A model for programming characteristics of Sonos type flash with high-kappa dielectrics

Varun Jain

University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Part of the Higher Education Commons

Repository Citation
https://digitalscholarship.unlv.edu/rtds/2172

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
A MODEL FOR PROGRAMMING CHARACTERISTICS OF SONOS TYPE FLASH
WITH HIGH-κ DIELECTRICS

by

Varun Jain

Bachelor of Science
Northern Arizona University
Flagstaff, Arizona
2004

A thesis submitted in partial fulfillment of the requirements for the

Master of Science Degree in Electrical Engineering
Department of Electrical and Computer Engineering
Howard R. Hughes College of Engineering

Graduate College
University of Nevada Las Vegas
August 2007
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
The Thesis prepared by

Varun Jain

Entitled

A Model for Programming Characteristics of

SONOS-Type FLASH with High-K Dielectrics

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative
ABSTRACT

A model for programming characteristics of SONOS type FLASH with high-κ dielectrics

by

Varun Jain

Dr. Rama Venkat, Examination Committee Chair
Professor, Electrical and Computer Engineering
University of Nevada, Las Vegas

Silicon Oxide Nitride Oxide Silicon (SONOS) FLASH memories have recently gained a lot of attention due to better retention and scaling opportunities over the conventional Floating Gate FLASH memories. The constant demand for device scaling, to attain higher density, higher performance, and low cost per bit, has posed charge leakage problems. SONOS type devices with high-κ storage layers and/or high-κ blocking oxide have been proposed to alleviate the demand for constant tunnel oxide scaling. In comparison to conventional FLASH, these devices operate at lower voltages, exhibit higher programming speeds, comparable retention times, less over-erase problem and better compatibility with low power CMOS logic.

The objective of this thesis is to develop a comprehensive model which can be used to obtain the programming characteristics, i.e., shift in threshold voltage vs. program time, for “trap-based” FLASH memories with high-κ dielectrics. The proposed model is used to obtain the programming characteristics for SONOS type devices. The results
from this model are compared with the experimental results and in general the agreement is good. For SONOS type devices with high-κ blocking oxides, the density of available nitride traps for charge storage is shown to have a linear dependence with the potential energy difference between the silicon substrate and the nitride storage for different gate biases. The model is also used to get an estimate of available trap energy levels in the nitride layer as a function of applied voltage.
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER 1  INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2  NON VOLATILE SEMICONDUCTOR MEMORY: FLASH</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Evolution of FLASH</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Basics of FLASH</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Programming Mechanisms of FLASH</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1 Charge Injection</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1.1 Channel Hot Electron (CHE)</td>
<td>14</td>
</tr>
<tr>
<td>2.3.1.2 Channel Initiated Secondary Electron (CHISEL)</td>
<td>15</td>
</tr>
<tr>
<td>2.3.2 Quantum Mechanical Tunneling</td>
<td>16</td>
</tr>
<tr>
<td>2.3.2.1 Fowler Nordheim Tunneling (FNT):</td>
<td>17</td>
</tr>
<tr>
<td>2.3.2.2 Direct Tunneling (DT) and Modified Fowler Nordheim Tunneling (MFNT):</td>
<td>19</td>
</tr>
<tr>
<td>2.4 Erasure Mechanism</td>
<td>20</td>
</tr>
<tr>
<td>2.5 Reliability</td>
<td>22</td>
</tr>
<tr>
<td>2.5.1 Retention</td>
<td>22</td>
</tr>
<tr>
<td>2.5.2 Endurance</td>
<td>23</td>
</tr>
<tr>
<td>2.5.3 Scaling</td>
<td>23</td>
</tr>
<tr>
<td>2.6 SONOS Literature Review</td>
<td>23</td>
</tr>
<tr>
<td>CHAPTER 3  PROPOSED MODEL</td>
<td>37</td>
</tr>
<tr>
<td>3.1 Motivation for Research</td>
<td>37</td>
</tr>
<tr>
<td>3.2 Device Architecture and Fabrication</td>
<td>38</td>
</tr>
<tr>
<td>3.3 Model</td>
<td>39</td>
</tr>
<tr>
<td>3.3.1 Electrostatics</td>
<td>41</td>
</tr>
<tr>
<td>3.3.2 Current Tunneling</td>
<td>45</td>
</tr>
<tr>
<td>3.3.3 Threshold Voltage Shift</td>
<td>48</td>
</tr>
<tr>
<td>3.3.4 Relative Dielectric Constant</td>
<td>49</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1.1: A schematic diagram depicting Non Volatile Semiconductor Memory (NVSM): FLASH .................................................................7

Figure 2.1.2: A schematic diagram depicting MIMIS memory pioneered by D. Kahng and S. M. Sze in 1967 [2] ........................................8

Figure 2.1.3: A cross sectional view of the MNOS FLASH [10] ..........8

Figure 2.2.1: A schematic diagram of FLASH cell with stored charges in the floating gate [2] .........................................................10

Figure 2.2.2: Threshold voltage shift: A schematic drain current (I_d) vs. (gate voltage) Vgs characteristics of a FLASH cell [3, 5] .........................11

Figure 2.2.3: Schematic diagrams depicting (a) FLASH cell, (b) spatial energy band diagram for erased state, (c) spatial energy band diagram for programmed state [5] ..................................................12

Figure 2.3.1.1: A schematic diagram depicting the Channel Hot Electron mechanism [5] .................................................................14

Figure 2.3.1.1.2: A schematic diagram depicting electron surmounting oxide potential energy barrier: CHE [2] ........................................15

Figure 2.3.1.2: A schematic diagram depicting the CHISEL injection mechanism [5] .................................................................16

Figure 2.3.2.1: A schematic energy band diagram depicting the Fowler Nordheim Tunneling process [8] ..............................................18

Figure 2.3.2.2.1: Schematic energy band vs. space diagrams for a Si3N4-SiO2-Si structure depicting (a) DT mechanism, (b) MFNT mechanism [2, 8] .20

Figure 2.4.1: A schematic spatial energy band diagram depicting FLASH erasing state via Fowler Nordheim Tunneling [3] .........................21

Figure 2.6.1: A schematic diagram depicting the cross sectional view of SONOS FLASH [7] .......................................................................24

Figure 2.6.2: A schematic energy band diagram of MNOS depicting oxide current, J_o, and nitride current, J_n, for a positive gate bias [10] .................26

Figure 2.6.3: A schematic band diagram representation of MNOS device with positive gate bias voltage [12] .................................................32

