Fig. 1.3a and b show the band diagrams of a cell with positive and negative charge stored in the FG, respectively. In the former case, the state of the memory is called *ERASED*, while in the latter case the state is called *PROGRAMMED*. The presence of electrons in the FG for the programmed state makes the electrostatic potential lower, asking for a higher bias at the CG in order to get the same voltage drop on the substrate. In this way, the device has a higher *threshold voltage (V_T)* and, by changing the amount of stored charge, it is possible to control the V_T of the transistor. Considering a memory cell with just one logic level, it is customary to associate the logic level 1 to the *erased* state, when V_T is low. The logic level 0 is instead associated to the *programmed* state, when the threshold is high. Furthermore, the presence or not of electrons inside the FG makes the I_d-V_{cg} curve, that is the drain current as function of the CG potential, shift rigidly by a quantity ΔV_T that is proportional to the amount of stored charge.

$$\Delta V_T = -\frac{q \cdot n_{fg}}{C_{pp}} \tag{1.1}$$

Eq. 1.1 states the relation between those two quantities, where q is the fundamental electron, n_{cg} is the charge density inside the *floating-gate*, while C_{pp} is the capacitance between the *control-gate* and the *floating-gate*. Thanks to this, by applying a V_{read} voltage to the CG and reading the current flowing at the drain, it is possible to determine the logic state of the cell.

1.2.2 Programming and erasing a memory cell

Once we realized that the presence of charge inside the FG is fundamental to determine the cell state, we need to understand how it is possible to move the electrons in and out the floating gate. For what concerns the NAND technology, the quantum tunneling is

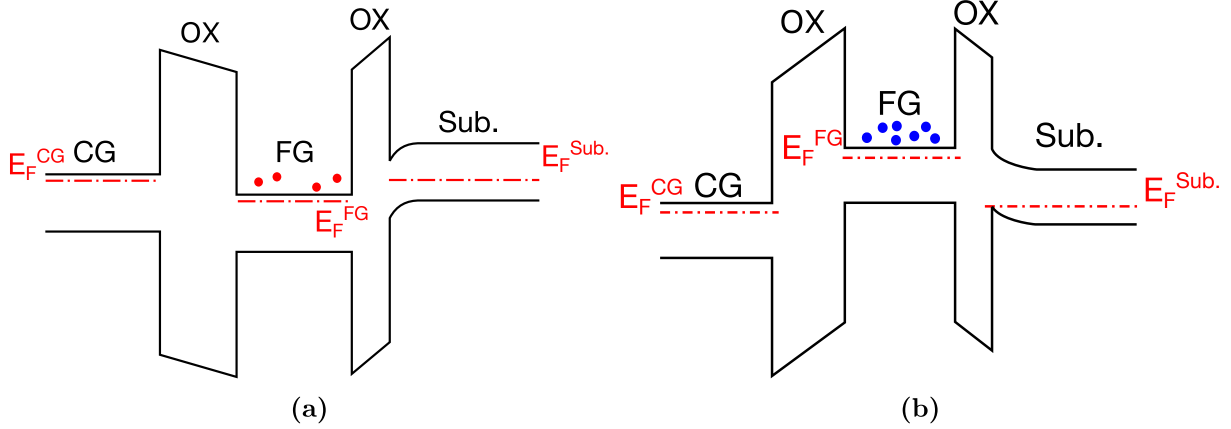


Figure 1.3: Band diagrams of a NAND Flash memory where (a) positive charge and (b) negative charge is stored.

exploited. If the tunnel oxide is sufficiently thin and V_{cg} is high enough, the band bending makes the barrier no longer rectangular but triangular and the barrier thickness that the electrons see is smaller. This condition takes the name of *Fowler-Nordheim (FN) regime* and it is characterised by a non-negligible probability that the tunneling takes place, granting the programming of the cell.

The electron flux J_{FN} just depends on the tunnel oxide field F_{ox} and on the barrier height ϕ_B with a relation described in the following equation:

$$J_{FN} = A_{FN} F_{ox}^2 e^{\frac{-B_{FN}}{F_{ox}}} \tag{1.2}$$

where

$$A_{FN} = \frac{q^3(2m_t + 4\sqrt{m_t m_l})}{16\pi^2 \hbar q \phi_B m_{ox}} \tag{1.3}$$

and

$$B_{FN} = \frac{4\sqrt{2m_{ox}}}{3\hbar q}(q\phi_B)^{\frac{3}{2}} \tag{1.4}$$

With m_{ox} we stated the tunneling effective mass, while m_t and m_l are the transversal and longitudinal effective mass, respectively. The same physical effect can be exploited both in the *programming phase* and in the *erasing phase* as depicted in Fig. 1.4a and b, respectively. In the former case V_{cg} must be positive, while in the latter the control gates must have a negative bias. These techniques are very efficient in power terms because, while the gate voltage controls the tunneling flux, source and drain are kept equipotential. So, there is no flowing current between these two regions and the power dissipation is kept at the minimum.

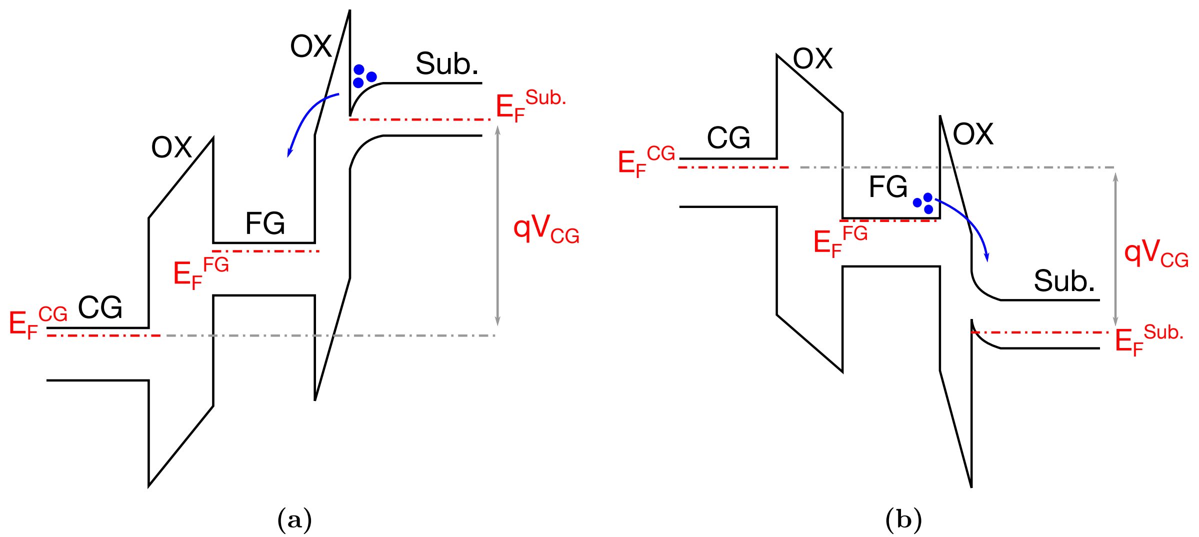


Figure 1.4: Band diagram for *2D NAND* memory during (a) the *programming* operation and (b) the *erasing* operation.

1.2.3 Array structure

A planar NAND Flash memory architecture usually is designed as in Fig. 1.5a, where we can see that the transistors are connected in series to form a string. Each string is connected to the respective *bit line (BL)*, while, on the other side, all the strings end up into a common *source line (SL)*. The control gates are driven by *word lines (WLs)* that run orthogonally with respect to the string direction. The *drain select line (DSL)* and *source select line (SSL)* are employed to select the desired string during the various operations that the memory bears. Finally, the cells adjacent to the DSL and SSL are considered dummy because they could behave differently with respect to the inner transistors. The source of this different behavior is the presence of fringing fields, i.e. peripheral fields, whose field lines come from adjacent cells, that produce some variations on the electrostatic potential. For this reason the dummy cells are not count as storing cells and their WLs are labeled by using the name *dummy lines*.

The cells belonging to one string are integrated on the same silicon stripe, while different strings are separated by *shallow trench isolations (STIs)*, as reported in Fig. 1.5b. The picture shows also the F size that in planar Flash memories is usually associated to the width of the WL. In the end, in Fig. 1.5c we can see a cross-section of the memory structure along the string direction.

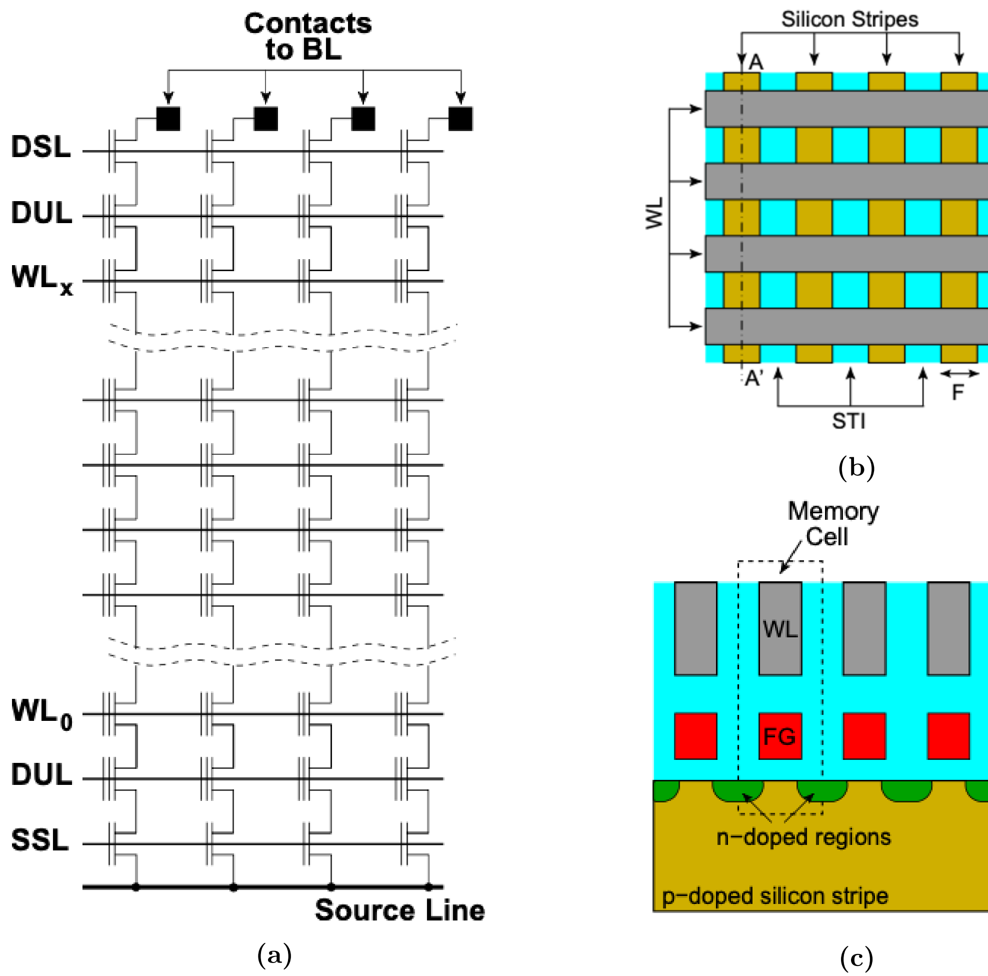


Figure 1.5: (a) Circuital scheme of a 2D NAND Flash array, highlighting the strings of floating-gate transistors driven by shared *WLs*. BL = *bit-line*; DSL = *drain-select line*; DUL = *dummy-line*; WL = *word-line*; SSL = *source-select line*; SL = *source-line*; (b)Planar layout of a 2D NAND Flash memory, (c) cross section of the array along the string direction (from [1]).

1.3 Array operations

As mentioned previously, the digital information is associated to the charge stored in the floating gate. Due to the linear relationship between the charge and the ΔV_T, it was realized that a single cell can store more than one bit. This can be done by discretizing the threshold voltage in 2^{BPC} levels, where BPC stands for *bits per cell*. In case of *single level cell* $BPC = 1$, for a *multi level cell* $BPC = 2$, and for a *triple level cell* $BPC = 3$. Fig. 1.6 depicts a schematic representation of the threshold voltages for the three technologies just spoken. For those memories with $BPC > 1$, all the states but the erased one are conventionally called *programmed* states.

Fig. 1.7 reports the bias scheme for the *read*, *program* and *erase* operations, that are

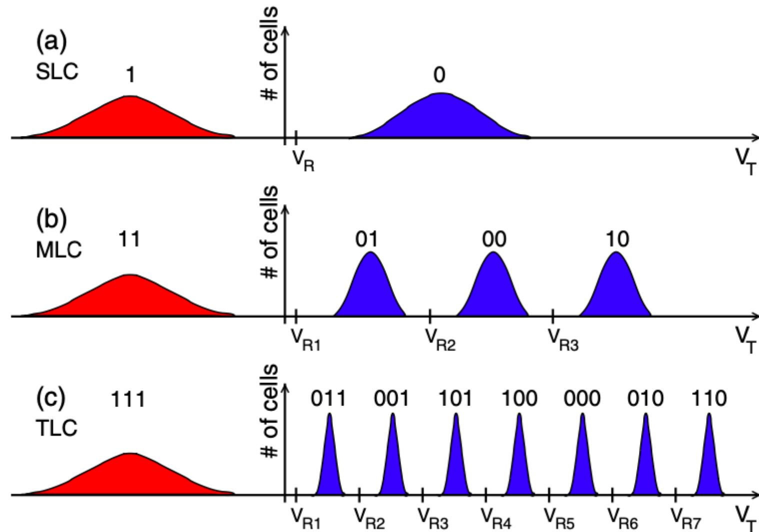


Figure 1.6: V_T distribution of cells in different states for (a) a *single level cell*, (b) a *multi level cell* and (c) *triple level cell* (from [1]).

discussed in detail in the following:

Read operation

The read operation consists in discriminating the V_T of a targeted cell irrespective of the threshold voltage of the other cells. In order to do so, the string must be connected to a sensing circuit where a sense amplifier allows to define if the selected V_T is lower or higher than V_{RX}, a reference voltage. The bias scheme is reported in Fig. 1.7a where we can see that the selected cell is biased to V_{RX}, while all the others are biased to V_{pass}^R, a voltage higher than the maximum V_T level, such that the unselected cells are in a high conductive state, irrespectively of their threshold voltage position. Moreover, the DSL and SSL are high as well as the BL, while the SL is kept to ground. In this way, a current can flow through the circuit and the threshold voltage can be sensed. We have to keep in mind that for the MLC and TLC, the read operations must be performed more than once to exactly define the position of V_T.

Program operation

As we said above, the FN tunneling is exploited to inject electrons into the FG and program the cell. With this scope, the string is biased as reported in Fig. 1.7b, in particular the selected cell sees a very high V_P voltage, while the unselected cells are biased to V_{pass}^P, different from the V_{pass}^R of the read operation. Moreover, the SL is disconnected from the string because the SSL is grounded, on the contrary the DSL is high and connects the grounded BL to the string. In the meanwhile, the

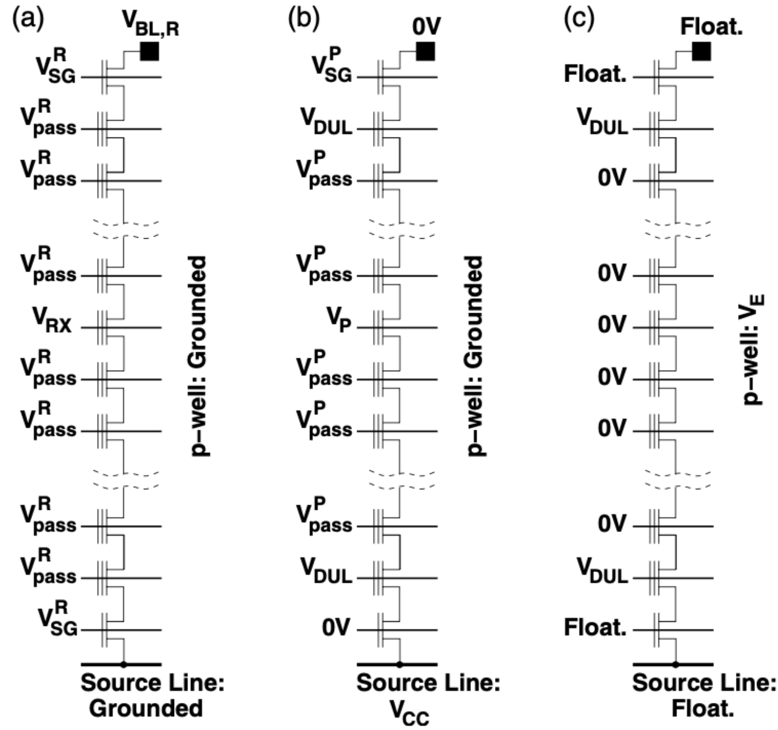


Figure 1.7: Array representation of a NAND string with bias scheme for (a) *read*, (b) *program* and *erase* operation (from [1]).

body of the string is kept to ground and the dummy lines are set to a voltage V_{DUL} lower that V_{pass}^{P} in order to decrease the electric field inside the dielectric material in the region between the DSL/SSL and the rest of the string. In this way, electrons can flow from the BL into the channel whose voltage is kept to **0**, then they can get to the FG thanks to the FN regime allowed by the high $V_P \approx 20\ V$. Since MLC and TLC memories require a fine placement of the V_T, a particular voltage pulses scheme for V_P is needed. This technique is called *Incremental Step Pulse Programming (ISPP)* method and consists of a fast sequence of pulses of increasing amplitude with a *verify*, as reported in Fig. 1.8. In such way, it is possible to better control the amount of charge injected into the floating gate and, in the end, have a threshold voltage very close to the desired level.

Erase operation

The erase operation is performed at the same time for all the cells belonging to one string. It exploits again the FN regime but, since high negative bias cannot be stood by the memory circuits, the bias scheme imposes that the WLs are grounded, while the p-well is set to a high positive voltage V_E. In this condition, the huge voltage drop between the channel and the gates makes the holes flow into the FG, exploiting once again the FN regime. The way the p-well voltage is driven to its final bias takes

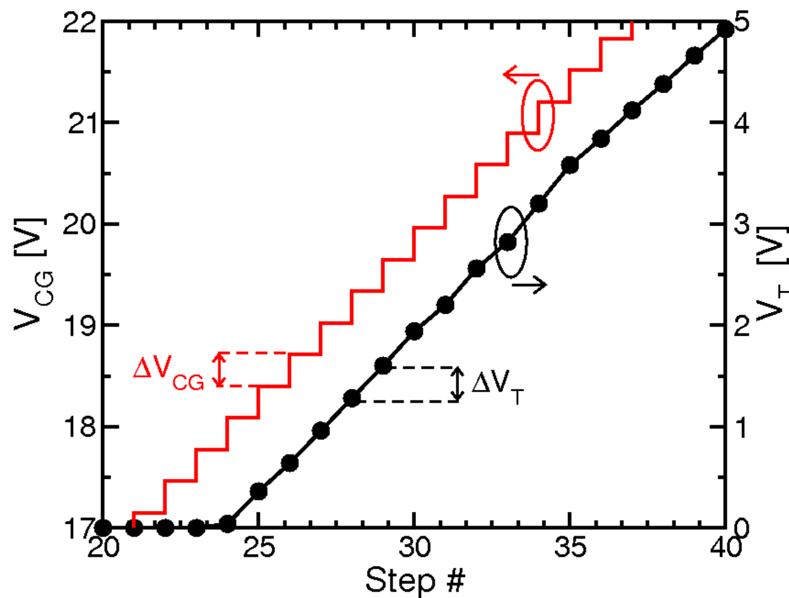


Figure 1.8: Example of the ISPP process. In red is depicted the *control gate* bias while in black is shown the corresponding programme V_T (from [2]).

place by means of fast voltage step pulses and the technique is called *incremental step-pulse erase (ISPE)*, like the programming technique. In the meanwhile, BL along with SL, DSL and SSL are left floating in order to keep as lowest as possible the electrical stress on the other parts of the circuit. The bias scheme is reported in Fig. 1.7c. At the end of this, as in the program phase, a verify operation occurs to confirm that all the cells actually have negative threshold voltage, otherwise another erase operation is performed.

1.4 Scaling issues in planar NAND memories

The reliability of nonvolatile memories represents the capability to store the data for a long time interval and to retrieve them intactly even if the array has performed some other operations. With the continuously growing demand for storing data, during the years the NAND technologies kept increasing the GBSD by enhancing the number of bits per cell and decreasing the *Feature* size. Unfortunately, this scaling down led to significant reliability issues that forced the introduction of *Error Correction Codes (ECCs)* in order to manage those errors that arise during the operations of the NAND memory. Errors occur when the detected V_T state of the targeted cell is changed from the desired V_T value that was set up. There are three main kinds of errors:

write errors: these are due to the inaccurate placement of V_T during the *program* operation. They usually consist in over-programming the cell and the errors mainly

arise from statistical fluctuations of the number of carrier undergoing the tunneling from the channel to the floating gate. An intrinsic fluctuation arises from the not completely stable amount of charge that during each step of the ISPP is injected through the tunnel oxide. This is what is is called *program noise*.

disturb errors: they are due to changes in the cell V_T while other operations are performed on the array. For instance, due to the high WL bias during the *read* operation, some electrons could be unintentionally injected into the FG. Another example of this kind of errors is the change of the V_T of unselected due to the change of the threshold in neighbors cells, that is also called *cell-to-cell electrostatic interference*.

data retention errors: they are instabilities in the V_T even if no operations are perturbing the memory state. One possible source of this disturb could be the *Stress-Induced Leakage Current (SILC)*, an undesired electron injection into the FG that takes place by means of the defects inside the tunnel-oxide layer. This phenomenon changes the *low-field conductivity* of the tunnel-oxide layer allowing the carriers passage even with almost no bias applied to the contacts. Defects in the tunnel-oxide layer can give rise to other V_T changes. Some examples of events belonging to this category of disturbs are the *Random Telegraph Noise (RTN)* and the *charge detrapping*.

We have just seen that there are several physical effects that impact negatively the data storage and retention of a memory but in order to better understand how the extreme scaling is impacting the reliability of the 2D NAND Flash memories it's better to further describe some of these issues.

1.4.1 Program noise

We have seen that the increase of the threshold voltage follows Eq. 1.1. This equation tells us that the threshold voltage is affected by the scaling. Indeed, the reduction of the cell area, coming along with the reduction of F, leads to a decrease of the cell capacitances that couples the floating gate to the floating channel and to the control gate. In turn, this is reflected on C_{pp} that becomes smaller. This means that single electrons have a greater impact on the cell threshold voltage the same amount of charge has a greater impact on the threshold voltage, i.e. the same variation of stored charge produces a larger shift in ΔV_T when C_{pp} is lower. Furthermore, the reduction of the characteristic feature size of the cell decreases the number of transferred electrons to/from the floating gate. This explain how a small difference in the amount of charge could produce a relevant variation of the performances of the cell. So, as the value of the C_{pp} has reached minimal value, the granular nature of the charge has became relevant. This fact leads to an important

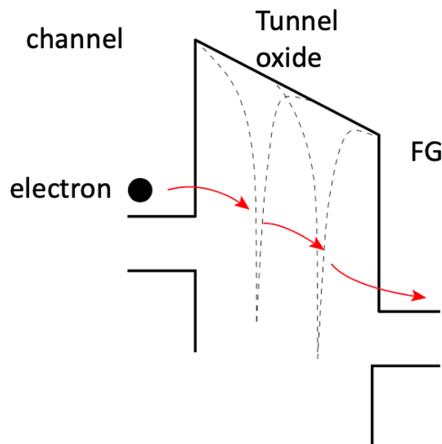


Figure 1.9: Pictorial representation of an electron undergoing the tunneling assisted by two traps states (from [3]).

consequence concerning the program process. In fact, due to the statistical nature of the electron injection during each pulse of the ISPP, the V_T no longer can be considered deterministic but becomes a statistical event that is affected by fluctuations. In case each ISPP pulse produces a small ΔV_T, the variability of this quantity can be found out considering a Poissonian distribution of the injected charge and resulting in the following equation:

$$\sigma_{\Delta V_T} \simeq \sqrt{\frac{q\langle \Delta V_T \rangle}{C_{pp}}} \tag{1.5}$$

1.4.2 Stress-Induced Leakage Current

In order to have a correct retention of the data stored the floating gate, just a minimum tunnel oxide layers of few nanometers is needed. Even though, in the last years, the thickness of the oxide has been not affected by the scaling because it was seen that the repetition of many *program/erase (P/E)* cycles compromises the quality of the dielectric adjacent to the FG. In fact, the repeated application of high electric field causes the onset of traps states in the oxide layer that leads to the problem called *Stress-Induced Leakage Current (SILC)*, i.e. the unwanted transfer of carriers from/to the FG. With the passing of time, the small thickness of the layer, together with the presence of defects, could bring to the unwanted transfer of carriers even if the gate voltage is low. This problem consists of an assisted tunneling injection of the electrons thanks to the presence of the defects in the oxide, as depicted in Fig. 1.9 where we can see the tunneling assisted by two traps close to each other. The picture shows a case where there are simultaneously two defects

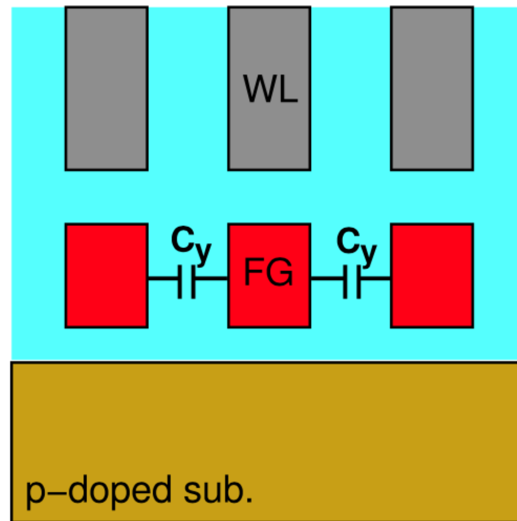
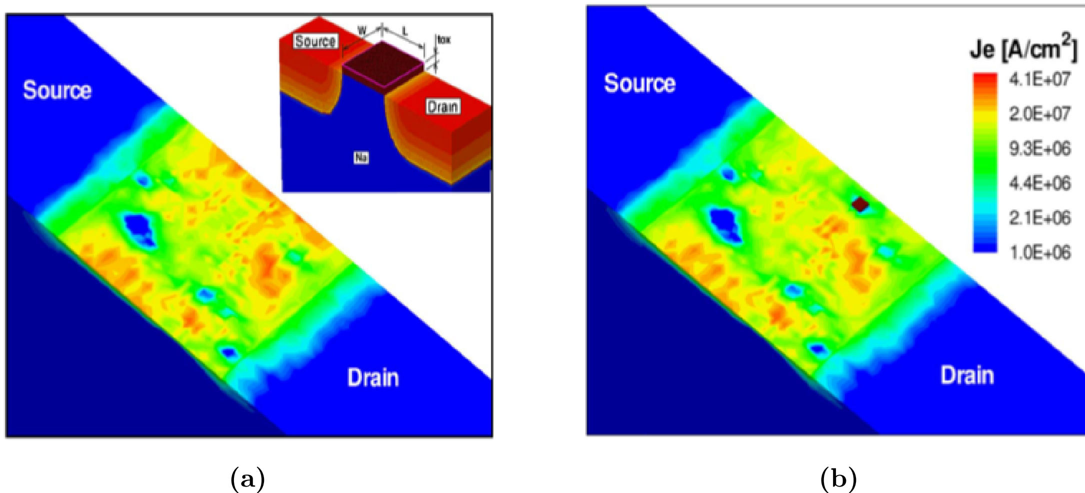


Figure 1.10: Schematic representation of the capacitances between first neighbor FGs (from [4]).

in the tunnel oxide layer. As we can seen, these isolated states modify the potential inside the oxide barrier. This modification consists in a lowering of the potential that in turn makes the *effective thickness* of the oxide smaller. Regarding the effective thickness, it is the thickness of the barrier seen by the electrons that accomplish the tunneling. This thickness is not the real spatial thickness of the oxide layer, actually these two values are independent. In fact, the barrier thickness is just determined by the electric field inside the oxide and it is precisely this value that affect the tunneling current as we have seen in Eq. 1.2. Even though, with the presence of defects, this barrier thickness can undergo variations of its value and this brings to a leakage current even when the electrostatic potential at the gates is low. In the end, this undesired current produces small ΔV_T that can become relevant in MLCs and TLCs memories where the separation between V_T levels is rather small and there is a reduced C_{pp}. All of this makes even an extremely small leakage current be a significant problem for the reliability of the memory.

1.4.3 Cell-to-cell electrostatic interference

Due to the reduction of the cell size, the distance between adjacent cells is reduced and consequently the electrostatic interference becomes more important. The Cell-to-cell electrostatic interference actually is the onset of a ΔV_T, usually positive, in a *victim* cell due to the change of the V_T state of other cells in the string. In particular the shift is higher if the level variation takes place in first neighbor cells, called *aggressor*. In planar MOSFET the problem can be associated to the capacitive coupling between the FG of the aggressor and the FG of the victim, as depicted in Fig. 1.10 where some FG-FG



(a) (b)

Figure 1.11: (a) Example of simulated current density for a MOSFET where atomistic doping is considered, (b) comparison of the density current in the case of a trapped electron (from [5]).

capacitances along the string direction are depicted. For this reason the phenomenon becomes dependent on the design of the string, in particular on the FG height and on the distance between FGs. The magnitude of the disturb can be found out with the following expression:

$$\Delta V_T^{vict} \simeq \frac{C_{FG}^{par}}{C_{FG}^{tot}} \cdot \Delta V_T^{agg} \tag{1.6}$$

where ΔV_T^{vict} and ΔV_T^{agg} are the threshold voltage shifts of the victim and aggressor, respectively. C_{FG}^{par} is the parasitic capacitance between neighbor FGs, while C_{FG}^{tot} is the total FG capacitance of the cell. Due to the fact that the scaling is not proportional for all the dimensions of a cell, the threshold voltage shift has become greater with the miniaturization of the cell. In fact, if on one hand the WL pitch is decreased, i.e. the width and the depth of a cell have been reduced, on the other hand the height of the floating gate has remained almost the same, such that the $C_{FG}^{par}/C_{FG}^{tot}$ ratio is increased with the scaling.

