Reliability Study of the Floating Gate Based Embedded Non-Volatile Memory (eNVM)

Author: Jindong Yang
NXP Supervisor: Dr. Ir. Guoqiao Tao
TU Delft Supervisor: Prof. Lis Nanver, TU Delft
Date: 01-04-2010

CONTENT

ACKNOWLEDGMENTS 6

1. INTRODUCTION 7

1.1 Working Environment 7

1.1.1 NXP Semiconductors 7
1.1.2 NXP Semiconductors, Nijmegen 7
1.1.3 Innovation department 8

1.2 Thesis task 8

1.3 Thesis outline 8

1.4 Keywords 9

2. OVERVIEW OF NON-VOLATILE MEMORIES 10

2.1 Introduction to semiconductor memories 10

2.2 Flash/EEPROM 11

2.3 Cell Structure 12

2.3.1 NXP eNVM device structure 12
2.3.2 2T-FN/FN-NOR cell 13

2.4 Principles 13

2.4.1 V_T change 13
2.4.2 P/E operation 15
2.4.3 Read operation 17

2.5 Cell array 18

2.5.1 Array structure 18
2.5.2 Array operation 19

2.6 Reliability aspects 19

2.6.1 Endurance 20
2.6.2 Retention 22
2.6.3 Disturbs 26

2.7 Summary 26

3. CHARACTERISTICS MEASUREMENT 28

3.1 Introduction 28

3.2 Folwer-Nodheim tunneling 28

3.3 V_T shift speed 29

3.4 Endurance 31

3.4.1 Cycling on RF 31

3.4.2 Cycling on FG 35

3.5 Data retention 38

3.6 Conclusion 39

4. COUPLING RATIO 40

4.1 Introduction 40

4.2 Methods 40

4.2.1 Gm method 41
4.2.2 S (Subthreshold) Method 42
4.2.3 Bulk bias (V_b) method 42
4.2.4 Matching factor 43

4.3 Measurements 44

4.3.1 Gm method 45
4.3.2 S method 45
4.3.3 Bulk bias method 46

4.4 Calculation 47

4.4.1 Analytic model 47
4.4.2 FEM Simulation 50

4.5 Aarray tests 53

4.5.1 Capacitor charging 53
4.5.2 Gm Method 55

4.6 Conclusion 56

4.6.1 Measured & simulated results 56
4.6.2 Analysis & discussions 56

4.7 Conclusion 58

5. V_T INSTABILITY 59

5.1 Introduction 59

5.2 Theory 60

5.2.1 V_T shift 60
5.2.2 Voltage partition 61

5.3 Electron displacement 62

5.4 Trap locations 64

5.4.1 Test on floating gate cells 64
5.4.2 Test on ONO layers 65

5.5 2.7Mbit test structure 67

5.6 V_T instability dependences 71

5.6.1 Electric field (E) 71
5.6.2 Temperature (T) 72
5.6.3 Pre-stress time (t) 74

5.6.4 Pre-stress voltage (V) 75
5.6.5 Conclusion 76
5.7 Multi-bit 77
5.7.1 Multi-bit capability[5.6] 77
5.7.2 V_T instability on multi-bits 78
5.8 Conclusion 78

6. CONCLUSION 80

6.1 Summary 80
6.2 Future work 80
6.3 Publication 80

REFERENCES 81

Acknowledgments

I would like to thank my supervisor, Dr. Guoqiao Tao, who is senior principal, domain leader of NVM Reliability Wafer Technology and Foundary organization (WT&FO)/Operations in NXP Semiconductors, Nijmegen. Dr. Tao gave me tremendous supports and precious advices on this master thesis study. I benefited a lot from his continuous encouragement and help. I am impressed by his diligence and enthusiasm on the scientific pursuit. I also appreciate the talks with him on morals of being a man. I feel very lucky to have him as my research supervisor on master thesis. His broad view of this world, way of thinking and behavior is admired.
I am also grateful to Professor Lis Nanver, for her sincere and generous helps on many issues. Lis Nanver is the best expert on semiconductor device research within TU Delft, and has a very distinguished reputation within the field. I am impressed by her hard work and actue insight. I also enjoyed following her at the first stage of the master project.
I would also like to thank Dr. Rene Van Swaaij for the help on course credits and thesis defense arrangement, and appreciate the interesting conversations with him.
I would like to thank Professor Guoping Ru in Fudan University, Shanghai, China for the discussion and collaboration, and providing this opportunity to study in the Netherlands.
I am grateful to my fellow Ngana Paul, Aftatb Qureshi, Andre Moreira for the friendly working environment in the office, wish them nothing but the best.
And I want to give my loyal gratitude and love to my girlfriend, she always accompany with me, especially during my great distress. I will love her forever.
Finally I would like to thank my family their love and continuous support, encouragement, especially to my father for the previous and helpful guidance during my growth. Without him, I could not have any achievement. Wish my father happy forever.

1. INTRODUCTION

This chapter gives a brief introduction to NXP Semiconductors and its manufacturing site in Nijmegen, the Netherlands. The major task and outline of this thesis is also followed.

1.1 Working Environment

1.1.1 NXP Semiconductors

This master thesis was written during my internship, at the reliability group of the innovation department, in NXP Semiconductors, Nijmegen, the Netherlands.
NXP Semiconductors was established in 2006 (formerly a division of Royal Philips Electronics), with more than 50 years of experience in semiconductors, and also one of the largest global electronics companies. Its business is divided into 5 units: automotive, high performance mixed signal, identification, standard products, and NXP software, which are used in a wide range of electronic equipment such as TV, mobile phone, car, identification applications, and so on. The company has approximately 29,000 employees distributed in more than 30 countries, and it has more than 35 direct customers including Nokia, Continental, Bosch, Philips, Samsung, Panasonic, Sony, Huawei, Dell and Asus. In 2008, the net sale is 5.4 billion US dollars.
NXP Semiconductors has 4,300 engineers, 20 R&D centers located in 14 countries, invests $1.2 billion in 2008 on R&D, holds more than 5400 patent families, owns 7 part-time professors, supports $3 million a year university programs, 20 PHD programs and 16 master courses annually.

1.1.2 NXP Semiconductors, Nijmegen

NXP Semiconductors, Nijmegen is the largest semiconductor manufacturing plant of NXP, responsible for 30% of NXP chip production. Manufacturing in Nijmegen is done in four wafer foundries named after the wafer size (in inches, 1 inch=25.4mm), namely ICN4, ICN5, ICN6, ICN8 (Innovation Center Nijmegen). In 1953 when it was found, the diameter of a wafer was 10mm, which has been ever-increasing. In 2009 the diameter of a wafer in ICN8 is 200mm, with a capability down to 0.13 μm CMOS technology, ICN8 is one of the most advanced mega wafer fabs within NXP Semiconductors. The current production of Nijmegen site is 1.3 million 6-inch or 725,000 8-inch equivalent wafers a year.
NXP Semiconductors, Nijmegen employs 4400 people from 60 different countries, 2700 working in wafer foundries, and the rest in innovation, marketing, logistics, management, and so on.

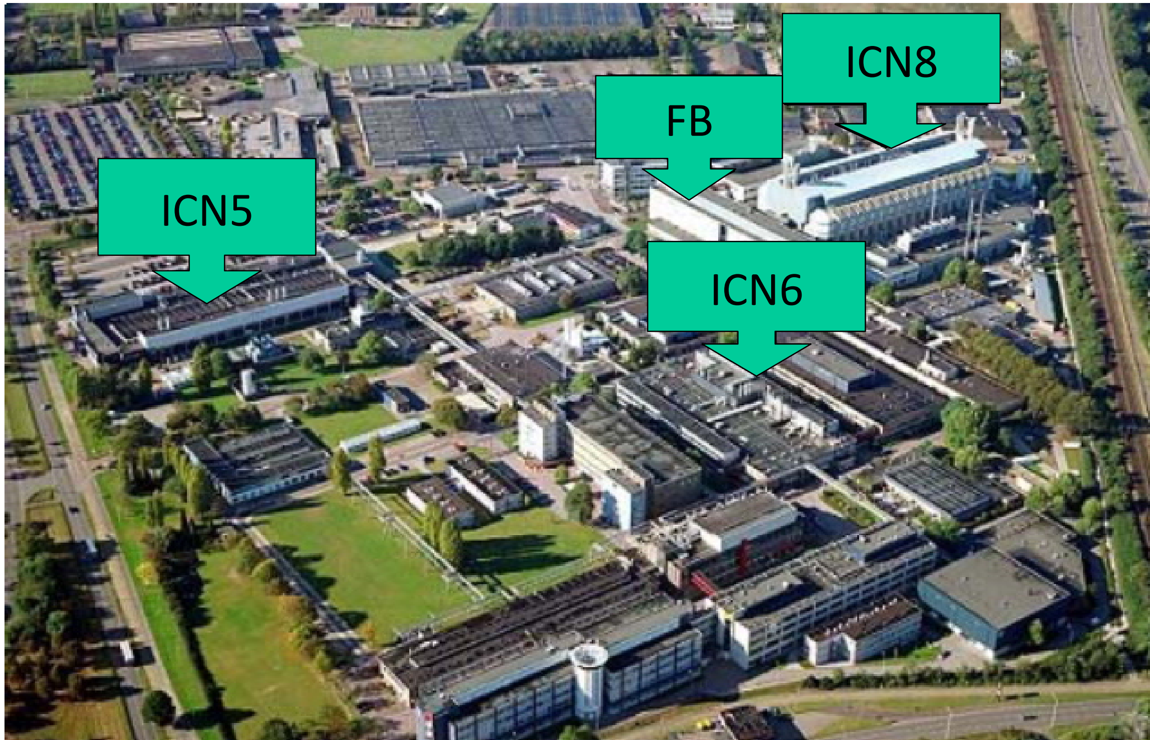

Figure 1. 1 The site of NXP Semiconductors, Nijmegen, two wafer foundries (ICN5, ICN6 and ICN8) are high-lighted, reliability group of the innovation department is located in building "FB".

1.1.3 Innovation department

The department of innovation is part of the ICN support organization, and is challenged to develop new process generations, improve processes in manufacturing high performance products. Now the section NVM development is specialized in developing embedded non-volatile memory technologies.

1.2 Thesis task

The task of this project is to investigate the reliability aspects of NXP 2T-FNFN-NOR embedded non-volatile memories (eNVM). The structure of the memory cell is presented and its principle of operations is described. The reliability aspects (mainly endurance & retention) of the floating gate memory cell are inspected. Methods are presented to extract important parameters related to this eNVM. A new influence (charge transport in the nitride) on V_T instability is addressed.