Figure 3.2.1: (a) SONOS/SOHOS device architecture with either Si3N4/HfAlO/HfO2/Al2O3 storage layer. (b) SONOS device architecture with high-k blocking oxide of either SiO2/Al2O3/HfO2/(HfO2)0.48(Al2O3)0.52/(HfO2)0.15(Al2O3)0.85 [7] ..........39

Figure 3.3.1.1: A schematic band diagram depicting SONOS flat band condition 41

Figure 3.3.1.2: A schematic band diagram depicting positive gate voltage band bending for SONOS structure.................................................42
Figure 3.3.1.3: Spatial energy band diagrams depicting flat band condition for SONOS device with (a) Si$_3$N$_4$, (b) HfAlO, and (c) Al$_2$O$_3$ storage layer [7].................................................................44

Figure 3.3.2.1: Schematic energy band diagrams depicting FNT for classical turning point $x_1$ [8].................................................................47

Figure 3.3.2.1: Schematic energy band diagrams depicting (i) DT, and (ii) MFNT for classical turning point $x_1$ [8]...........................................48

Figure 4.3.1: A schematic representation for the architecture of SONOS type devices with various storage layers, namely Si$_3$N$_4$, Al$_2$O$_3$, and HfAlO [7]..............................................................................................................53

Figure 4.3.2: Programming characteristics ($\Delta V_{th}(t)$ vs. program time) of SONOS FLASH with storage layer (i) HfAlO, (ii) Si$_3$N$_4$, and (iii) Al$_2$O$_3$....56

Figure 4.4.1: A schematic representation for the architecture of SONOS type devices with various blocking oxides, namely SiO$_2$, Al$_2$O$_3$, HfO$_2$, HfAlO (with 15% HfO$_2$ concentration), and HfAlO (with 48% HfO$_2$ concentration) [7].................................................................57

Figure 4.4.2: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS device with SiO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$.................................................................59

Figure 4.4.3: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with Al$_2$O$_3$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$.................................................................60

Figure 4.4.4: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$.................................................................61

Figure 4.4.5: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with HfAlO$_{15\%}$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$.................................................................62

Figure 4.4.6: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with HfAlO$_{48\%}$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$.................................................................63

Figure 4.5.1: A schematic spatial energy band diagram depicting conduction band energy differences under applied bias.................................................64

Figure 4.5.2: Density of traps, $N_T$ (m$^{-2}$), vs. energy difference, $E_{diff}$ (eV), for all the investigated blocking oxide dielectrics and a linear approximation...65

Figure 4.5.3: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with Al$_2$O$_3$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$ .................................................................66

Figure 4.5.4: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$ .................................................................67

Figure 4.5.5: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with HfAlO blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$ .................................................................68
silicon nitride, with HfAlO_{15\%}\text{HfO}_2 blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6\,V$, 7\,V, and 9\,V..............................................................68

Figure 4.5.6: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with HfAlO_{48\%}\text{HfO}_2 blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6\,V$, 7\,V, and 9\,V..............................................................69

Figure 4.5.7: Programming transients ($\Delta V_{th}(t)$ vs. program time) for blocking oxides Al_2O_3, HfO_2, HfAlO_{15\%}\text{HfO}_2, and HfAlO_{48\%}\text{HfO}_2 at program voltage of $V_g - V_{fb} = 8\,V$.........................................................................................70

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF TABLES

Table 1.1: Classification of semiconductor memories [1].................................2
Table 3.3.4.1: Dielectric constants for the material used in our study ..............50
Table 4.3.1: List of parameters used in the model to study SONOS programming characteristics with various storage layers.......................................55
Table 4.4.1: List of parameters used in the model to study SONOS programming characteristics with high-κ blocking oxide........................................59
Table 4.5.1: A comparison of the density of traps, $N_{T1}$, obtained from the model with the densities obtained from the linear approximation given by equation 4.5.2. $N_{T1}$ is calculated as a function of $E_{diff}$ for all cases of blocking oxide including SiO$_2$, Al$_2$O$_3$, HfO$_2$, HfAlO$_{15\%}$ HfO$_2$, and HfAlO$_{48\%}$ HfO$_2$ and programming voltages of 6V, 7V, and 9V ..........66
Table 4.5.2: A comparison between the final value of threshold voltage shifts from the experimental results with the shifts obtained from the model with linear approximation fit for blocking oxides Al$_2$O$_3$, HfO$_2$, HfAlO$_{15\%}$ HfO$_2$, and HfAlO$_{48\%}$ HfO$_2$ and programming voltages of 6V, 7V, and 9V ........................................70

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
ACKNOWLEDGEMENT

I would like to dedicate this thesis to my parent Mr. Naresh Kumar Jain and Mrs. Sudha Jain, my sister Hina Jain and my brother Tarun Jain. I would like to thank them for everything they have done for me and for always standing by my side. I would like to thank Dr. Rama Venkat for being my advisor and making this thesis possible. His ingenuity provided motivation and made this learning experience an enjoyable one.

I would also like to thank Dr. Henry Selvaraj, Dr. Paolo Ginobbi, and Dr. Ajit Roy for their cooperation, kindness and serving in my thesis committee. A would also like to make a special mention for my friends Pavan Singarju and Priya Raman for always being there to help. Thanks for all your help. I am highly obliged. Thanks to God for helping me through this thesis.
CHAPTER 1

INTRODUCTION

Memory plays an integral part in the computing industry. Without memory, the advancements in the digital computer industry would not be feasible. Majority of the silicon area in many digital designs is occupied by memory. This trend is going to continue as the demand for storing data grows in consumer applications. Traditionally, memories are categorized according to functionality, access pattern, and the storage mechanism [1]. The two most popular categories of memories are as follows: Non-Volatile Read Only Memory (NVROM) and Random Access Memory (RAM).

The difference between these memories can be attributed to their ability to retain data in conjunction with power supply. Specifically, NVROM is termed non-volatile for its ability to retain data even when the power supply is removed. RAM memories use active circuitry to store information and lose data once the power supply has been removed, hence the term volatile memory. On the other hand, Read-only memories encode the information into the circuit topology. It should be noted that RAM memories provide the flexibility of reading and writing to memory or in other words, re-programmability and can also be classified as Read Write Memory (RWM). Even though ROM also provides
the ability of random access, RAM is reserved for volatile memory for historical reasons only [1].

The conception of the idea that the Metal Oxide Semiconductor structure can be used as memory has posed problem of volatility as in the case of SRAM and DRAM [2]. Using the well developed MOS process provides significant advantages in terms of process integration, density, speed, reliability, and cost. FLASH is a semiconductor based memory cell which uses a modified MOS stack to counteract the volatility problem of RAM while still providing re-programmability. In essence, FLASH memories provide the non-volatility of ROM, re-programmability of RAM and high densities with moderate access times. Classifications of semiconductor memories are listed in Table 1.1.