1.4.4 Random Telegraph Noise

Defects in the tunnel oxide layer are of different kinds and they give rise not just to write errors but also to retention errors. For instance, those defects localized at the interface between channel and oxide are the source of the *RTN*. This effect is essentially a two-level fluctuation of the current read at the drain. The fluctuation is generated by the event of capture or release of an electron from the defect at the interface. The event becomes particularly significant when the scaling of the cells is rather big. In

fact, the interplay of the localized nature of the defect and the percolative nature of the channel conduction could give rise to large variation of drain current and, in turn, to huge ΔV_T shifts. Fig.1.11a and Fig.1.11b shows two examples simulated current density for a deca-nanometer MOSFET. In this high scaled MOSFET, the atomistic nature of the doping leads to a statistical variation in number and position of the dopants. This doping condition is called *Random Dopant Fluctuation (RDF)* and it is the source of a particular non-uniform conduction in the channel. Indeed, the random disposition of the dopants gives rise to a filamentary current flow, i.e. percolative paths, where the conductivity is higher, are established. Then, if a defect is present right above a percolative path, the capture of just one electron is sufficient to produces a huge variation in the conductivity of the channel because it means that through that path the flow of carriers will be lower. The decrease of the conductivity brings to a reduction of I_D that, in turn, can be associated to an increase of the threshold voltage.

1.5 3D NAND Flash memories

In the last decade, it was seen that the scaling of the device was reaching the technology limit and so new solutions were necessary to keep up with the GBSD trend requested by the Moore's law. For this reason, many efforts were driven to the study of the 3D NAND Flash technologies because with that structure is possible to increment the GBSD, keeping a quite large F. The main idea of the 3D NAND technologies is based on the exploitation of the vertical dimension to get a higher bit density in the same wafer area. Even though, it was seen that just a replica of the transistors along the **z** axis was not a convenient way to keep with the evolution of the memory but a rather complex redesign of the structure was necessary. From this geometry revolution, two main groups of 3D memories came out: the *Vertical channel memories* and the *vertical gate memories*. In this work the attention will be focused on the former one due to its relevance in the market.

1.5.1 Vertical channel structure

In Fig. 1.12, a representation of a *vertical channel memory* is reported. We can see that the WLs are no longer parallel lines but parallel planes that surround several vertical pillars. These pillars carry out the counterpart of the channel for the planar MOSFETs, so from this point on we will use the word channel to refer the whole pillar. Each intersection between a pillar and a WL corresponds to a single cell and the set of cells belonging to one pillar constitutes one string. The string is connected at the bottom to the *source line*, that is common to all the channels, while at the top, different rows of pillars are connected

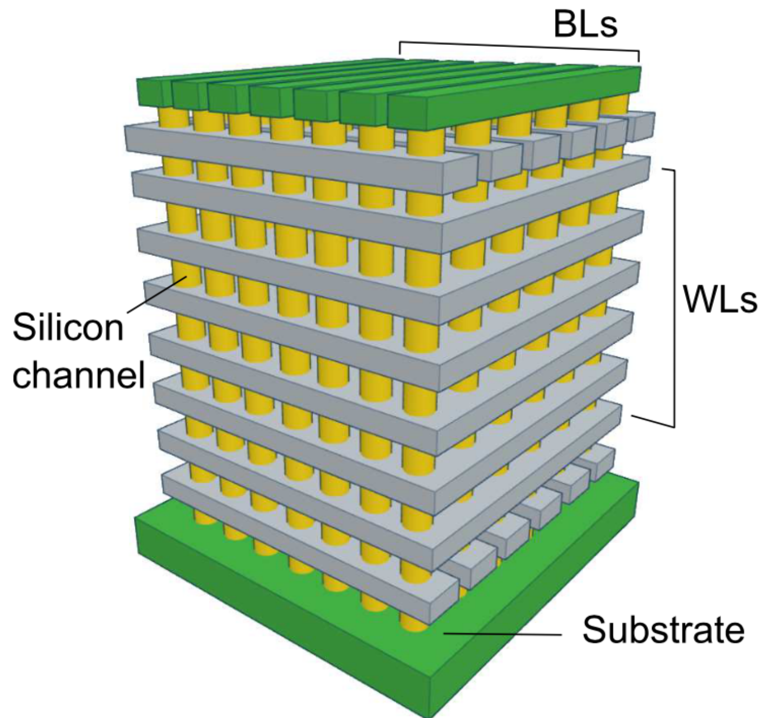


Figure 1.12: Simple representation of a vertical channel memory (from [1]).

to different *bit lines*. Furthermore, the first and the last plane of WL are etched in order to create rows perpendicular to the *BLs* such that the single string could be selected.

Fig. 1.13a depicts a simple representation of a *Gate All Around (GAA)* cell, where we can see clearly that the gate contact is on the whole lateral surface of the pillar, allowing a better control of the electrostatics inside the channel. We notice that in the image there is a sequence of three different dielectrics. Indeed, in 3D NAND memories the substitution of the floating gate with a dielectric material, having a high density of traps, is a quite efficient solution. The reason for adopting this different design lies in the fact that localized traps make the trapping layer more immune to *leakage* problems. In particular, for the GAA geometry, the leakage involves not just the unwanted passage of charge through the tunnel layer but also the carrier flow through the blocking layer. This happens due to the high coupling that the cylindrical geometry actually grants to the control gates. So the solution was adopting the *Oxide-Nitride-Oxide (ONO)* stack made of SiO_2-Si_3N_4-SiO_2. The Si_3N_4 dielectric layer is the layer where the carriers are stored. The storing takes place exploiting the capture of the carriers from the many traps present in the layer that, for this reason, is also called *trapping layer*. The WLs are composed of heavily doped polysilicon and some insulating layers are positioned between the control gates in order to avoid unwanted currents flow. Fig. 1.13b shows a vertical section of this structure.

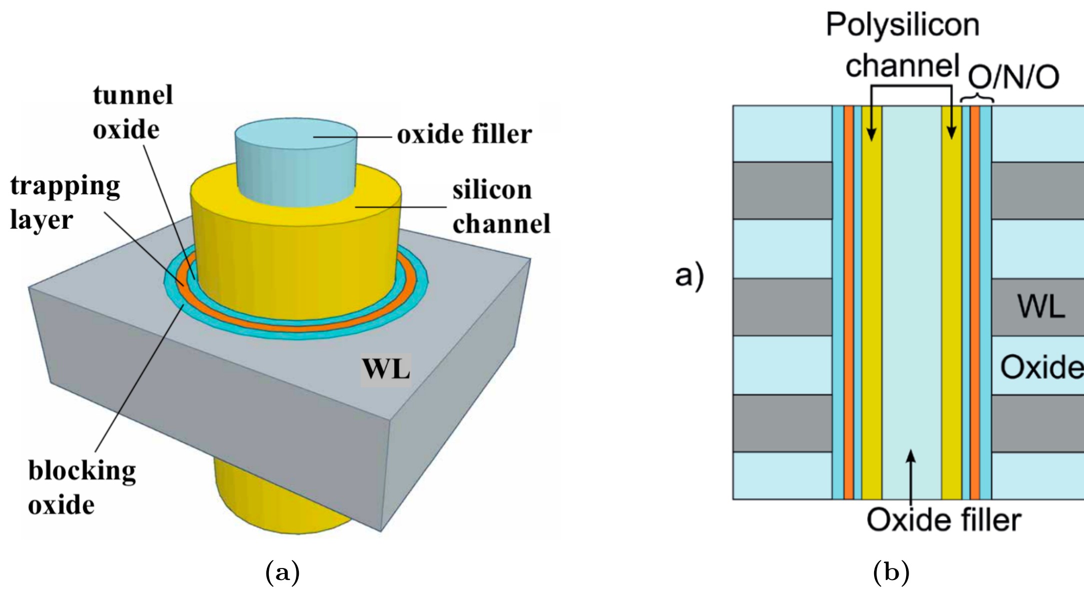


Figure 1.13: (a) Schematic view of a single cell for a GAA memory structure (b) Vertical section of a GAA 3D NAND memoriy(from [1]).

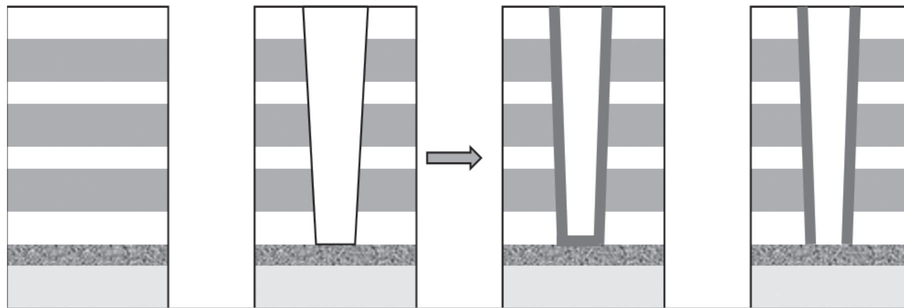


Figure 1.14: Manufacturing process flow (from [6]).

For what concern the production of this kind of structure, Fig. 1.14 shows the main steps of the manufacturing process. In particular we can see that the pillars are created after an etching operation that destroys the sacrificial columns and leave holes in the stack of the *WLs*. This hole is then filled with the desired layer in the correct order. First comes the deposition of the *Oxide-Nitride-Oxide* sequence, next the deposition of the silicon and finally the dielectric filler Due to the fashion in which the deposition takes place, a crystalline silicon cannot be used for the channel, but that is made up of polycrystalline silicon. This material is composed by several *grains* whose boundaries are characterized by the presence of many dangling bonds that acts like traps for the electrons. In order to have the least number of defects in the channel, the dielectric filler is adopted. Indeed, the filling of the centre part of the hole with the dielectric filler is convenient for two main reason. First, the grains tend to meet each other at the centre of the hole and in that region the concentration of trap states is higher. Second, the reduction of the

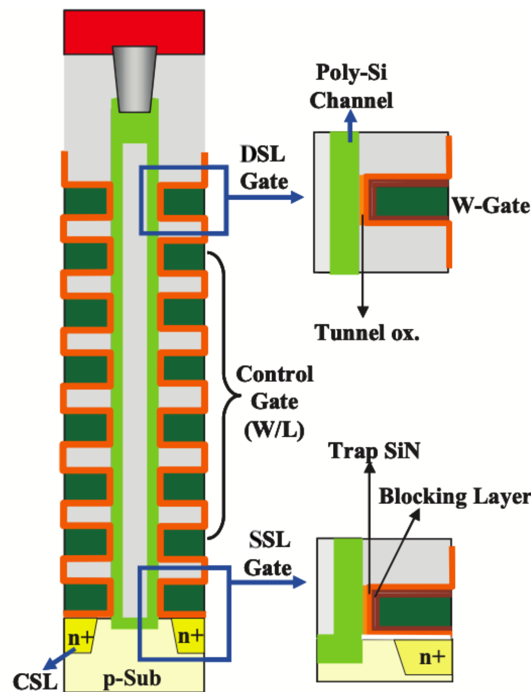


Figure 1.15: Vertical section of a *TCAT* flash memory (from [7]).

silicon thickness decreases the drawbacks given by the short channel effect. This means that, without the filler, if the length of the WL is too short, the field inside the channel becomes intrinsically two-dimensional, i.e. not just the longitudinal component of the field is relevant but also the radial one and this brings to a degradation of the system characteristics.

The whole manufacturing process described so far is called *gate first* and the particular kind of vertical memory that has been just described goes with the name of *Bit-Cost Scalable (BICS)* memory. There is another manufacturing process flow, the *gate last* process flow, that is used for the creation of the *Terabit Cell Array Transistor (TCAT)* memories. Fig. 1.15 shows a representation of this kind of memory. This technology has some differences with respect the former one. First, the *WLs* are made of metal and not of highly doped polysilicon and they do not have cylindrical symmetry. Second, in the case of *gate first* the substrate consists of just a unique region doped $\mathbf{n}^+$, instead in the case of *gate last* the channel is linked to a **p** doped substrate with the presence of lateral wells doped $\mathbf{n}^+$ that work as *source* regions. A common characteristic of these two structures is the absence of the body contact. This constitutes a main difference with respect to the planar MOSFETs and it brings to the onset of new phenomenon during the operations of the memory. In fact, the design of the vertical channels force the carriers to move towards either the BL or the SL in order to get in/out the string. When this movement is impeded, the channel enters in a floating condition and new effect arises.

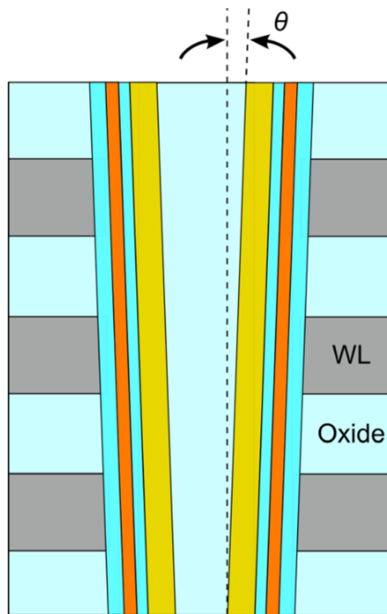


Figure 1.16: Schematic vertical section of a 3D NAND Flash memory highlighting the non-uniform pillar radius in the vertical direction (from [1]).

1.6 Advantages and issues of 3D memories

3D NAND Flash memories offer the main advantage of reducing those problems reported in section 1.4. This kind of memory has the possibility of increasing its GBSD even if the F of the memory is kept relatively large. The fact that the characteristic dimension of the cell is not so small mitigates the weight that each carrier has in determining the threshold voltage. Indeed, having a quite large F means that the variability of the V_T due to the program noise and to the RTN decreases because the statistics describing these effects is based on a higher number of events. This brings to a thinner ΔV_T distribution and allows the achievement of a better endurance for 3D technologies. Then, the GAA structure produces other advantages. In particular, it reduces the *cell-to-cell* electrostatic interference between WLs. In fact, in vertical channel structure the interference is restricted just between first neighbors. Moreover, the better control of the cylindrical structure on the electrostatic of the channel reduces the short channel problems and allows the speed-up of the programming schemes.

Even though, the transition to vertical arrays is not occurring without the onset of new issues. The main drawbacks lie in the high density of defects in the polysilicon. Indeed, the presence of grain boundaries comes along with the presence of energy barriers in the channel that makes the RTN dependent on the lattice temperature and the channel resistance higher. This in turn brings to a reduction of the read current and the need of the development of better sensing circuits. Moreover, the manufacturing process itself is

not immune from problems. In fact the manufacturing technology is not able to produce pillars with constant radius along the whole height. What happens is that the pillars is thinner at the base, where it is connected to the SL, and larger at the top, as reported in Fig. 1.16. The tilting of the pillar produces a variability between cells belonging to the same string and in turn gives rise to a reliability problem. This issue is linked to the engineering problem of the progress of the 3D NAND Flash technology. In fact, due to the narrowing of the pillar at positions close to the SL, the diameter of the pillar cannot be scaled too much and so the first solution, in order to increase the GBSD, is enhancing the number of WLs. This is an attractive challenge because the enhancing of the number of layer leads to higher pillars and, in turn, to a higher channel resistance. This means that the read current will be progressively lower and further improvements in the sensing circuits are needed.

For what concerns the dynamic of the operations that occur in the 3D NAND Flash memories, they are very similar to what we have presented in the previous section for planar MOSFET. The main difference lies in the absence of the body contact for the 3D NAND Flash structures. This brings to the onset of new effects in order that the memory could carry out the different operations, an example is the exploitation of the GIDL during the erase operation in BiCS memory. On the other hand, due to the fact that the channel easily enters in a floating state without the body contact, new disturbs arises as well. One of them derives from what is called Down-Coupling Phenomenon (DCP) that will be discussed in detail in the remaining part of this work.

1.7 Thesis work goals

As we have mentioned, the transition from planar memories to a vertical technology has brought to a huge re-design of the string. Even if the functioning of the memories has not changed a lot, the differences in the structure make the inspection of the operations necessary. In particular, the absence of a direct connection of the BODY contact with the channel led to the onset of new phenomena related to the floating condition of the channel during some operations. The down-coupling phenomenon is one of this and in the following chapter this phenomenon will be described in detail. The goal of this work is actually understand the physics of this effect in order to be able of solving the related disturbs. At first, we will study the effect in the BiCS structure due to the simplicity of dealing with this geometry, simplicity that comes from the symmetry. Then, we will analyze the DCP in a TCAT structure in order to see if the presence of a **p** doped substrate could bring variations in the effect and, in turn, mitigate the disturbs associated to the phenomenon.

Chapter 2

Gate-induced drain leakage effect and down-coupling phenomenon

In this chapter some important physical effects characterising the 3D NAND Flash technology will be analyzed. The gate-induced drain leakage effect will be described in more detail because it plays an important role in the neutralization of a disturb that takes place just in 3D NAND Flash memories during the program operation. This disturb can be generated when the down-coupling phenomenon and the self-boosting effect occur one after the other in a short time interval. So, in this chapter ,these two effects will be presented in order to give the basis to understand the work presented in the following part of this work.

2.1 Gate-induced drain leakage

In order to understand the results of the following chapters we need to describe some effects that take place during the operation of the NAND memory. An important phenomenon that we encountered during our simulation is a particular kind of carrier flow that involve the passage of the charge through the energy gap of the semiconductor and it takes the name of *Gate Induced Drain Leakage (GIDL)*. It is a particular phenomenon of *Band-To-Band Tunneling (B2BT)* that it could take place when the bands are so bent that there are state in the valence band and in the conduction band sharing the same energy. In order to give a better explanation, we start with describing the phenomenon when it occurs in planar MOSFETs.

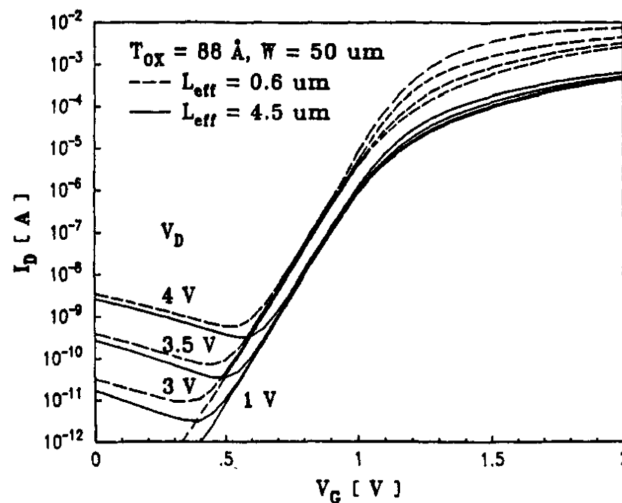


Figure 2.1: Subthreshold characteristic of planar *n-MOSFETs* showing a significant drain leakage current when V_D is high (from [8]).

2.1.1 Physical origin of GIDL

The GIDL was seen for the first time in the 2D planar MOSFETs during the analysis of the OFF state current in sub-micrometer channel devices. The study showed that the OFF state current was higher than the expected sub-threshold current and soon it was understood that there was the presence of a leakage contribution. Fig. 2.1 shows the comparison of several transcharacteristics of planar MOSFETs when different voltages are applied at the drain. It can be seen that the current does not drop down to the null value when the gate voltage tends to $V_G = 0$ and that this current is larger as the drain voltage is higher.

The origin of this effect is attributed to the *Band-to-band Tunneling* (B2BT) at the junction between the drain and the body of the transistor when the junction is in a condition of reverse bias. In particular the phenomenon corresponds to the tunneling of electrons from the *valence band* (VB) to the *conduction band* (CB) of the n^+ region of the drain where then the carriers flow through the contact due to the positive bias applied to the drain. On the other hand, as the electrons perform the tunneling, it leave empty states in the VB and so holes are generated at the VB side of the barrier. When holes appears in the VB, they feel the electric field established at the junction between the drain and the substrate so that they are pushed towards the substrate generating a hole current that flows at the BODY contact. Fig. 2.2a shows a schematic representation of the carrier displacement. From the picture we notice that the effect takes place in the zone of the drain superposed with the gate. In this region the positive voltage applied at the drain and the low or even negative voltage at the gate tend to push away the electrons and to leave positive charge at the surface underneath the gate. In the drain, in

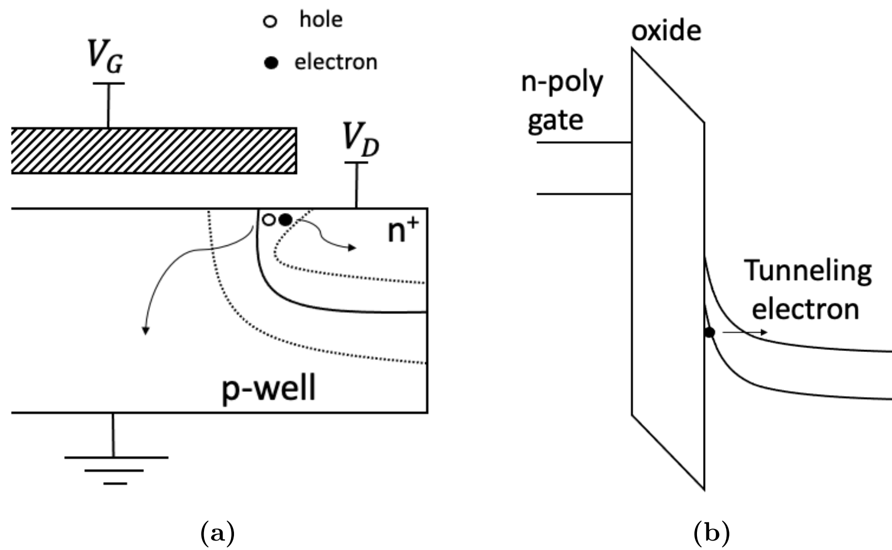


Figure 2.2: (a) Schematic representation of the carrier movement during the B2BT and (b) representation of the band bending and tunneling of the electron (from [8]).

the zone superposed with the gate, a band bending, like that one depicted in Fig. 2.2b, is created. We see in the picture, that the band bending can be so high that some states of the VB end up having the same energy of the CB. This fact generates the possibility that the tunneling mechanism takes place at the band gap. It means that, without a collision event, an electron in the VB can disappear, leaving in its position a hole, and instantaneously an electron appears in a state of the CB with the same energy of the disappeared electron. We also notice from the picture that the barrier thickness that the electron sees to perform the tunneling is determined by the tilting of the bands. In fact the thickness of the barrier depends on the electric field present at the position where the effect occurs. The problem can be traced back to the case of a triangular barrier where just the height of the barrier, i.e. the energy band gap, and the electric field determine the tunneling flux. Considering that, the tunneling current can be described with an expression alike the Eq. 1.2 that we have seen in the previous chapter to describe the FN regime.

2.1.2 GIDL in BiCS technologies

The GIDL is also present in 3D NAND memories where actually the phenomenon is not seen only in a negative way, but in some structures, like the BiCS memories, it is exploited during the erase operation. This operation in 3D memories has a dynamic very similar to the comparable operation in planar MOSFET. The bias scheme expect that the WLs are grounded while the channel is brought to a high electrostatic potential so that a huge voltage drop could establish the FN regime and inject holes into the trapping layer.

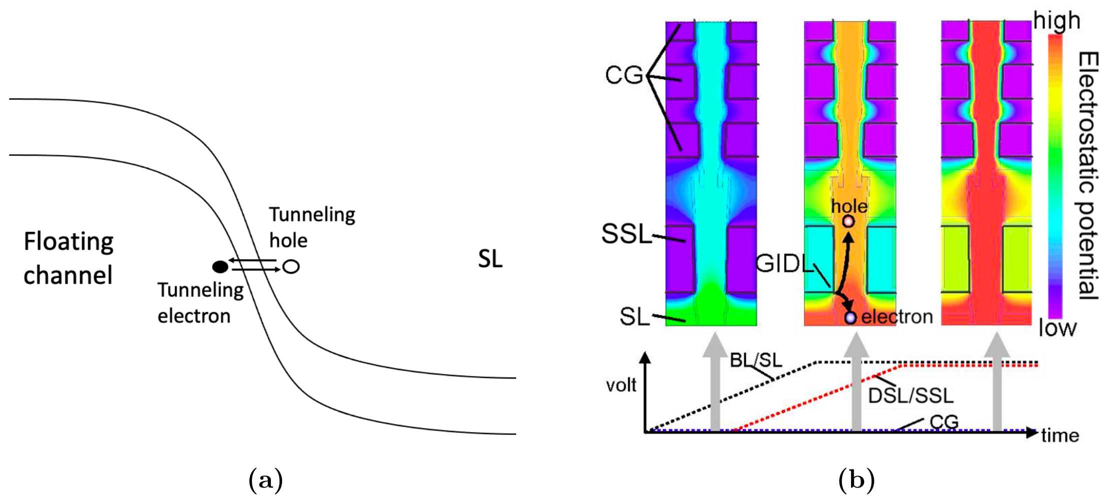


Figure 2.3: (a) Schematic representation of the band-to-band tunneling in BiCS memories, (b) Simulated result for a *erase* operation (from [9]).

But in case of 3D NAND Flash technology, the channel is not directly connected to the body contact so it must be brought to high positive voltage in another way. The solution is to create a floating channel condition where the V_{WL} is kept to ground and the same is done with DSL and SSL so that all the transistors are turned OFF. In the meanwhile, the BL and the SL are biased at high voltage in order to push holes inside the channel. Even though, BiCS memories have the BL and the SL composed by a n-doped silicon, so they do not have a positive carrier supply. At this point the GIDL gets into action because the high voltage drop between the SL and the channel, or between the BL and the channel, makes the band-to-band tunneling possible. Fig. 2.3a shows a schematic representation of the band diagram. In particular the representations corresponds to a section parallel to the channel direction. The picture shows how the band bending present at the edge of the floating channel allows the tunneling of the electron. As the electron is generated at the SL side, a hole is generated inside the channel and in this way it is possible to inject positive carriers into the string and increase the voltage of the channel. Fig. 2.3b shows the temporal evolution of what we have just described. We can see that at the beginning, despite of the the increase of the BL and SL potential, the WLs are turned off. This brings to a floating condition of the floating channel that, being capacitively coupled with the CGs, remains at a low electrostatic potential. Then, the holes start being injected into the channel that increases its potential until it reaches the same value of the the BL and the SL. As we will see in the remaining part of this work, this kind of current flow is not exclusively associated to the BiCS memories or to the erase operation, but it can take place when in 3D NAND Flash memories the channel is in a floating state and there is a significant voltage drop with respect to the BL and the SL. In fact we will encountered

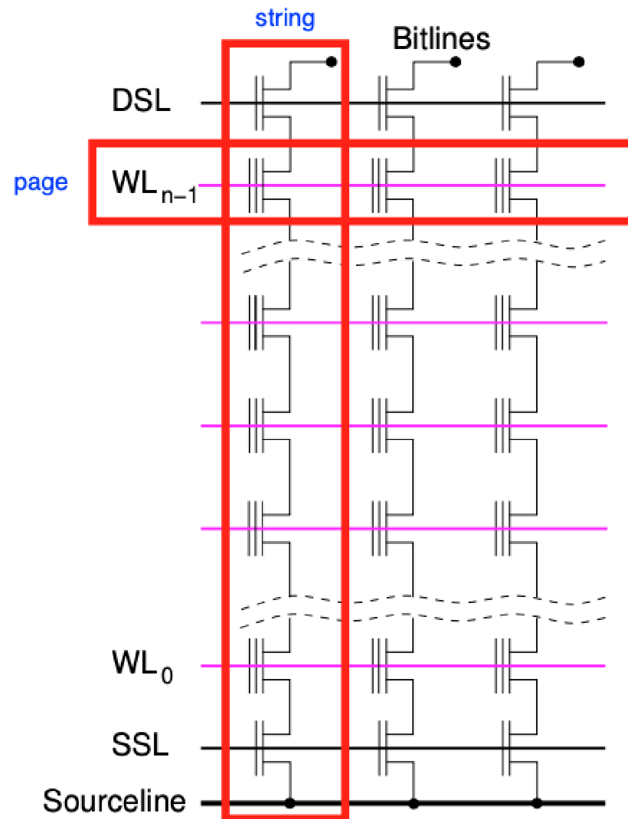


Figure 2.4: Schematic representation of a NAND Flash array where the various connection are highlighted (from [10]).

the GIDL during the program operation.

2.2 Disturbs related to the program operation

In this section we want to better describe the disturbs related to the program operation in order to have a better picture of the topic of the next chapters. To do so we want to remind the composition of the memory that is composed by a series of connections forming a matrix structure. Fig. 2.5 shows in detail how the array is subdivided. The transistors connected in series to one BL form what is called *string*. A control gate, connected across the strings, takes the name of *word line (WL)* and the set of transistors connected to the same WL, but belonging to different strings, form a *page*. This implies that when a certain operation is performed on the array, the bias scheme must ensure that unselected cells on the string are not affected by the operation. In the same time it must be ensured that unselected cells on the same page of the selected cell, but on different strings, maintain their original state as well. In the previous chapter we have showed Fig. 1.7 that reports the bias scheme related to the program operation in the selected string. In this case, we

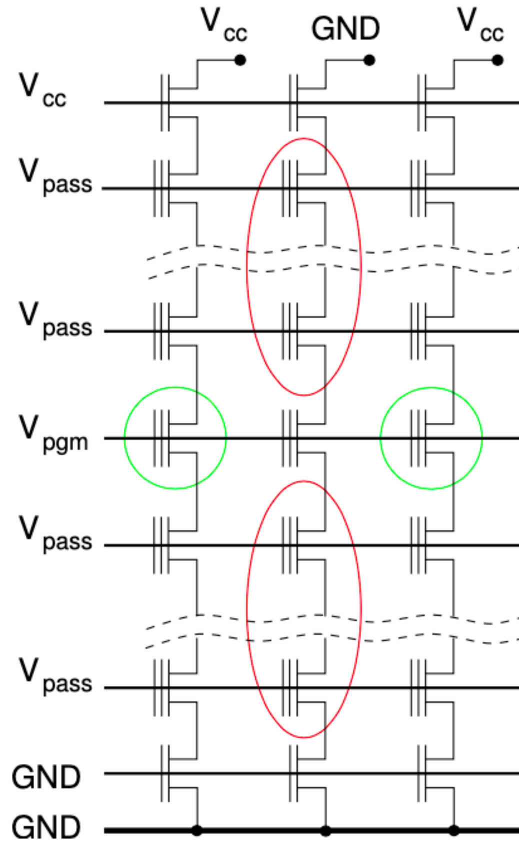


Figure 2.5: Schematic representation of a program operation in a 2D NAND Flash array where in red are circled the cells subjected to pass disturbs and in green the cells subjected to program disturbs (from [10]).

have seen that the unselected WLs have a V_{pass} voltage that is a lot lower than $V_P \approx 20\ V$ applied to the selected WL. Due to the electrostatic interference between WLs, it is not possible to set $V_P \approx 20\ V$ and at the same time $V_{pass} \approx 0$. On the other hand, if V_{pass} is too large the voltage drop between the channel and the unselected cells can be sufficient to establish the FN regime and produce a *soft-programming*. This unwanted injection of electrons in unselected cells produces shifts in the V_T of the cells and it is called *pass disturb*. For these reason usually the value of V_{pass} is set to approximately $8\ V$ for 3D NAND Flash memories. Fig. represents the NAND array and in red highlights the cells subjected to pass disturb.