1.3 Thesis outline

Chapter 2 gives an overview of the non-volatile memory.
Chapter 3 inspects the reliability aspects of the floating gate memory devices
Chapter 4 proposes methods to determine the coupling ratio of floating gate memory devices.
Chapter 5 focuses on V_T instability induced by charge transport in the nitride layer.

1.4 Keywords

Flash/EEPROM, Fowler-Nordheim tunneling, Endurance, Retention, Floating gate memory, Programming, Erasing, Coupling ratio, V_T shift, Poole-Frenkel, SILC, Interface states, Trapped charge, Multi-level-bit, stacked layer, neutral V_T (V_{Tn}), . Hopping conduction (HC), Trap assisted tunneling (TAT), Extrapolated V_T, fixed current V_T, Floating gate, Control gate.

2. OVERVIEW OF NON-VOLATILE MEMORIES

In this chapter, an overview of non-volatile memory is presented. Non-volatile memory is widely used in electronic applications, especially EEPROM and Flash memories. A NXP floating gate based embedded non-volatile memory cell on 90nm technology node is introduced, including structure, operations. The reliability aspects related to this memory cell is also presented.

2.1 Introduction to semiconductor memories

Semiconductor memory has become an indispensable component of modern information processing system, and its application is expected to increase as growing device capabilities. The semiconductor memory can be divided into two main categories: random access memories (RAM), which are volatile, and read-only memories (ROM), which are non-volatile. Both types of memories can be subdivided into different groups, based on the functions, as shown in figure 2.1.

- The volatile memories lose the stored information as soon as the power supply is shut off, so a continuous power must be supplied to keep the data available for read-out. The most common types of volatile memories are SRAM (Static Random Access Memory) and DRAM (Dynamic Random Access Memory).
- The non-volatile memories are able to keep the stored information for a long time even if the power supply is switched off. Normally for application in the products, a data retention time of 10 years is required.

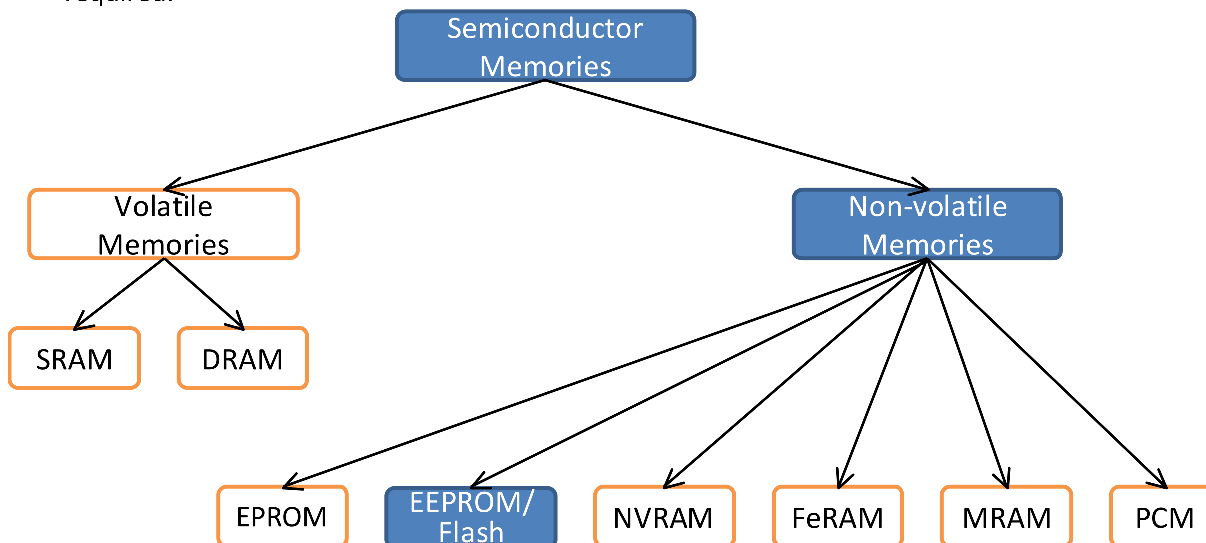

Figure 2. 1 Overview of different groups of semiconductor memories.

Non-volatile memory sector takes up one of the largest portion of the semiconductor market, and its market share has been continuously growing in the past years, and further growth in the close future is foreseen. The application of non-volatile memories lies widely in personal computers, cellular phone, digital cameras, smart media, portable storage, networks, automotive system, and global positioning system. Figure 2.2 lists some of the common products and applications, which can be found in our daily life.

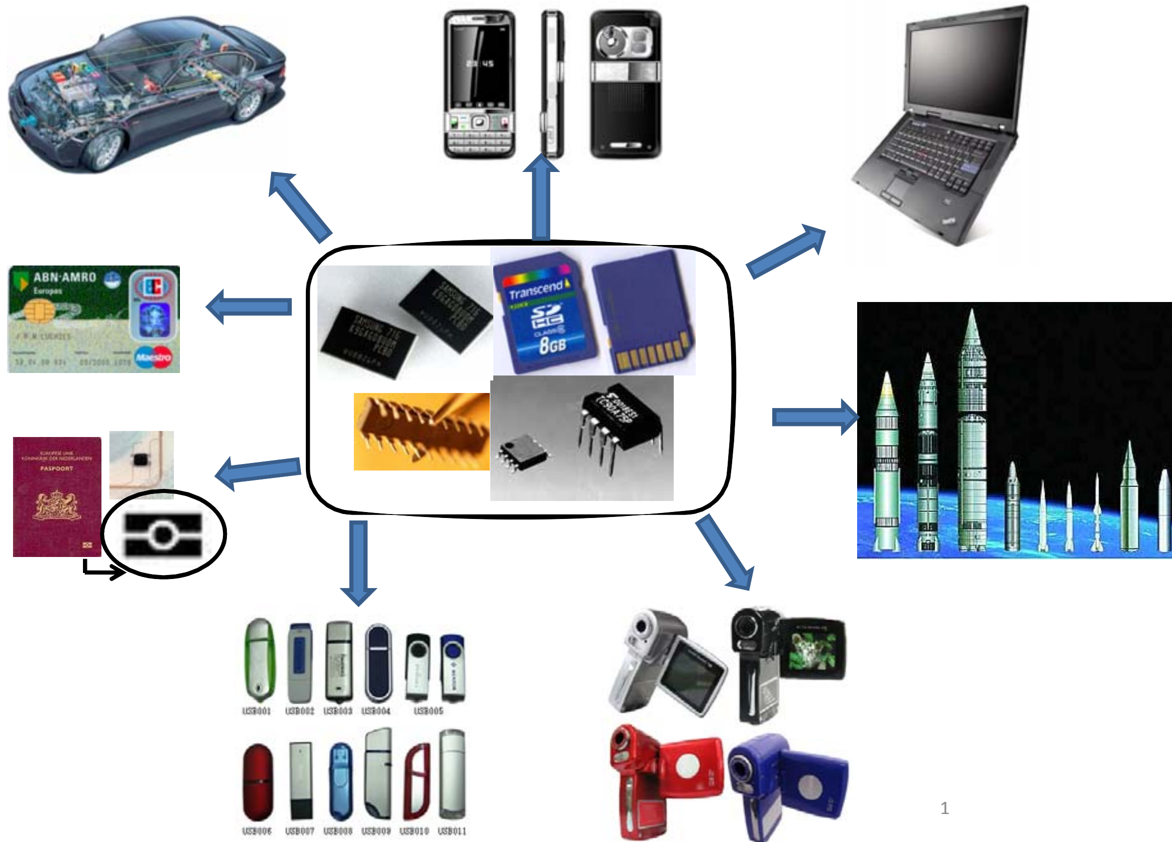

Figure 2. 2 Common products and application of non-volatile memories.

2.2 Flash/EEPROM

This thesis focuses on most commonly used non-volatile memories, EEPROM and Flash memories. The EEPROM's is Electrically Erasable Programmable Read Only Memories, in which all operations are controlled by electrical signals. EEPROM's are electrically erasable and programmable per single bit, manufactured for specific application only. These memories use large area, and are expensive. The term Flash is actually a specific type of EEPROM that the whole memory array (in large blocks, or sector) is erased and programmed in one step at the same time. Technically the term "EEPROM" generally refers specifically to the non-flash memories that are erasable in small blocks, or even bytes. Since the erase cycles are slow, the large block sizes used in flash memory has a significant advantage on speed over EEPROM when large amount of data is being written. Figure 2.3 shows the function of embedded Flash & EEPROM in electronic applications.

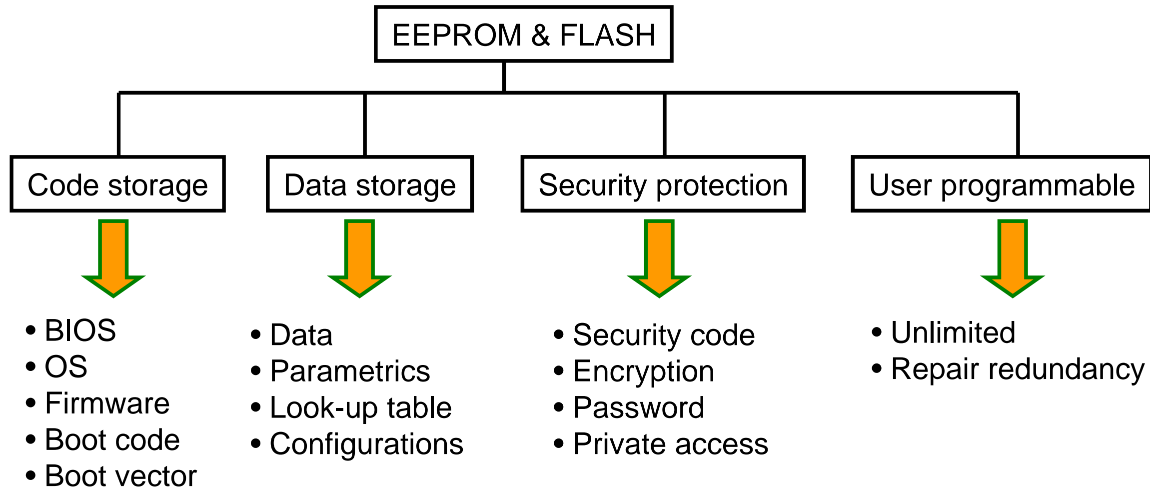

Figure 2. 3 Functions of embedded Flash&EEPROM.