<table>
<thead>
<tr>
<th>RWM</th>
<th>NVRWM</th>
<th>ROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Access</td>
<td>Non-Random Access</td>
<td>EPROM E²PROM Mask-programmed programmable (PROM)</td>
</tr>
<tr>
<td>SRAM, DRAM</td>
<td>FIFO LIFO Shift register CAM</td>
<td>FLASH</td>
</tr>
</tbody>
</table>

FLASH memory cell is formed by modifying the conventional Metal Oxide Semiconductor (MOS) stack by introducing an additional layer known as the Floating Gate (FG). The floating gate is formed by embedding a silicon layer which acts as a charge storage layer for threshold modification within the oxide. The top oxide (TO) and the bottom oxide (BO) isolate the storage layer from the polysilicon control gate and the
channel in the silicon substrate, respectively. Currently, there are two distinct architecture in existence namely, continuous floating gate and the split gate for FLASH memory [3].

FLASH memory is a direct consequence of the Metal Nitride Oxide Semiconductor (MNOS) device, introduced in the mid-1960s [4]. The usefulness of FLASH as memory lies on the detectable change in the threshold voltage of the FLASH transistor due to presence/absence of charges in the floating gate. The insertion of floating gate in the MOS stack results in higher equivalent oxide thickness and hence, higher threshold voltage and lower transconductance. The threshold voltage of FLASH cell is modified by storing charges via tunneling or Si/SiO₂ potential barrier surmounting from the channel to the floating gate. A logic value of ‘1’ or ‘0’ can be assigned to FLASH cell depending upon whether the threshold voltage has been modified or not [1].

FLASH memories have applications in the telecommunication, computer, automotive, and consumer electronic industries. With the introduction of portable electronics such as MP3 players, applications in the multimedia market demands ever increasing size and higher performance [5]. FLASH memories that are based on uniform layer floating gate pose problems with scaling. Scaling is desired to obtain higher densities as the demand for larger chip memories grows with the advancements in semiconductor technology. Specifically, scaling the bottom oxide also known as the tunnel oxide becomes difficult beyond a certain limit as the reliable operation and charge retention of the device cannot be guaranteed over the device’s lifetime.

For ultra-thin tunnel oxide, any defect present in the oxide has an exaggerated effect on device’s reliability. These defects provide a path for stored charges to leak out from the floating gate and render the device useless. Silicon Oxide Nitride Oxide Silicon
SONOS) FLASH memory reduces the charge leakage/retention problem related to the defects in ultra-thin tunnel/bottom oxides. The principle of operation for SONOS is the same as the conventional FG FLASH except, the injected charges from the inverted channel are trapped in the defect states in the nitride instead of uniformly distributed in the conduction band of silicon floating gate. The leakage current from the silicon nitride floating gate is considerably reduced as the charges need additional activation energy to free themselves from the traps as the traps do not interact with each other.

To keep up with the scaling in CMOS where the minimum feature size is nearing 15 nm and equivalent oxide thickness (EOT) is getting less than 1 nm, further advancements in FLASH are needed. This scaling ensures better integration between CMOS logic circuitry and FLASH memory. Requirements for the use of high supply voltages such as 10V, for FLASH operations also needs to be addressed to ensure compatibility with CMOS logic which is operating with lower than 1V power supplies [6].

Tan et al. [7] proposed replacing silicon nitride in the SONOS type structure with high-κ dielectric material and/or using high-κ dielectric materials as the blocking oxide (top oxide). Specifically, Hafnium Aluminum Oxide with 10% Al₂O₃ concentration is proposed as a replacement for silicon nitride storage layer. Also, two different alloys of Hafnium Aluminum Oxide are proposed to be used as the top oxide in the SONOS structure. Tan et al. [7] observed higher programming speeds, comparable retention times, and less over-erase problem with the proposed SONOS type FLASH devices. These memories also ease the requirements for tunnel oxide scaling and high operation voltages.
The objective of this thesis is to develop a comprehensive model which can be used for “trap-based” FLASH memories. This model is used to study the programming characteristics specifically, the shift in the threshold voltage vs. program time for SONOS devices proposed by Tan et al. [7]. This procedure provides insight into potential benefits of replacing silicon nitride in FLASH memories. The other approach to solving the power and scaling issues is to replace the blocking oxide with high-κ dielectric material. For this case, the shift in threshold voltage vs. program time is studied for three different gate voltages. In addition, the model confirms that by using high-κ material the capacitive coupling between the floating gate and the control gate can be modified in such a way that the requirements on the FLASH programming voltages can be eased. The model is also used to get an estimate for the energy distribution of the traps in the silicon nitride. The density of nitride traps available for charge storage shows a linear dependence on the energy difference between the conduction band of the silicon substrate and the conduction band of the silicon nitride (at the blocking oxide edge) when the device is under applied gate bias.
CHAPTER 2

NON VOLATILE SEMICONDUCTOR MEMORY: FLASH

An ideal memory chip offers high-density, low cost, random access, electrical bit alterability, short access and cycle times, and excellent retention, reliability and endurance. Although such a memory chip is still a dream, FLASH memory comes close by offering many of the desired characteristics of an ideal memory chip [5, 13].

2.1 Evolution of FLASH

Ever since the inception of the idea that semiconductor technology be used as memory, these memory devices have been plagued with the problem of volatility. The first semiconductor based memories which solved the problem of volatility were proposed in 1967. These memories consisted of Floating Gate and the Metal Nitride Oxide Semiconductor (MNOS) trap based devices [2].

Non Volatile Semiconductor Memories (NVSM) such as floating gate and trap based devices are similar to normal Metal Oxide Semiconductor Field Effect Transistor (MOSFET). The difference is in the MOS stack. For NVSM devices, a storage layer is sandwiched between oxide layers and this sandwich provides isolation between the polysilicon gate and the channel as depicted in Figure 2.1.1.
Figure 2.1.1: A schematic diagram depicting Non Volatile Semiconductor Memory (NVSM): FLASH

Figure 2.1.2 shows the MIMIS (Metal Insulator Metal Insulator Silicon) Floating gate memory device, first proposed by D. Kahng and S. M. Sze in 1967 [14]. The storage layer used in the device is a uniform layer of silicon. This storage layer does not have any electrical connections (hence the name floating gate) and the voltage drop is established via capacitive coupling. Direct Tunneling mechanism is used to program and erase the device which is facilitated by the thin tunnel oxide (<5nm).