Furthermore, we are also interested on what happens in unselected strings. We have to point out that the only difference in the scheme bias between a selected string and an inhibited one lies in V_{BL}. In fact, the BL is not grounded but it has a voltage V_{CC} equal to that one applied to the DSL. In this way, when also the channel reaches about the electrostatic potential of V_{CC}, the DSL is turned OFF and the channel enter in a floating condition. Then, when the pass voltage is applied to the WLs, thanks to the capacitive

coupling established between the floating channel and the WLs, the channel potential is *boosted* to a value high enough to avoid the FN regime. This process is called *self boosting* and it is used to avoid the injection of electrons into the FG of cells belonging to the selected page but related to inhibited BLs. This particular injection gives rise to what is called *program disturb* and the cells affected by this disturb are circled in green in Fig. In order to better understand how this last disturb can be avoided let us analyze in detail the self boosting effect.

2.3 Self-Boosting effect

Already in the middle of the 90s, the *self-boosting* was exploited as inhibition program technique. It was rather promising because it allowed the improvement of the scalability of the memory. Indeed, in the years before the advent of this improvement, in order to avoid the programming of unselected cells in inhibited strings, the BL was driven to a high voltage. This inhibition scheme gave some problems because a quite big amount of time was needed to boost the BLs and a lot of power was required as well. In the *self-boosting*, it was decided to not involve too much the highly-capacitive *BL* but exploit the *Gate-channel* capacitance. The inhibition scheme expects, also reported in that the SSL is kept to ground like the SL, while the DSL, right before the program pulse, is biased to V_{cc}. The same V_{cc} potential is applied to the BL, in this way the channel is precharged to a voltage $V_{cc} - V_T^{DSL}$, [11]. By doing so, the DSL turns off, leaving the string in a *floating* state. This condition allows the temporarily increase of the channel potential by means of the capacitive coupling with the WLs. The final potential of the boosted channel can be modeled in this way [12]:

$$\Delta V_{CH} = \frac{C_{WL-CH}}{NC_{tot}} \left(N-1\right) V_{pass} + \frac{C_{WL-CH}}{NC_{tot}} V_{pgm} \qquad (2.1)$$

Where C_{WL-CH} is the capacitance between the gate and the channel; C_{tot} is the total unit cell capacitance including other parasitic components; N is the total number of *WLs* in the string. This makes the voltage drop between the gates and the channel lower and the tunnels oxide field is lower as well, preventing the electron injection into unselected cells. However, for the 2D case there is still an electron flow coming from the *p-well* that eventually lowers the channel potential. Indeed, it's quite difficult completely decoupling the channel from the substrate. From this point of view, this technique found a further development with the advent of the 3D NAND memories. Due to the singular geometry of the three dimensional structure, the process of cutting off the channel from the BL is quite more efficient than the planar counterpart. In vertical arrays, indeed,

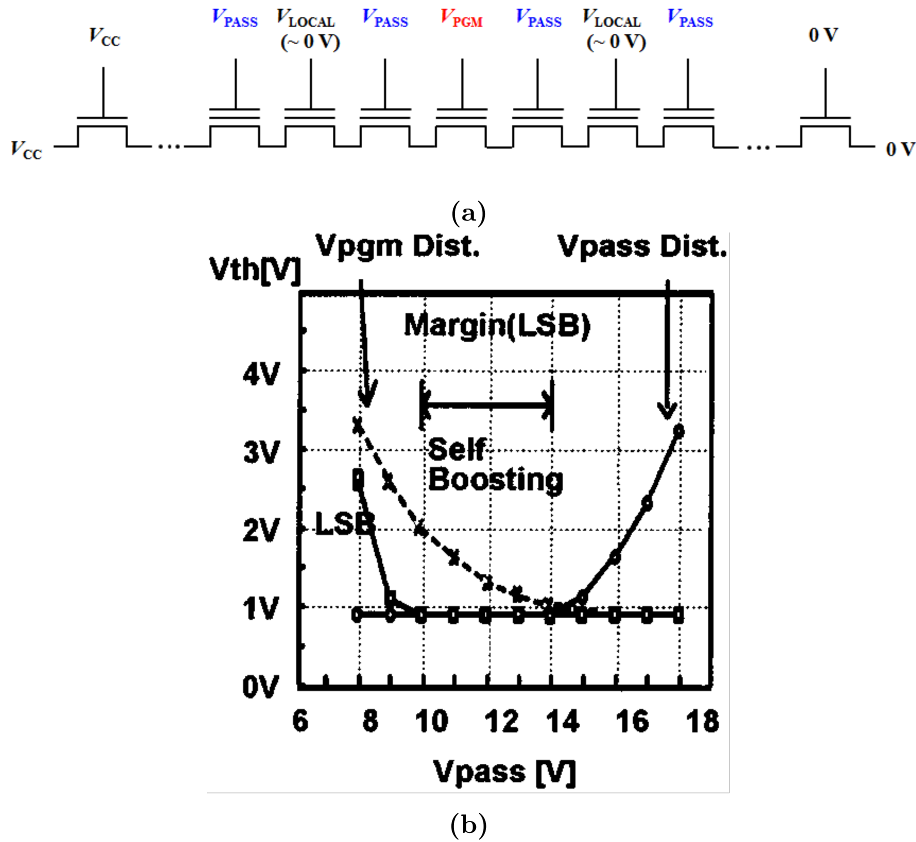


Figure 2.6: (a) Bias scheme representation of the *local self-boosting effect* (from [12]) and (b) threshold voltage shift due program and pass disturb for a 2D NAND Flash memory (from [11]).

the channel does not lie directly upon the substrate but it is just connected with silicon regions at the boundaries. So, if the path is blocked towards the BL and SL, the electrons cannot come from other components of the structure. As a consequence, in the vertical structure, the disturb errors related to the program phase occur much less. But, even if it is easier decoupling the channel from BL and SL, the voltage difference with respect the V_{pgm} remains quite high and program disturbs can still occur. For this reason, some improvements at the bias scheme were needed.

2.3.1 Local Self-Boosting effect

During the write operation, all the *WLs* but one are biased to V_{pass}, the remaining one is at the voltage V_{pgm}. By considering that, we can say that V_{pass} is the dominant parameter in determining the potential of the *floating channel*. Regarding this fact, the higher V_{pass} the better the boosting effect, however, as we have said, a high V_{pass} results in pass disturbs of unselected cells in the selected string.

On the other hand, due to the fact that the boosted channel is inversely proportional

to the number of WLs, in 3D memories with a large N, the drop voltage, between the channel of inhibited string and the FG belonging to the selected WL, still remains too high to completely avoid the tunneling through the oxide. For this reason a further solution is needed. The solution is found in the *Local Self-Boosting* effect that localizes the *self-boosted* channel underneath the selected *WL* by applying a $V_{local} = 0V$ to two *WLs* around the selected one, as reported in Fig. 2.6a. In this way the word lines at V_{local} are turned off as the gates begin to rise in voltage and the localized channel is decoupled from the rest of the string. Consequently the boosted channel potential can be expressed as:

$$\Delta V_{CH} = \frac{C_{WL-CH}}{NC_{tot}}\left(N-1\right)V_{pass} + \frac{C_{WL-CH}}{MC_{tot}}V_{pgm} \tag{2.2}$$

Where in this case M is the number of isolated *WLs* by V_{local}, that it is $M = 3$, considering Fig. 2.6a.

The result is that LSB is a good solution to reduce program disturbs in multilevel NAND Flash memories. On the same time, thanks to this achievement, it is possible to lower V_{pass} and in turn decrease the pass disturbs as well. In the end we can say that the LSB provides a much wider V_{pass} *window* as depicted in Fig. 2.6b. Anyway, also in the 3D configuration there are further program disturbs always related to the bias scheme, as we will show in the next section.

2.4 Down-Coupling Phenomenon

Few years ago it was discovered that 3D NAND Flash memories suffer of another program disturb that was called *Down-Coupling Phenomenon (DCP)*. In order to understand how this disturb is generated, we have to further specify some characteristics of the bias scheme used for the program operation. By neglecting the presence of the dummy lines for simplicity, Fig. 2.7 shows the bias timeline for the main contacts of a realistic program operation, taking into account also the presence of the voltages applied to inhibited strings. We can see that the program operation is composed by a verify phase and a program phase that are separated by a certain interval of time in which all voltages are kept to ground. We have to point out that the verify phase is very similar to the read operation in which the SL is grounded while the BL is high, both for selected and unselected strings. The DSL and SSL are high in order to allow the flow of carriers through the string. In the meanwhile the unselected WLs are brought to a *pass voltage* higher than the maximum V_T, and in the end the selected WL has a voltage V_{read} that can take different values in order to discern the V_T of the cell. Once the verify phase ends,

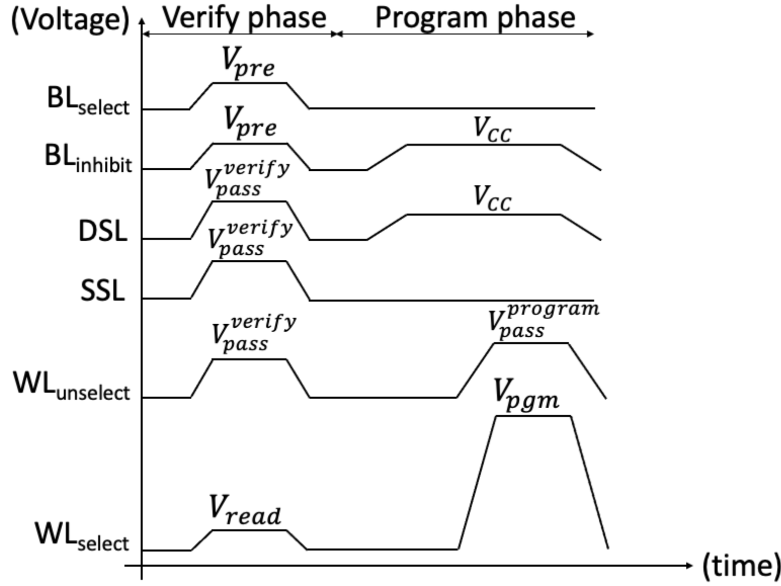


Figure 2.7: Array bias scheme for the program operation.

a small time interval pass before the program phase begins in order that the system could return to the stationary condition. As we can see in Fig. 2.7, in the last part of the verify phase there is a simultaneous drop of the various contacts voltages. In particular, if we consider V_{pass}^{verify} and V_{read}, it happens that this voltages move from their high value to 0. At a certain point of the falling edge, the CGs voltages becomes lower than V_T of the various transistors. When this happens, the transistors switch off and the BL and the SL are disconnected from the channel that, in turn, enter into a floating state. This condition establishes a capacitive coupling between the channel and the control gates such that, as V_{pass}^{verify} and V_{read} drop below V_T, the channel potential falls down alongside them and reaches a negative value. Fig. 2.8 shows the time evolution of the simulation results for the DCP and we can see that for the 3D NAND case the DCP is occurred, while in the 2D NAND, that has the body contact underneath the channel, the DCP does not take place.

For what concerns the relation between the amount of the channel potential drop and the threshold voltage of the cells, we have to specify that the potential drop of of a certain region of the channel is determined by the threshold voltage of the neighbors cells. As reported in [13], the amount of *DCP voltage* can be modeled as follow:

$$\Delta V_{down-coupling} \approx V_{T,neighbors} \tag{2.3}$$

and in Fig. 2.9 we can see the effects on the channel potential in the case of a memory with 16 *WLs*. From the picture we can notice that the higher is the program level of the outer cells the larger is the amount of voltage drop due to the DCP.

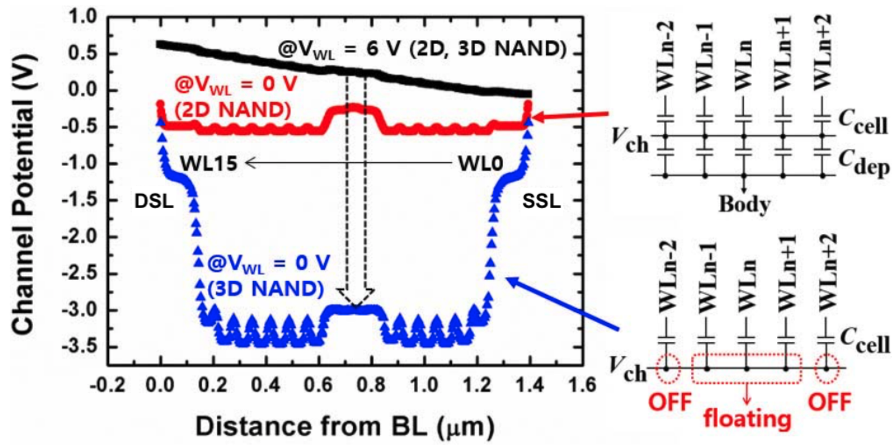


Figure 2.8: Simulation results of DCP occurring during a verify operation (from [13]).

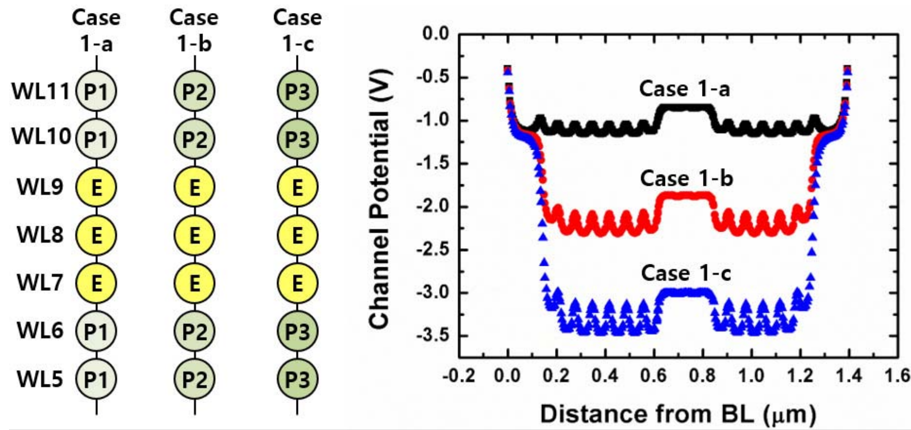


Figure 2.9: Simulation results of DCP for different V_T on neighbors cell (from [13]).

As we will see in the remaining part of this work, this phenomenon can be a bit of different depending on the process used for the production of the memories. Indeed, what affects the most this effect is the doping of the substrate. As we said in the previous chapter, the substrate doping is different between a BICS and a TCAT technology. For the former structure, the substrate is doped $\mathbf{n}^+$ and this means that, when the channel becomes floating, the substrate cannot generate a sufficient hole current to restore the equilibrium, unless the GIDL takes place. For the TCAT structure, instead, there is a p-doped region that can supply the channel with holes. We will see in the next chapter if this supply of holes is sufficient to allow the system to bring back the channel potential to the stationary condition. In fact, the importance of the analysis of this phenomenon lies in understanding how much the DCP could contribute to the onset of program disturbs. This phenomenon can become very troublesome because, once the verify operations is finished, after a brief instant the program operation begins. Usually, the time interval that split the two phase is of the order of μs. In this interval all the contacts are kept to ground

so that the system could go back to the equilibrium state. Even though, the recover of the equilibrium could take a rather long time and what happens next is that, during the program operation, the boosting effect starts with a negative bias of the channel and reaches a lower final state voltage, giving rise to a higher probability of having program disturbs. We are particularly interested in analyzing what happens during the return to the equilibrium state, how much time it takes in order to have a safe program operation and which physical effects are dominant in this transient. By means of the technology computer-aided design (TCAD), we performed some simulations in order to study in more detail the phenomenon.

2.5 Conclusions

In this chapter we have discussed of the GIDL that in BiCS memories is exploited for the correct functioning of the erase operation. Even though, we have seen that this effect is not associated just to this particular operation but it can be established all the times that the channel enter in a floating condition with a sufficient bending of the bands. For this reason the GIDL becomes a relevant effect to exploit as possible solution in other disturbs. One of these disturbs is the program disturb that consists in the undesired injection of electrons in the trapping layer in inhibited strings. The main solution to this problem is the self boosting effect, i.e. the boosting of the channel electrostatic potential of unselected strings thanks to the capacitive coupling established between the floating channel and the CGs. But we have seen also that this solution can be degraded by the onset of the down-coupling phenomenon, an effect that lowers the channel potential during the verify phase right before the beginning of the programming of the memory. This discussion has provided the necessary background to face the analysis that will be presented in the following chapters.

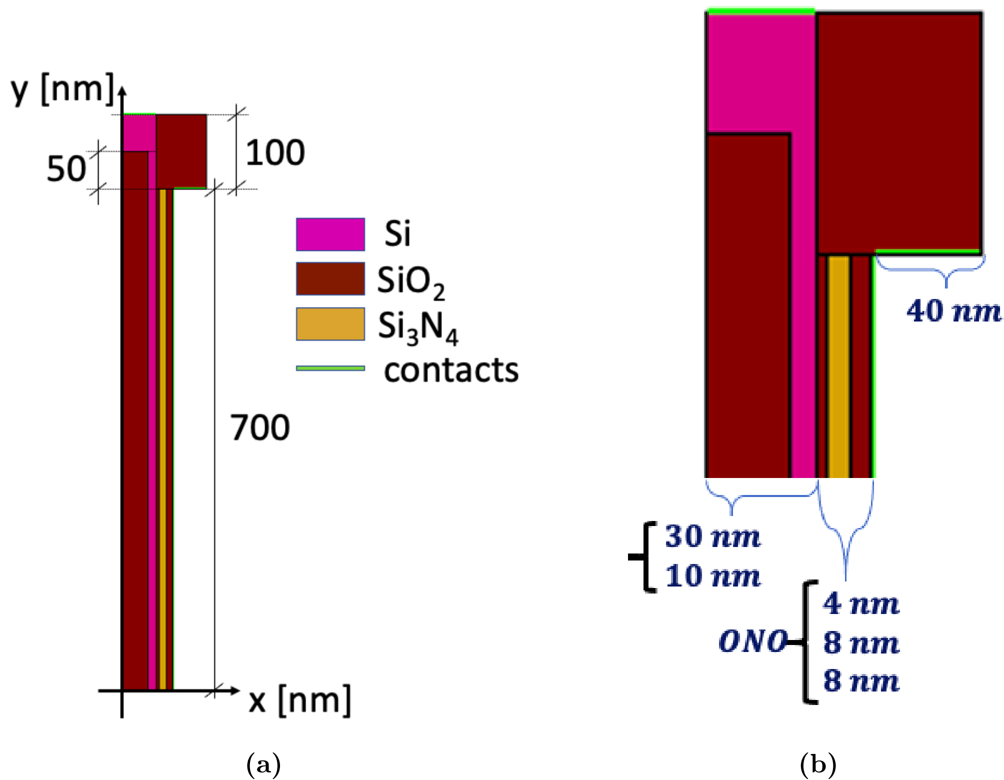
Chapter 3

TCAD simulation about the DCP

In this chapter the analysis of the down-coupling phenomenon will be carried out. AT first, the simulation environment will be presented, then the structure implemented in the simulation, together with the physical models considered to reproduce the phenomenon we are interested in, will be described. In particular, the structure considered in this chapter is a simplified BiCS structure that we have implemented in order that we could focus our attention on the basic physic events that take place during the down-coupling phenomenon. A discussion about the obtained results will be presented and at the end a brief introduction of a preliminary compact model of the phenomenon will be submitted.

3.1 Simulation environment

In order to perform the analysis of the DCP, we have used a *TCAD software* called *Sentaurus* produced by *Synopsys*. The simulation environment is composed of several tools that accomplish different tasks. Firstly, the *Sentaurus Structure Editor* is needed to define the geometry and the materials of the 3D memory that is considered. Next, *Sentaurus Device* defines the numerical equations allowing to perform the simulation of the phenomenon. Alongside with the equations definition, the software is able to account for a vast amount of physical processes but we decided to includ just the *Drift-Diffusion* transport, the *B2BT* (Band-to-band tunneling) and the *SRH* (Shockley-Read-Hall) generation-recombination processes. The inclusion of more advanced processes does not bring a significant improvement of the solution but it just makes the time taken by the simulation longer. The processes not considered in our analysis are those of the second order, so we expect that their contributions are small enough to consider our solution very close to the realistic one.



(a) (b)

Figure 3.1: a) half geometry of the vertical-channel memory with the longitudinal lengths of the structure, b) radial lengths of the memory.

3.2 Simulation definition

3.2.1 Geometry of a single cell string

For the sake of simplicity, a *GAA* structure was defined so that the cylindrical symmetry could be exploited. This fact is very useful because it gives the opportunity of dealing with a 2D cross section of the string structure, instead of the three-dimensional geometry. At the beginning, a structure resembling that one of *BiCS* memories was defined because it has also a longitudinal symmetry. Indeed, it is possible to create a structure in which BL and SL are interchangeable because they are defined in the same way and linked to the same central common structure. Due to that, it is possible to focus our attention on just half of the structure since the other half behaves in the same way. At the beginning, we created a simplified string with just one cell whose length was very big. We have defined a single WL structure in order to better analyze the physics of the *DCP*. In particular, this structure does not include those fringing fields that otherwise would be present in between adjacent control gates. Once we obtained a satisfactory understandings of the physics we moved to a geometry with several *WLs*, specifically 16 *WLs*. Even though, we still wanted the two geometry to be comparable, so we decided to make the channel of

Table 3.1: Geometry parameters used in the simulation for the *single WL* string.

quantity	symbol	value
Gate length	L_{WL}	1.4 μm
Gate thickness	t_{WL}	40 nm
Channel length	L_{ch}	1.4 μm
tunneling oxide layer thickness	t_{tOx}	4 nm
silicon nitride layer thickness	t_n	8 nm
blocking oxide layer thickness	t_{bOx}	8 nm
Filler radius	R_{fil}	30 nm
Channel radius	R_{ch}	40 nm
Polysilicon channel thickness	t_{ch}	10 nm

the simplified structure as long as the channel of the structure with 16 *WLs*.

Fig. 3.1a shows the upper part of the string that was created with the structure editor of the simulator. The material used in the purple area is silicon, the brown regions are made of silicon oxide. The brown internal region constitutes the oxide filler, while on the other hand, in the upper-east part, there is the oxide layer that isolates the WL from the contact of the BL. Finally, the yellow layer is made of silicon nitride where the carriers are trapped, while the contacts are drawn in green. Table 3.1 shows the main parameters of the geometry where we can notice that the length of the WL is equal to the length of the channel in order to have the best electrostatic control of this region.

For what concerns the ONO layer, by going from left to right, it is composed of a tunneling layer of silicon oxide, a silicon nitride layer and another oxide layer, the blocking oxide layer. The radial structure is depicted in detail in Fig. 3.1b where the radial lengths are also reported.

Fig. 3.2 shows the doping profile of the structure. Considering just the silicon regions, the red area is doped n^+ with Arsenic atoms, where the dopant atoms concentration is $N_{BL} = 5 \times 10^{19}\ cm^{-3}$, and it represents either the *BL* region or *SL* region cause they are equal. For what concerns the blue area, it corresponds to the channel and it is doped again with Arsenic atoms but this time the amount of doping is quite lower, $N_{channel} = 10^{15}\ cm^{-3}$. For what concerns the trapping layer, we have defined that it is filled with a fixed charge. The value of this charge density is variable depending on the level of the cells that we want to replicate. For the sake of simplicity, we have neglected all the possible causes that can modify the density of trapped charge, this means that effects like the SILC or the RTN are not considered in the simulations.

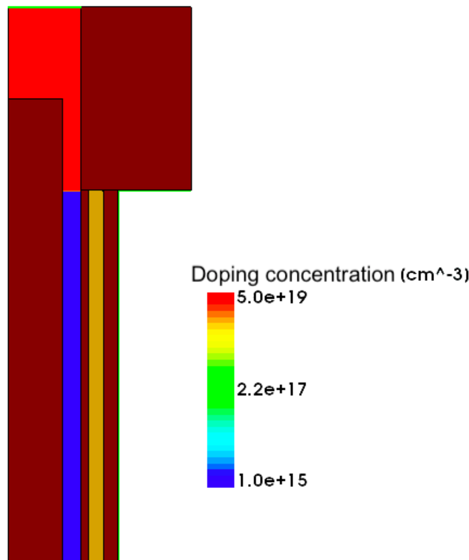


Figure 3.2: Doping profile used in the simulations

Table 3.2: Modified parameters for the *Masetti* model.

	electrons	**holes**	**unit**
μ_{const}	141.7	47.05	$cm^2/(V \cdot s)$
μ_{min1}	5.22	4.49	$cm^2/(V \cdot s)$
μ_{min2}	5.22	0	$cm^2/(V \cdot s)$
μ_1	4.34	2.90	$cm^2/(V \cdot s)$
P_c	0	$9.23 \cdot 10^{16}$	cm^3
C_r	$9.68 \cdot 10^{16}$	$2.23 \cdot 10^{17}$	cm^3
C_s	$3.43 \cdot 10^{20}$	$6.10 \cdot 10^{20}$	cm^3
α	0.680	0.719	–
β	2.0	2.0	–

3.2.2 Physical models included in the simulations

In this subsection, the physical models considered in the simulations are reported. We chose to keep just the essential models in order to describe correctly the effects taking place during the *verify* operation but to keep the computational cost at the minimum as well. Carrier transport is described by the *Drift-Diffusion* equations (*DD* model). In this model, the most important parameter is the mobility. In our model, the acoustic collisions of the carriers were considered not sufficient to well describe the electrons transport inside the string so we added two factors that deteriorate the mobility of the carriers. These factors are the collisions with the impurities due to the doping and the collisions at the sillicon-oxide interface. The former contribution is described by the *Masetti* model [14], that use an empiric equation to explain the contribution of the dopants to the mobility of

Table 3.3: Modified parameters for the *Enhanced Lombardi* model.

	electrons	**holes**	**unit**
B	$4.75 \cdot 10^7$	$9.925 \cdot 10^6$	cm/s
C	$5.80 \cdot 10^2$	$2.947 \cdot 10^3$	$cm^{5/3} \cdot V^{-2/3} \cdot s^{-1}$
N_0	1	1	cm^{-3}
N_2	1	1	cm^{-3}
λ	0.125	0.0317	$-$
δ	$1.455 \cdot 10^{14}$	$0.51365 \cdot 10^{14}$	$cm^2/(V \cdot s)$
A	2	2	$-$
η	$1.455 \cdot 10^{30}$	$0.51365 \cdot 10^{30}$	$V^2/(cm \cdot s)$
F_{ref}	1	1	V/m

the electrons. By considering the table 3.2 we can obtain the aforementioned equation:

$$\mu_{dop} = \mu_{min1} e^{\frac{P_c}{N_{donor}}} + \frac{\mu_{const} - \mu_{min2}}{1 + (N_{donor}/C_r)^{\alpha}} - \frac{\mu_1}{1 + (C_s/N_{donor})^{\beta}} \tag{3.1}$$

Where μ_{const} is the initial value of the mobility that considers just the interactions of the carriers with *phonons* in *bulk* regions.

The second factor consider the roughness of the interface and the possible interaction of the carriers with the oxide atoms. Actually, in our case it is quite important include this degradation process because in some cases, when the longitudinal electric field is not so large, the radial component of the field can become relevant. When this happens, the carriers not only move along the channel direction but, due to the perpendicular field, they steer out of their path and collide with the oxide walls. In order to include this process inside the simulation we added a corrective contribution to the mobility by implementing the *Enhanced Lombardi* model. This model considers the mobility given by two principal components: μ_{ac} that takes into account the scattering coming from the phonons of the surface and μ_{sr} that describes the scattering contributions due to the surface roughness. These values are obtained in a semi-empirical way and they can be expressed by the following set of equation, as reported in [15] and [16]:

$$\begin{cases} \mu_{ac} = \frac{B}{F_{\perp}} + \frac{C\left(\frac{N_{donor}+N_2}{N_0}\right)^{\lambda}}{F_{\perp}^{1/3}} \\ \mu_{sr} = \left(\frac{\left(\frac{F_{\perp}}{F_{ref}}\right)^{A}}{\delta} + \frac{F_{\perp}^3}{\eta} \right)^{-1} \end{cases} \tag{3.2}$$

where $F_{\perp}$ is the perpendicular electric field to the oxide surface, while the other parameters are reported in table 3.3.