2.3 Cell Structure

One of the memory device classes is specially designed to be embedded in the EEPROM and Flash memories for charge storage, the floating gate devices. These devices are stacked-layer MOS transistor with additional floating gate as charge trapping location, as shown in figure 2.4. The first gate is floating gate, electrically isolated by dielectric. The second gate is control gate, acting as an external gate of the memory transistor. The charge injected from channel region through tunnel oxide will be stored in the conductive layer (floating gate) at programming state, and will tunnel out into channel region at erasing state.

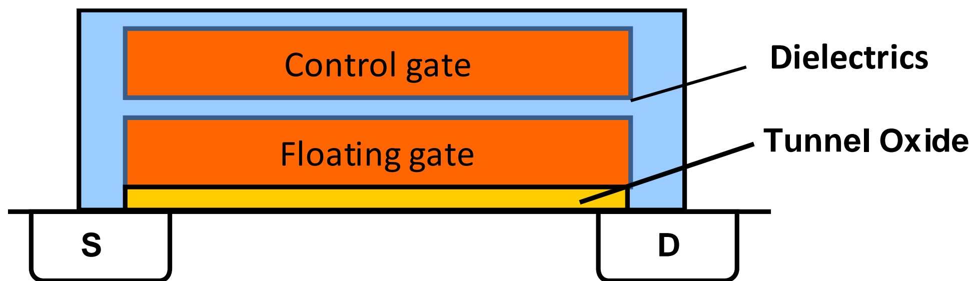

Figure 2. 4 Floating gate device structure used in EEPROM's and Flash Memories.

2.3.1 NXP eNVM device structure

Figure 2.5 (a) shows a schematic cross section of the NXP eNVM device. It has a same structure as a MOS transistor, but with an additional gate (floating gate, FG) and a stacked dielectric layer (inter-poly-dielectric, IPD), both of which are sandwiched between control gate (CG) and tunnel oxide (SiO2). The dielectric layers (tunnel oxide, IPD) are used as charge blocking barrier, isolating the floating gate where the charges are stored. In our sample, the IPD layer is Oxide-Nitride-Oxide (ONO) stacked layers.
The floating gate is electrically isolated, and can't be accessed directly; its voltage is controlled by capacitive coupling. The capacitive coupling model is shown in figure 2.5 (b). With the capacitive coupling, the voltage applied on the control gate can be coupled to floating gate, controlling the voltage on the tunnel oxide, thus the current through tunnel oxide.

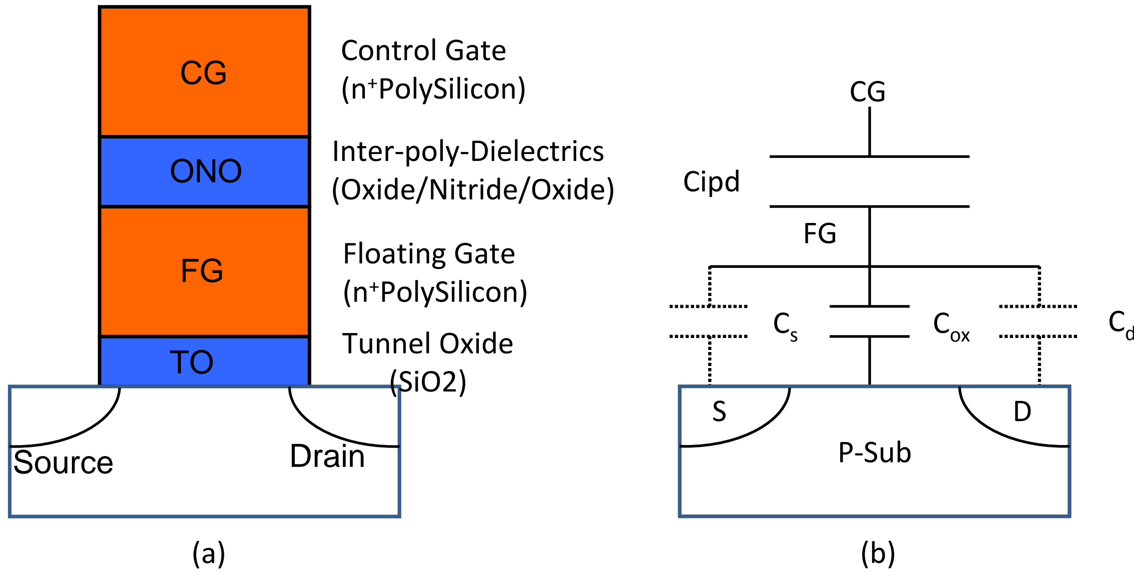

Figure 2. 5 *The schematic cross section of a (a) floating gate transistor and (b) its capacitive coupling model.*

2.3.2 2T-FN/FN-NOR cell

The memory cell unit studied is NXP 2T-FN/FN NOR device based on 90nm technology node. The schematic cross section of a 2-transistor (2T) is presented in figure 2.6 together with its TEM photo. As shown the cell consists of a floating gate device, together with an access gate device at the source side. These two devices are fabricated in isolated-Pwell, which is embedded in deep-Nwell buried placed in the P-substrate.
In this memory cell, both program and erase are performed by Fowler-Nordheim (FN) tunneling mechanism on the tunnel oxide, and a select device (access gate transistor) is put in series with the floating gate transistor, leading to the so called two-transistor FN/FN memory cell (2T-FN/FN-NOR).

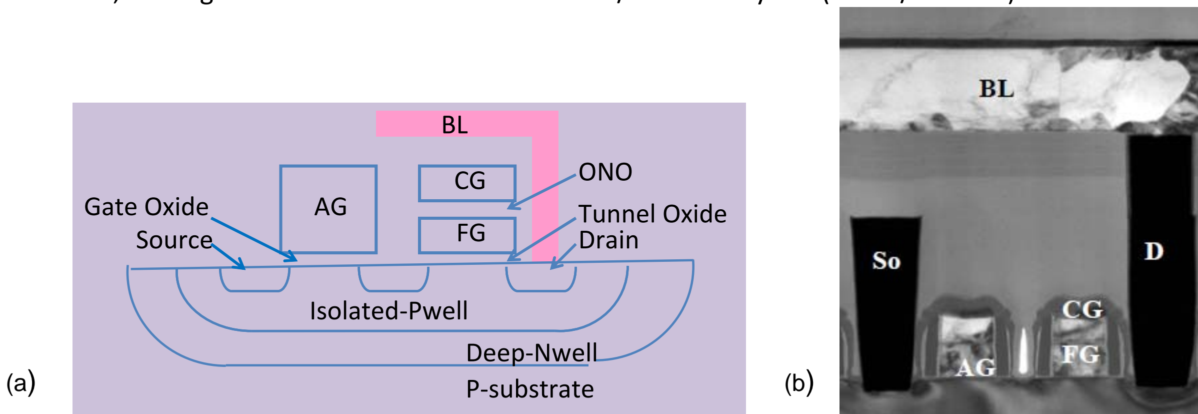

Figure 2. 6 *Schematic cross section of a 2T-FN/FN-NOR device (the shaded part), and its TEM photo, (AG=access gate for select transistor, BL=bit line for drain)*

2.4 Principles

2.4.1 V_T change

The basic operating principle of such a floating gate device is by changing the amounts of stored charge in the floating gate, the threshold voltage (V_T) can be modified to switch between two distinct values, defined as programmed state sand erased state.

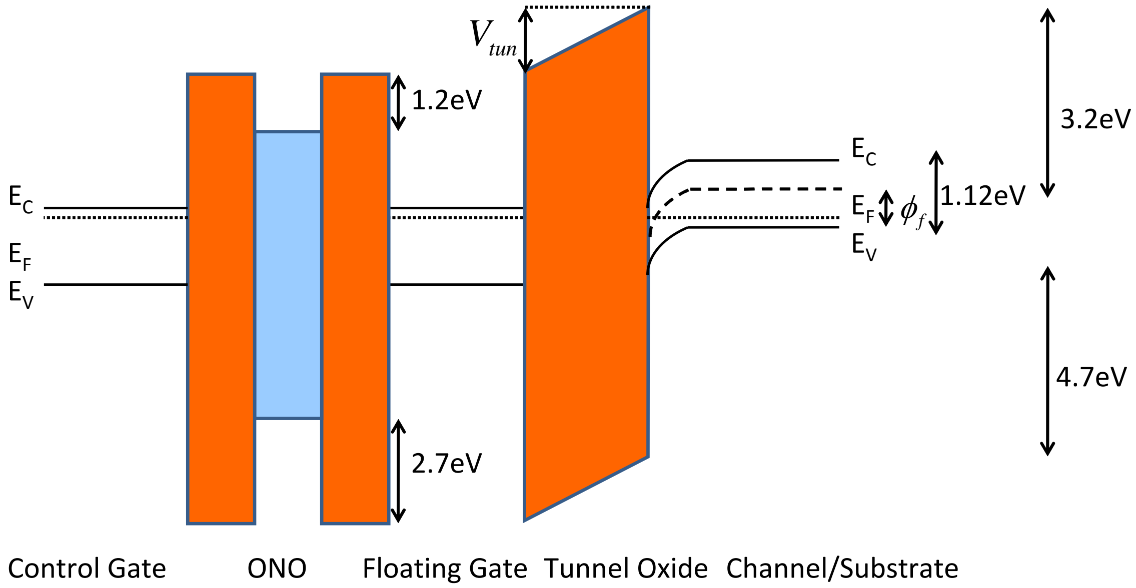

Figure 2. 7 Energy band diagram of the floating gate device, without charge on the floating gate, and voltage applied on control gate. Work function difference between silicon in the channel and polysilicon in the floating gate is indicated.

The energy band diagram of the floating gate cell without charge on the floating gate and voltage on the control gate is shown in figure 2.7. As shown, the floating gate acts as a potential well, if a charge is forced into the well, the charge has very little probablity leaking out of the floating gate without applying external force (external voltage). Due to work function difference between floating gate and channel/substrate (P-type substrate and N+poly gate), the channel is depleted and voltage on the tunnel oxide (V_{tun}) is originally induced.

Starting from basic principle of conventional MOS transistor, the threshold voltage of a floating gate cell seen from floating gate instead of control gate is give by:

$$V_{T,FG} = \phi_{ms} + 2\phi_f - \frac{Q_d + Q_{ox} + Q_{FG}}{C_{ox}} \quad 2.1$$

Where ϕ_{ms} is the work function difference between gate and bulk substrate (~0.9eV);

ϕ_f is Fermi potential at the surface of the channel (~0.3eV);

Q_d is the charge in the depletion region;

Q_{ox} is the charge trapped in the tunnel oxide;

Q_{FG} is the charge trapped in the floating gate.