The floating gate based FLASH memory pose problems as it requires thin dielectrics oxides between the floating gate and the p-silicon substrate for tunneling electrons and storing data. It is hard to achieve high quality, defect free tunnel oxides. The defects in thin oxide poses problem with data retention by providing current leakage paths for the stored charges. The solution to this problem is either to have a thicker oxide and use different programming mechanism and/or use different floating gate material [2].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 2.1.2: A schematic diagram depicting MIMIS memory pioneered by D. Kahng and S. M. Sze in 1967 [2]

Around the same time as the first floating gate device was proposed, Wegener et al. proposed Metal Nitride Oxide Silicon (MNOS) FLASH [15]. MNOS utilized silicon nitride as the storage layer. Unlike the silicon floating gate which captures electrons in the conduction band, silicon nitride captures/traps charges in deep level nitride traps. Since the traps do not electrically interact, pinholes in the oxide do not lead to significant charge loss [2]. A cross sectional view of the MNOS FLASH is depicted in Figure 2.1.3.

Figure 2.1.3: A cross sectional view of the MNOS FLASH [10]
Several devices have since come to realization and have received commercial attention for various applications. Two of these early incarnations of FLASH are Electrically Programmable Read Only Memory (EPROM) and Electrically Erasable Programmable Read Only Memory (EEPROM).

EPROM cell is based on the floating gate FLASH concept and is the first nonvolatile memory that provided the ability of user programmability. The major drawback of EPROM memory is that it requires UV light for erasing. Moreover, the whole memory needs to be erased and reprogrammed even if only one bit needs to be altered. Since erasure cannot be performed electrically, EPROM memory requires expensive packaging and presents inconvenience relating to taking the chip out of the circuit. This problem has led to the use of EPROM memory as One Time Programmable (OTP) memory. EPROM does offer the advantage of high density as the memory cell consists of only one transistor [2, 5].

EEPROM memories were introduced to tackle the problem with the erase operation of EPROM. EEPROM offers the ability of erasing the memory by electrical means. Moreover, erasing and programming can be performed in a byte addressable format. However, the EEPROM cell requires two transistors: a select transistor and a memory transistor and thus offers lower densities in comparison to EPROM [2].

2.2 Basics of FLASH

The term "flash" for FLASH memory is used due to the device’s ability to erase large number of cells, blocks, sectors or pages, at fast speeds. FLASH memory provides high densities of EPROM and electrical erase-ability of EEPROM [13]. The basic working
principle of FLASH memory depends upon a detectable change in the threshold voltage by storing charges in the gate insulator of the MOSFET [2]. (Figure 2.2.1)

![Figure 2.2.1: A schematic diagram of FLASH cell with stored charges in the floating gate](image)

Conventionally, a FLASH memory cell with nominal/low threshold voltage or without stored charges is given a logic value of "0" which is also known as the erased state. Logic value of "1" or the programmed state is assigned to the cell when charges have been stored on the gate insulator [5]. Threshold voltage for MOSFET is defined as the voltage applied on the gate to create a conducting channel under the oxide in the substrate. From the basics of MOSFET theory, the equation for threshold voltage can be written as follows [2]:

$$V_{TH} = 2\phi_F + \phi_{MS} - \frac{Q_I}{C_I} - \frac{Q_D}{C_I} - \frac{Q_T}{\varepsilon_I} d_I$$  \hspace{1cm} (2.2.1)
where \( \Phi_F \) is the Fermi potential of the semiconductor at the surface, \( \Phi_{MS} \) is the work function difference between the gate and the Si substrate, \( Q_I \) is the fixed charge at the silicon-oxide interface, \( Q_D \) is the depletion layer charge, \( Q_T \) is the stored charge in the gate insulator at distance \( d_l \) from the gate, \( C_I \) is the capacitance of the insulating layer, and \( \varepsilon_I \) is the dielectric constant of the insulating layer.

From equation 2.2.1, the threshold voltage shift due to stored charges can be calculated as:

\[
\Delta V_{TH} = \frac{Q_T}{\varepsilon_I} d_I
\]  

\[ \quad \text{...(2.2.2)} \]

![Threshold voltage shift diagram](image_url)

Figure 2.2.2: Threshold voltage shift: A schematic drain current (\( I_D \)) vs. (gate voltage) \( V_{gs} \) characteristics of a FLASH cell [3, 5]
As shown in Figure 2.2.2, a positive threshold shift is obtained if electrons are stored in the storage layer. To read the cell, a read voltage whose value is between the nominal...
and the modified threshold voltages ($V_{TH}$) is applied to the control gate. The magnitude of drain current ($I_{ds}$) indicates whether the cell is programmed or not. If an appreciable $I_{ds}$ is sensed for read voltages, then the cell is in "0" or the erased state otherwise the cell is in "1" or the programmed state [3, 5]. Figure 2.2.3 shows the band diagrams for the erased and programmed state of the FLASH cell. Voltage drops on the blocking and tunnel oxides appear when electrons are stored in the storage layer. Electrons in the conduction band of the floating gate see a barrier with the blocking and tunnel oxides. Once the programming voltage is removed, the floating gate/SiO$_2$ potential energy barrier prevents the charges from leaking out of the floating gate and thus, providing the memory functions.

2.3 Programming Mechanisms of FLASH

As described in Section 2.2, the ability of FLASH to store logic values is due to the fact that the threshold voltage of FLASH MOS cell can be modified by storing charges in the storage layer. Following are the two mechanisms used to program a FLASH cell:

(i) "Charge Injection" in which the carriers surmount the tunnel oxide barrier.

(ii) "Quantum Mechanical Tunneling" in which the carriers tunnel through the tunnel oxide.

2.3.1 Charge Injection

In this mechanism, carriers gain kinetic energy from the large electric field present in the depletion region at the drain end and are injected over the silicon dioxide (tunnel oxide) barrier into the storage layer (floating gate). There are two main types of charge
injection mechanisms: Channel Hot Electron (CHE) mechanism and CHannel Initiated Secondary ELection (CHISEL) mechanism [3, 5].

2.3.1.1 Channel Hot Electron (CHE)

For CHE mechanism, the FLASH cell is biased in the saturation region of MOS I-V characteristics. High voltages and hence, high electric fields for both the gate and the drain ends are required for this mechanism. The cell is biased to create a pinch-off region near the drain end. In moderate electric fields, less than 100 KV/cm ([5, 16]), minority carrier electrons in the n-channel MOS lose energy by acoustic and phonon scattering and stay in thermodynamic equilibrium with the lattice. At high lateral fields, electrons gain more energy than they can lose by scattering with the lattice and become "hot". A small fraction of these electrons gain enough energy to surmount the 3.1 eV silicon dioxide potential energy barrier of the tunnel oxide. These electrons can be stored in the storage layer if the vertical oxide field favors this type of injection [2]. Figure 2.3.1.1.1 and Figure 2.3.1.1.2 depicts the CHE mechanism.