Next, the mobility of the bulk is combined with the interface contributions by means of the *Matthiessen's rule*:

$$\mu_{int} = \left(\frac{1}{\mu_{bulk}} + \frac{D}{\mu_{ac}} + \frac{D}{\mu_{sr}} \right)^{-1} \tag{3.3}$$

D is the factor that holds in consideration the distance between the carrier and the interface, $D = e^{\frac{-d}{l_{critic}}}$ with d the physical distance and $l_{critic} = 100nm$, a characteristic length of the silicon. The D value is null when the carriers are far away from the surface and, in that situation, just the bulk behavior is considered. Finally, the mobility value that takes into account of both the interface and the impurities is calculated again with the *Matthiessen's rule*:

$$\mu_{tot} = \left(\frac{1}{\mu_{dop}} + \frac{1}{\mu_{int}} \right)^{-1} \tag{3.4}$$

We have to point out that, for those regions where the charge transport occurs, the material used in the manufacturing process is not silicon but polysilicon. As we have said, this material is composed by many grains whose size and orientation is utterly random. Problems arises at the boundaries of these grains where the disordered and incomplete atomic bonds give rise to a huge amount of defects. These defects acts like traps for the electrons, producing both a decrease of the number of free carriers and non-uniformities in the channel electrostatic, [17]. Due to these factors, the mobility in polysilicon is degraded with respect to the crystalline structure. Despite the modeling of the transport in polysilicon is out of the scope of this work, we have decided to consider this degradation modifying the parameters and adapting them to the case of the polysilicon. We did so by exploiting the results of [18] and [19] where it was found out that the mobility in polysilicon is about one order of magnitude lower than in crystal silicon. The mobility values reported in table 3.2 have already taken into account of this correction.

3.3 Generation and recombination processes

We have included two physical models describing the generation and recombination processes: the B2BT and the *Shockley-Read-Hall (SRH)*. Regarding the B2BT, we decided to implement the *non local* model. In this model the hole and the electron are generated at the same energy in two different spatial points at the two opposite sides of the potential barrier. What the simulator does at each step is searching a tunneling path according to the fact that it must be a straight line oriented towards the opposite direction of the valence band gradient. Furthermore, it is assumed that the path starts

at the valence band, that it is also the point where the gradient is calculated, and stops at the conduction band. The holes are generated at the valence band side, while the electrons at the conduction band side. The expression of the generation rate implemented in the model is quite laborious, but assuming the electric field as constant along the whole tunneling path, the expression can be simplified in the following one:

$$G_{B2BT} = A \left(\frac{F}{F_0} \right)^p e^{-\frac{B}{F}} \tag{3.5}$$

where F is the electric field, $F_0 = 1\ V/cm$, $p = 2.5$, while A and B are two constants whose values are respectively $4 \cdot 10^{14}\ cm^{-3}/s$ and $1.9 \cdot 10^7\ V/cm$.

For what concerns the SRH, it describes the mechanisms of generation and recombination assisted by defects. Since in the simulator the Fermi statistics is implemented, the SRH rate as the following equation:

$$R_{net}^{SRH} = \frac{np - \gamma_n \gamma_p n_i^2}{\tau_p \left(n + \gamma_n n_1\right) + \tau_n \left(p + \gamma_p p_1\right)} \tag{3.6}$$

With $n_1 = n_i e^{\left(\frac{E_{trap}}{KT}\right)}$ and $p_1 = n_i e^{\left(\frac{-E_{trap}}{KT}\right)}$, while E_{trap} is the difference in energy between the defects levels and the intrinsic level, that we decided to set equal to 0 eV for simplicity. γ_n and γ_p are the correction parameters in case the Fermi statistic is used, n_i is the intrinsic density of carriers for the intrinsic silicon, while τ_n and τ_p are the average life times of the electrons and holes, respectively. These two values are not fixed for the whole structure but they depend on the doping value of the silicon region in which the carriers exist. The relation of the average life time is the following:

$$\tau_{n,p} = \frac{\tau_{max}^{n,p}}{1 + \frac{N_{D,A}}{N_{ref}}} \tag{3.7}$$

where $N_{ref} = 10^{16}\ cm^{-3}$. About the $\tau_{max}^{n,p}$, we have to point out that, in order to partially take into account the presence of defects at the grain boundaries, we have reduced the value of a factor 10^3 both for electrons and holes, such that $\tau_{max}^n = 10^{-8}\ s$ and $\tau_{max}^p = 3 \cdot 10^{-9}\ s$.

3.4 Threshold extrapolation

Our use of the TCAD software is oriented to study the down-coupling phenomenon and its consequences on the program operation. In order to be able to see the DCP, the cell must not be in the *ERASED* state but it must have a positive V_T. In fact, just in this case, the cut off of the channel during the *verify* phase occurs. So, the first thing

to do was setting the V_T by implanting the right density n_t of trapped charge inside the *trapping* layer. It is possible to find out the wanted density value by means of the relation:

$$C_{NG} = -\frac{Q}{\Delta V_T} \tag{3.8}$$

First of all, we have to find ΔV_T. This is the threshold voltage shift due to the presence of the charge in the trapping layer and it can be calculated by solving the Poisson equation in cylindrical coordinates. If we take the null reference voltage at the gate, we obtain:

$$\begin{aligned} \Delta V_T &= -\left(\frac{qn_t}{2}\right) \cdot \\ & \left[\left(r_n^2 - r_{tOx}^2\right)\left(\frac{1}{2\varepsilon_0\varepsilon_n} + \frac{1}{\varepsilon_0\varepsilon_{ox}}\right)\ln\left(\frac{r_{ONO}}{r_n}\right) - \frac{r_{tunOx}^2}{\varepsilon_0\varepsilon_n}\ln\left(\frac{r_n}{r_{tunOx}}\right)\right] \end{aligned} \tag{3.9}$$

where

$$r_{tOx} = R_{ch} + t_{tOx} \tag{3.10}$$

$$r_n = r_{tOx} + t_n \tag{3.11}$$

$$r_{ONO} = r_n + t_{bOx} \tag{3.12}$$

Instead, the charge is found out just by calculating the volume integral of the charge density in the silicon nitride layer:

$$Q = qn_t\pi L_{WL}(r_n^2 - r_{tOx}^2) \tag{3.13}$$

If we substitute Eq. 3.9 and Eq. 3.13 in Eq. 3.8, the charge density n_t is simplified and we can calculate the capacitance. At this point, we can find the target charge by using the wanted ΔV_T. This value is the voltage shift between the threshold voltage for $n_t = 0\ cm^{-3}$ and the desired threshold voltage of the defined programmed state. Once we have found the total charge, we can calculate the density by inverting Eq. 3.13.

In our model we supposed of dealing with an MLC memory, so the cell can store two bits. This means that the cells has to be programmed to one out of four defined V_T levels. These levels were defined for convenience in the following way: $V_T^E = -0.4\ V$ for the *ERASED (E)* state, while $V_T^{P1} = 1.6\ V$, $V_T^{P2} = 2.6\ V$ and $V_T^{P3} = 3.6\ V$ for the *PROGRAMMED P1*, *P2* and *P3* states, respectively. The choice of these values assumes as reference voltage the threshold voltage $V_T\ |_{n_t=0} \approx 0.6\ V$ that we have extracted empirically from the transcharacteristic of the structure as the larger voltage before the transistor exits from the sub-threshold regime. If we consider the V_T values that we have just listed for denoting the levels of the cell, than the threshold voltage shift between

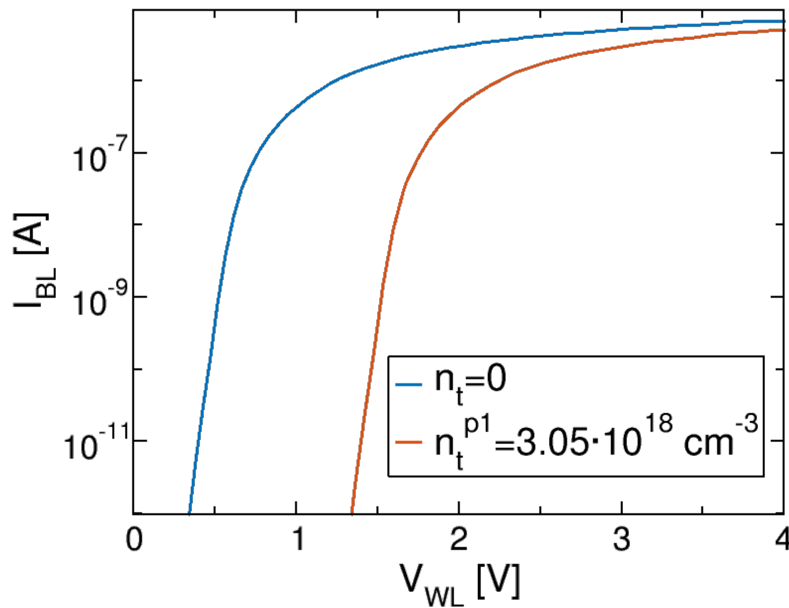


Figure 3.3: *BL* current comparison between a memory without trapped charge in the cell and a *P1* programmed memory.

neighbor levels is defined as $\Delta V_T = 1\ V$. At this point, we can define the density of trapped charge that correspond to the different levels. We have obtained: $n_t^E = 3.05 \cdot 10^{18}\ cm^{-3}$ with a positive sign of the charge, $n_t^{P1} = 3.05 \cdot 10^{18}\ cm^{-3}$, $n_t^{P2} = 6.1 \cdot 10^{18}\ cm^{-3}$ and $n_t^{P3} = 9.15 \cdot 10^{18}\ cm^{-3}$ with a negative charge for the programmed levels.

Fig. 3.3 shows a comparison of the transcharacteristic for two structures having $n_t = 0$ and n_t^{P1} as density of charge implanted inside the *trapping* layer. We can notice that actually the transcharacteristic of the string shifts rigidly due to the trapped charge and the value of the shift is $\Delta V_T \approx 1\ V$

3.5 Transient simulation definition

As we said in the previous chapter, when we talked about the DCP, the amount of potential drop between the channel potential and the potential at BL/SL depends on the threshold voltage of the *WL*. The recovery time of the equilibrium condition must be shorter than the interval time passing before the *program* phase begins, otherwise, the *boosting* effect will start from a negative condition, bringing to a worsening of the program disturbs. We wanted to study the interval time between the end of the *verify* phase and the beginning of the *program* phase in order to understand, first, if the system can restore the steady-state condition in such time interval, second, which physical factor is determinant in the return to the equilibrium.

The dynamics of the DCP was studied by performing a transient simulation that takes

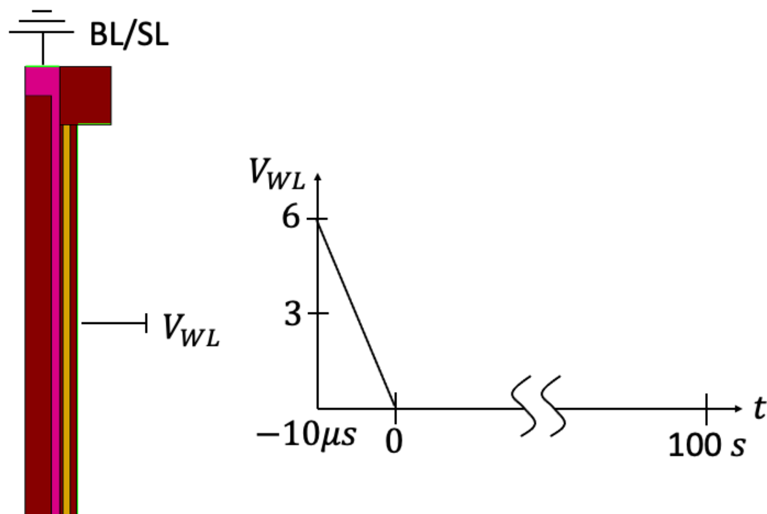
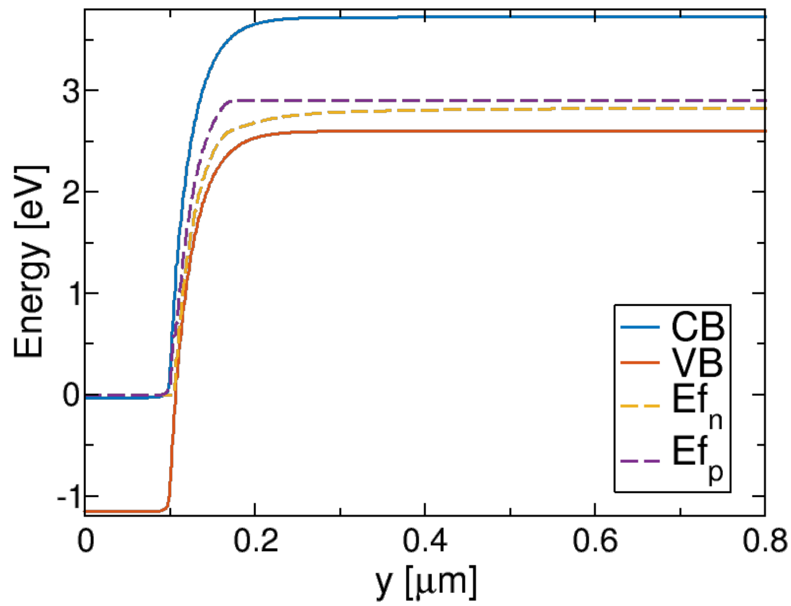
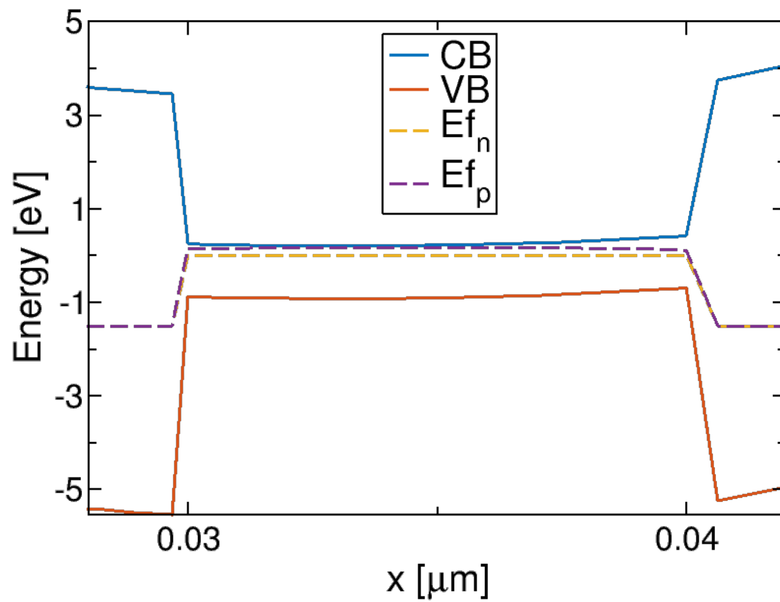


Figure 3.4: Bias scheme of the simulation.

into account the B2BT and the SRH processes described in the previous sections. In our scheme bias, as reported in Fig. 3.4, the WL voltage is moved from the value of $V_{pass} = 6\ V$ to $0\ V$ within a time interval of $10\ \mu s$. After that, the WL is kept to ground for $100\ s$. Both the SL and the BL are grounded in order to simplify the analysis of the results because in this way we are sure that there are no drift current contributions due to the voltage applied at BL and SL. The cell was defined to be in the P3 level, that means that a fixed and uniform density $n_t^{P3} = 9.15 \cdot 10^{18}\ cm^{-3}$ of trapped electrons was set for the whole simulation.



(a)



(b)

Figure 3.5: Band diagram at $t = 0$ s in (a) longitudinal direction at 0.25 nm from the silicon-oxide interface and (b) radial direction at the point of maximum Electric field located at the junction between the BL and the channel

3.6 DCP simulation results

3.6.1 Analysis of the CB energy time evolution

As first thing we have checked if the DCP was correctly simulated. Fig. 3.5a reports the band diagram as a function of the y coordinate. This coordinate represents the

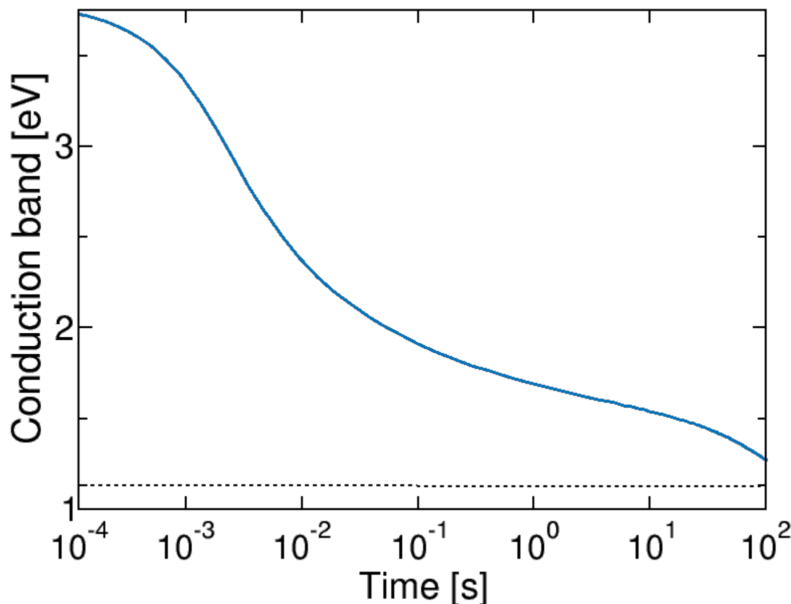


Figure 3.6: Time evolution of the conduction band energy at the middle of the string, since the end of the WL falling edge. The dashed black line represent the equilibrium value.

longitudinal coordinate, this means that it corresponds to the direction parallel to the channel, see Fig. 3.1a. The cut is taken at 0.25 nm from the silicon-oxide interface at the end of the falling edge of V_{WL}, i.e. at the end of the verify phase that we have defined as $t = 0$. For what concerns the silicon-oxide interface, we are referring to the surface between the channel and the tunnel oxide of the ONO stack. From this point on, we will always refer to this interface unless it will be specified differently. From the picture, we can see that actually the energy of the channel is rather high, pointing out that the DCP is occurred.

Instead, Fig. 3.5b depicts a radial section of the band diagram. The cut is taken at the junction between the SL and the channel, again at $t = 0\ s$. In this case the x coordinate represent the direction perpendicular to the channel, see again Fig. 3.1a. We want to point out that this definition of the cartesian axes will be used throughout this work. We can notice from the picture that the band bending in the radial direction is small enough such that the B2BT can be supposed to null along the x-direction. On the other hand, along the vertical direction, at the lateral edges of the channel, a strong vertical field is present such that the bands tilt a lot and there are many states available for the tunneling. Thanks to these observations, we can conclude that in our simulations the B2BT is present and the flow of the injected carriers is mainly longitudinal.

What happens after the falling edge is depicted in Fig. 3.6, where we can see the time evolution of the conduction band energy. It reaches a positive value of $E_{CB}^{ch} \approx 3.72\ eV$. If we consider that at the beginning, right before the channel enters in a floating state,

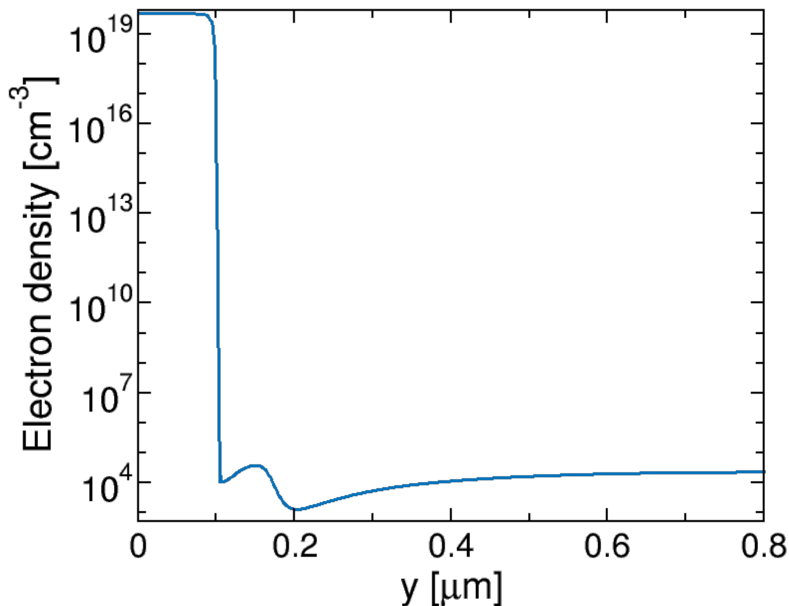


Figure 3.7: Longitudinal section of the electron density. The cut is taken at 0.25 nm from the silicon-oxide interface and at $t = 0$.

the CB energy is $E_{CB}^{ch} \approx 0.16\ eV$, we can trace back the threshold voltage shift that is actually $\Delta V_T \approx 3.56\ V$, very close to the value that we have supposed for the given trapped charge. We can notice from the picture that the return to the equilibrium does not follow a unique trend but it undergoes some changes during the evolution of the transitory. Always from Fig. 3.6, we notice that in 100 s the system is not went back to the stationary value, represented by the dashed line. In order to explain this behavior we have to point what happens to the carrier concentration inside the channel.

3.6.2 Time evolution of the carriers concentration

For the way we have defined the simulation, at the initial condition the string is characterised by a strong inversion of the channel. In fact, the electrons are driven inside the channel by the rather high voltage value $V_{WLs} = 6\ V$. As soon as this voltage decreases, the electrons starts to feel the field produced by the stored charge and they are pushed outside the channel. As long as V_{WL} remains above the threshold voltage, the transistor is turned on and the electrons can quickly exit from the channel. In this configuration, when the transistor switches off, the electron density is already rather small. The rest of the electrons still inside of the channel is driven from the middle of the string towards the edges by a diffusion process, since at the centre the electron concentration is higher. At the edges, the electrons feel the electric field established at the junction and, by drift, they can exit from the channel. We are not interested in quantifying the exiting current during the transition of V_{WL}, even though we are interested in the remaining

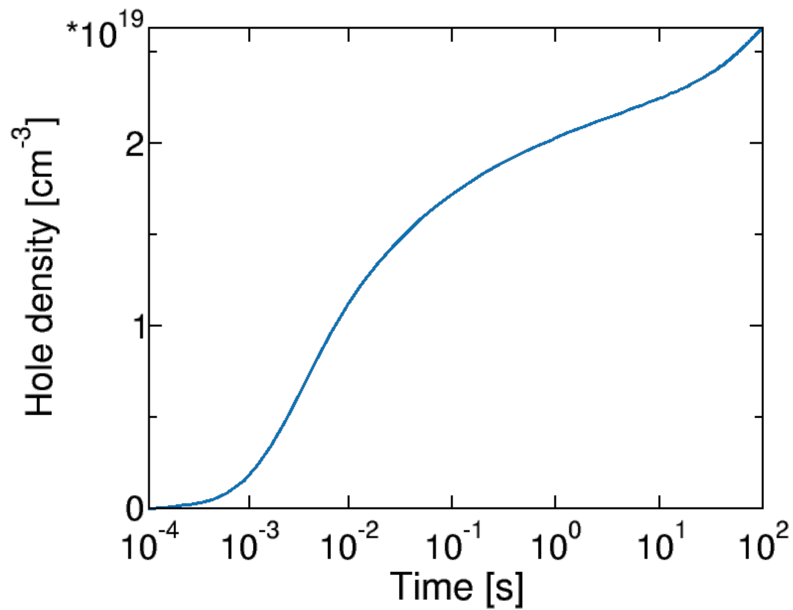


Figure 3.8: Time evolution of the hole density at the middle of the string since the end of the WL falling edge.

electron density inside the channel when $V_{WL} = 0\ V$. Fig. 3.7 shows two longitudinal section of the electron density at the end of the falling edge of V_{WL}. The picture confirms that the density is extremely small when $V_{WL} = 0V$ and this make us suppose that the electrostatic of the system is no longer subjected to changes of the electron density but it is driven just by the holes that can enter into the channel.

Fig. 3.8 depicts the time evolution of the hole density at the centre of the channel with a logarithmic scale just for the time axes. We can see that the hole density trend follows quite well that one of the conduction band energy. In the first ms, the growth of the hole density is not visible due to the linear scale of the y axes. Despite this inconvenient, it is actually this hole concentration that, being not big enough, does not produce significant variations on the electrostatic potential of the channel. Indeed, we can compare Fig. 3.8 with Fig. 3.6 and notice that both curves does not present significant changes of their respective values in the first part of the transition. After few ms the hole density reaches a value comparable with that one of the trapped charge, this means that the density becomes relevant for the electrostatic of the system. In fact we can notice that when this happens, the energy of the conduction band starts decreasing. Now, we can understand that in order to justify the curve of the conduction band energy, we have to motivate the trend of the hole density inside the channel.

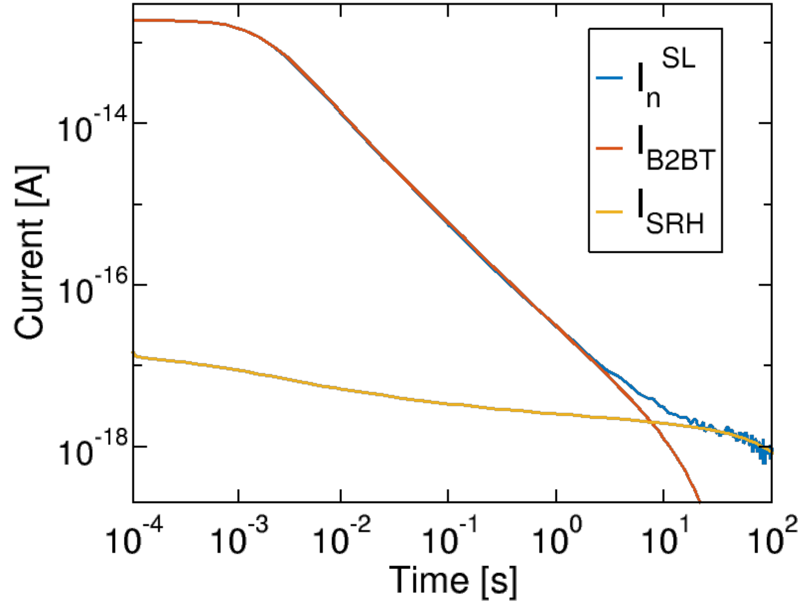


Figure 3.9: Comparison of the electron flow at the SL, blue curve, with respect the current due to the rate generation of the B2BT and SRH, red and yellow curves, respectively.

3.6.3 Time evolution of the current

In order to understand how the hole concentration changes with the passing time, we have analyzed the time evolution of the current that flows through the string. Fig. 3.9 reports the flow of electrons at the source contact and shows that it corresponds with the GIDL current. The current component generated by SRH, instead, is smaller than the GIDL current except for the end of the transitory when it becomes dominant. For the sake of clarity, the GIDL and SRH currents were calculated by integrating for each time instant the density of holes generated per unit time inside the half of the whole string, as reported in the following equation:

$$I_{GIDL} = \frac{q}{2} \int\int\int G_{B2BT} dV \tag{3.14}$$

We did the same thing for the SRH generated current, as written in the following equation:

$$I_{SRH} = \frac{q}{2} \int\int\int R_{SRH} dV \tag{3.15}$$

In the integrals we took just one half of the total value since, due to the symmetry, just half of I_{GIDL} and I_{SRH} heads towards the SL, the other half flows through the BL.

As we have said, the holes are generated at the edges of the channel thanks to the tunneling. In that position, they feel the longitudinal electric field of the junction that pushes them towards the centre of the string. The holes, even far away from the junction keep moving by drift towards the middle of the channel where the attraction due to the

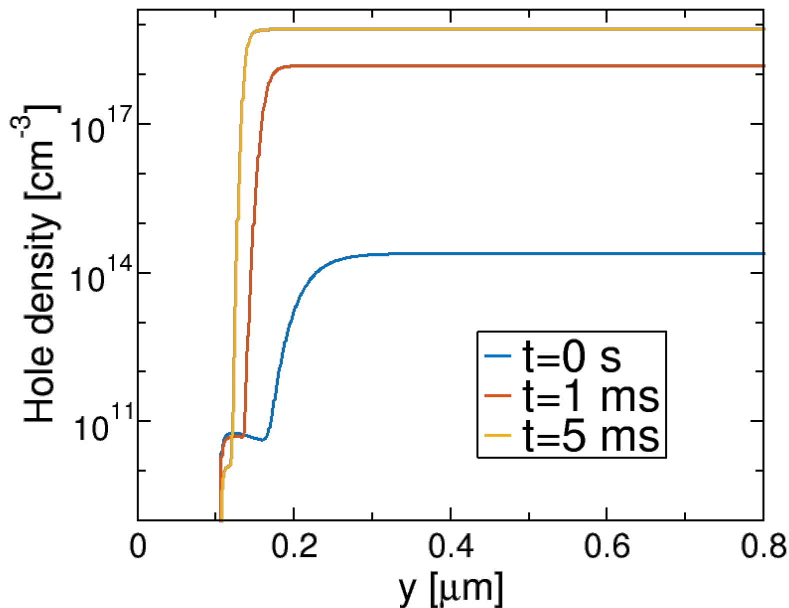


Figure 3.10: Longitudinal section of the hole density taken at 0.25 nm from the silicon-oxide interface for three different instants of the transitory.

negative stored charge is stronger. For this reason, the hole concentration increases faster at the centre of the string than at the edges. Fig. 3.10 shows the longitudinal section of the hole concentration at different instants of the transitory and supports what we have just said. Indeed, we can see that at first the density becomes bigger at the centre and then it tries to widening itself towards the edges. This last behaviour of the hole density is fundamental to understand the trend of the red curve in Fig. 3.9, that represents the GIDL current.