The capacitor C_s and C_d shown in figure 2.5 (b) is assumed to be negligible compared to C_{ox}, which simplifies the capacitor coupling model in figure 2.8. When an external voltage is applied at the control gate, due to the capacitive coupling, the voltage residing on the floating gate (V_{FG}) is expressed as a function of the capacitance and the charge stored in the floating gate (Q_{FG}) (superimpose of both capacitive coupling influence and charge influence):

$$V_{FG} = \frac{C_{ipd}}{C_{ipd} + C_{ox}} V_{CG} + \frac{Q_{FG}}{C_{ipd} + C_{ox}} \quad 2.2$$

Where C_{ipd} is the inter-poly-capacitor between the control gate and floating gate.

The floating gate transistor threshold voltage V_{Tn} (neutral V_T, without charge on floating gate) seen from control gate is given as a function of $V_{T,FG}$ (threshold voltage of a conventional MOS transistor) and capacitive coupling factor:

$$V_{Tn} = V_{T,FG} / K_{ipd} \quad 2.3$$

Where Kipd is the coupling ratio will be discussed in the chapter 4, equals to $C_{ipd}/(C_{ipd}+C_{ox})$.
When there is charge stored in the floating gate, the cell threshold voltage V_T seen from control gate can be written from equation 2.3 to be:

$$V_T = V_{Tn} - \frac{Q_{FG}}{C_{ipd}} \quad 2.4$$

Thus, the threshold voltage (V_T) shift seen from control gate is induced only by the change in storage of charge ΔQ_{FG} , and expressed as:

$$\Delta V_T = -\frac{\Delta Q_{FG}}{C_{ipd}} \quad 2.5$$

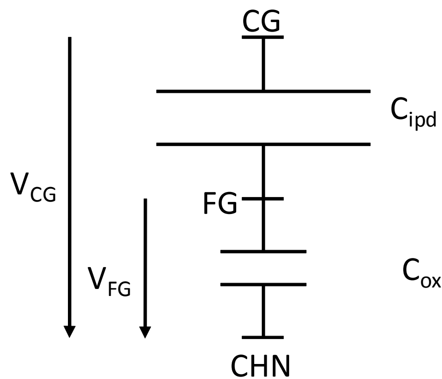

Figure 2. 8 Simplified capacitor coupling model

Figure 2.9 shows the information content varying in change of stored charge on the floating gate. The two distinctive threshold voltages (V_{Te}, V_{Tp}) represent two states. In one state, the cell with the control gate at read voltage V_{read} conducts current over I_{RF} (low V_T), recognized as "1", while, In the other state, the cell is switched off (high V_T) at read voltage V_{read} , no current above I_{RF} flows through the cell, recognized as "0".

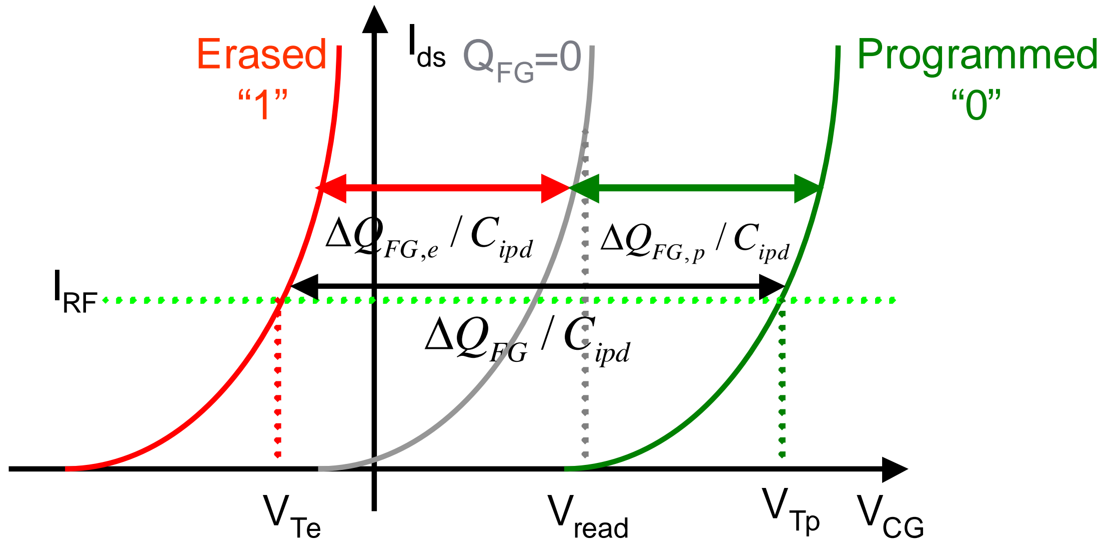

Figure 2. 9 Influence of charge in the floating gate on threshold voltage, at programmed state ("0") and erased state ("1"). I_{RF} is reference current, which is used to define the critical V_T of an I-V curve.

2.4.2 P/E operation

In our sample, channel/substrate is p-type, making the memory transistor NMOS. Floating gate and control gate is fabricated by N+polysilicon, which is sufficient with electrons. During the operations, the electron dominates the charge movement.
The operation “program” is to force electron into floating gate, thus threshold voltage upward shift. The operation “erase” transfers the electron out of the floating gate, thus threshold voltage downward shift. The transfer of electron into (P) or out (E) of floating gate in our sample is realized by FN (Fowler-Nordheim) tunneling mechanism.
FN tunneling is one of the most important injection mechanisms and widely used in non-volatile memory, which is actually a field-assisted tunneling mechanism. When an electron in the conduction band of silicon wants to get across the oxide, it will encounter a barrier (the height is determined by semiconductor material and band structure of Oxide) as indicated in figure 2.7. But when the thin oxide layer suffers a high electric field, the energy band diagram of the oxide is very steep and the width of the barrier is so narrow that electron tunneling high probability has been increased. The process of FN tunneling during programming is schematically illustrated in figure 2.10. During erasing, the FN tunneling mechanism is also used to push the electrons from the floating gate back into the substrate, by applying a high negative voltage on the control gate. The advantage of this method is that tunnel oxide can be thick that oxide degradation by P/E operation can be reduced, but high voltage must be used.

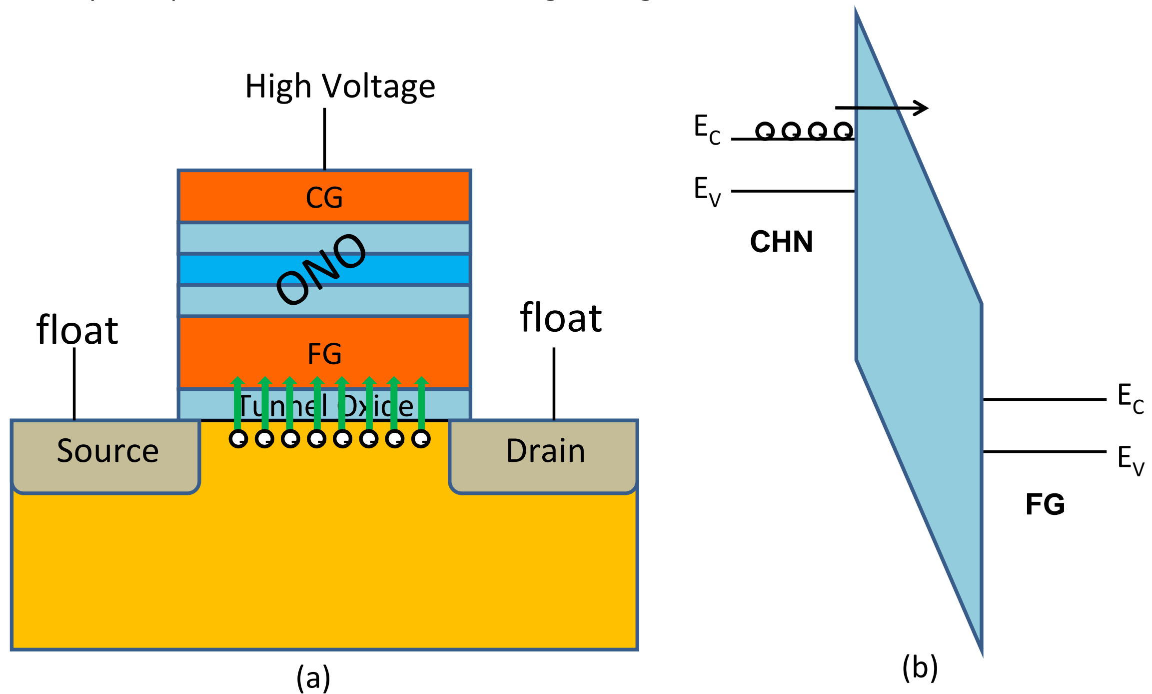

Figure 2. 10 The schematic cross section of a floating gate cell during programming by FN tunneling mechanism, and energy band diagram representation of electron tunneling through oxide at high electric field.

The probability of electron tunneling through a triangle barrier depends on distribution of occupied states in the injecting material (Silicon), and the shape, height, and width of the barrier. Using a free-gas model and WKB approximation, Fowler-Nordheim current density is given by [2.1]:

$$J_{FN} = \alpha_{FN} \cdot E_{OX}^2 \cdot \exp(-\beta_{FN} / E_{OX}) \qquad 2.6$$

With the FN tunneling parameters:

$$\alpha_{FN} = \frac{q^3}{8\pi h \phi_b} \cdot \frac{m}{m^*} \quad \textit{and} \quad \beta_{FN} = 4\sqrt{2m^*} \cdot \frac{\phi_b^{3/2}}{3\hbar q}$$

where ϕ_b is the energy barrier at the injecting interface (4.2 eV for Si-SiO_2)
E_{OX} is the electric field at the injecting interface, roughly applied voltage divided by oxide thickness
q is the charge of a single electron (1.6×10^{-19} C)

m is the mass of a free electron (9.1×10^{-31} Kg)
*m** is the effective mass of an electron in the band gap of SiO_2 (0.42*m*[2.2])
$\hbar$ equals to $h/2\pi$, with Planck's constant (6.63×10^{-34} J.s)

With the value of these parameters, the value of the two tunneling parameters is

$\alpha_{FN} = 1.14 \times 10^{-6}$ A/V^2

β_{FN} = 253 MV/cm

Equation 2.6 is the fundamental expression for FN tunneling, and is quite well for the use in fitting the experimental data of current through tunnel oxide under voltage stress. The other two second-order effects have been taken into the equation 2.6 to make a complete expression for the tunneling current density.