Figure 2.3.1.1.1: A schematic diagram depicting the Channel Hot Electron mechanism [5].
CHE is a power consuming mechanism as it requires high gate and drain voltages and drain currents. Other detrimental effects include drain current reduction, small signal performance degradation and sub-threshold slope lowering. Repeated programming of the cell can lead to permanent device failure due to oxide breakdown by high electric fields [3, 5].

2.3.1.2 CHannel Initiated Secondary ELectron (CHISEL)

CHISEL mechanism compensates for the high voltages and high currents involved in the CHE injection mechanism. Moreover, the injection efficiency of electrons in comparison to CHE is also increased. The CHISEL mechanism employs negative biasing of the body (silicon substrate) to create the "hot carrier" phenomena.
The channel electrons, denoted as $e_1$, are heated by the lateral field and injected into the drain where they impact ionize to create the electron-hole ($e_2-h_2$) pairs (Figure 2.3.1.2.1). The electrons, $e_2$, are collected at the drain, whereas the holes, $h_2$, move towards the substrate. Due to negative biasing at the gate, holes $h_2$ ionize again in the depletion region and create new electron-hole ($e_3-h_3$) pairs. Holes, $h_3$, exit through the substrate whereas, electrons, $e_3$, move towards the gate where the fields are favorable for them to inject into the storage layer [3, 5].

2.3.2 Quantum Mechanical Tunneling

Carrier tunneling is used to program and/or erase FLASH memory. Tunneling is the process of transporting electrons through a barrier via quantum mechanical process. Quantum mechanical tunneling is due to the wave nature of the electrons. Tunneling allows for the electrons to cross the classically forbidden barriers, i.e., when the barrier potential energy ($U$) is greater than the kinetic energy ($E$) of the electron [13].
Wentzel Kramers Brillouin (WKB) approximation provides a first order solution to the Schrödinger's equation for spatially varying potential energy barrier and the transmission probability of incident wave propagation through the classical potential barrier. WKB method is extensively used to calculate tunnel current used for programming/erasing FLASH cells. Tunneling depends on the distribution of occupied states in the injected material, and on the shape, height and width of the potential barrier [5, 13].

2.3.2.1 Fowler Nordheim Tunneling (FNT):

The number of injected charges captured in the storage layer depends on the electric field in the tunnel oxide, number of available energy states and duration of programming [2 - 3, 5, 13]. It is possible to bend the energy bands of the polysilicon-SiO₂-Si (MOS) structure in such a way that the voltage drop on the oxide exceeds the Si-SiO₂ barrier of 3.1 V for electrons i.e. \( V_{\text{ox}} > \Phi_{B} \) (Figure 2.3.2.1.1).

The condition for current tunneling when the oxide voltage drop is greater than 3.1 V is called FNT. The electrons are subjected to a triangular barrier where the width of the barrier is significantly reduced. Electrostatic voltage drop on the tunnel oxide of FLASH cell is obtained via capacitive coupling between the floating gate and the control gate. The electric field over the oxide for FNT is of the order of 10 MV/cm [2, 5]. This high electric field causes some reliability and retention issues as the oxide needs to be able to withstand high electric fields without breakdown and still acts as an insulator to keep stored charges from leaking once the programming is finished. This is especially a problem for ultra-thin oxides, which are prone to defects during processing [3].
Figure 2.3.2.1.1: A schematic energy band diagram depicting the Fowler Nordheim Tunneling process [8]

Following are the three main reasons for the popularity of FNT in FLASH memories: (i) internal generation of the power supply is possible because low current levels are involved; (ii) the tunneling process is purely electrically controlled; (iii) time to program (<1 ms) is considerably shorter than retention time (>10 years) [5]. The current density due to FNT, J\text{FNT}, can be obtained from the following expression [2, 8]:

\[ J_{\text{FNT}} = \frac{q^3 m_o}{8\pi \hbar (q\phi_B) m_{ox}} E_{ox}^2 \exp \left( -\frac{8\pi \sqrt{2m_{ox} (q\phi_B)^3}}{3q\hbar E_{ox}} \right) \]  

... (2.3.2.1.1)
where q is the charge of single electron, $\Phi_B$ is the conduction band energy barrier between Si-SiO$_2$ interface, h is the Planck’s constant, $E_{ox}$ is the electric field on the oxide, $m_0$ is the mass of free electron, and $m_{ox}$ is the effective mass of electron in SiO$_2$.

### 2.3.2.2 Direct Tunneling (DT) and Modified Fowler Nordheim Tunneling (MFNT):

For trap-based memory, such as SONOS with ultra-thin tunnel oxide (< 3nm), the current injection mechanism can be Direct Tunneling or Modified Fowler Nordheim Tunneling. The tunneling mechanism strongly depends upon the electric field on the tunnel oxide and the thickness of the tunnel oxide. In Direct Tunneling (DT), electrons tunnel directly from the conduction band of the Si to the conduction band of silicon nitride through the oxide barrier only. DT happens when $[\Phi_B - \Phi_N] < E_{ox} < [\Phi_B]$ and the electrons see a trapezoidal barrier (Figure 2.3.2.2.1(a)). For MFNT mechanism, in addition to the trapezoidal oxide barrier of DT, electrons also see a nitride barrier as shown in Figure 2.3.2.2.1(b). The condition for MFNT is $[(\Phi_B - (\Phi_B - \Phi_N)) > E_{ox} < [\Phi_B - \Phi_N]]$ [2, 8].