We know from Eq. 3.5 that I_{GIDL} just depends on the electric field at the junction. This, in turn, is affected by the presence of free charge inside the channel. Indeed, the field generated by the accumulation of holes goes against the field of the junction. So, the net longitudinal field at the junction becomes smaller with the increasing hole density. Fig. 3.11 shows the time evolution of the longitudinal component of the electric field calculated at the edge of the channel. We can confirm that the trend of the field follows that one of the hole density. Both these parameters are related to the electrostatic potential of the channel and, like in a *closed loop*, they affect each other. To sum up, once the hole density gets big enough, it contributes to decrease both the channel energy and the electric field at the junction. In turn, as the electric field decreases, the I_{GIDL} decreases as well. In fact, if we focus our attention on the red curve in Fig. 3.9, we can notice that the current start decreasing about 1 ms from the start. From that moment on, the I_{GIDL} keeps getting smaller until the band bending is so small that there are very few states available for tunneling. Indeed, in the last part of the transitory, the current

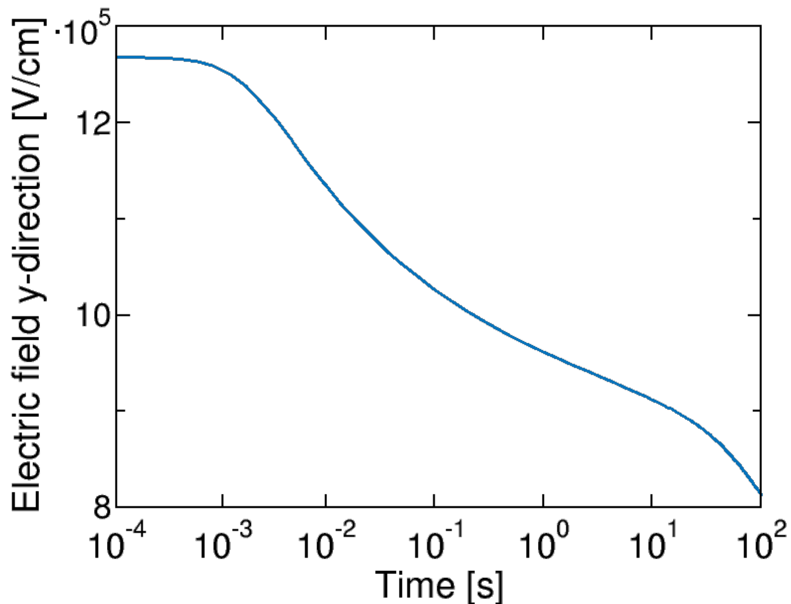


Figure 3.11: Time evolution of the longitudinal component of the electric field that is present at the junction between the SL and the channel since the end of the WL falling edge.

due to the tunneling becomes smaller than the current generated by SRH.

For what concerns I_{SRH}, Fig. 3.9 shows that its value decreases much more slowly than I_{GIDL}. In order to understand this behaviour, we can analyse how the SRH rate changes in time and space during the transitory. Fig. 3.12 shows longitudinal section of the SRH rate generation taken at three different instants during the transitory. We notice that the generation inside the channel decreases a lot with the passing time. This is due to the fact that inside the string the hole density is already quite big at few ms after the end of the falling edge of V_{WL}. Considering Eq. 3.6, we know that, as the product np approaches to the value of n_i^2, the denominator tends to zero and the same does the recombination rate. This is what happens in the middle of the string, as the hole density increases, getting closer to the stationary value, the generation rates drops down. Even though, we have seen in Fig. 3.10 that the hole density keeps being rather small at the edge of the channel and it is actually in this zone where the generation rate is higher. Indeed, I_{SRH} is given by the spatial integral of the generation rate and, as we can see in Fig. 3.12, the major contribute comes from the generation at the lateral depletion regions. It is true that the depletion region shrinks with the widening of the hole concentration, but it is also true that the peak of the generation rate, located at the junction with the SL, remains almost constant over time. For this reason I_{SRH} does not decrease too much and there is a moment at the end of the transitory in which I_{SRH} becomes dominant for the return to the equilibrium. The moment, when the SRH current becomes higher than the B2BT current, is also recognizable in Fig. 3.6, where we can notice a change of the

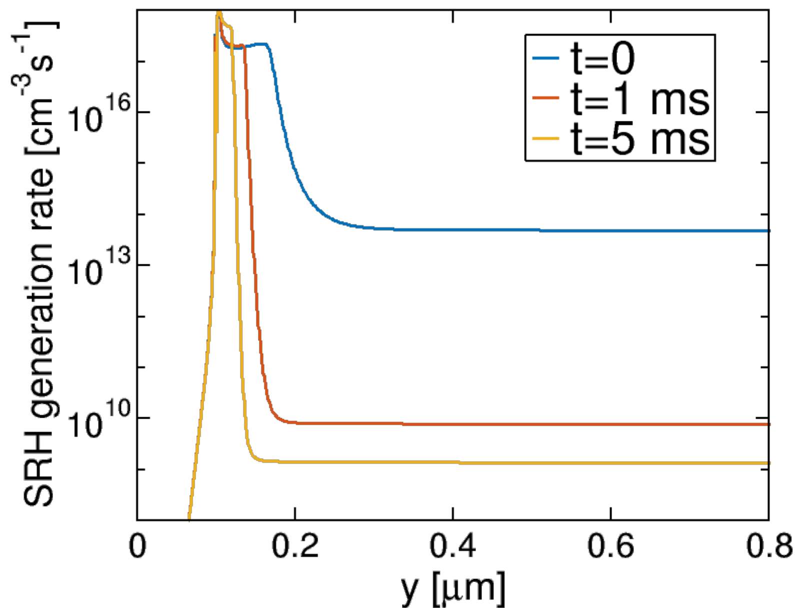


Figure 3.12: Longitudinal section of the SRH rate generation taken at 0.25 nm from the silicon-oxide interface for three different instants of the transitory.

energy trend at long times.

3.7 Dependence on parameters

In order to further understand how much the B2BT and the SRH are relevant in the return to the equilibrium of the system. We decided to perform a simulation where the B2BT was not accounted in the models. By switching off the tunneling, it is the SRH generation to determine the return to the equilibrium. For what we have seen so far, this kind of generation is rather smaller than the band-to-band tunneling, so we expect that the return to the steady-state condition takes a bigger time interval. Fig. 3.13 shows the time evolution of the conduction band energy at the centre of the string in case the B2BT is turned off. We can notice that the conduction energy remains high for a longer period due to the fact that the filling of the channel with holes is a lot slower. That means that the hole concentration takes almost $t = 1$ s to become relevant for the electrostatic of the system. We have to notice that the switching off of the tunneling in the simulation affects the generation along the channel, as we can see in Fig. 3.14a.

The picture shows a comparison of the SRH generation rate in the case of active B2BT and not active B2BT. Without tunneling the generation inside the channel is higher with respect to the case that includes the B2BT. The reason of this higher generation can be found again in Eq. 3.6 where we have to consider that the hole density in the present case is lower than the previous one. This because the B2BT starts injecting holes inside

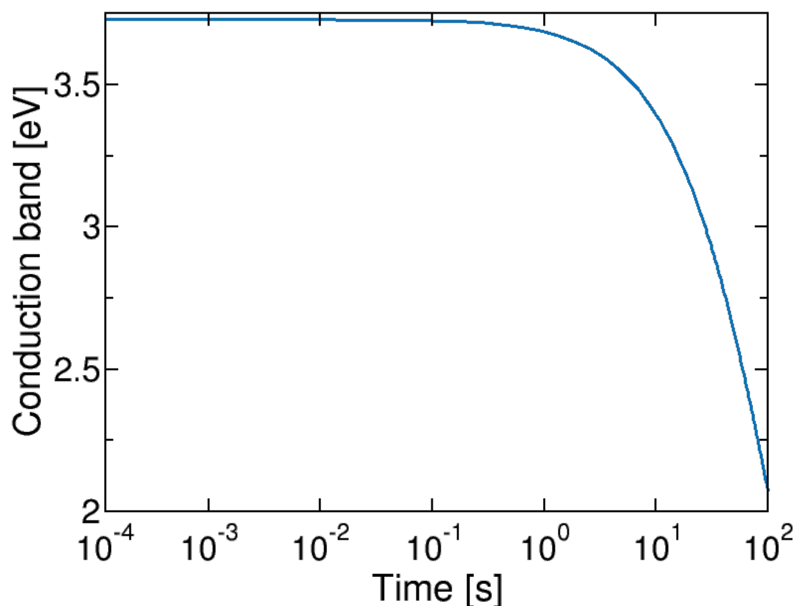
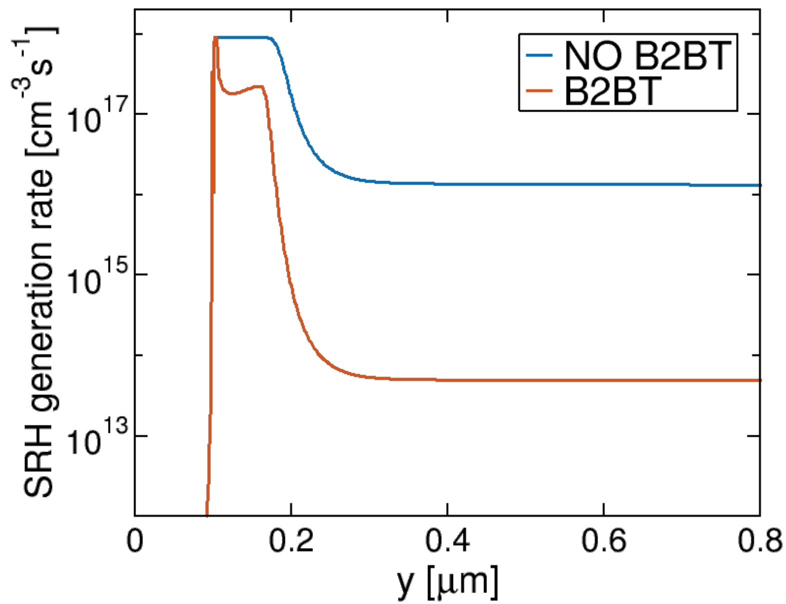


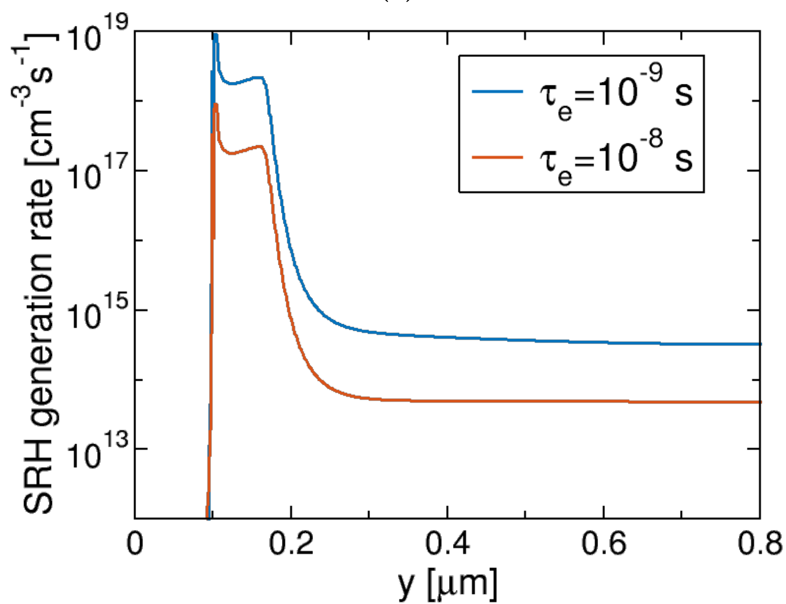
Figure 3.13: Time evolution of the conduction band energy at the middle of the string, since the end of the WL falling edge.

the channel few instants before the ending of the falling edge of V_{WL}. Another important difference lies in the generation at the junction, in particular in the channel side of the junction. The SRH generation, in case of B2BT activated, is lower actually due to the presence of the charge generated by the tunneling. This charge is subjected to the field of the junction and move towards the middle of the string but this displacement does not happen instantly. So it means again that the population of positive carriers, in case of B2BT activated, is higher in the channel side of the junction and in turn the generation is lower.

In the end, we tried also to analyze how much the SRH is determinant in the transient of the system. In order to do that, we simulated the phenomenon keeping activated both the B2BT and the SRH but we changed the life-time parameters of the carriers. Specifically, in the present simulation we have used $\tau^n_{max} = 10^{-9}$ s and $\tau^p_{max} = 3 \cdot 10^{-10}$ s. We have reduced the life-time of the carriers of one order of magnitude with respect to the settings used so far. As we haves seen in Eq. 3.6, these parameters are inversely proportional to the SRH rate, so we expect that it is higher than the rate we have seen in the previous sections.



(a)



(b)

Figure 3.14: (a) Comparison of the SRH rate generation between the case of not implemented B2BT (blue curve) and implemented B2BT (dashed red curve). (b) Comparison of the SRH rate generation between the case where $\tau^n_{max} = 10^{-9}\ s$ (blue curve) and the case where $\tau^n_{max} = 10^{-8}\ s$ (red curve). The longitudinal cuts are taken at $0.25\ nm$ from the silicon-oxide interface at the end of the verify phase.

Fig. 3.14b shows the comparison of the SRH generation rate as function of the longitudinal coordinate, where the blue curve represents the result of the simulation with $\tau^n_{max} = 10^{-9}\ s$, while the red curve represents the simulation with $\tau^n_{max} = 10^{-8}\ s$. The picture confirms what we have supposed, the generation increases exactly one order of magnitude. Despite of this generation enhancement, we have to keep in mind that it

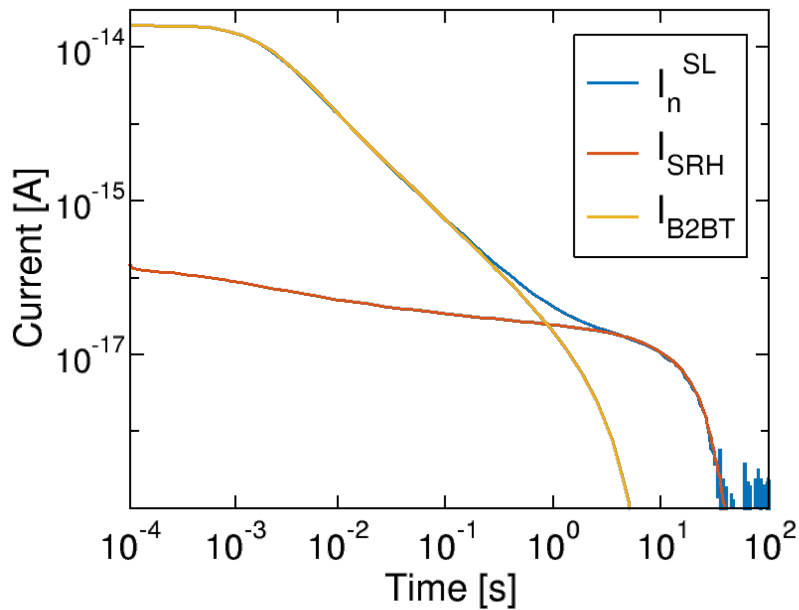


Figure 3.15: Comparison of the electron flow, blue curve, with respect the current due to the rate generation of the B2BT and SRH, red and yellow curves, respectively. The plot is related to the simulation where $\tau^n_{max} = 10^{-9}$ s.

is the B2BT the dominant effect to bring holes inside the channel and the difference in magnitude between I_{GIDL} and I_{SRH} is greater than one order, mainly at short times, as we have see in Fig. 3.9. Considering that, we can conclude that at the beginning of the transient, the enhancement of the generation does not bring any advantage and there are no evident differences in the hole concentration inside the channel, since it is controlled by the tunneling injection. Even though, for long times we know that I_{SRH} gets dominant and so we expect that the hole density reaches the steady-state condition more quickly. This is confirmed by Fig.Fig. 3.15 where we can see again the time evolution of the different kinds of current. We notice that I_{SRH} becomes greater than I_{GIDL} at a previous instant than the case studied before. Finally we see that also the I_{SRH} has a drop in the final part of the transient. This behaviour means that the system has reached the equilibrium

3.8 Preliminary model of the return to equilibrium of the DCP

In order to better understand understand the effects caused by the DCP and describe the phenomenology, a preliminary compact model was created. This model include just the key elements that characterize the return to the equilibrium of the DCP. The main aspects that we have considered are: i) the electrostatic at the edges of the channel where

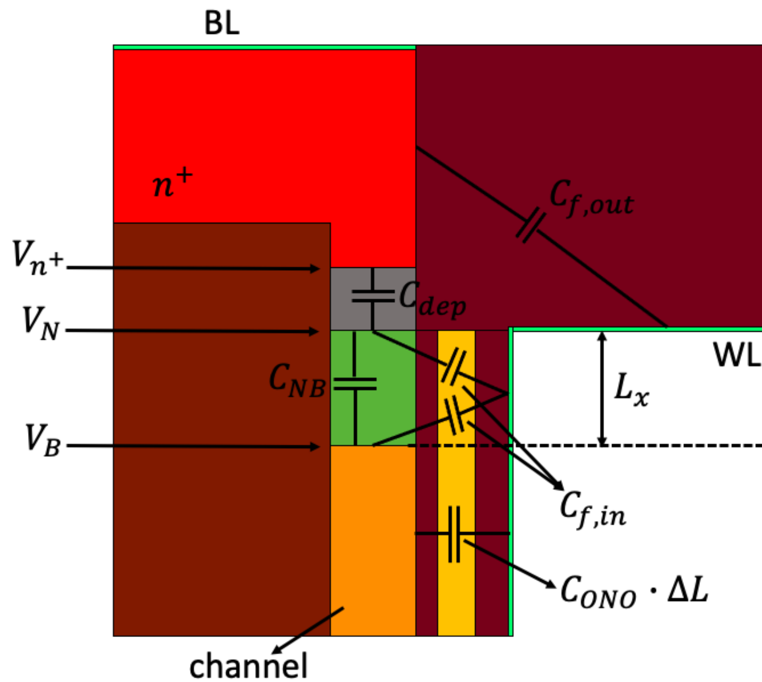


Figure 3.16: Schematic representation of the capacitive couplings in the string at the edge of the channel.

a junction between the channel and the n^+ region of the BL/SL is present; ii) the I_{GIDL} and the I_{SRH}; iii) the time evolution of the hole concentration inside the string.

3.8.1 Description of the compact model

We know from the analysis that we have just done, that when the verify phase ends, at the edges of the channel a junction is establish. This means that in the n^+ zone, a depleted region is created where the electrons are pushed away towards the BL contact and positive charge is left uncovered. This charge produces an electric field that is mainly directed towards the centre of the string along the longitudinal direction, as we were able to confirm by Fig. 3.5a and Fig. 3.5b. At the same time we have seen in Fig. 3.10 that the holes accumulates first at the centre of the channel and then the concentration becomes wider expanding itself towards the outer regions. This condition left empty of holes the area of the channel close to the boundaries with BL/SL and so a depleted region is established also at the channel side of the junction. Considering that, in order to reproduce the dynamic of the string, we have taken into account the capacitive couplings between these regions and we have implemented a circuital model to simulate the transient. Fig. 3.16 shows the various capacitances considered in the model, in particular we can distinguish in gray the depleted region at the n^+ side, while in green is represented the depleted region at the channel side. The line separating these two areas corresponds to the physical edge

of the channel.

In order to calculate the values of the capacitances, we have exploited the results obtained in [20] where a similar problem was analyzed. In that work a compact model was created to simulate the injection of electrons by GIDL for a BiCS structure during the erase operation. In our case, the condition that we want to simulate is not so different since, in both the analysis, the channel is in a floating condition and the capacitances are the same except for some missing capacitive couplings due to the differences in the structure. In particular we have exploited the results just of those capacitances that are included in our geometry.

Referring to Fig. 3.16 we have calculate $C_{f,out}$ that is the capacitive coupling between the transverse face of the CG and the longitudinal face of the n^+ region. In order to find out the value of this capacitance, the approach followed in [21] and [22] has been used, and in the end we obtained:

$$C_{f,out} = \frac{16}{\pi}\varepsilon_0\varepsilon_{ox}t_{WL}\left(3 - \frac{1}{\sqrt{2}}\right)\eta \tag{3.16}$$

where T_{WL} is taken by table 3.1, instead η is:

$$\eta = \sqrt{\frac{2\pi R_{ch}\left[\sqrt{2}t_{WL} + \left(\sqrt{2}-1\right)\left(R_{ch} + t_{eq,ONO}\right)\right]}{4\left(R_{ch} + t_{eq,ONO} + t_{WL}\right)^2 - \pi\left(R_{ch} + t_{eq,ONO}\right)^2}} \tag{3.17}$$

In the equation above $t_{eq,ONO} = \varepsilon_{ox}\left(t_{tOx}/\varepsilon_{ox} + t_n/\varepsilon_n + t_{bOx}/\varepsilon_{ox}\right)$ represents the equivalent thickness of the ONO stack, i.e. the thickness that the ONO stack would have if all the dielectrics in the stack have the same dielectric relative constant ε_{ox}.

For what concerns the series of C_{dep} and $C_{f,in}$, the former capacitance is the capacitance of the depletion layer at the at the n^+ side of the junction, the latter capacitance, instead, represents the coupling between the edge of the channel and the longitudinal face of the CG. In order to calculate C_{dep} we have assumed that the potential is constant along the radial direction such that the edges of the depletion regions, both in the n^+ side and in the channel side, are parallel to the edge of the channel. The value of C_{dep} can be calculated considering the voltage drop in the depleted region at the n^+ side as the voltage drop present in the substrate of a planar MOSFET. So we obtained:

$$C_{dep} = \frac{\varepsilon_0\varepsilon_{Si}}{W_{dep}}A_{tr,chan} \tag{3.18}$$

where

$$W_{dep} = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}}{qN_{BL}^{donor}}\left(V_{n^+} - V_N\right)} \tag{3.19}$$

Instead, $A_{tr,chan}$ is the transverse area of the channel that can be easily found by means of the equation:

$$A_{tr,chan} = \pi \left(R_{ch}^2 - R_{fil}^2 \right) \tag{3.20}$$

The values R_{ch} and R_{fil} are once again taken by the table 3.1.

On the other hand, $C_{f,in}$ represents two equivalent couplings. The first one is between the edge of the channel and the longitudinal face of the CG, the second one is between the CG and the edge of the depletion region at the channel side, see Fig. 3.16. We have to point out that the portion of CG to which we are referring is the portion whose lenght corresponds to the length of the depletion region at the channel side. We have called this length L_x. C_{NB} is the capacitance of the depletion region at the channel side. In order to find out the values of $C_{f,in}$, C_{NB} and L_x, we have solved in cylindrical coordinates the Laplace equation $\bigtriangledown^2 \phi(z, r) = 0$ in the depleted region at the channel side. Thanks to the approximation in which the potential does not depend on the radial coordinate, we could set the following conditions for the potential:

$$\phi(0, r) = V_N \tag{3.21}$$

and

$$\phi(L_x, r) = V_B \tag{3.22}$$

where $z = 0$ correspond to the physical edge of the channel. The remaining calculations to find $C_{f,in}$, C_{NB} and L_x are very laborious and, if the reader is interested, the results are reported in the Appendix of [20].

C_{ONO} represents the capacitance per unit length of the ONO stack. It can be calculate exploiting Eq. 3.9 in the following relation:

$$C_{ONO} = \frac{q n_t \pi \left(r_n^2 - r_{tOx}^2 \right)}{\Delta V_T} \tag{3.23}$$

In the end, $C_{ONO} \cdot \Delta L$ is the remaining capacitance between the WL and the channel filled by holes where, considering the geometry of our interest, that has just one big WL, we can set $\Delta L = \frac{1}{2} L_{WL} - L_x$.

3.8.2 Circuital model

Starting from their initial values, the evolution of V_N and V_B was calculated by performing a transient simulation of the circuit reported in Fig. 3.17. We can notice from the picture that the voltage drop between the channel and the BL is initialized by increasing

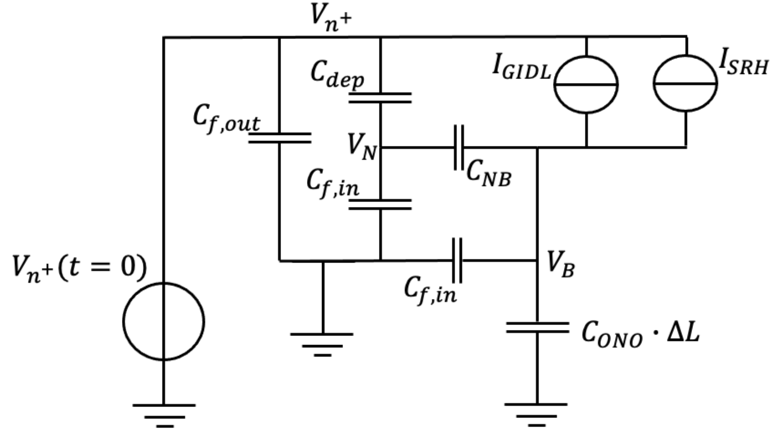


Figure 3.17: Schematic representation of the circuital model.

the potential of the BL instead of lowering the potential of the channel. Despite of this, the physics during the transient does not change. We also notice that the inclusion of the generation by B2BT and SRH is implemented by means of two current generators. For what concerns the SRH generation we have neglected the generation component that takes place in the middle of the string because it is several orders of magnitude lower than the generation at the junction, as we have see in Fig. 3.12. Regarding I_{GIDL}, it has been calculated as the product between the GIDL current density, J_{GIDL}, and the transverse area of the channel $A_{tr,chan}$. In the specific for the current density:

$$J_{GIDL} = qA_{tr,chan}\left(V_{n^+} - V_B - E_G/q\right) F_{max}^{3/2} e^{-F_0/F_{max}} \tag{3.24}$$

where A is the same coefficient used in the TCAD simulation $A = 4 \cdot 10^{14}\ cm^{-3}/s$, while $F_0 = 3.1 \cdot 10^7\ V/cm$. E_G is the energy gap of the silicon while F_{max} is the the electric field calculated at the edge of the channel as:

$$F_{max} = 2\frac{V_N - V_{n^+}}{W_{dep}} \tag{3.25}$$

For what concerns I_{SRH}, we have used an approximated formula of the Eq. 3.6 and the expressions becomes:

$$R_{approx}^{SRH} = \frac{n_i}{2\left(\tau_p + \tau_n\right)} \tag{3.26}$$

where n_i is the intrinsic electron density, while $\tau_n = 10^{-8}\ s$ and $\tau_p = 3 \cdot 10^{-9}\ s$ are the life times of the carriers. In the end, the SRH current is described by the following equation:

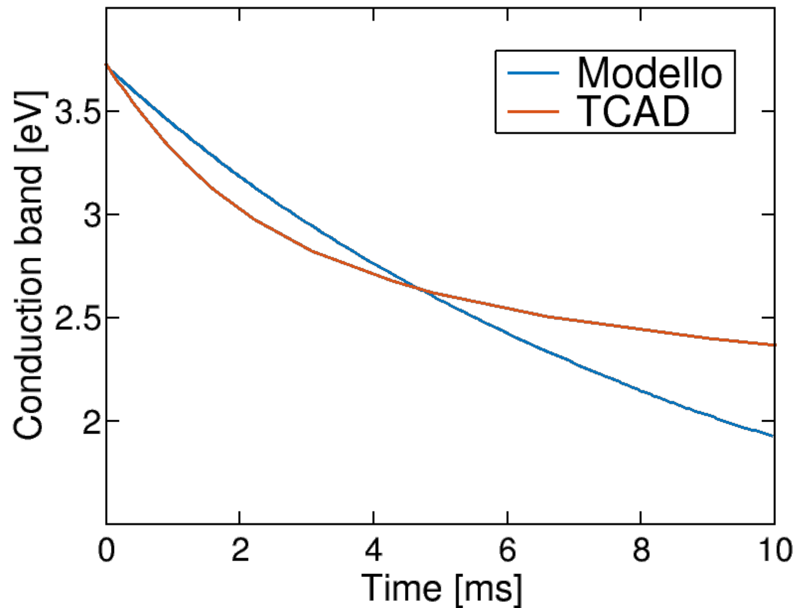


Figure 3.18: Comparison of the time evolution of the CB energy between the compact model result (blue curve) and the Sentaurus TCAD result (red curve).

$$I_{SRH}^{model} = qA_{tr,chan}\,(L_x + W_{dep}) \cdot R_{approx}^{SRH} \tag{3.27}$$

3.8.3 Model results

Fig. 3.18 shows the comparison of the time evolution of the CB energy between what we have obtained in the compact model simulations and the results obtained in Sentaurus TCAD. We can see that at short times the drop of the two curve is quite similar. At the beginning the energy drops more quickly for the TCAD result, instead for longer times the trend is the opposite. Actually the compact model must be reviewed because it reaches the equilibrium condition much earlier than the case of the TCAD simulations and when this happens the simulation breaks. For this reason at the moment we are able to plot a transient of just 10 ms.

The reason of this faster transient lies in the magnitude of the calculated currents in the compact model. Fig. 3.19 shows the I_{GIDL} and I_{SRH} for both the compact model and the TCAD simulation. We notice that, as we have supposed, the difference in the CB energy transient can be traced back in the difference between the I_{GIDL} of the two cases. At the beginning, the GIDL current of the TCAD simulation is higher, in fact the energy lowers more quickly. Then this current becomes smaller than the analogous of the compact model and for this reason, in the compact model, the transient takes less time to return to the stationary condition. Regarding the SRH current, the two curves are not so distant but remains the fact that I_{SRH}^{TCAD} decreases more than I_{SRG}^{model}. Probably these

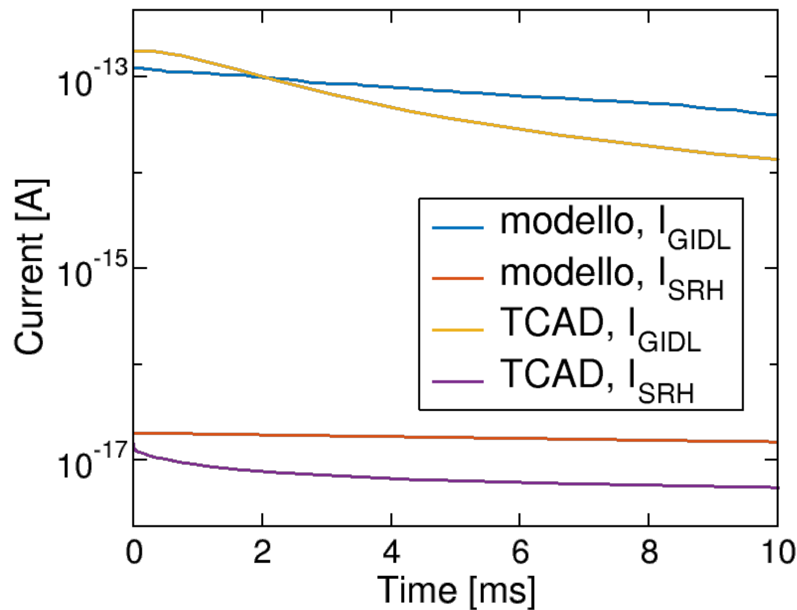


Figure 3.19: Comparison of the time evolution of the currents between the compact model result and the Sentaurus TCAD result.

differences are related to the calculation of the field at the junction but we cannot be sure until a deeper analysis of the compact model is performed. At the end, we can conclude that the results, as order of magnitude, are not so distant from what we have obtained in the TCAD simulations but, in order to have a higher agreement, further improvements of the compact model must be done.