- The image force barrier lowering:

As the charge approaches a dielectric, an opposite charge on the other end of dielectric will be induced, which causes electrical image force. Due to the electrostatic influence an electron close to injecting interface, the image force lowers the effective barrier height ($\Delta\phi_b$).

$$\Delta\phi_b = \frac{1}{\phi_b}\sqrt{\frac{q^3 E_{OX}}{4\pi\varepsilon_{ox}}}$$

- Temperature dependence:

The number of electrons in the conduction band, available for tunneling, increases with the temperature, which leads to a correction factor.

$$f(T) = \frac{\pi.k.T.c}{\sin(\pi.k.T.c)}$$

In literature [2.3], the problems have been discussed, and the tunneling parameters have been modified to be

$\alpha_{FN} = 1.88 \times 10^{-6}$ A/V^2

β_{FN} = 255 MV/cm

There are two major problems in determining the tunneling parameters by measurement:

- The accuracy in determining the electric field.
- Influence of trapped charge in the oxide.

The dependence of the FN tunneling current density on the electric field is plotted in figure 2.11 (a), and also log (J/E^2) versus 1/E is plotted in figure 2.11 (b), yielding a straight line with its slope proportional to the tunneling barrier.

As shown in the figure 2.11, the electric field at 10MV/cm causes a current density of 10^{-3}A/cm^2, which is the order that is required for programming and erasing the cell through the tunnel oxide. Due to the limitation on the value of applied voltage, to achieve such high electric field, very thin oxide thickness is used for memory cells. By applying on the oxide, the tunnel oxide thickness should be thinner than 10mn, but a thickness of 6nm is the minimum limit for good data retention [2.4].

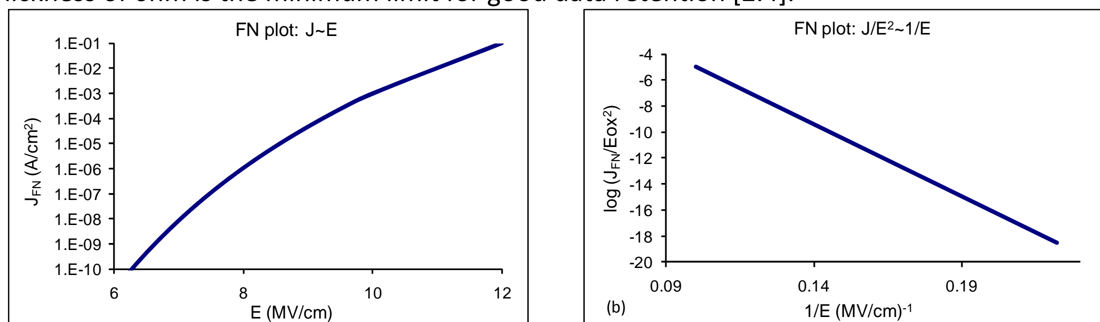

Figure 2. 11 Fowler-Nordheim tunneling plot: (a) J as a function of E, (b) J/E^2 as a function of 1/E.

2.4.3 Read operation

Figure 2.12 shows the energy band diagram after programming and erasing. The information content stored on the memory can be detected by applying a read-out voltage on the control gate, meanwhile the drain-source is biased. The read-out voltage is so low that charge stored on the floating gate can not overcome the barrier, and information content is able to be kept after reading. As shown in figure 2.9, such a read-out voltage is between the two distinctive threshold voltages, namely programmed $V_{T,p}$ and erased $V_{T,e}$. At erase state ("1"), the read-out voltage (V_{read}) is sufficient to obtain a current over the reference current I_{RF}, while at program state ("0"), the read-out voltage (V_{read}) is not sufficient to allow a current over I_{RF} passing through the channel.

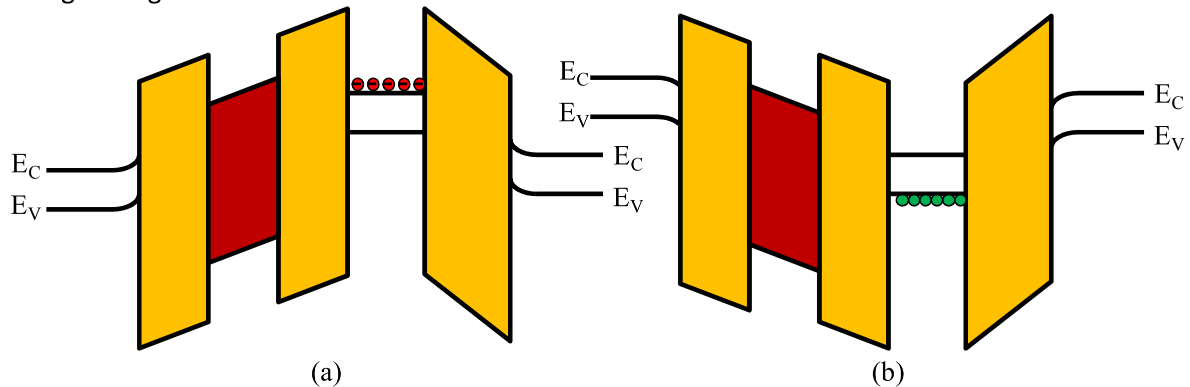

Figure 2. 12 Energy band diagram of a floating gate transistor at (a) programmed state and (b) erased state.

2.5 Cell array

2.5.1 Array structure

In a memory product, the floating gate memory cells need to be incorporated into large-scale memory array. Figure 2.13 indicates the organization of 4×4 2T-FN/FN-NOR cells (one cell is indicated in blue circle), consisting of bitlines (BT) in columns and wordlines (WL) in rows. A bitline is the electric connection to the drain of the cell (bit), which can be used to detect drain-source current I_{ds}. A wordline is connected to the gate (AG) of access devices, which can be used to active read-out in selected row, and switch off the unselected row. The control gates line (CG) is parallel to worlline (WL). All the source lines of the cells are connected to a common source.

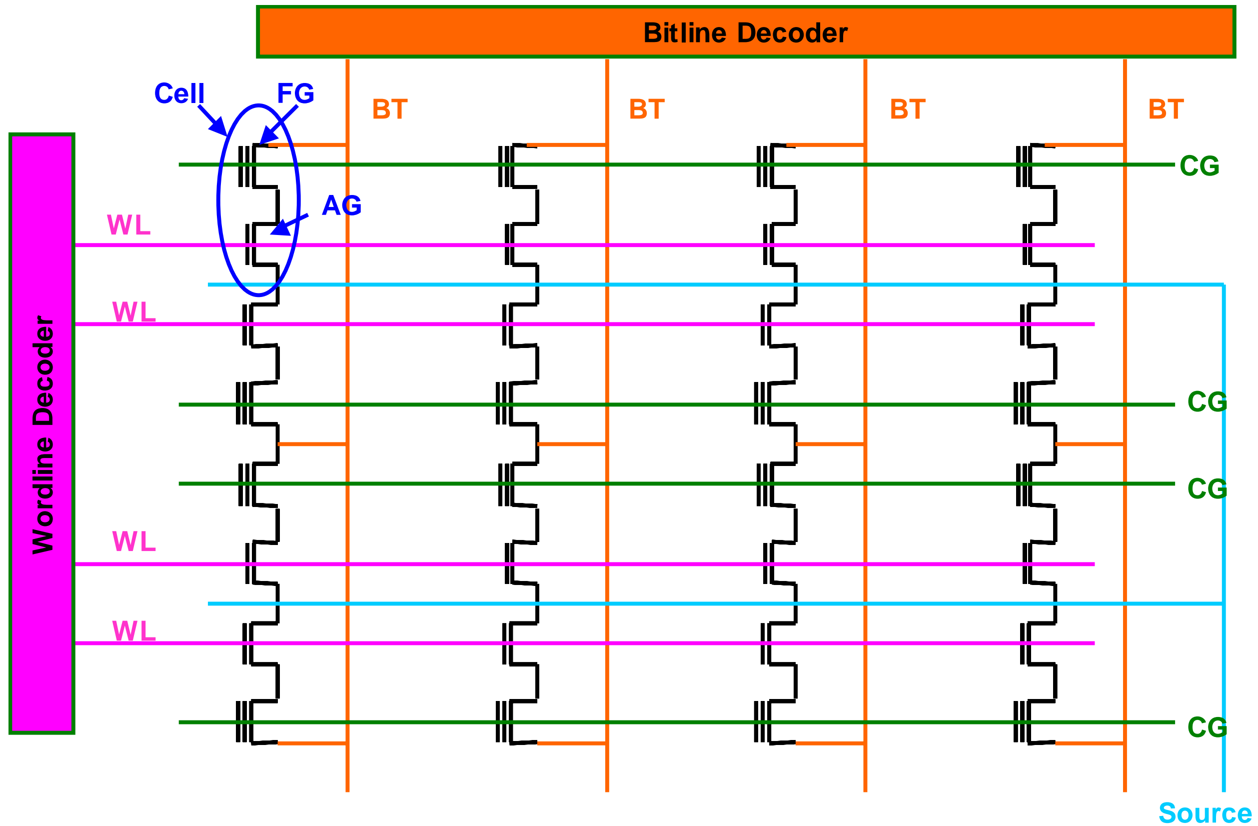

Figure 2. 13 4×4 2T-FN/FN-NOR cell array structure.

2.5.2 Array operation

The operation conditions of the 2T-FN/FN-NOR cell array are listed in the table 2.1. In the table, IPW and DNW represents the isolated Pwell and deep Nwell respectively, indicated in figure 2.6 (a). V_{hp} is high positive voltage (typically 10V) for programming. V_{pi} is intermediate positive voltage (typically 6V). V_{pn} is the negative voltage(typically -5V) used during programming and erasing. The program and erase operations are carried out in a row, while the read-out operation can be carried out cell by cell.

Table 2. 1 Operation conditions of 2T-FN/FN-NOR cell array

	BL	WL	CG	S	IPW	DNW
Program						
Selected	V_{pn}	V_{pn}	V_{hp}	Float	V_{pn}	V_{hp}
Unselected	Float	V_{pn}	0	Float	V_{pn}	V_{hp}
Erase						
Selected	Float	V_{pi}	V_{pn}	Float	V_{hp}	V_{hp}
Unselected	Float	V_{pi}	V_{pi}	Float	V_{hp}	V_{hp}
Read						
Selected	0.5	V_{dd}	1.2	0	0	1.2
Unselected	0	0	1.2	0	0	1.2

2.6 Reliability aspects

In real memory products, millions of cells are fabricated. Not a single cell is identical to another, during operations like programming, erasing, and reading, the characteristics of cells are different, which probably drive one cell or more out of order. During operations, the cells are degrading, at certain critical point, the product doesn't work normally any more. The information content stored in the cell also can be lost, even without any operation on the cells. The task of reliability research is to investigate the characteristics of the entire population.