Current density expressions for DT, $J_{DT}$, and MFNT, $J_{MFNT}$, are obtained using free-electron gas model for metal and the WKB approximation for tunneling probabilities and are given by [2, 5, 8]:

$$J_{DT} = \frac{AE_{ox}^2}{\left(1 - \frac{V_{ox}}{\Phi_B}\right)^2} \exp \left(-\frac{B}{E_{ox}} \left(1 - \left(1 - \frac{V_{ox}}{\Phi_B}\right)^{1.5}\right)\right)$$

.... (2.3.2.2.1)
\[
J_{\text{MFNT}} = \frac{AE_{\text{ox}}^2}{\left(1 - \sqrt{1 - \frac{V_{\text{ox}}}{\phi_B}} \frac{\varepsilon_N}{\varepsilon_{\text{ox}}} \sqrt{\frac{m_N}{m_{\text{ox}}}} \left(1 - \frac{\phi_N + V_{\text{ox}}}{\phi_B}ight)^2 \right)^2}
\]

and,

\[
X \exp\left(-\frac{B}{E_{\text{ox}}} \left(1 - \left(1 - \frac{V_{\text{ox}}}{\phi_B}\right)^{1.5} \frac{\varepsilon_N}{\varepsilon_{\text{ox}}} \sqrt{\frac{m_N}{m_{\text{ox}}}} \left(1 - \frac{\phi_N + V_{\text{ox}}}{\phi_B}\right)^{1.5}\right)\right)
\]

...(2.3.2.2.2)

where \( A = \frac{q^3 m_o}{8 \pi \hbar (q \phi_B) m_{\text{ox}}} \), \( B = \frac{8 \pi \sqrt{2 m_{\text{ox}} (q \phi_B)^3}}{3 \hbar} \), \( V_{\text{ox}} \) is the voltage drop on the oxide, \( m_N \) is the effective mass of electron in the nitride, and \( \varepsilon_N \) and \( \varepsilon_{\text{ox}} \) are the dielectric constants of the nitride and the oxide, respectively.

![Schematic energy band vs. space diagrams for a Si₃N₄-SiO₂-Si structure](image)

---

Figure 2.3.2.2.1: Schematic energy band vs. space diagrams for a Si₃N₄-SiO₂-Si structure depicting (a) DT mechanism, (b) MFNT mechanism [2, 8]

2.4 Erasure Mechanism

Erasure of FLASH cell is performed to remove the stored charges from the storage layer and return the device to its original un-programmed state. Conventionally, the
erased state of low threshold voltage represents the logic value of "0". FNT is one of the most popular mechanisms to erase the FLASH cell. A large negative voltage is applied on the control gate and once again the capacitive coupling between the control gate and the floating gate is used to bend the conduction energy bands to create favorable condition for electrons to tunnel out of the storage layer (Figure 2.4.1) [3]. Other than applying large negative voltage on the control gate, large positive voltage on the body can also be applied to facilitate electron tunneling from the floating gate to the Si substrate. Moreover, if high positive voltage is applied to the drain, electrons tunnel into the drain [3].

![Figure 2.4.1: A schematic spatial energy band diagram depicting FLASH erasing state via Fowler Nordheim Tunneling](image)

Figure 2.4.1: A schematic spatial energy band diagram depicting FLASH erasing state via Fowler Nordheim Tunneling [3]
2.5 Reliability

Reliable working of memory is one of the most desirable characteristic for commercial applications. FLASH memory faces retention and endurance reliability issues mainly due to high voltages and currents involved for device operation. These problems are accentuated due to continuous scaling of FLASH devices. Scaling is the driving force of the semiconductor industry as it enables higher densities and faster speeds at lower cost per bit. FLASH scaling is necessary to lower operation voltages and power supply requirements. By lowering operation voltages, it is feasible to build internal power supplies in the chip and have FLASH compatible with low power CMOS [3, 5 - 6].

2.5.1 Retention

Retention for FLASH refers to the ability of the memory to retain the information once the cell has been programmed. The desired retention for FLASH devices is about 10 years [5]. The problem regarding retention results from thin oxides. Simply put, it is hard to obtain high quality defect free ultra thin oxides. Defects in thin oxide, compromise the insulating nature of the oxide. Defects provide leakage paths for electrons to tunnel out of the storage layer and hence render the device useless. Trap based FLASH devices, such as SONOS; provide a possible solution for this problem. In SONOS memory, charges are traps in isolated deep level bulk traps. For the charges to leak out of storage layer, additional activation energy is needed. Also, the traps do not interact electrically, minimizing movement of trapped electrons [2, 3].
2.5.2 Endurance

Endurance in FLASH technology stands for the ability of the device to withstand electrical stress and still be operational. Common endurance requirement for FLASH is to be reliably operational for at least 10,000 program/erase cycles. Over the device lifetime, stress induced defects leads to permanent damage and leakage current such as, SILC (Stressed Induced Leakage Current) and TAT (Trap Assisted Tunneling) [5].

2.5.3 Scaling

As mentioned earlier, FLASH scaling is desired to attain higher density and performance. Carrier injection efficiency can be increased by decreasing $L_{\text{EFF}}$ and thus increasing speed, but with the trade-off of enhancing punch-through and drain turn-on due to the increased capacitive coupling between the floating gate and the drain. To attain lower voltages for program and erase, tunnel and interpoly dielectric thickness needs to be reduced. This poses problem with charge retention. Silicon Oxide Nitride Oxide Silicon (SONOS) memory and SONOS type memory with high-\( \kappa \) dielectric materials presents a viable solution to this problem [5].

2.6 SONOS Literature Review

SONOS is a derivative of MNOS memory and has received a lot of attention recently due to superior reliability, retention, and endurance characteristics over Floating Gate Non Volatile Semiconductor Memories. In SONOS devices, a silicon nitride layer replaces the silicon floating gate as the storage element. Injected charges are trapped in the nitride which leads to threshold modification. Moreover, SONOS type FLASH with high-\( \kappa \) storage layer and/or blocking oxide provides higher performance and lower
operation voltages while minimizing the scaling requirements. The following is a literature review and summary of work done on SONOS.

The trapping characteristics of the SONOS memory have been explained mainly by following two models [9]:

(i) charge trapping models
(ii) current continuity models

In the “charge trapping” models, knowledge of trap density, trap capture cross section area, actual trap energy, and trap distribution in the nitride are important. Current continuity models on the other hand, do not require specific information regarding trap distribution. Trapped charge in the nitride is determined by balancing currents until continuity is achieved [9]. Several extensions and modifications of the above mentioned models have been proposed to account for program/erase: Current (I) vs. Voltage (V), Current (I) vs. Time (t), and retention characteristics of SONOS. Cross sectional view of SONOS is shown in Figure 2.6.1.

![Figure 2.6.1: A schematic diagram depicting the cross sectional view of SONOS FLASH][7].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Frohman-Bentchkowsky et al. [10] proposed a current continuity model to determine the charge accumulation at the nitride-oxide interface under steady state condition for the MNOS device. The current continuity model is based on the fact that steady state current flows through the MNOS structure and due to the difference between the current-field characteristics of the nitride and oxide dielectrics, the accumulated charge tends to change the fields until current continuity is achieved.