3.9 conclusions

In this chapter we have introduced the Sentaurus TCAD software that was used to carry out the simulations of the DCP in 3D NAND Flash memories. At the beginning, we have presented the simplified BiCS geometry implemented in the simulation, a single WL string used to simplify the analysis of the results. Next we have described the physical models included in the simulation and then we have discussed the obtained results. In particular we have showed that for the present geometry, whose cell was programmed P3, the DCP brings the system out of equilibrium and this equilibrium cannot be restored in short times. The GIDL current is the main supply of holes for the channel, even though the magnitude of this current is not sufficient to increase the channel potential before the program phase begins. This is due to the relation between the GIDL current and the field present at the edge of the channel. In fact, as the field decreases, the band bending does the same and the generation by tunneling turns down. For this reason, at long times, the SRH generation becomes dominant to bring back the system to the equilibrium. In the

end, we have introduced a preliminary compact model that has the purpose of simulating the DCP neutralization process. We have showed the first results of the model, they are close to the results of the TCAD simulations but not in complete agreement, for this reason an improvement of the model is needed.

Chapter 4

Simulations of BiCS and TCAT structures

In this chapter will be discussed the TCAD results about the DCP simulations for the two realistic structures BiCS and TCAT. In particular, the effects related to the presence of several WLs instead of just one global WL will be discussed. At first the transient in the BiCS structure will be analyzed, then we will move to analyze the TCAT structure. The differences in the DCP between the two geometry will be showed. In the end, the simulation results for a differently programmed TCAT string will be discussed.

4.1 DCP in BiCS memories

4.1.1 Geometry and characteristics of a 16 WLs NAND string

The structure considered in the previous chapter is a simplified one, with just one gate electrode. This simplification is useful for assessing the main physical phenomena that take place inside the device, but does not represent any real memory. This Chapter addresses instead the DCP in a NAND string, discussing the peculiar differences with respect to the previous one. In the current case, the control gate is split in 16 WLs plus an SSL and a DSL, to simulate a realistic NAND cell string. The control gates are separated by oxide spacers having the same size as the WLs, while the other parameters retain the same values of the single cell string, as reported in table 4.1.

Fig. 4.1 shows half of the BiCS structure that we have defined. In the present case, the set of control gates (drawn in green), separated by spacer regions having the same size as the WLs, is no longer equivalent to a single electrode spanning over the entire channel. From an electrostatic point of view, the capacitive coupling between the CGs and the channel is not equivalent to a single capacitor but the coupling is better described by a

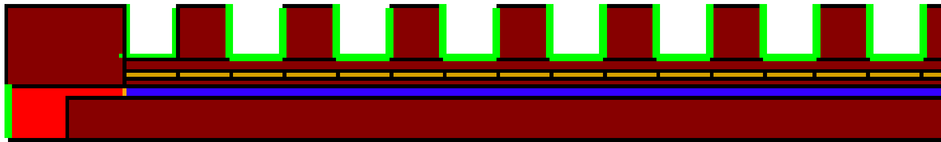


Figure 4.1: Half geometry of the BiCS structure.

Table 4.1: Geometry parameters used in the simulation for the *16 WL* string.

quantity	symbol	value
Gate length	L_{WL}	40 nm
spacer length	L_{sp}	40 nm
Gate thickness	t_{WL}	40 nm
Channel length	L_{ch}	1.4 μm
tunneling oxide layer thickness	t_{tOx}	4 nm
silicon nitride layer thickness	t_n	8 nm
blocking oxide layer thickness	t_{bOx}	8 nm
Macaroni radius	R_{mac}	30 nm
Channel radius	R_{ch}	40 nm
Polysilicon channel thickness	t_{ch}	10 nm

set of smaller capacitors positioned in parallel. This means that there are fringing fields between adjacent gates. Actually, these fringing fields control the electron charge in the channel regions underneath the oxide spacers and so we can conclude that the control exerted by the gates is less than the case described in the previous chapter.

4.1.2 Physical differences between single WL structure and 16 WLs structure

Since we are now simulating a realistic NAND string, we must account for the fact that the electron charge stored in the nitride is not uniformly distributed along the channel, but it is placed just in correspondence of the gates, where the FN tunneling takes place. In order to implement this non-uniformity of the trapped charge, the nitride layer has been broken into several regions, as we can see in Fig. 4.1. The regions are in correspondence of the gates and constitute the areas where the fixed charge density was placed to simulate the program operation. Please note that the definition of those limited zones is necessary just to define the regions where the charge is positioned but it does not reflect any physical boundaries within the nitride layer. For what concerns the charge, it was placed the same density of charge used in the previous chapter, i.e. $n_t = 9.15\ cm^{-3}$ that is the charge

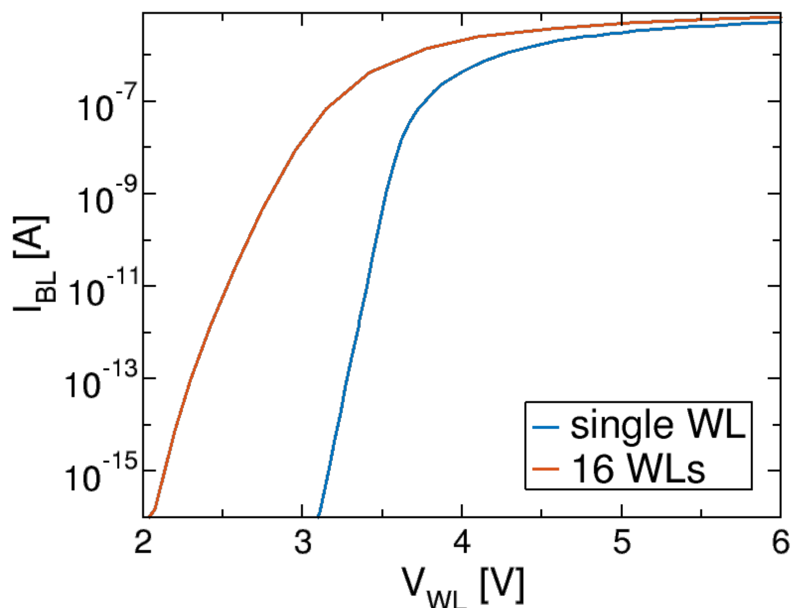


Figure 4.2: Logarithmic comparison of the transcharacteristics belonging to a structure with 16 WLs (blue curve) and the structure with just one WL (red curve).

defined to have the cells programmed $P3$. However, this means that, for the this given density, the total trapped charge in the present case is half of the previous one. Moreover, in the current geometry not only the coupling between the channel and the control gates undergoes a variation, but the coupling between the trapping layer and the channel is modified as well. This is very important from the viewpoint of NAND operation. In fact, despite the trapped charge is concentrated just in those region in correspondence of the gates, it has the role of inverting the whole channel. The main consequence of such difference is visible in the transcharacteristic of the cell. Indeed, as reported in Fig. 4.2, the current flowing through the string is rather different. From the picture we notice that not just the V_T moves to a lower value, but also the *Sub-Threshold slope (STS)* degrades as well. Since the total trapped charge is half of the one implanted in the single-cell geometry, we would expect the new ΔV_T^{16WLs} to be half of ΔV_T^{1WL}. Unluckily, the relation is not so straightforward, but we have to take into account the fringing capacitance in order to justify the value of the shift in the threshold voltage. By considering that, with respect to the case of the previous chapter, in the present case there is half of the total trapped charge but the capacitance is higher than half of the capacitance found in Eq. 3.8. So, we expect that the $\frac{1}{2}\Delta V_T^{singleWL} < \Delta V_T^{16WL} < \Delta V_T^{singleWL}$. In order to find out this value, first we have to quantify the total capacitance that links the trapping layer to the channel. We call this capacitance C_N and we suppose that it is the same of the capacitance related to each trapping region, that in turn are built up by different components. These components are: the capacitance related to the coupling of parallel surfaces, calculated

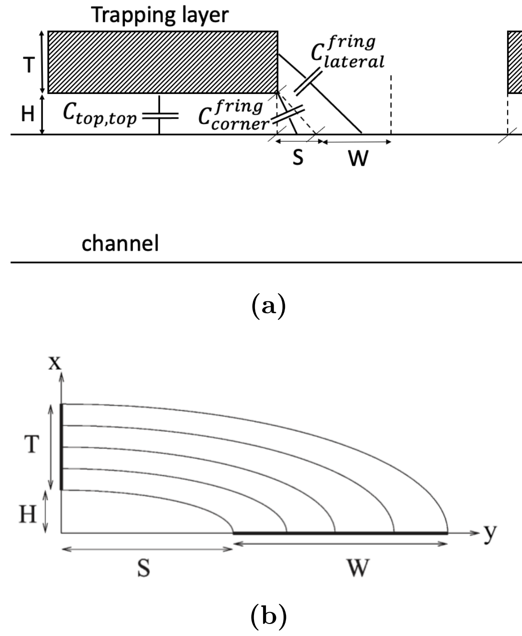


Figure 4.3: (a) Schematic representation of the contributions to the capacitance between the trapping layer and the channel and (b) pictorial view of the field lines in an elliptical system (from [23]).

as described in the previous chapter in Eq. 3.8, and the fringing capacitance that couples the lateral surface of the trapping zone with the channel. In the computation of this latter component, we considered it subdivided in two terms that are: the C_{corner}^{fring} and the $C_{lateral}^{fring}$.

Fig. 4.3a reports schematically this capacitance subdivision, where the $C_{top,top}$ is the component calculated in the Eq. 3.8, but in the current case it addresses just one trapping region. The other capacitances are calculated by exploiting a conformal mapping, i.e. the problem becomes finding the electric field in a quadrant of an ellipses system, as depicted in Fig. 4.3b. From this system it is possible to make a transformation of the coordinates and move to a parallel-plate capacitance problem. The calculations are present in [23] where at the end an equation for $C_{lateral}^{fring}$ is found:

$$C_{lateral}^{fring} = \frac{2\varepsilon_0\varepsilon_{ox}}{\pi} \ln\left(\frac{H + T' + \sqrt{S^2 + T'^2 + 2HT'}}{S + H}\right) \tag{4.1}$$

where $H = t_{tOx}$, $S = H$ and $W = \frac{1}{2}L_{sp} - S$. This value of W can be justified by the fact that just one half of the channel underneath the oxide spacer is controlled by one trapping region, the other half of the channel is controlled by the next trapping zone. In the end $T' = \eta T$ is the equivalent thickness of the trapping region in the elliptical system. η is an empirical parameter, but to understand the relation of this coefficient

with geometry it can be empirically represented as:

$$\eta = e^{\frac{W+S-\sqrt{S^2+T^2+2HT}}{\tau W}} \tag{4.2}$$

where $\tau = 3.7$ and it is determined to match the model with the computational results. For what concerns the C_{corner}^{fring} it can be found with the equation:

$$C_{corner}^{fring} = \frac{\varepsilon_0 \varepsilon_{ox}}{\pi} \sqrt{\frac{HS}{H^2 + S^2}} \tag{4.3}$$

What we have found so far is a capacitance per unit space, to get the total capacitance of one trapping region we approximated the cylindrical problem to a rectangular one where the characteristic length of the capacitor is equal to the circumference of the channel. Since C_{corner}^{fring} and $C_{lateral}^{fring}$ are in parallel, we can calculate the fringing capacitance of one trapping region in the following way:

$$C_{tot}^{fring} = 2\pi R_{ch} \left(C_{corner}^{fring} + C_{lateral}^{fring} \right) \tag{4.4}$$

In the end, to find the total C_N^{tot} we have to sum the contributes coming from each trapping region. We have to consider that each trapping region underneath a WL contributes with two radial surfaces a double contribute, while those one corresponding to the SSL and the DSL have just one lateral surface connected to the channel, so we have obtained:

$$\begin{aligned} C_N^{tot} &= 16 \left[2 \left(2\pi R_{ch} \right) \left(C_{corner}^{fring} + C_{lateral}^{fring} \right) + C_{top,top} \right] + \\ &\quad 2 \left[\left(2\pi R_{ch} \right) \left(C_{corner}^{fring} + C_{lateral}^{fring} \right) + C_{top,top} \right] \end{aligned} \tag{4.5}$$

With this value of the total capacitance of the nitride layer we can determine the new ΔV_T that came out to be $\Delta V_T^{P3} \approx 2.38\ V$. We have extrapolated the transcharacteristic for the 16 WLs geometry in case of no trapped charge and we obtained $V_T\ |_{n_t=0} \approx 0.66\ V$. At the end, we have calculated $V_T^{P3} \approx 3.04\ V$ that is very close to the value reported in Fig. 4.4. In the picture it is reported the linear trend of the transcharacteristics for the case of 16WLs and a single WL. It is more useful using this kind of plot because, due to the degradation of the STS, it is rather difficult extrapolate the threshold from Fig. 4.2. In Fig. 4.4 the threshold corresponds to the lowest voltage value for which the current is different from zero. As we can notice, the extrapolation of the threshold from the picture gives us $V_T^{P3} \approx 3\ V$ that is quite close to the value we have calculated. At the end, thanks to what we have seen about the threshold voltage of the new structure, we can come to

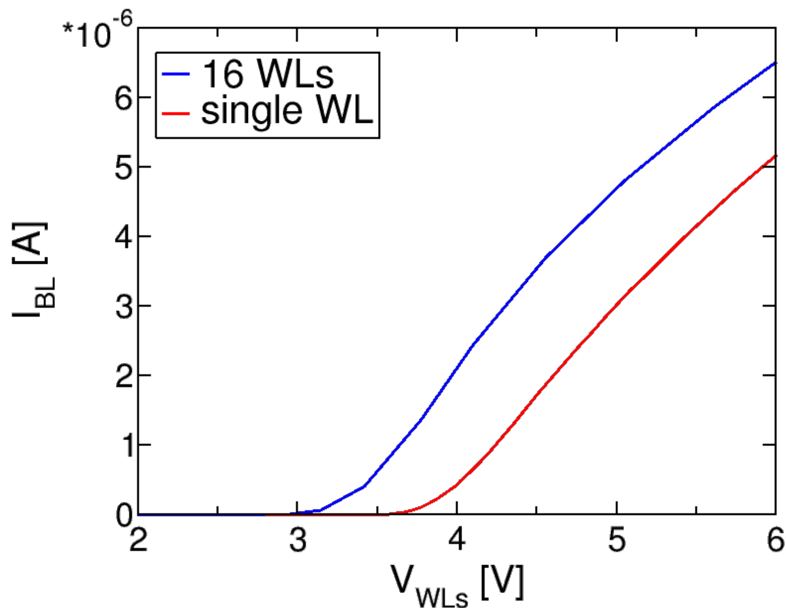


Figure 4.4: Linear comparison of the transcharacteristics belonging to a structure with 16 WLs (blue curve) and the structure with just one WL (red curve).

the conclusion that this system will have a lower boosting effect.

4.1.3 Time evolution of the conduction band energy in BiCS

For what concerns the simulations, we have included the same physical models reported in the previous chapter. Fig. 4.5 shows the comparison of the conduction band energy between the current structure and the single-cell geometry at the end of the verify phase. We can notice that the maximum value of the CB energy does not reach even 3 eV, while we have just said in the previous section that the extrapolated threshold is $V_T^{P3} \approx 3\ V$. This incongruence is due to the degradation of the STS that makes the turning off of the transistors not so sharp. In turn, this degradation affects the final value of the channel potential that becomes:

$$\Delta V_{down-coupling} \lesssim V_{T,neighbors} \tag{4.6}$$

In order to understand how much the STS degradation impacts on the DCP, further investigations must be accomplished. In our work we have decided to give more relevance to the return of the equilibrium. From Fig. 4.5, we can also see that now the CB energy is not constant along the channel but it is modulated periodically by the alternated presence of gates and intercell oxide spacers. In correspondence with the gates, the channel energy is higher due to the negative charge stored in the nitride layer. The presence of this modulation of the energy affects mostly the way in which the electrons get out from the

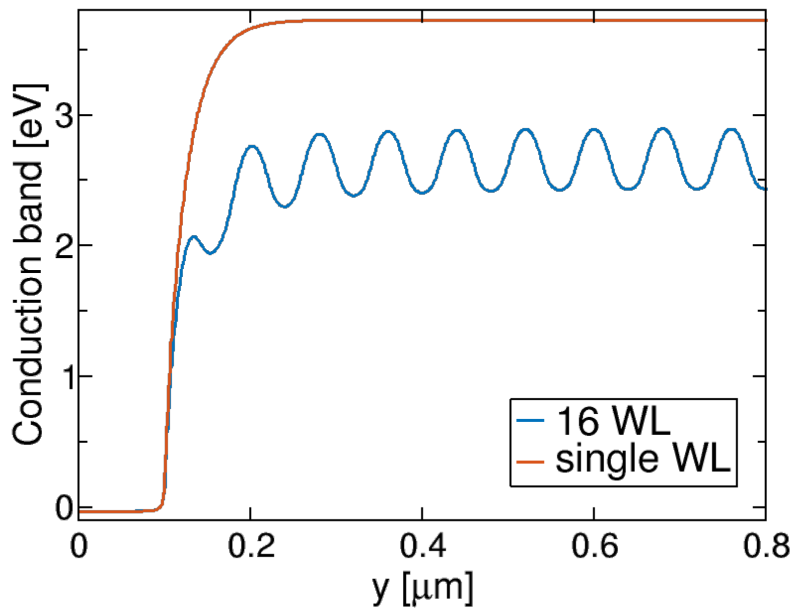


Figure 4.5: Comparison of the CB energy between the structure with 16 WLs (blue curve) and the structure with one single WL (red curve). The longitudinal section is taken at 0.25 nm from the silicon-oxide and $t = 0$.

channel. What happens is that when the V_{WL} becomes lower than the threshold voltage, the channel is no longer well controlled by the CGs and the field created by the stored charge becomes dominant. The channel areas underneath the trapping regions, that for the sake of simplicity we call *Control Gate Zones (CGZs)*, are the first areas where the CGs loose the control. So, in correspondence of the CGZs, depleted regions are created because the electron are pushed away by the electric field of the negative stored charge. This displacement of the electrons towards the channel regions underneath the spacers, that we call *Spacer Zones (SZs)*, creates energy peaks in the CB energy, as we have seen in Fig. 4.5. During the falling edge of V_{WL}, the electrons tends to flow towards the BL and the SL. Despite of the electrostatic force that pushes the electron outside the string, locally the negative carriers have to overcome the potential barriers in the CGZs and this makes them slow down in their travel towards the BL and the SL. In the end, the electrons are shut in between the potential barriers in the CGZs and their density remains rather high.

Fig. 4.6 shows the time evolution of the CB energy in a CGZ together with the steady-state value of the conduction band for the same position in the channel. Focusing our attention on the blue curve, there is an initial small lowering of the energy, then the curve remains almost constant till $t \approx 3\ ms$. This is the main difference with respect to the case of the single WL, that we have seen in the previous chapter. The similarities, instead, lie in the last part of the trend where we can see that also this structure cannot reach the

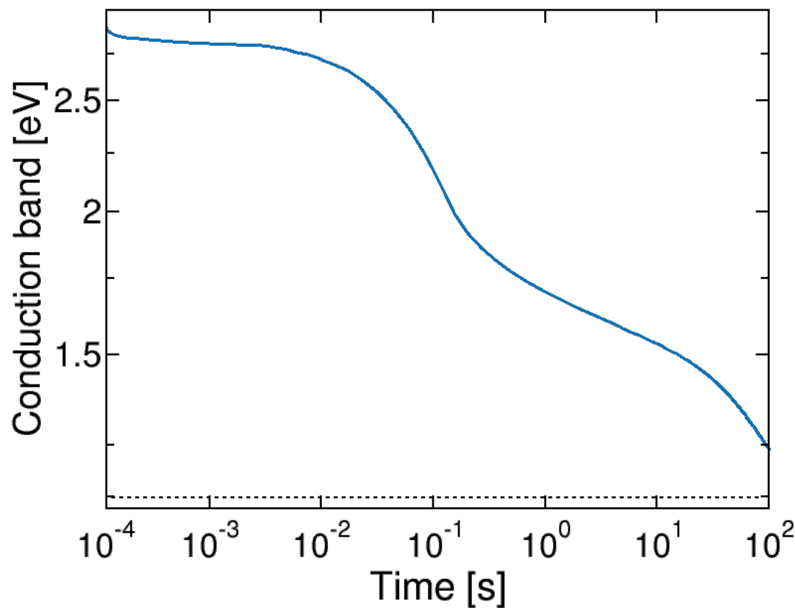


Figure 4.6: Time evolution of the conduction band energy underneath a trapping region in the middle of the string, since the end of the WL falling edge. The dashed black line represent the equilibrium value.

equilibrium value in $t = 100$ s. Again, to understand the trend of the CB energy, we first have analyzed the carrier concentration inside the string.

4.1.4 Carriers concentration in BiCS

Fig. 4.7 reports the comparison of the electron density along the longitudinal section of the string at $t = 0$ s. We can notice that the electron density is still very high inside the channel, in particular in correspondence of those channel regions underneath the oxide spacers. The density is so high that it is still comparable with the trapped charge density and so it is a bit relevant for the electrostatic of the system. This can be confirmed in Fig. 4.6 where at very small times there is a small drop of the CB energy. This tells us that the system, in order to go back to the equilibrium, not just has to inject the holes into the string but it has to pull out the electrons as well. There are two ways for the system to restore the correct electron density inside the channel: the first possibility is the recombination of electrons and the second is making the electrons flow outside the channel by means of drift-diffusion processes. We will see that both options take place simultaneously because just one out of the two processes is not strong enough to bring the system at the equilibrium. In fact, the recombination, in order to have a rate sufficiently large, needs an appropriate high density of holes. On the other hand, considering the drift-diffusion process, we must account for the potential barriers that must be not too high to allow the electrons to overcome them. These two alternatives about the re-establishment

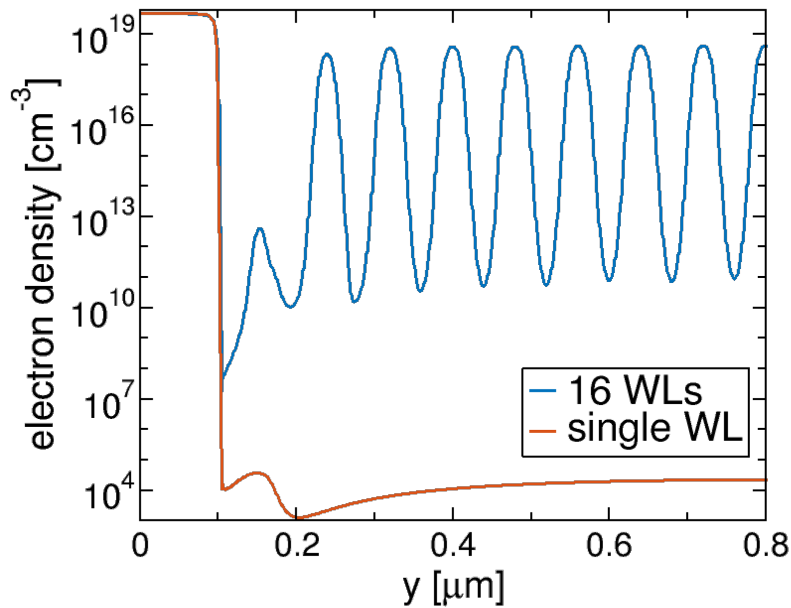


Figure 4.7: Comparison of the electron density between the structure with 16 WLs (blue curve) and the structure with one single WL (red curve). The longitudinal section is taken at 0.25 *nm* from the silicon-oxide interface taken at $t = 0$.

of the equilibrium electron concentration have in common the needs of the presence of holes in order for them to occur.

Fig. 4.8 reports the comparison of the hole density as a function of the longitudinal coordinate at $t = 0$. We can notice that for the 16 WLs string, the hole density is almost ever lower that the case of the previous chapter. The only position where the density is higher is underneath the first WL adjacent to the SSL. Indeed, the holes that get inside the channel begin to accumulate under the third gate that they meet, i.e. WL^1. The holes does not stop neither at SSL nor at WL^0 because they still feel the longitudinal electric filed that makes them move towards the centre of the string. Even though, at WL^1 the longitudinal field is not so strong and the holes are attracted by the negative negative stored charge and so they starts accumulating underneath WL^1. This accumulation does not bring the local electrostatic potential to the equilibrium value because the hole density does not become too large. Even though, the hole concentration starts to screen the electric field, bringing to a small lowering of the barrier potential seen by the electrons. A lowering of this barrier allows the electrons, accumulated in the next SZ, to move towards the SL. Moreover, the increasing availability of holes underneath the WLs makes the recombination between holes and electrons feasible. In this way, the cooperation between the exiting flux of electrons and the recombination reduces the electron density in the region adjacent to the gate. When the hole density in this CGZ is large enough to screen properly the electric field of the stored charge, the holes coming from the junction can

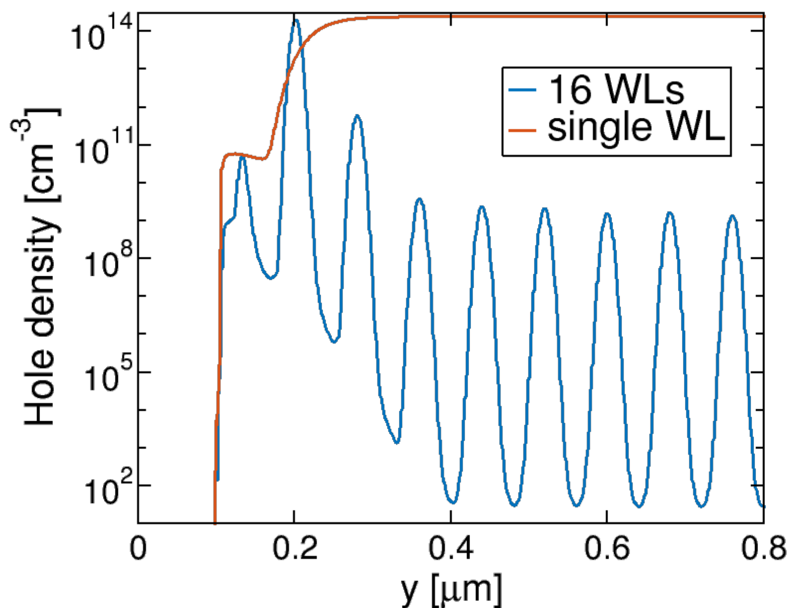
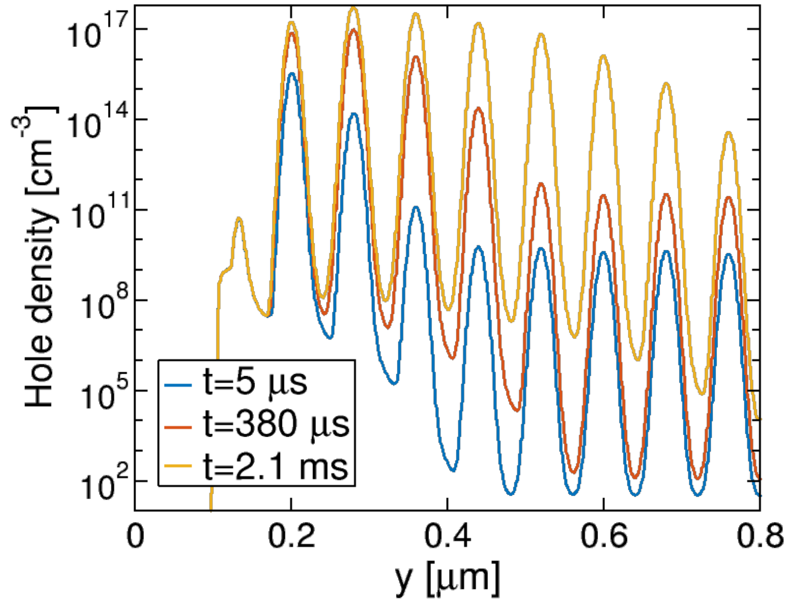


Figure 4.8: Comparison of the hole density between the structure with 16 WLs (blue curve) and the structure with one single WL (red curve). The longitudinal section is taken at 0.25 nm from the silicon-oxide interface taken at $t = 0$.

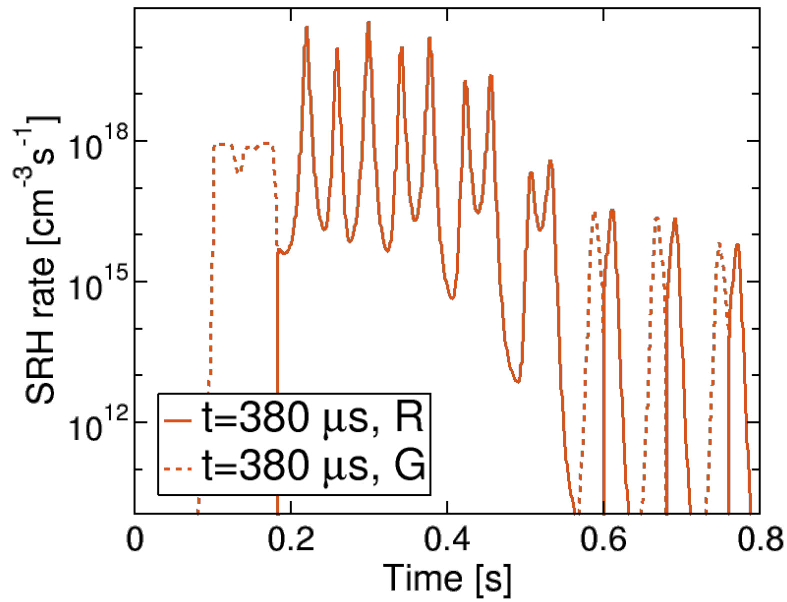
maintain their kinetic energy, cope the energy barrier in correspondence of the SZ, and start accumulating in the following CGZ.