2.6.1 Endurance

Endurance is referred to the number of P/E cycles that a memory can withstand. Each time when P or E operation is performed on the memory cells, some permanent damages are introduced into the cells. In the product, the number of the cycling times for a memory must be above the defined end-of-life endurance (normally 10^6 cycles).
The typical endurance test of a memory array with single shot programming and erasing is presented in figure 2.14. The threshold voltage at programmed state (V_{Tp}) and erased state (V_{Te}) shift upward with cycles. The variations of program and erase threshold voltage reflects the oxide aging. The failure of a cell occurs at the point when V_{Te} crosses the V_R after certain P/E cycles.

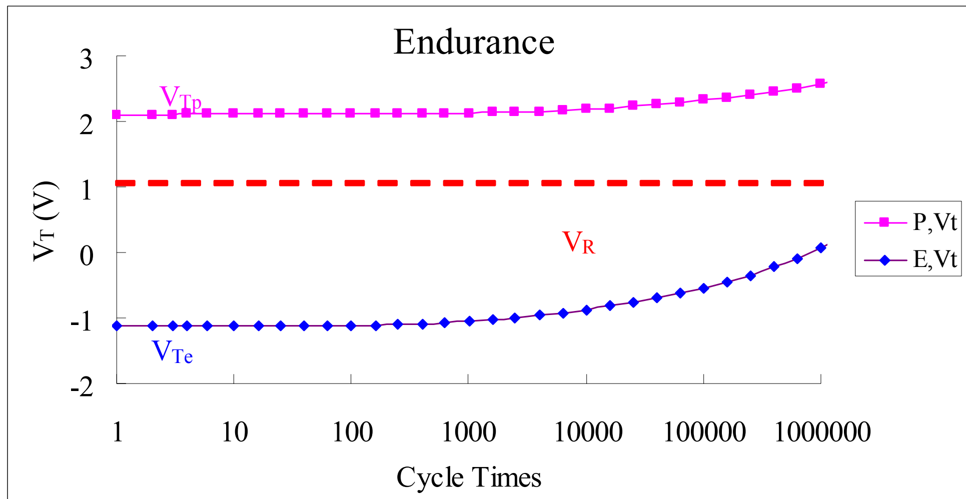

Figure 2. 14 Endurance characteristics of a memory array (minimum, median, maximum V_T, 64kB cells) in terms of cycle times. The blue represents V_R, the read voltage.

If the x-axis of figure 2.14 is changed to be cubic root of cycles, the V_T can be plot as a linear function of cubic root of cycles, shownin the figure 2.15, which provides a way to predict the V_T shift with P/E cycles.

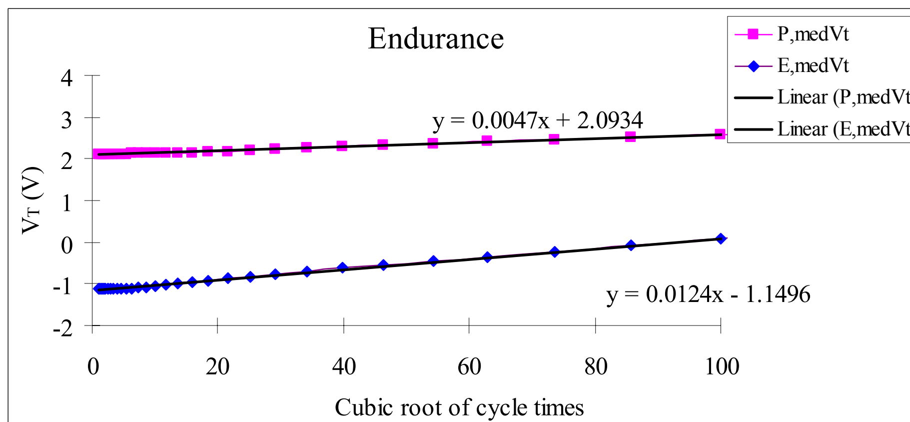

Figure 2. 15 V_T plot as a function of cubic root of cycle times

Two factors of the endurance are introduced to reflect the V_T change with cycle times, V_T mid-window (V_T-mid), and V_T window-closure (V_T-win). The shift of V_T-mid and V_T-win with cycles is plot in the figure 2.16. Empirically, ΔV_T-mid shift correlates the shift of V_{Tn}[2.5]. And since V_T-win must be large enough to distinguish between programmed and erased states, the shift of V_T-win reveals the ability to identify the "1" and "0" state. In the figure 2.16, the V_T-mid shifts down, while V_T-win shifts up.

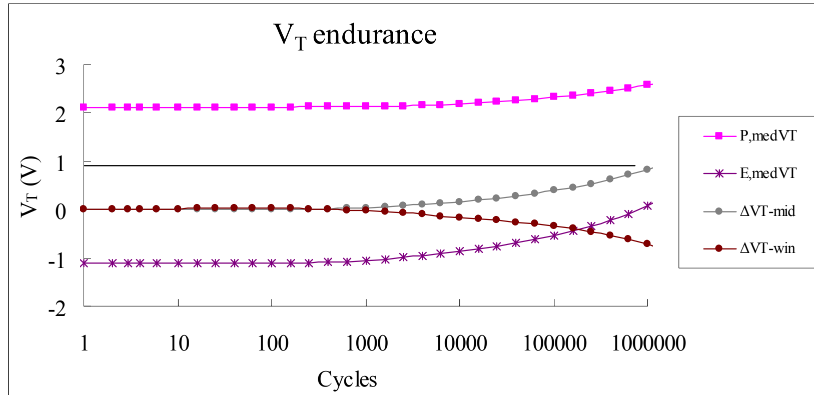

Figure 2. 16 A typical endurance plot with programmed and erased V_T, ΔV_T mid-window, and ΔV_T window-closure.

Cycling induced damage to the tunnel oxide results in the endurance problem, mainly two mechanisms are involved [2.5]:

- Charge trapping: During programming and erasing, electrons tunnel through tunnel oxide by FN tunneling mechanism, which introduces trapped charge into the tunnel oxide. These trapped charges change electric field in the tunnel oxide, thus reducing the charge probability of tunneling through tunnel oxide. According to the following analysis and literature [2.6], the trapped charge is believed to be negative, which causes decrease in V_T window closure (V_{Tp}- V_{Te}, V_{Tp} goes down, and V_{Te} goes up). Also the trapped charges increase the V_{Tn} (the threshold voltage without charge on the floating gate), thus V_T. In figure 2.17, the IV curve shifts right.
- Interface state generation: The unsaturated dangling bonds at the channel/tunnel oxide interface are referred as interface states. Electron tunneling through the interface will break the saturated bond, increases the number of interface states, which cause V_{Tp} and V_{Te} increase (reduced carrier mobility). Trans-conductance of I-V characteristics in the triode regime can be used as an indication for mobility degradation, as indicated in figure 2.17.

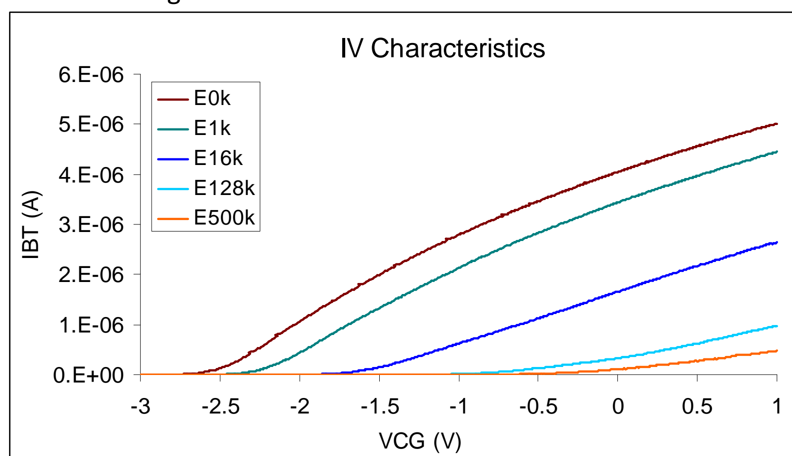

Figure 2. 17 IV characteristics of a floating gate cell at erased state after various P/E cycles (test on single cell of BT1, wafer P9A723 W09).

Method is proposed to investigate the cycling induced trapped charge and interface states. Figure 2.18 plots the IV curve of a fresh cell and a cycled cell. For the conventional MOS transistor, the threshold voltage (by IV curve extrapolate method, shown in figure 2.18 (a)) can be expressed as:

$$V_T = V_{FB} - \frac{Q_{SD} + Q_{ox}}{C_{ox}} + 2\phi_f \qquad 2.7$$

Where V_{FB} flat band voltage, Q_{SD} is the space charge in the channel, Q_{ox} is the increase trapped charge in the tunnel oxide, ϕ_f is the Fermi potential.

When the trapped charge is increased by the cycling, the threshold voltage shifts right (if trapped charge is negative), the ΔV_T can be expressed as

$$\Delta V_T = -\frac{Q_{ox}}{C_{ox}} \qquad 2.8$$

So the amount of increased trapped charge by cycling can be estimated by equation 2.8.

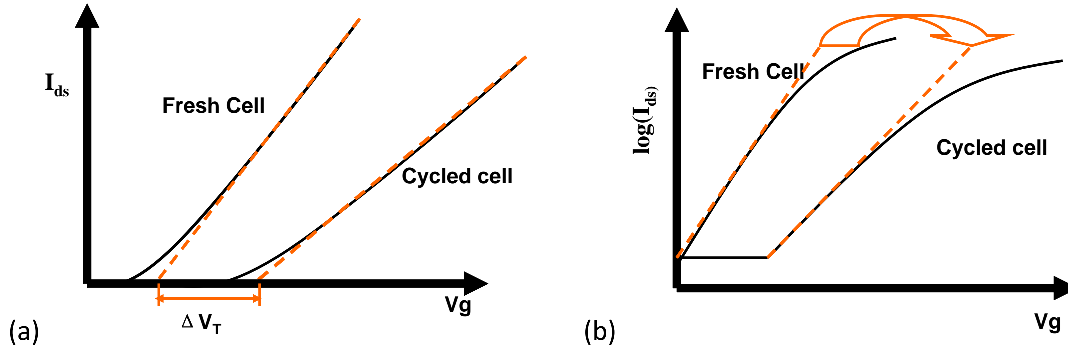

Figure 2. 18 IV plot of fresh cell and cycled cell, y-axis in (a) linear value and (b) in log value.

Figure 2.18 (b) shows log Ids-Vg characteristics, from which the subthreshold slope (S) can be extracted by fitting the linear region, defined as

$$S = \ln 10 \frac{dV_g}{d(\ln I_{ds})} \qquad 2.9$$

The subthreshold slope can be expressed as a function of capacitances [2.7]:

$$S = \ln 10 \frac{kT}{q}\left(1 + \frac{C_D + C_{it}}{C_{ox}}\right) \qquad 2.10$$

Where C_D is the depletion layer capacitance in weak inversion region. C_{it} (=Q_{it}) is an equivalent capacitance for interface states.