Electrode limited current mechanism of Fowler Nordheim tunneling is used to model the current through the oxide (> 5 nm). Concepts from the Metal Nitride Silicon (MNS) structure are utilized to model the nitride current. Bulk limited current transport constitutes of currents due to Schottky emission from the traps above room temperature, field emission of electrons from the occupied traps into the conduction band of the nitride, and low field ohmic current. Silicon nitride current density can be written as follows:

\[
J_n = J_{n1} + J_{n2} + J_{n3} \quad \quad \text{(2.6.1)}
\]

The \( J_{n1} \) component represents the Poole-Frenkel effect and is given by:

\[
J_{n1} = C_1 E_n \exp \left( -\frac{q \phi_1}{kT} \right) \exp \left( \frac{q}{kT} \left( \frac{\beta E_n}{\sqrt{2}} \right) \right) \quad \text{(2.6.2)}
\]

where \( q \) is the unit charge of the electron, \( k \) is the Boltzmann constant, \( T \) is the temperature, \( E_n \) is the nitride electric field, \( \phi_1 \) is the potential well depth, and \( C_1 \) and \( \beta \)
are constants whose values depend on the trap level and dielectric constant. The \( J_{n2} \)
component represents the field emission of electrons from the traps into the conduction
band of the nitride and is given by:

\[
J_{n2} = C_2 E_n^2 \exp \left( -\frac{E_2}{E_n} \right)
\] .... (2.6.3)

where \( C_2 \) and \( E_2 \) are constants whose value depend upon the trap level. Thermally excited
electrons move between isolated traps and results in \( J_{n3} \).

\[
J_{n3} = C_3 E_n \exp \left( -\frac{q\phi_3}{kT} \right)
\] .... (2.6.4)

where \( J_{n3} \) is the thermal activation energy. Current densities \( J_0 \) and \( J_n \) for MNOS devices
under positive gate bias are shown in Figure 2.6.2.

![Figure 2.6.2: A schematic energy band diagram of MNOS depicting oxide current, \( J_0 \), and nitride current, \( J_n \), for a positive gate bias [10]](image)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Under steady state, when a gate voltage is applied, current continuity is achieved when the oxide current density \( (J_o) \) balances the nitride current \( (J_n) \), i.e.,

\[
J_o = J_n \quad .... (2.6.5)
\]

Oxide electric field \( (E_o) \) and nitride electric field \( (E_n) \) can be related to the accumulated interface charge \( (Q_i) \) via electric flux continuity as follows:

\[
K_o E_o - K_n E_n = \frac{Q_i}{\varepsilon_0} \quad .... (2.6.6)
\]

where \( K_o \) and \( K_n \) are the relative dielectric constants of oxide and nitride, respectively, and \( \varepsilon_0 \) is permittivity of free space.

The gate voltage can also be obtained by summing the voltage drops on the two dielectrics, i.e.,

\[
V_G = E_o x_o + E_n x_n \quad .... (2.6.7)
\]

where \( x_o \) and \( x_n \) are the thickness of oxide and nitride respectively. Equations (2.6.1 - 2.6.7) along with the current density equation of Fowler Nordheim tunneling for oxide current can be solved to obtain the following expressions for accumulated interface charge, \( Q_i \), and steady state current density for low operating temperatures:
where \( C_0 \) is a pre-factor from the FNT expression.

Assumptions made to derive the expressions for \( Q_i \) and the steady state current density equations are:

(i) \( E_2 X_n \gg E_1 X_o \) .... (2.6.10)

(ii) \( C_0 \geq C_2 \) .... (2.6.11)

(iii) \( \ln(E_o^2/E_n^2) \ll \ln(C_0/C_2) \) .... (2.6.12)

Note that equation 2.6.10 states that the voltage drop on the nitride is far greater than on the oxide, equation 2.6.11 signifies that the \( C_0 \) constant from the FNT expression is larger than or equal to the pre-factor \( C_2 \) in the field emission current component of the nitride current, and equation 2.6.11 states that the ratio of the electric fields in the oxide and nitride is much smaller than the ratio of \( C_0 \) and \( C_2 \) (field emission current component of the nitride current). It should also be noted that the equations (2.6.8 - 9) were derived for low operating temperature. For moderate to high temperature, at high voltages silicon nitride current is dominated by \( J_{n1} \), (Poole Frenkel equation 2.6.2) whereas, at low voltages \( J_{n3} \) (ohmic component equation 2.6.4) dominates. Silicon dioxide current also increases with the increase in temperature. However, the increase in the nitride current is prominent as it has exponential dependence on the temperature.
Under transient analysis, the difference between the current densities of oxide and nitride can be used to obtain the rate of change of interface charge as follows:

\[
\frac{\partial Q_i(V_G, t)}{\partial t} = J_o(V_G, t) - J_n(V_G, t)
\]  

It was found that the charging time response is few orders of magnitude faster than the discharging time response under the same conditions. The vast difference in the time response of the charging and discharging phases is what renders the "trap-based" MNOS device useful as memory. In summary, the following key points can be identified:

(i) Accumulated interface charge depends on the current-field characteristics and the thickness of the oxide/nitride dielectrics.

(ii) The time required to reach steady state is a strong function of applied gate voltage due to the exponential nature of current-field characteristics.

(iii) The concept that charge accumulation leads to current continuity is valid for any two-layer dielectric structure.

De Salvo et al. [11] proposed a trap-based approach to model the steady and dynamic charging of nitride trap based memory. The model is based on the Shockley-Reed-Hall statistics [18]. In this model, knowledge of the trap density, trap capture cross section area, actual trap energy and trap distribution in the nitride is important. The probability of a trap is occupied, at energy level \( E_t \), can be calculated as follows:

\[
\frac{df_i(t, E_t)}{dt} = c_n + e_p - (e_n + c_n + e_p + c_p)f_i
\]  

\[\text{(2.6.14)}\]
where $c_n$, $c_p$, $e_n$, and $e_p$ are capture and emission rate for electrons and holes. Capture and emission rate for electrons and holes depend on trap characteristics such as cross section area, density, energy and depth. Electric field between the traps and the interface also influence $c_n$, $c_p$, $e_n$, and $e_p$. Surface potential, $\psi_S$, depends upon the gate voltage and can be used to obtain the carrier densities as follows:

$$V_{CG}(t) = V_{FB} - \psi_S(t) - \frac{Q_{SC}(\psi, t)}{C_{ox-eq}} + \frac{Q_{ot}(t)}{C_2} \quad \ldots (2.6.15)$$

where $V_{CG}(t)$ is the time varying gate voltage, $V_{FB}$ is the flat band voltage of the device without any stored charges, $Q_{SC}$, is the semiconductor charge in the bulk, $C_2$ is the control gate coupling capacitor, $Q_{ot}$ represents oxide trapped charge, and $C_{ox-eq}$ is the equivalent MOS capacitance given by $(C_1C_2)/(C_1+C_2)$ ($C_1$ is substrate coupling capacitor).