Fig. 4.9a shows the hole density for three different instants of the transitory. The picture is useful to visualize the sequential filling of the CGZ. In fact, the hole density underneath the inner CGZ begins to increase at higher instants than the regions underneath WLs closer to the edge of the channel. If we focus our attention to the red curve of Fig. 4.9a, that is a snapshot of the hole concentration at $t = 0.380$ ms, we can notice that the hole density is increased underneath four lateral WLs, that correspond to $WL^{1,2,3,4}$, and have started to increase underneath WL^5. Now if we compare this curve with the curve in Fig. 4.9b, that depicts the longitudinal section of the SRH rate at $t = 0.38$ ms, we can confirm what we have said before, i.e. the recombination in a certain region of the channel becomes relevant just when the holes starts accumulating in the given CGZ. In fact we can see in Fig. 4.9b that in the region of $0.2 \lesssim y \lesssim 0.6$, that is the region underneath the WLs previously mentioned, the recombination has a rather large value. In the inner part of the string, instead, there is an alternation of recombination and generation. The recombination is present where the electrons are accumulated and there is generation in the depleted region underneath the control gate. Between this two opposite processes, it is the generation to be a bit dominant with respect to the recombination. In fact, the cooperation between this small net generation and the coming of a small of number holes from outer regions produces the small increase of the hole density in the middle of the

string, as we can see in Fig. 4.9a. In the end we can notice that at the edge of the channel a net generation is present, as we have also seen in the single WL string case. In order to further understand the evolution of the carrier density we have studied the time evolution of the current.



(a)



(b)

Figure 4.9: Longitudinal section of the string at 0.25 nm from the silicon-oxide. The figure reports: (a) the hole density comparison for $t = 0.005, 0.380, 2.1$ ms (b) the SRH rate at $t = 0.380$ ms.

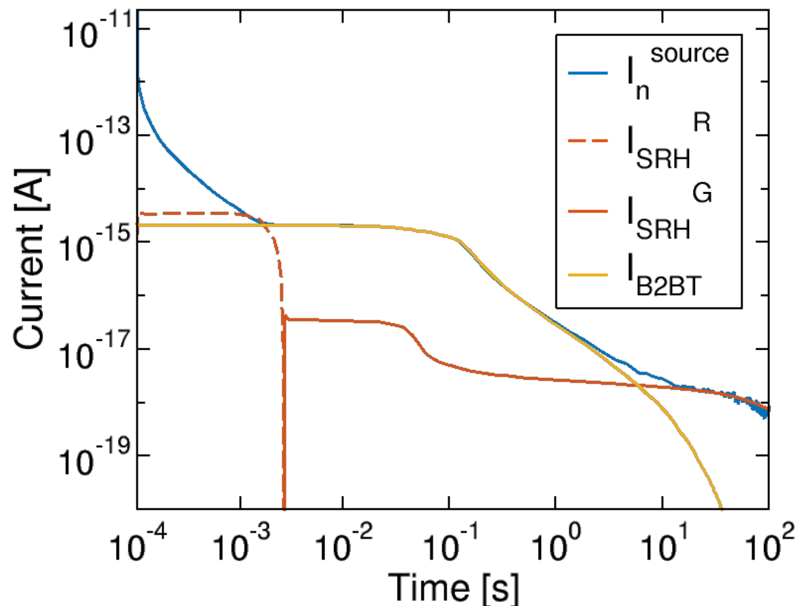


Figure 4.10: Comparison of the electron flow(blue curve) with respect to the current due to the rate generation of the B2BT and SRH (red and yellow curves, respectively). For what concerns I_{SRH}, the dashed part comes from a recombination rate, instead the solid line comes from the generation rate.

4.1.5 Time evolution of the current in BiCS

Fig. 4.10 shows the time evolution of the main current components involved in the transition. We can see that at very short times the blue curve, corresponding to the current of negative carriers that flow through the SL, is higher than the other two currents, that are I_{SRH} (red curve) and I_{GIDL} (yellow curve). This means that the number of electrons passing at the contact is higher than the number of electron generated by the tunneling and SRH. Indeed the channel, after the end of the falling edge of V_{WL}, is still quite full of electrons. A certain amount of this density recombines, while the remaining can flow outside the channel and it is added to the electrons generated by tunneling. About the recombination we have to specify some concepts in order to not bring to misunderstanding. From the discussion we have done in the previous section, we know that in those CGZs, where the holes are present in large quantity, there is a net recombination. On the other hand, in the middle of the string, where the holes are not come yet, there is a small net generation. The I_{SRH} reported in red in Fig. 4.10 comes from the spatial integral of the SRH rate over the whole string. This statement implies that the value of the recombination is higher in the first ms of the transitory, as we can confirm by looking at the dashed red curve. We see also that at the beginning the number of electrons that can exit from the channel is rather big, for this reason the total I_n is greater than I_{GIDL}, but in $t \approx 3\ ms$ the electron density reaches a sufficiently small value that the number of electrons flowing

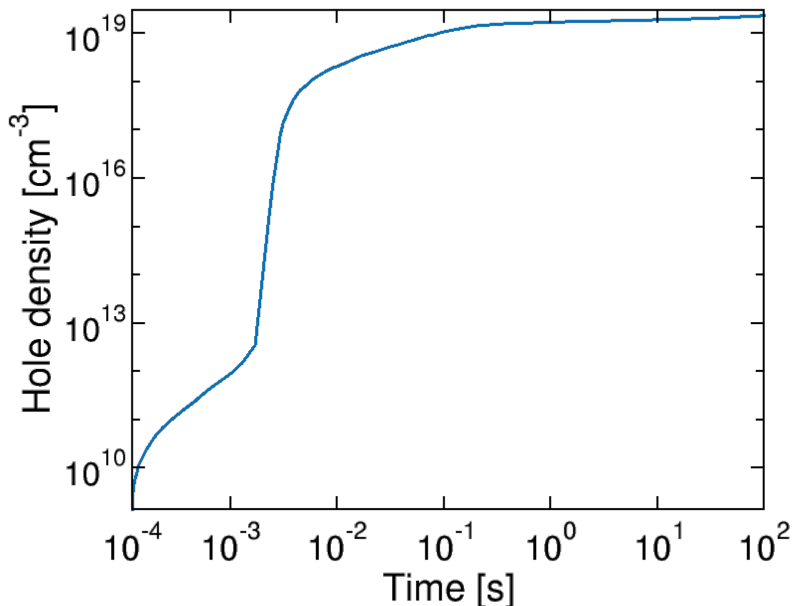


Figure 4.11: Time evolution of the hole density underneath WL^7.

at the SL becomes completely determined by the B2BT generation. This fact temporally coincide with two events: the coming of holes at the centre of the string and the switch of the SRH from a net recombination to a net generation. The former statement is confirmed by Fig. 4.11 where we can see the time evolution of the hole concentration underneath WL^7. This conclusion is in agreement with Fig. 4.6 where we can notice that at $t \approx 3\ ms$ the CB energy starts decreasing, the proof that the hole density is became relevant for the electrostatic of the system. The transition from recombination to generation, instead, can be verified by looking at the red curve in Fig. 4.10.

From this time instant on, the current flowing at the contacts corresponds to the superposition of the currents generated by B2BT and SRH generation at the edges of the channel. For what concerns the SRH rate, the generation becomes relevant just in the depleted region at the edge of the channel, in the rest of the string the generation rate is so small that it does not affect the final value of I_{SRH}. The hole density keeps growing thanks to the holes entering from the junction and at a certain instant the hole concentration is so high that holes are pushed towards the edge of the channel where they starts accumulating even underneath WL^0 and the first SZ, i.e. the region between SSL and WL^0. Fig. 4.12a depicts the time evolution of the hole density in the first SZ and we can notice this sudden growth of the concentration. This growth leads to a decrease of the generation in the depleted region, as depicted in Fig. 4.12b. In this picture we can see the comparison of the SRH rate as a function of the longitudinal coordinate for three different instants.

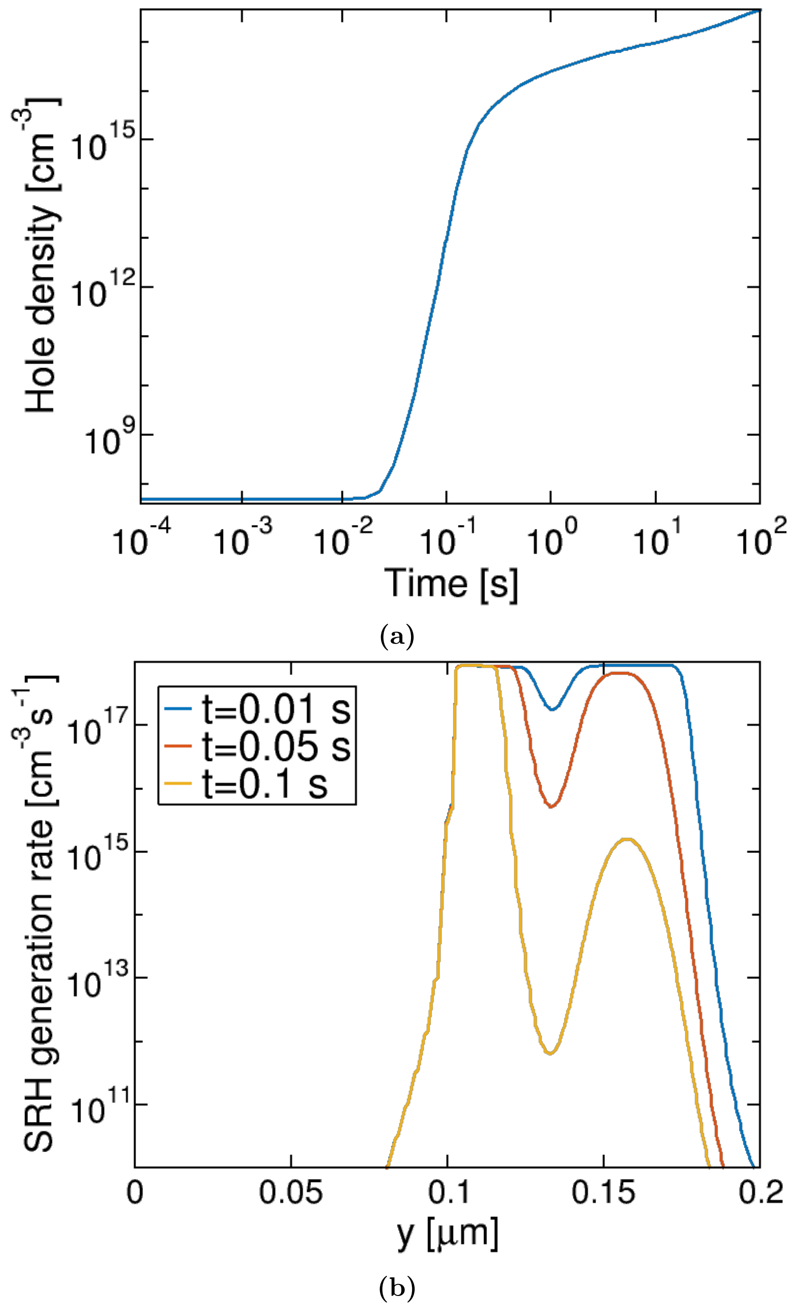


Figure 4.12: (a) Time evolution of the hole density underneath the spacer between the SSL and WL^0, (b) SRH generation rate for three different time instants. The longitudinal section is taken at 0.25 nm from the silicon-oxide.

The relation of this two phenomena can be explained by considering the numerator of Eq. 3.6 that relates the SRH rate with the carrier density. The increase of the hole concentration makes the value np closer to n_i^2 and so the numerator decreases. In turn, the value of I_{SRH} lowers as well, as we can notice from the red curve in Fig. 4.10 at $t \approx 0.1$ s. Even though, the SRH generation is decreasing not in all the depleted region but just in one part. Indeed, we can see that the lowering of I_{SRH} is limited to one decade

because the peak of the generation rate at the edge of the channel remains relevant for the integral of the SRH current.

I_{SRH} is not the only current component that decreases. The holes accumulation between the SSL and WL^0 makes the longitudinal field at lateral junction decrease as well. This leads to a drop of I_{GIDL}, as we can see from the yellow curve in Fig. 4.10. We have seen in Eq. 3.5 that the band generation depends exponentially from the field at the junction. For this reason, a small variation of the field at the junction produces a huge variation in I_{GIDL} with a large decrease of the number of holes injected into the string. This drop of the injection can be related to the slowing down of the CB energy decrease that is depicted in Fig. 4.6. In the end, I_{GIDL} becomes smaller than I_{SRH} and the return to the equilibrium speeds up again, even if the stationary condition cannot be reached in $t = 100\ s$.

4.2 Dependence on parameters for the BiCS structure

As we have done in the previous chapter, we wanted to understand how relevant are the B2BT and the SRH models. In the current structure, the simulation that does not include the B2BT is very different with respect to the case just analyzed. The study of this simulation results quite difficult. In the present work we have decided to skip this analysis that require several efforts, moreover it is not so interesting due the fact that in real case the tunneling cannot be switched off, while it is possible to act upon the life-time of the carriers by changing the characteristics of the silicon. For this reason we have decided to focus our analysis just on the possible variations of the transient due to the changing in the SRH rate. As we have done in the previous chapter, we have changed the life-time parameters of the carriers that are became: $\tau^n_{max} = 10^{-9}\ s$ and $\tau^p_{max} = 3 \cdot 10^{-10}\ s$.

We have found that the effect on the transient is almost the same as for the case of single WL, i.e. the generation rate is higher and I_{SRH} becomes greater than I_{GIDL} in a shorter time. This leads to the fact that the system can keep a higher injection in long times and so it can reach the steady-state condition earlier. But there is another small difference with respect the case of the previous chapter. At the beginning of the transient when the hole density is increasing in the various CGZs, the enhancement of the SRH rate has an opposite effect and it slows down the filling of the channel with holes. Indeed, we have to keep in mind that at the beginning there is no a net generation inside the string but the opposite, a net recombination. The positive carriers, that tends to move towards the centre of the string, recombine with the electrons accumulated in the SZs and for this reason the hole concentration needs more time to reach high enough value in each CGZ.

Fig. 4.13 shows the comparison of the hole density along a longitudinal section ex-

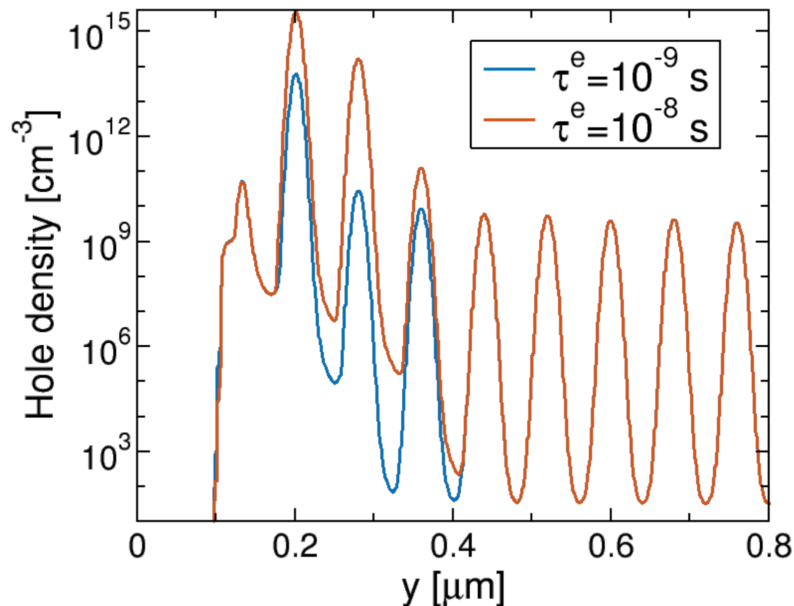


Figure 4.13: (a) Longitudinal section of the string at 0.25 nm from the silicon-oxide. The figure reports the hole density comparison at $t = 5\ \mu s$ between the simulation where was set $\tau^{n}_{max} = 10^{-9}\ s$ (blue curve) and the simulation with $\tau^{p}_{max} = 3 \cdot 10^{-10}\ s$ (red curve).

tracted at $t = 5\ \mu s$ and confirms the discussion we have just done. Even though, we have to keep in mind that at the beginning the B2BT is dominant, so, with a long time transient, the system can in any way fill the channel of positive carriers and invert the SRH rate. In fact at the end, the decrease of the life-time leads to a reduction of the time that the system needs to go back to the stationary condition.

4.3 DCP in TCAT memories

Up to now we have described structures where the channel is not linked to a tank of holes but the substrates were doped n^{+}. In this condition, the only effect that quickly and significantly can supply holes to the string is the I_{GIDL} current. On the other hand, if the substrate is composed by a p-doped region and a n-well, both kinds of carrier can be supplied to the channel. First of all, we have to point out that, what we call TCAT structure is not not exactly a realistic TCAT structure but there are small differences with our geometry. We did not use metal gates and we did not change the geometry of the ONO stack, like we should have done in order to respect the gate last manufacturing process, see Fig. 1.15. We choose to not apply this changes in order to maintain the cylindrical symmetry and keep the computational cost to a minimum. After said that, Fig. 4.14 shows the substrate region of the new geometry.

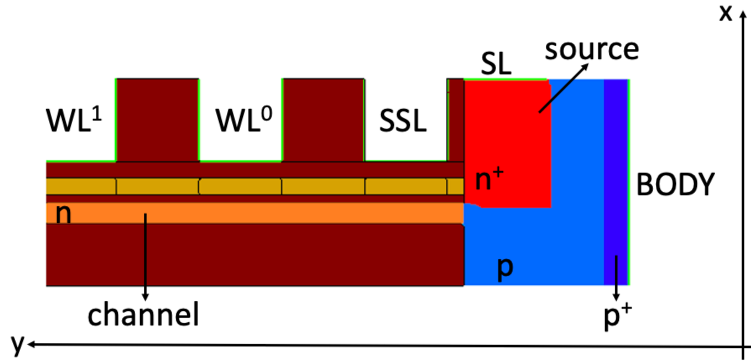


Figure 4.14: Representation of the substrate region for a TCAT geometry.

4.3.1 TCAT geometry

The few modifications are done in this substrate region of the memory. Now the silicon oxide does not reach the bottom of the string but it is deleted in order to make room at the n-well. The only part that we have left of SiO_2 is that area between the SSL and the SOURCE such that an unwanted current does not flow between these two contacts. The substrate is entirely made of silicon and it is doped in different ways in order to allow the creation of the contacts. The silicon close to the *BODY* contact has a higher doping concentration in order to avoid abnormal recombination effects at the contact. This area, depicted in dark blue in Fig. 4.14, is doped with *Boron* atoms and the doping value is $N_a^{BODY} = 5 \cdot 10^{19}\ cm^{-3}$. The light blue zone is still doped with Boron but the doping concentration is lower $N_a^{sub} = 5 \cdot 10^{18}\ cm^{-3}$. In the end, we can see the red area that acts as n-well. This zone is doped with *Arsenic* atoms, with a concentration of $N_d^{source} = 5 \cdot 10^{19}\ cm^{-3}$. The rest of the structure does not change with respect to the BiCS case. We wanted to remind here the not changed parameters: the channel, in orange, is doped $N_d^{channel} = 10^{15}\ cm^{-3}$ while the BL has the same doping of the SOURCE, $N_d^{BL} = 5 \cdot 10^{19}\ cm^{-3}$, but without any p contact.

The definition of a correct substrate doping is very important for the correct functioning of the memory. In fact, we wanted the substrate to accomplish two duty: the p-doped substrate must be able to supply a sufficient amount of holes to the channel in order to avoid the DCP, while the n-well must have an optimal conductivity such that the electrons can flow trough the SL and allow the correct read operation in the string. For what concerns the hole supply, we have to point out that the holes flux from the BODY to the channel is not sufficient for small value of N_a^{sub}. On the other hand, the source must be positioned quite close to the boundary of the channel in order that the resistance is not too high to compromise the operation of the string. Regarding this latter condition, we have found out that setting the extension of the source as depicted in Fig. 4.14, where

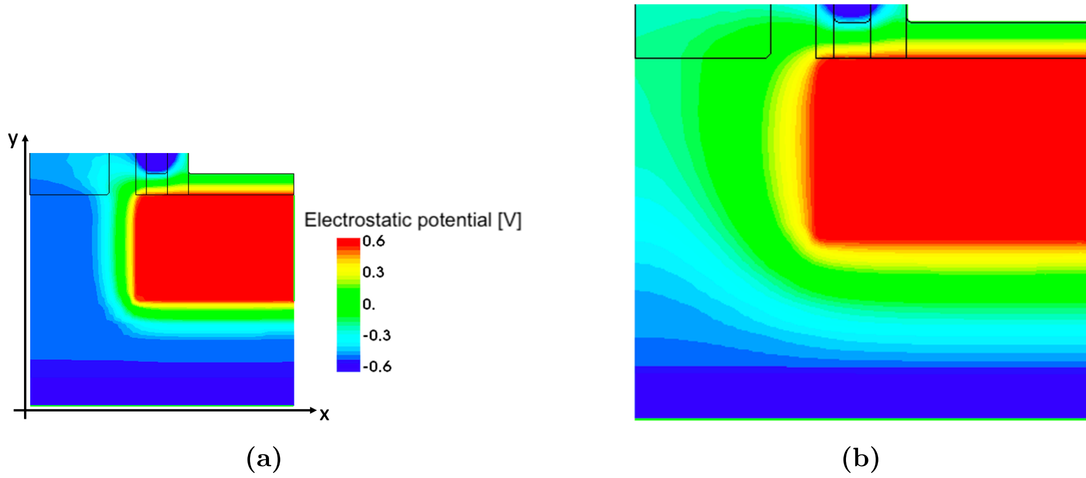


Figure 4.15: Electrostatic potential profile of the substrate for (a) $N_a^{sub} = 5 \cdot 10^{18}[cm^{-3}]$ and (b) $N_a^{sub} = 1 \cdot 10^{18}[cm^{-3}]$.

the boundary of the source is aligned with the silicon-oxide interface, is a good solution for what concerns the correct behavior of the string. Regarding the supply of holes, N_a^{sub} must be not too small due to the presence of the depletion region established between the substrate and the source well. The less N_a^{sub} the more extended will be the depleted region. If the depleted region is too wide it goes to cover the entrance of the channel such that the holes, in order to get into the channel, have to cope a potential barrier created by the opposite electric field of the p-n junction. As represented in Fig. 4.15a, with $N_a^{sub} = 5 \cdot 10^{18}$ cm^{-3}, the drop of the electrostatic potential, related to the depletion region, ends very close to the channel. So, the holes have to travel a small path with an opposite electric field and their kinetic energy allows them to enter the channel. Instead, Fig. 4.15b reports the electrostatic potential for the case of $N_a^{sub} = 1 \cdot 10^{18}$ cm^{-3}. We can see from the picture that the drops of the potential covers a wider zone. This means that there is a higher electric field at the entrance of the channel and the positive carriers cannot proceed towards the middle of the string.

4.4 DCP results for a completely programmed TCAT structure

Fig. 4.16 shows the comparison of three longitudinal sections of the conduction band energy for a string with all the cells programmed. Note that here we are showing the entire string and not just half of it, as source and drain junctions are not equal and the structure is not strictly symmetrical. The solid blue line depicts the CB energy at the end of the falling edge of $V_W L$, while the red one shows the conduction band at $t = 5$ μs after

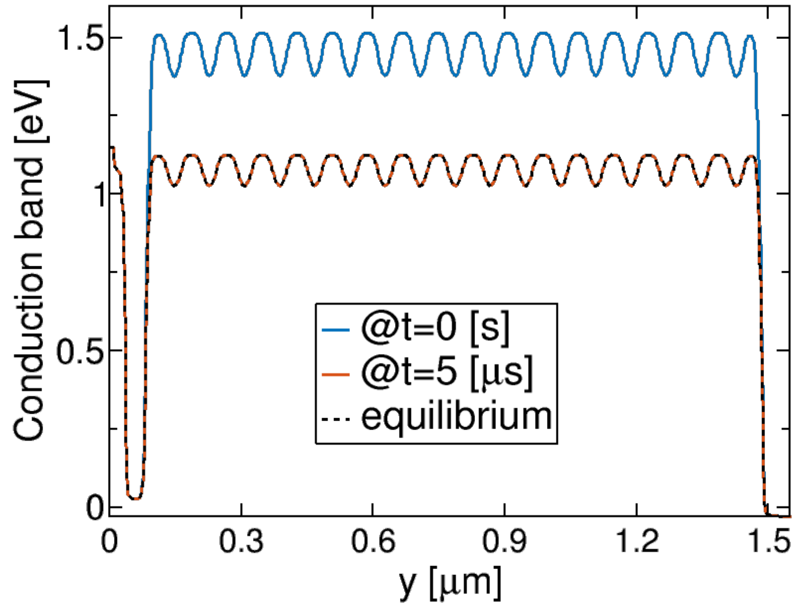


Figure 4.16: Longitudinal section of the CB energy at different time instants. The cut is of taken at 0.25 nm from the silicon-oxide interface. The dashed line shows the equilibrium value.

the end of the verify phase, finally the dashed black line represents the CB energy at the equilibrium for the same memory configuration. We have to specify what the minimum of the CB energy curve, in the left part of the plot, represents. The longitudinal section was taken very close to the silicon-oxide interface and the cut begins at the BODY contact. So in the plot is depicted also the CB energy of the substrate. We have to remind that the source region is extended till the beginning of the ONO stack, as depicted in Fig. 4.14. We know that there is a p-n junction between the source and the substrate and the drop of the energy before $y = 0.1\ \mu m$ is due to the presence of the junction.

Focusing on the rest of the plot, we can notice that the DCP is almost not present for this TCAT configuration. This is due to the fact that the holes can enter into the channel in a huge number. We have to keep in mind that, even if the channel potential is almost at the equilibrium already at the end of the V_{WL} falling edge, this does not mean that the DCP did not take place at all. Like the BiCS structure, the channel enters in a floating state as soon as the V_{WLs} becomes lower than the threshold voltage, but in the present case the system can go back to the equilibrium within a small time interval after the end of the CG bias transition. This is possible thanks to the holes coming from the BODY contact such that the hole concentration inside the string reaches very quickly high enough values to affect the electrostatic of the system.

Fig. 4.17a shows the holes current density at $t = 1.5\ \mu s$ before the end of the falling edge of the V_{WLs}. The picture shows that a huge number of holes moves in the substrate region toward the channel. Holes can enter into the channel at the side of the macaroni

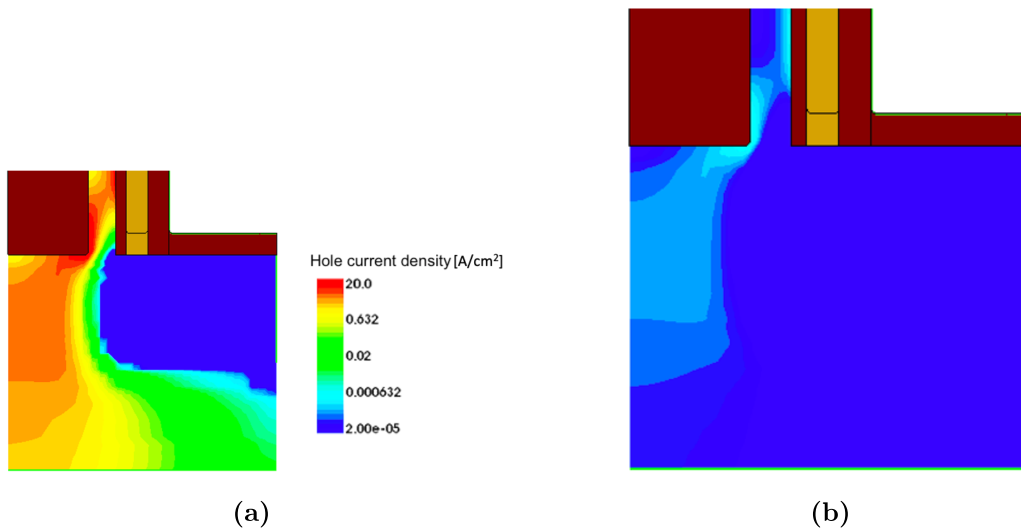


Figure 4.17: holes current density profile at (a) $t = 1.5[\mu s]$ before the end of the falling edge and (b) $t = 10[\mu s]$ after the falling edge of WLs.

where the electric field established by the p-n junction is lower such that the positive carriers have enough kinetic energy to overcome the energy barrier and are able to start accumulating in correspondence of the gates region. On the other hand, Fig. 4.17b shows that, already at $t = 10$ μs after the end of the verify phase, the holes current density is very low. This current density is not responsible of some changes in the energy of the system but it is just due to the displacement of the carriers under the presence of electric field established at the junction. Indeed the channel has already reached the equilibrium as we can see in Fig. 4.18. This picture reports the time evolution of the CB energy in the channel in a point close to the silicon-oxide interface. In particular, the energy is taken underneath WL^{15}, the last WL before the DSL. We have made this choice because in this structure, the middle of the string is no longer the last region to be filled with holes. In a TCAT structure the holes enter in the channel at the SSL side and reach the DSL before the holes injected by the BL could become relevant for the electrostatic of the system. This means that in the present case, the growth of the hole concentration in each CGZ takes place sequentially for all the WL until WL^{15}. From Fig. 4.18 we can see that the system reaches the equilibrium within $t = 5$ μs. This is a good result by considering the program operation, in fact the time constant of the phenomenon is smaller than the time interval that separates the end of the verify operation and the beginning of the program operation. In the end, we can conclude that for the TCAT structure, in case of a uniformly programmed string, the DCP does not produce any kind of program disturb.

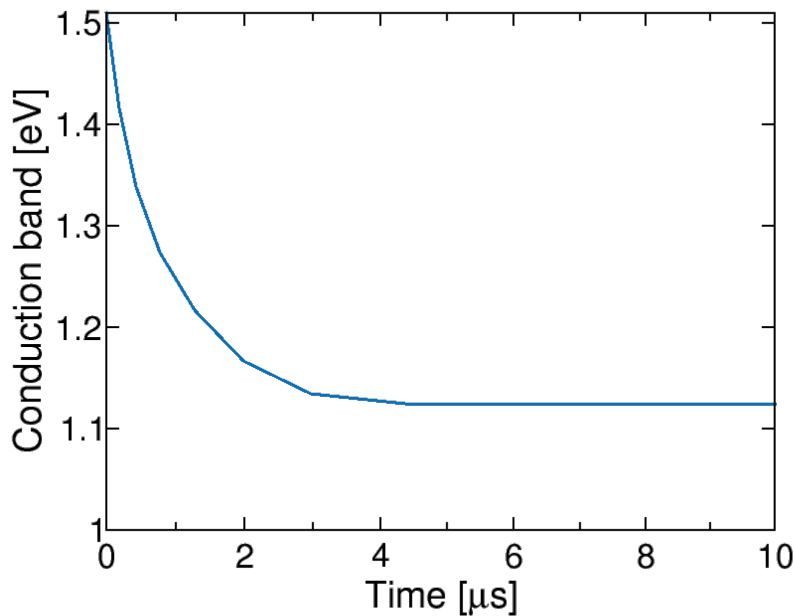


Figure 4.18: Time evolution of the conduction band energy of the channel underneath WL^{15}. The time origin corresponds to the end of the falling edge of the verify operation.