From the IV curve, the subthreshold slope increases with P/E cycles. From equation 2.9 and 2.10, the amount of increased effective interface states by cycling can be extracted by

$$\Delta Q_{it} = \left(\ln 10 \frac{kT}{q}\right)^{-1} \cdot \Delta S \cdot C_{ox} / q \qquad 2.11$$

2.6.2 Retention

Retention is to investigate the maximum duration that a memory is able to retain its stored information, without power supply or at reading operations. The charge loss from the floating gate contributes to the retention problem. Though the charges are trapped in the floating gate, still they have probability leaking out, which causes failure for the cells to keep information content. In our samples, the inter-poly capacitance is approximately 0.3fF, a loss of 0.3fC will cause V_T shift 1V. Typically, for products, 10-year life time is required, which means a loss above 0.5 electrons per day can not be tolerated (1V shift).

Mainly two mechanisms lead to the charge loss [2.8].

- Intrinsic mechanism: field-assisted electron emission, thermionic emission, and electron detrapping.
    - Field assisted electron emission: For a programmed cell, the electrons stored in the floating gate can migrate to the interface at the oxide, and from there tunnel into substrate. For an erased cell in

a low threshold voltage, the opposite leakage can happen. The leakage current depends exponentially on the electric field. The electric field at the oxide close to the floating gate is

$$E = \frac{Q}{2\varepsilon_{ox} A_{ipd}} \qquad 2.12$$

where Q is charge stored on floating gate, A_{ipd} is floating gate area, ε_{ox} is silicon dioxide dielectric constant.

- Thermionic emission: Carriers overcome the potential barrier with thermal energy. At room temperature, this mechanism is negligible, but becomes significant at high temperatures.
- Electron detrapping: Detrapping of electrons happens in the tunnel oxide. The trapped charge in the tunnel oxide has probability to detrap from the original location, cause shift in threshold voltage.

➢ Extrinsic mechanism: Oxide defects (increase on decreasing oxide thickness), and ionic contamination.

- Oxide defects: At high electric field, the oxide defects in the thin oxide can induce conductive paths between channel and floating gate.
- Ionic contamination: Ions, usually positively charged, are attracted to the floating gate which is negatively charged, inducing charge loss.

Since at normal operation conditions, it takes very long time for the cells to fail (more than 1 year), temperature and voltage factors are commonly used to accelerate the charge loss, thus to estimate data retention characteristics in a short time.

➢ Temperature accelerating: High temperature gives electron thermal energy, which increases the probability for electrons surmounting the potential barrier. Figure 2.19 shows that bit failure fraction (BFR) of a cell array develops with time at different baking temperatures. At certain criterion of bit failure fraction, the bake time is extracted and plot with the respective temperature in figure 2.20. Then extrapolate the line to the operating temperature 25 ℃ (1/KT=38.5/eV), finding the retention time at operating temperature. However it is found that the slopes of straight lines in figure 2.20 are different, when the criterion changes, the predicted result also will be changed.

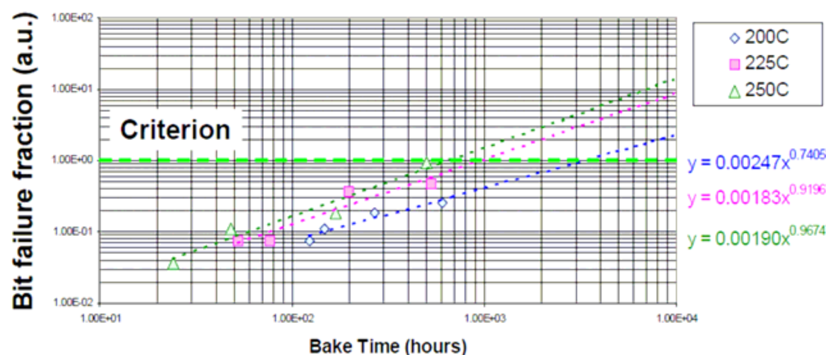

Figure 2. 19 Cell BFR (bit failure fraction) as a function of bake time at various temperatures [2.9].

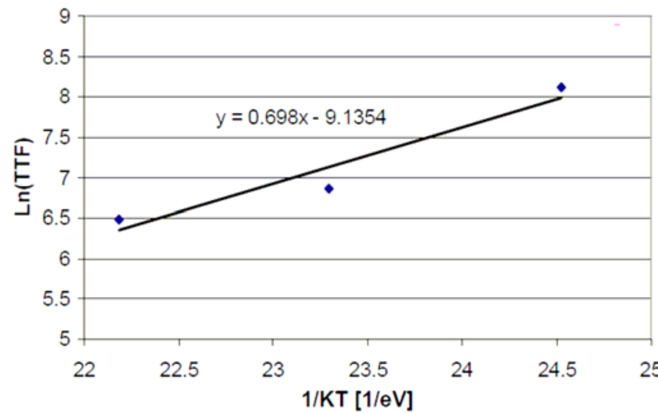

Figure 2. 20 Arrhenius plot of the failure time reaching criterion with thermal energy (in eV)[2.9]

- Voltage stress accelerating: The increased tunneling current at low bias is called SILC (stress induced leakage current), mainly responsible for the retention problem. SILC is independent of temperature, but highly dependent of electric field, which will enhance the charge loss. Figure 2.21 plot the SILC of a cycled cell as a function of oxide electric field, compared with FN tunneling of a normal cell. The figure indicate that electric field below 7MV/cm is preferred to investigate SILC, to eliminate the FN tunneling current.

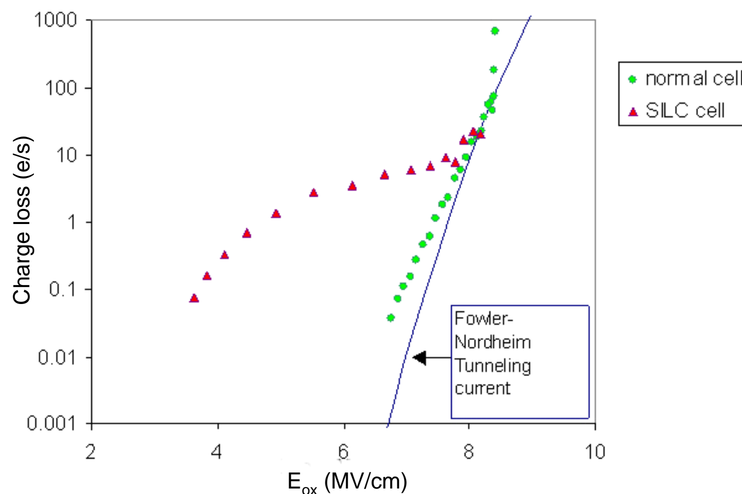

Figure 2. 21 Charge loss due to SILC as a function of electric field, compared with FN tunneling induced charge loss for the normal cell.

Models are proposed to explain the charge loss in the memory cell: TAT (trap-assisted-tunneling), HCM (hopping conduction model), PF (Poole-Frenkel mechanism), FN (Fowler-Nordheim tunneling)

- TAT: tunneling probability of carriers is, to first order, exponentially dependent on the potential barrier at the interface. However, the presence of traps in the oxide splits the band diagram into two pieces (single trap mode), shown in figure 2.22, thus enhances tunneling probability at certain electric field [2.10]. The leakage current density can be expressed by

$$J = \alpha_{TAT} \times E_{OX}^2 \times \exp(\beta_{TAT} \times E_{OX}) \quad (2.13)$$

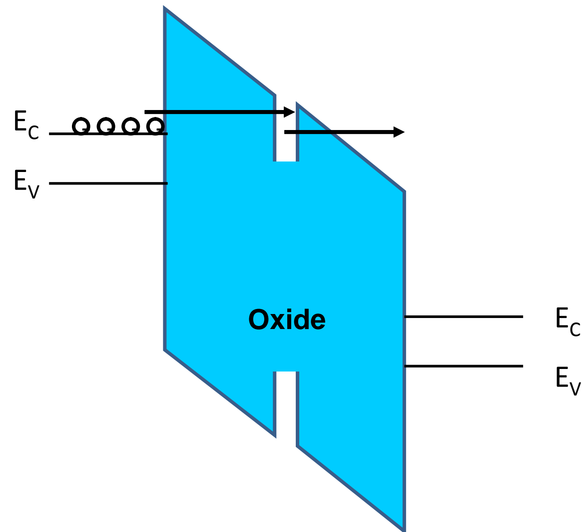

Figure 2. 22 Band diagram of a silicon/oxide/silicon structure with presence of the traps inside the oxide, single trap-assisted tunneling model.

- HCM: Hopping conduction mechanism is thermally excited electrons hopping from one isolated state to the next in the oxide, process indicated in figure 2.23 [2.11].

$$J = \alpha_{HC} \times E_{OX} \times \exp(\beta_{HC} \times E_{OX}) \quad (2.14)$$

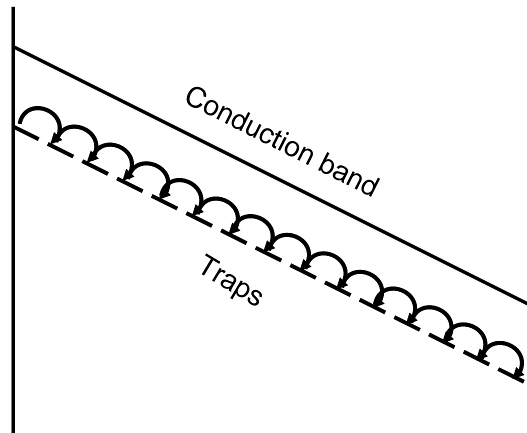

Figure 2. 23 Schematic of charge transport by hopping conduction mechanism in the oxide.

- Poole-Frenkel: Poole-Frenkel conduction is a field-enhanced excitation of the trapped electrons from the trapping center into the conduction band of oxide, process indicated in figure 2.24 [2.12].

$$J = \alpha_{PF} \times E_{OX} \times \exp(\beta_{PF} \times \sqrt{E_{OX}}) \quad (2.15)$$

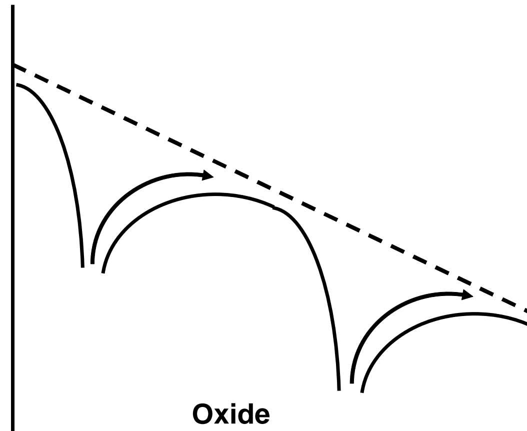

Figure 2. 24 Schematic of charge transport by Poole-Frenkel mechanism in the oxide.

2.6.3 Disturbs

The reliability of a single cell is investigated, but the cells should be organized in array shown in figure 2.13 . The operation conditions are listed in table 2.1. During operation of cell array, new reliability issue “disturbs” is introduced. Any operation with specific stress on one cell, block, or sector causes alternation on other cells. Figure 2.25 schematically illustrates disturbs mechanism during programming. When program the cell 1 and cell 3, 10V is applied on the control gate, gate-drain voltatge (VGD) is 15V, but for cell 2 and cell 4, the VGD is 10V, which may induce tunneling current. And at cell 5 and cell 6, the VGD is 5V, even when the voltage on control gate is 0V. Due to this effect, cells will be programmed accidentally. The disturb problem will become even more significant after many P/E cycles. Some technical solutions have to be implemented to reduce disturbs during operations. Anyway this will not be addressed in this thesis.

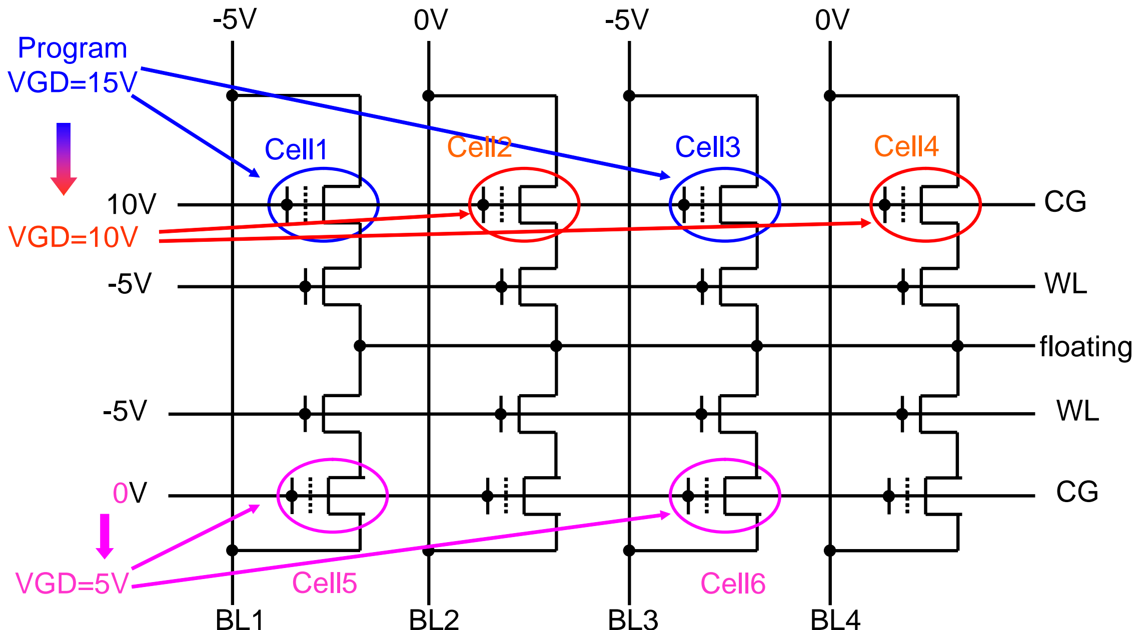

Figure 2. 25 Program disturbs in a 2T-FN/FN-NOR memory array

2.7 Summary

In this chapter, an overview of non-volatile memory is presented, and this thesis focuses on EEPROM and Flash memories.

Floating gate transistor is used in our example. The structure of the floating gate cell is stacked layers from bottom to top (substrate/tunnel oxide/n^+poly-floating gate/inter-poly-dielectrics/n^+poly-control gate), with its floating gate electrically isolated by dielectrics. The distinguish between programmed state and erased state is judged by the value of the threshold voltage (fixed current method), thus read voltage (V_{Read}) is applied on the gate to conduct drain-source current I_{ds}. Then bit-decorder will compare the detected current to the reference current (previously defined). The threshold voltage shift between two states is induced by charge transfer between channel and floating gate through tunnel oxide by Fowler-Nordheim tunneling mechanism.
In order to eliminate over-erase influence on reading, an access transistor (select transistor) is embedded in the isolated Pwell in series with the floating gate transistor, forming a 2T-FN/FN memory cell, which switch off the current of unselected transistor.
Reliability aspects of the floating gate memory such as V_T distribution in matrix, endurance, data retention, and disturbs during operations, are explained in this chapter, and methods are introduced to estimate the reliability problems.

Chapter 3 to Chapter 5 are censored by NXP (for confidential reasons).
For detailed info please contact the supervisor at NXP:

Guoqiao Tao
NXP Semiconductors
Rm FB2.020
Gerstweg 2
6534 AE Nijmegen

Guoqiao.Tao@nxp.com

6. CONCLUSION

6.1 Summary

In this project done at NXP Seminconductors, Nijmegen, the reliability aspects of the NXP floating gate based embedded 90 nm non-volatile memory (eNVM) cell is investigated.
The main reliability aspects of the floating gate non-volatile memory cell are related to endurance and retention. This thesis is able to plot the increased trapped charge and interface states which are believed to result in the endurance problem. Effort is needed to reduce the cycling induced trapped charge in the tunnel oxide, and low reference current should be used to improve the endurance. Related to retention, models are presented to predict the data rentention time. However the conditions for the use of these models is needed to investigate for accurate prediction.
Methods are introduced to extract one of the most important parameters in the this floating gate non-volatile memory. And the result extracted by Gm method is used in the other analysis. Further work can be put on test more cells to find a distribution, which is more reasonable.
The electron displacement in ONO layer is revealed in this project. Gate stress experiments are carried out to evaluate the electron displacement influence on the V_T. Possible electron displacement dependences are investigated. The Vt instability is not a problem to 2bit/cell application, but may become a problem to multi-bit/cell application.

6.2 Future work

Process improvement may be needed to reduce cycling induced trapped charge.
More dummy cells can be test, and find the coupling ratio distribution.
Different ONO layer structures can be tested to find the suitable structure for multi-bit-per cell application.

6.3 Publication

"VT instability in High Bit Count per Cell Floating Gate Non Volatile Memories", on IIRW 2009.

References

[2.1] M. Lenzlinger and E. H. Snow, “Fowler-Nordheim tunneling in thermally grown SiO2”, J. Appl. Phys., vol. 40, p. 278, 1969.

[2.2] M. Kamiya, Y. Kojima, Y. Kato, K. Tanaka, and Y. Hayashi, “EPROM cell with high gate injection efficiency,” IEEE IEDM Tech. Dig., p. 584, 1986.

[2.3] A. Kolodny, S. T. K. Nieh, B. Eitan, “Analysis and Modeling of Floating-Gate EEPROM Cells”, IEEE Trans. Electron Devices, vol. ED-33, no. 6, p. 835-844, 1986.

[2.4] G. Tao, Tutorial “Reliability of Floating Gate Based Embedded Non-Volatile Memories”, Procesdings of IEEE international Reliability Physics Symposium (IRPS), 2007.

[2.5] G. Tao “A simple and accurate method to extract neutral threshold voltage of floating gate flash devices and its application to flash reliability charazterization”, IRW 2007, 2007.

[2.6] Sze SM. Physics of Semiconductor Devices. 2nd ed. New York: John Wiley and Sons, 1981

[2.7] A. Shappir, “Subthreshold slope degradation model for localized-charge-trapping based non-volatile memory devices”, Solid-State Electronics, Vol. 47 (2003), p.937-941.

[2.8] P. Pavan, “Flash Memory Cells-An overview”, Proc. IEEE, Vol. 85, NO. 8, Aug 1997.

[2.9] A. Scarpa, "Reliability implications in advanced embedded 2T-FN-NOR flash memory devices", Solid-State Electronics, Vol. 46 (2002), p.1765-1773.

[2.10] M. P. Hound, “Current transport mechanism in trapped oxides: a generalized trap-assisted tunneling model”, J. Appl. Phys. 86 (3) (1999), p.1448-1491)

[2.11] R. Perera, “Trap assisted leakage current conduction in thin silicon oxynitride films grown by rapid thermal oxidation combined microwave excited plasma nitridation”, Microelectronic Engineering, Vol. 65, (2003), p.357-370.

[2.12] J. R. Yeargan, “The Poole-Frenkel effect with compensation present”, J. Appl. Phys, Vol. 39, 1968.

[3.1] A. Wu, “A novel high-speed, 5-V programming EPROM structure with source-side injection”, IEEE IEDM Tech. Dig., p.584, 1986.

[3.2] L. Schrader, “The influence of the interface states on dynamic transconductance of MISFETs”, Solid-State Electronics, Vol. 20, p. 671-674.

[4.1] R. Duane, M. Florian Beug, and A. Mathewson, “Novel Capacitance Coupling Coefficient Measurement Methodology for Floating Gate Nonvolatile Memory Devices”, IEEE Electr Dev Lett, VOL. 26, NO. 7, 2005.

[4.2] A. Moreira, Guoqiao Tao, “Influence of Parasitic Capacitance In Modeling and Analysis of Advanced Floating Gate Memory Devices”, Physical and Failure Analysis of Integrated Circuits, 2009, P.302-306.

[4.3] www.synopsys.com/Tools/TCAD/DeviceSimulation/Pages/SentaurusDevice.aspx

[5.1] S. M. Sze, J. Appl. Phys. 38, 2951 (1967).

[5.2] D. Frohman-Bentchkowsky and M. Lenzlinger, J. Appl. Phys., Vol.40, 3307, 1969.

[5.3] T. Ishida, “Characterization Of Charge Traps in MONOS Structures for Embedded Flash Memories”, IEEE, 44th Annual International Reliability Physics Symposium, San Joes, 2006.

[5.4] G. Tao, “A low voltage , low power, highly reliable, multi-purpose, cost-competitive embedded non-volatile memory in 90nm node”, ICMTD 2007, p.113-115.

[5.5] G. Tao, “Reliability of Advanced Embedded Non-Volatile Meories: the 2T-FN/FN device”, ICICDT Proc, 2008.

[5.6] J.Yang, G. Tao "VT instability in High Bit Count per Cell Floating Gate Non Volatile Memories", IIRW 2009.