4.5 DCP results for a not completely programmed TCAT structure

According to what we have said up to now, the TCAT structure seems significantly better than the BICS structure from the viewpoint of the DCP. However, this is only partially true, in fact there are situations in which not even the TCAT memory is able to go back to a stationary condition before the program operation begins. The dominant factor about the return to the equilibrium of the system is the presence of erased cells. Indeed, underneath an erased cell, the channel is in an inversion condition also when no bias voltage is applied at the control gates. The presence of this charge prevent the normal flux of holes from substrate towards the WLs with higher index than the index of the erased cell, reminding that the indexing increases going from the SSL to the DSL, see Fig. 4.14. In this way, the part of the channel between the erased cell and the BL does not have a holes supply and it cannot restore the equilibrium quickly. With this geometry, the DCP phenomenon becomes dependent on the presence of erased cells along the string. Due to the fact that the erased cell in part block the holes flux, the return to the equilibrium is a bit different with respect to the BiCS case. In this simulation we have defined the program level as $P3$ for all the cells, included SSL e DSL, except for WL^8 whose level is E. In this way we expect that the system can go back to the stationary condition in the part of the string connected to the body, while the other part remains in a floating condition due to the presence of the erased WL^8.

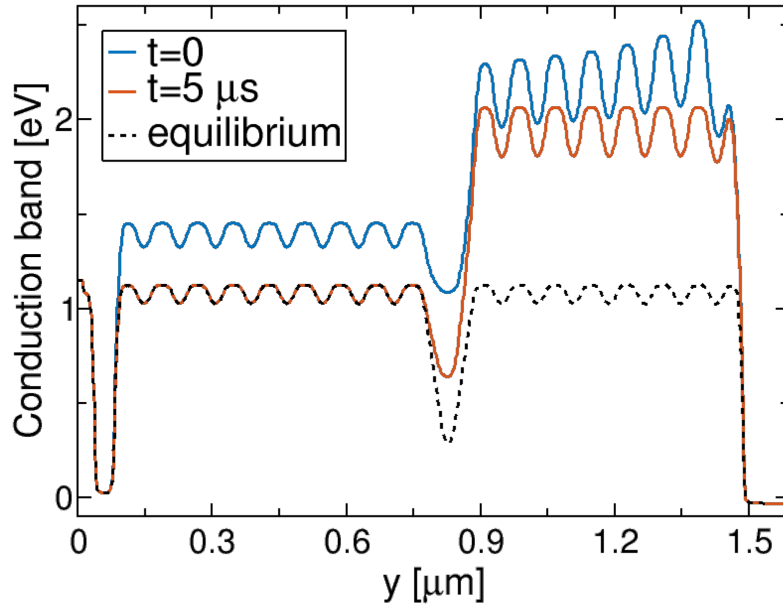


Figure 4.19: Comparison of the CB energy as a function of the position for different time instants. The longitudinal section is taken at 0.25 *nm* from the silicon-oxide interface in a string where WL^8 is erased. The dashed line shows the equilibrium value.

Fig. 4.19 shows the longitudinal section of the conduction band energy taken along a cut close to the silicon-oxide interface. The different curves represents the CB energy at different times, while the dashed line shows the equilibrium value. We can see in the picture that at the centre of the string there is a depression in the CB energy. That valley is due to the presence of positive charge trapped in WL^8 that increases the electrostatic potential of the channel. For simplicity, we have called *source-line side (SLS)* region the channel region underneath the set of cells between the central cell and the SL the. While we have given the name of *bit-line side (BLS)* region to the other part of the channel between the WL^8 and the BL. We can notice in Fig. 4.19 that it is in agreement to what we have supposed before, i.e. the SLS region behaves like the case of uniform programmed string because we can see that the CB energy move from the blue curve to the red curve and this one is superposed to the curve describing the stationary condition. The only difference that we could expect with respect to the uniform programmed string lies in the time constant. In fact, we could suppose time constant to be a bit smaller than the result of the completely programmed TCAT string, because in the present case we are considering just half of the WLs. On the other hand, the BLS region is in a floating state as we have supposed. Although we could think that these two channel regions are not related for what concerns the return to the stationary condition, actually they are not completely independent and the BLS region does not follow exactly the transient that we have seen for the BiCS structure.

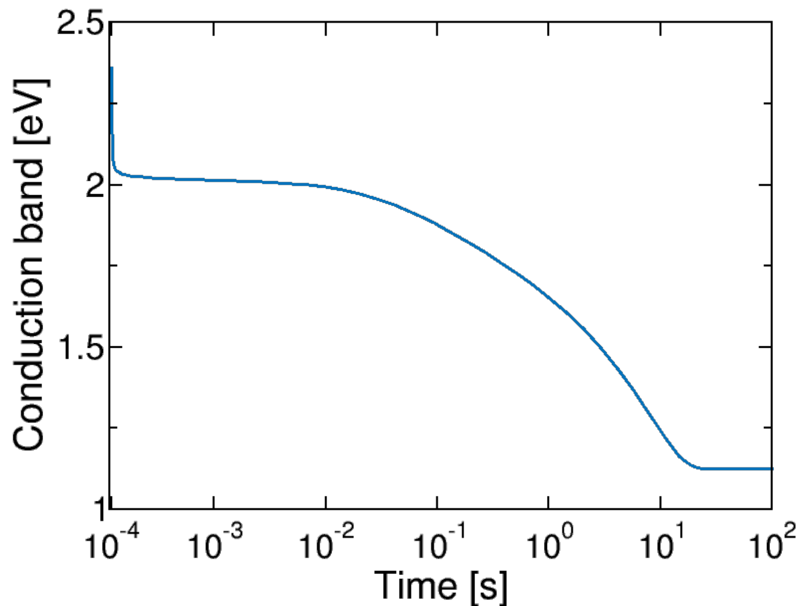


Figure 4.20: Time evolution of the conduction band energy in a point of the channel underneath WL^{12} that is the middle cell in the BLS region.

From this point on we will analyze the transient of the BLS region in order to find out if there are some differences with respect to the case of the BiCS structure. Fig. 4.20 shows the time evolution of the CB energy in a point of the channel underneath WL^{12} that is the middle cell in the BLS region. We can see that, at short times, the conduction band has a sharp decrease of about $0.5\ eV$, then it changes its trend and drops more slowly until $t \approx 20\ s$. In the last tens of second the curve remains constant underling that the system has reached the equilibrium condition. In order to explain the trend of the energy we have analyzed the carriers concentration inside the channel.

4.5.1 Carriers concentration analysis in a not completely programmed TCAT structure

In order to understand how the BLS region could go back to the stationary condition it is important pointing out a particular effect related to the other part of the channel. Indeed, during the return to the equilibrium of the SLS region, a certain amount of holes can overcome the potential barrier created by the erased cell and flow into the floating part. The amount of positive carriers that is able to enter in the floating part of the channel depends on how many erased cells build up the potential barrier. Indeed, the higher the number of erased cell the greater the electron accumulation underneath those cells and, in turn, the greater the zone where the recombination takes place. The supposition we have just done is confirmed by Fig. 4.21 that shows the longitudinal section of the hole

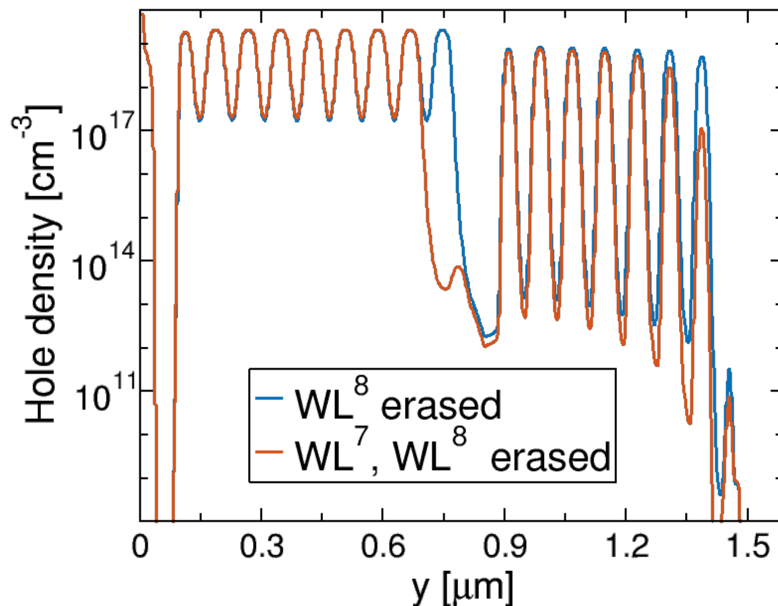


Figure 4.21: Representation of the hole density as a function of the longitudinal coordinate. The cut is taken at 0.25 nm from the silicon-oxide interface and at $t = 0$ s. The blue curve it is related to a string where just WL^7 is ERASED, the red curve represents the hole density section for a string with WL^7 and WL^8 ERASED.

density at $t = 0$ s for two different configuration of the string. The blue curve corresponds to the configuration with which we have dealt so far, i.e. all the cells are programmed $P3$ except for WL^8 that is erased. The red curve, instead, corresponds the hole density in a string where WL^7 and WL^8 have a E level, while the other cells are programmed $P3$. We can see in the left part of the plot, a huge drop of the density for both the curve that is once again due to the nearness of the cut to the source region. We notice that at the BLS region the hole density is a bit smaller. This is due to the greater loss of carriers that occurs underneath the erased cells for the SRH recombination. Furthermore, it is also important pay attention to the fact that the amount of holes in the floating part is rather high for both the curves.

Returning on focusing upon the structure with just WL^8 erased, we can say that the hole concentration has already reached a sufficient high value that a further change of the density affects significantly the energy of the channel. Actually the hole concentration is rather high also in the spacer zone between the WL^{15} and the DSL. Fig. 4.22 shows the time evolution of the positive carriers concentration in the channel region underneath the closest spacer to the BL. With respect to the Fig. 4.12a, we can notice that the concentration is much higher already at small instants such that the generation taking place at the boundary of the channel has already decreased. Indeed, the effect that we have described for the red curve of Fig. 4.12b, here it is already occurred after $t = 5$ μs. In order to correctly understand the time evolution of the hole density in the channel,

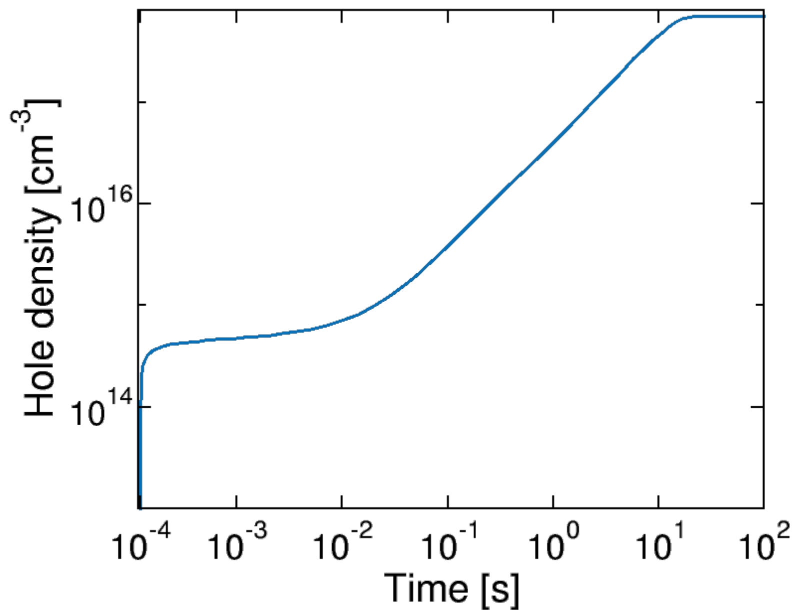


Figure 4.22: Time evolution of the hole density underneath the spacer between the DSL and WL^{15}.

also for this case we have analyzed the current flowing through the string.

4.5.2 Current time evolution in a not completely programmed TCAT structure

Fig. 4.23 shows the comparison between the electron current at the BL and the calculated I_{SRH} and I_{B2BT}. Moreover, in green is reported I_p^{BODY} that is the holes current coming from the body and that reaches the BLS region overcoming the potential barrier at WL^8. We have to specify that I_p^{BODY} is not exactly the current that flows at the contact but, to that value, it was subtracted the integral of the SRH recombination rate that takes place in the SLS region. In the brief time interval after the falling edge of the V_{WL}, we can see a peculiarity of I_n^{BL} with respect to what we have seen in Fig. 4.23. From the picture we notice that the electron current heads towards the channel instead of pointing in the opposite direction. This fact can be explained if we consider the blue curve of Fig. 4.19 where we can see that the energy is higher in the outer part of the BLS region, while it is lower close the erased WL. Considering that the hole concentration is rather high since the beginning of the transient, we conclude that the holes already produce a good screening of the field coming from the stored charge such that the electrons can overcome the potential barriers and have a sufficient mobility to move. But, since the electrostatic potential is higher close the erased WL, the electrons in the BLS region tend to move towards the SL instead to flow outside the channel into the BL. For this

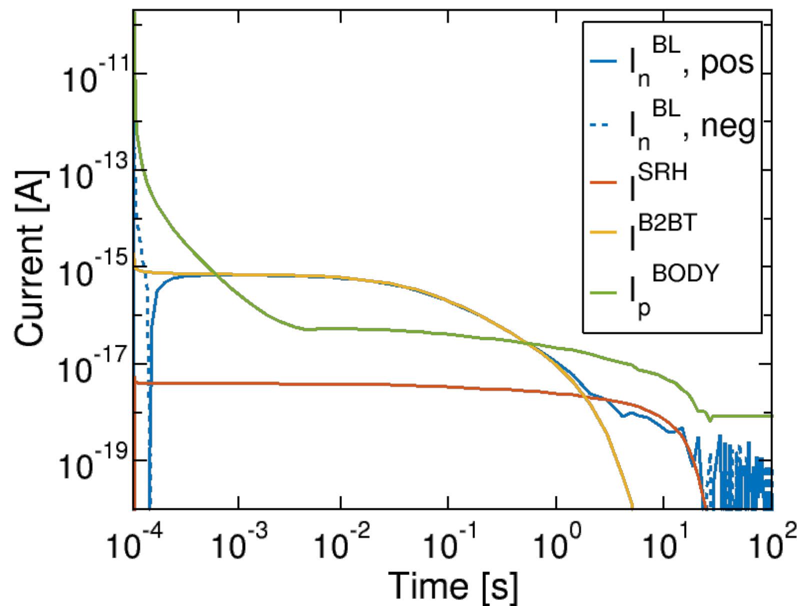


Figure 4.23: Comparison of the electron flow at the BL (blue curve) with respect to the currents calculate by the SRH and B2BT generation rate (red and yellow curves, respectively) and the hole current coming from the BODY contact (green curve).

reason at the beginning we see an opposite sign of the electron current at the BL. This current is reversed when the electron concentration becomes small because the negative carrier move towards the SL. This drop of the electron concentration also coincide with a uniform arrangement of the holes in the BLS region. This take arrangement place few μs after the end of the verify phase, as we can state by looking at the red curve of Fig. 4.19 where we can see that the CB energy in the BLS region is not tilted. This arrangement of the holes takes place because the amount of positive carriers coming from the SLS region decreases quickly after the end of the falling edge and the incoming concentration is no longer sufficient to determine a variation of the potential. In this way, the holes already present in the BLS re-arrange and the CB energy becomes flat.

This behavior is confirmed by the green curve of Fig. 4.23 that denotes how, at the beginning of the transient, the hole flux reaching the BLS region, overcoming the potential barrier, is rather big. Next, we notice that this current drops down quickly because in few instants the potential barrier present at the CGZ of WL^8 barrier becomes much higher and the number of positive carriers, that can pass over it, decreases. Fig. 4.24 shows the temporal evolution of the barrier height that we have calculated as the difference between the CB energy in CGZ correspondent to WL^7 and the CB energy in the CGZ correspondent to the erased WL^8. In order to not produce misunderstandings, we remind that the channel region underneath WL^7 belongs to the channel part that is returned to the equilibrium. We can notice that the curve in Fig. 4.24 and the green curve of

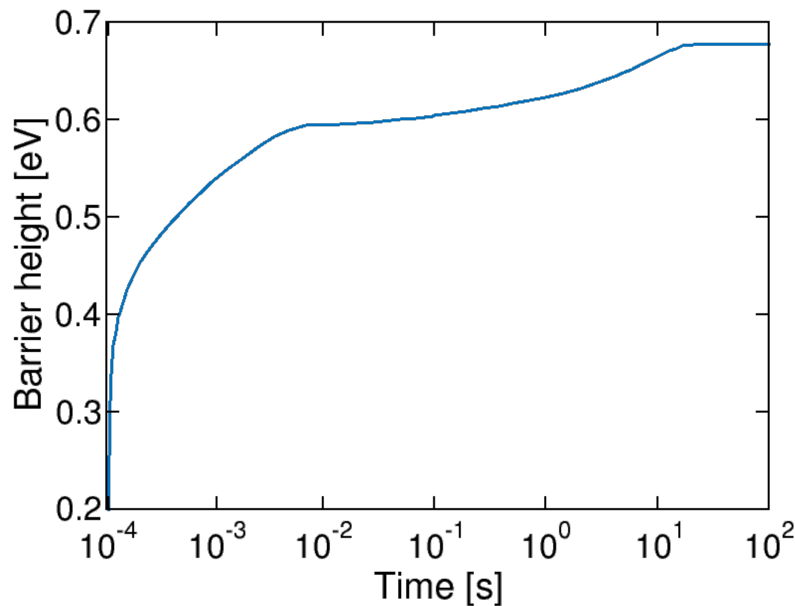


Figure 4.24: Time evolution of the barrier height present in the channel position correspondent to WL^8.

Fig. 4.23 have a similar temporal evolution. In fact, there is an exponential relation that relates this two quantity. As we have said, at the beginning of the transient the barrier quickly becomes higher because the accumulated electrons tend to both flow towards the SL and recombine with the holes coming from the BODY. For this reason, the electric field produced by the positive stored charge is less screened with the passing of time and the electrostatic potential increases. The barrier height reaches a rather high value in less than 1 ms such that the I_p^{BODY} that flows into the BLS region becomes lower than I_{B2BT}. At this point, the transient becomes quite similar to the case seen for the BiCS where the tunneling is dominant for the return to the equilibrium. Even though it is a bit different for what concerns the current coming from the SRH rate. If we focus on the red curve, that represents I_{SRH}, we see that for the whole transient there is a net generation. This is quite different with respect to what we have seen for the BiCS structure. In fact, in this case, I_{SRH} it is constant for the whole transition without showing neither an initial period with net recombination nor a decrease in the rate as reported in Fig. 4.10. For what concerns the first difference, we have already specified that also at small times the electron concentration is small in the floating region. This implies that in the system, that is not in an accumulation condition, a net generation is present from the beginning of the transient. For the second difference, we have seen in Fig. 4.22 that the hole density has reached from the beginning a high value also in the SZ between WL^{15} and DSL and what we have seen in Fig. 4.12b, in this case happens almost instantly. At long times, the electric field at the junction with the BL decreases and I_{B2BT} does the same. At this

point, it is not the generation current that brings the BLS region to the equilibrium but it is the current of holes coming from the body. In fact, even if the barrier potential is rather high, a small amount of holes can pass the barrier and this amount is larger than the generated one.

It is difficult to make a statement about the time needed from the TCAT system to neutralize the DCP. In fact, it depends by the number of the erased cell or even by the position of the erased cell. At the moment we can say that, due to the randomness of the levels of the cells, it is not predictable to know if the DCP will take place or not and if it will return to the stationary condition in a sufficient short time. For this reason, at the end we cannot say that the TCAT structure leads to an improvement of the memory characteristics to counter the DCP.

4.6 Conclusions

In this chapter, as first thing we have studied the down-coupling phenomenon in a realistic and completely programmed BiCS structure. In particular we have seen that this structure have not a sufficient supply of holes because both the generation for SRH and the generation for tunneling are too small to increase the electrostatic potential. Moreover, due to the non-uniform presence of stored charge in the trapping layer, that produces a non-uniform electric field in the channel, the electrons are stacked inside the channel and they need a relative long time to go outside the string. So, we have seen that the DCP actually is a possible source of program disturbs in BiCS memory. Then, we have studied the DCP in the TCAT structure, hoping that the presence of the p-doped substrate could bring a solution to this problem. At first we have seen that, for the same cells configuration of the string as in the BiCS case, i.e. all the cells programmed, the TCAT structure actually is able to fill the channel of holes within a short times. Such time interval is sufficiently small that the channel potential can be brought to the stationary condition before the beginning of the program phase. Even though, if there is even just one erased level, there is no the certainty that this structure is able to restore the correct condition of the channel. For this reason, we can say that the TCAT structure is not a reliable solution to suppress the disturbs related to the DCP, but further efforts are necessary in order to find out some ways to solve this problem.

Conclusions

In this work we have presented the first thorough analysis of the *Down-Coupling Phenomenon* and its impact on the reliable operation of realistic 3D NAND Flash structures. The aim of the work was to employ numerical simulations to understand what are the main physical effects that play a role in restoring the equilibrium potential of the NAND string once it was broken due to the DCP. The work started with the identification of the essential physical models, like electron/hole generation by tunneling or SRH effect, needed to obtain a realistic description of the phenomenon. In implementing these models, we have taken into account the fact that we are dealing with devices whose conductive channel is made of polysilicon instead of crystalline silicon.

Preliminary work was devoted to the analysis of the relation between the charge stored in the cells and the threshold voltage of the string, to calibrate the cell V_T condition. Then, the analysis was carried out first in a simplified string structure, having only one cell (i.e., one control gate), but with the same size as a full string, set to a high program level. The analysis confirmed the importance of the DCP, and showed that the main mechanism responsible for the return to the equilibrium condition is *Band-to-band Tunneling.* However, the results clearly show that the time needed to restore equilibrium is unacceptably long, because such current exponentially decreases as the string bias is lowered. A simplified analytical model has been also developed to describe the DCP transient.

We then moved to the analysis of a realistic BiCS structure, where several control gates separated by oxide spacers are placed along the string. As for the case of the single-cell structure, the problem of restoring the correct channel potential is determined by the too slow injection of holes. However, the physics here is complicated by the existence of the cells: their charge affects the channel regions underneath the oxide spacers, slowing down the outflow of the electrons from the channel. This results in a process in which injected holes accumulate sequentially underneath each gate, screening the electric field of the stored charge and leading to the release of the electrons.

Finally, in order to solve the problem of the too slow return to the equilibrium, we moved to the analysis of a TCAT structure, where an appropriate *p*-doping of the substrate provides a hole reservoir. Our analysis shows that now the string can restore the

equilibrium in few microseconds, a time interval sufficiently small to avoid errors during the program phase. Despite of this, the DCP in this structure can still occur if along the string there are some erased cells that block the flow of electrons. In fact, erased cells remain in inversion during the transient, blocking the passage of holes. This condition has been verified and the detailed physical mechanisms analyzed.

At the end of this analysis, we can conclude that both string structures currently adopted in 3D NAND Flash suffer from DCP. The BiCS structure is intrinsically affected by the DCP every time the program level of the cells demand for the accumulation of holes. The TCAT structure, usually considered to be superior in this respect, is also not immune, as the hole flux can be blocked by the presence of erased cells. For both cases, we have quantitatively investigated the DCP transients and assessed the role of the different physical processes. The present work opens new perspectives for the evaluation of the impact of the DCP on cell reliability and for the tailoring of improved string designs for 3D NAND Flash.

Bibliography

[1] C. Monzio Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaita, and A. Visconti, "Reviewing the evolution of the nand flash technology," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1609–1633, 2017. 7, 8, 9, 16, 17, 19

[2] C. M. Compagnoni, A. S. Spinelli, R. Gusmeroli, A. L. Lacaita, S. Beltrami, A. Ghetti, and A. Visconti, "First evidence for injection statistics accuracy limitations in nand flash constant-current fowler-nordheim programming," in *2007 IEEE International Electron Devices Meeting*, pp. 165–168, 2007. 10

[3] D. Ielmini, A. Spinelli, A. Lacaita, and A. Modelli, "A new two-trap tunneling model for the anomalous stress-induced leakage current (SILC) in Flash memories," *Microelectronic engineering*, vol. 59, no. 1, pp. 189–195, 2001. 12

[4] C. Monzio Compagnoni and A. S. Spinelli, "Reliability of nand flash arrays: A review of what the 2-d–to–3-d transition meant," *IEEE Transactions on Electron Devices*, vol. 66, no. 11, pp. 4504–4516, 2019. 13

[5] A. Ghetti, C. Monzio Compagnoni, A. S. Spinelli, and A. Visconti, "Comprehensive analysis of random telegraph noise instability and its scaling in deca–nanometer flash memories," *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1746–1752, 2009. 14

[6] Y. Nishi, *Advances in Non-volatile Memory and Storage Technology*, ch. Developments in 3D-NAND Flash technology. Woodhead Publishing Series in Electronic and Optical Materials, Elsevier Science, 2014. 17

[7] J. Jang, H. Kim, W. Cho, H. Cho, Jinho Kim, S. I. Shim, Younggoan, J. Jeong, B. Son, D. W. Kim, Kihyun, J. Shim, J. S. Lim, K. Kim, S. Y. Yi, J. Lim, D. Chung, H. Moon, Sungmin Hwang, J. Lee, Y. Son, U. Chung, and W. Lee, "Vertical cell array using tcat(terabit cell array transistor) technology for ultra high density nand flash memory," in *2009 Symposium on VLSI Technology*, pp. 192–193, 2009. 18

[8] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on mosfet scaling," in *1987 International Electron Devices Meeting*, pp. 718–721, 1987. 22, 23

[9] Y. Fukuzumi, R. Katsumata, M. Kito, M. Kido, M. Sato, H. Tanaka, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory," in *2007 IEEE International Electron Devices Meeting*, pp. 449–452, 2007. 24

[10] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of nand flash memories: Planar cells and emerging issues in 3d devices," *Computers*, vol. 6, no. 2, 2017. 25, 26

[11] Tae-Sung Jung, Young-Joon Choi, Kang-Deog Suh, Byung-Hoon Suh, Jin-Ki Kim, Young-Ho Lim, Yong-Nam Koh, Jong-Wook Park, Ki-Jong Lee, Jung-Hoon Park, Kee-Tae Park, Jhang-Rae Kim, Jeong-Hyong Yi, and Hyung-Kyu Lim, "A 117-mm/sup 2/ 3.3-v only 128-mb multilevel nand flash memory for mass storage applications," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1575–1583, 1996. 27, 28

[12] M. Kang and Y. Kim, "Natural local self-boosting effect in 3d nand flash memory," *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1236–1239, 2017. 27, 28

[13] Y. Kim and M. Kang, "Down-coupling phenomenon of floating channel in 3d nand flash memory," *IEEE Electron Device Letters*, vol. 37, no. 12, pp. 1566–1569, 2016. 30, 31

[14] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-doped silicon," *IEEE Transactions on Electron Devices*, vol. 30, no. 7, pp. 764–769, 1983. 36

[15] C. Sah, T. Ning, and L. Tschopp, "The scattering of electrons by surface oxide charges and by lattice vibrations at the silicon-silicon dioxide interface," *Surface Science*, vol. 32, no. 3, pp. 561 – 575, 1972. 37

[16] A. Hartstein, T. Ning, and A. Fowler, "Electron scattering in silicon inversion layers by oxide and surface roughness," *Surface Science*, vol. 58, no. 1, pp. 178 – 181, 1976. 37

[17] A. Mannara, A. S. Spinelli, A. L. Lacaita, and C. Monzio Compagnoni, "Current transport in polysilicon-channel gaa mosfets: A modeling perspective," in *ESSDERC*

2019 - 49th European Solid-State Device Research Conference (ESSDERC), pp. 222–225, 2019. 38

[18] D. M. Kim, A. N. Khondker, S. S. Ahmed, and R. R. Shah, "Theory of conduction in polysilicon: Drift-diffusion approach in crystalline-amorphous-crystalline semiconductor system—part i: Small signal theory," *IEEE Transactions on Electron Devices*, vol. 31, no. 4, pp. 480–493, 1984. 38

[19] N. C. . Lu, L. Gerzberg, Chih-Yuan Lu, and J. D. Meindl, "Modeling and optimization of monolithic polycrystalline silicon resistors," *IEEE Transactions on Electron Devices*, vol. 28, no. 7, pp. 818–830, 1981. 38

[20] G. Malavena, A. L. Lacaita, A. S. Spinelli, and C. Monzio Compagnoni, "Investigation and compact modeling of the time dynamics of the gidl-assisted increase of the string potential in 3-d nand flash arrays," *IEEE Transactions on Electron Devices*, vol. 65, no. 7, pp. 2804–2811, 2018. 55, 56

[21] Jibin Zou, Qiumin Xu, Jieying Luo, Runsheng Wang, Ru Huang, and Yangyuan Wang, "Predictive 3-d modeling of parasitic gate capacitance in gate-all-around cylindrical silicon nanowire mosfets," *IEEE Transactions on Electron Devices*, vol. 58, no. 10, pp. 3379–3387, 2011. 55

[22] J. Zou, Q. Xu, J. Luo, R. Wang, R. Huang, and Y. Wang, "Corrections to "predictive 3-d modeling of parasitic gate capacitance in gate-all-around cylindrical silicon nanowire mosfets" [oct 11 3379-3387]," *IEEE Transactions on Electron Devices*, vol. 59, no. 3, pp. 867–867, 2012. 55

[23] A. Bansal, B. C. Paul, and K. Roy, "An analytical fringe capacitance model for interconnects using conformal mapping," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 12, pp. 2765–2774, 2006. 64