Equations (2.6.14 - 15) can be solved to obtain the I-V curves and the time evolution of trapped charge for the SONOS cell when a time varying voltage is applied on the control gate. It can be seen that the effect of control gate currents are ignored from the model. As a result, steady state is achieved when the current through the substrate balances i.e. current "in" equals the current "out". In summary, the following key points can be identified:

(i) Shockley-Reed-Hall statistics and surface potential is used to model the FLASH charging behavior.

(ii) Information regarding trap energy and trap distribution is required. This information is hard to obtain and still not clearly understood.

(iii) Forward and reverse currents through the control gate are ignored.
Lundstrom et al. [12] in 1972 studied nitride-oxide layer in MNOS devices with oxide thickness less than 50 Å. An extension to current continuity model of Lenzlinger et al. [10] was proposed to predict the charging behavior of the MNOS devices. Specifically, Modified Fowler Nordheim Tunneling (MFNT) current is proposed as the mechanism for charge storage when a positive gate voltage is applied to the MNOS structure. In the case of tunnel oxide, electrons see a trapezoidal oxide barrier and an additional nitride barrier during the charging of the nitride layer (Figure 2.6.3). The proposed equation for MFNT is as follows:

\[ J_{ox} = C_{FN} E_{ox}^2 P_{ox} P_N \] .... (2.6.16)

where \( C_{FN} \) is a constant from the Fowler Nordheim Tunneling current equation for electrons tunneling from silicon into silicon dioxide, \( E_{ox} \) is the electric field on the tunnel oxide, and \( P_{ox} \) and \( P_N \) are calculated using the WKB approximation to find the electron transmission probabilities through the oxide and nitride respectively. Equations for \( P_{ox} \) and \( P_N \) are as follows:

\[ P_{ox} = \exp \left\{ -\frac{4}{3h} \sqrt{2q m_{ox}} \left[ \phi_{1}^{3/2} - \left( \phi_{1} - E_{ox} W_{ox} \right)^{3/2} \right] \right\} \] .... (2.6.17)

\[ P_{N} = \exp \left\{ -\frac{4}{3h} \sqrt{2q m_N} \left[ \left( \phi_{1} - \phi_{2} - E_{ox} W_{ox} \right)^{3/2} \right] \right\} \] .... (2.6.18)
where $E_N$ is the nitride electric field, $m_{ox}$ and $m_N$ are effective tunnel mass of electron in the oxide and nitride, respectively, $\Phi_1$, and $\Phi_2$ are the potential energy barrier heights seen by the electron in the conduction bands of oxide and nitride, respectively, $W_{ox}$ is the thickness of the oxide, and $h = h / (2\pi)$ with $h$ being the Planck's constant.

![Figure 2.6.3: A schematic band diagram representation of MNOS device with positive gate bias voltage [12]](image)

If $\Phi_1 \leq (\Phi_2 - E_{ox}W_{ox})$, $P_N$ is equal to 1 in equation (2.6.16) and the electrons tunnel through a trapezoidal oxide barrier only. This condition is known as Direct Tunneling. If $\Phi_1 \leq E_{ox}W_{ox}$, the equation 2.6.16 becomes equivalent to that of Fowler Nordheim Tunneling and $P_N$ is neglected. In the case of FNT, electrons tunnel through a triangular oxide barrier only. The distance, $X_N$, electrons are injected into the nitride for the case of MFNT is given as:

$$X_N = \frac{E_N}{\epsilon_{ox}} \left( \frac{\phi_1 - \phi_2 - E_{ox}W_{ox}}{E_{ox}} \right)$$

\[ (2.6.19) \]
where $\varepsilon_n$ and $\varepsilon_{ox}$ are relative dielectric constant for nitride and silicon dioxide, respectively.

While using the proposed model to predict the MNOS behavior, Lundstrom et al. found disagreement with the experimental results for low nitride fields and thin oxides. Good agreement with experimental results was obtained with thicker oxides. The disagreement at low nitride fields and thin oxide was attributed to extra current due to direct tunneling of electrons into the forbidden gap of the nitride at low nitride fields. This direct tunneling current of electrons from the silicon conduction band into the forbidden gap of the nitride traps is not presents at high oxide voltage drops because the traps are located below the silicon conduction bands. In the case of thick oxide, oxide repulsion is enough to inhibit the excess current at low fields. This excess current is given as follows:

$$J_N = C_N \exp \left( G \sqrt{E_N} \right) \quad \ldots (2.6.20)$$

where $C_N = 2.5 \times 10^{-16}$ A/m\(^2\) and $G = 1.5 \times 10^{-8}$ (m/V\(^{1/2}\)). In summary, the following key points can be identified:

(i) Extension to Lenzlinger et al. [10] current continuity model to predict MNOS behavior was successfully demonstrated with the introduction of Modified Fowler Nordheim Tunneling for tunnel oxide thinner than 50 Å.

(ii) Disagreement between the model and the experiment for thin oxide at low nitride fields was attributed to an excess current due to direct tunneling into the forbidden gap nitride traps.
De Salvo et al. [9] studied charge conduction through nitride/oxide dual-layer and oxide/nitride/oxide tri-layer films. Role of each single layer dielectric in multilayer structure conduction mechanism was studied by analyzing the experimental data. The authors proposed that the current conduction through multilayer structures can be described by convoluting current conduction mechanisms of single-layer layer dielectrics.

The current continuity model proposed that the conduction of electrons in nitride/oxide dual-layer is from the interface trap states in the nitride forbidden gap through the oxide into the silicon conduction band. This model assumes that a trap level at the oxide/nitride interface is at an energy $\Phi_t (3.5 \text{ eV})$ below the conduction band of the silicon dioxide. Current emission from discrete traps can be calculated as follows:

$$J_{\text{SiO}_2-\text{SiN}} = q f_0 N_{st} \Theta(\phi_t) \quad \ldots (2.6.21)$$

where $f_0$ is the electron escape frequency, $N_{st}$ is the trap density with the value $10^{12} \text{ /cm}^2$, and $\Theta(\Phi_t)$ is the transmission coefficient of the trap barrier. The transmission coefficient is obtained from the WKB approximation as follows:

$$\Theta(\phi_t) = \exp \left( -\frac{8\pi \sqrt{2m_{\text{ox}}} \phi_t^{3/2}}{3hq} \frac{\phi_t}{F} \right) \quad \ldots (2.6.22)$$

where $m_{\text{ox}}$ is the effective mass of electron in the oxide band gap, and $F$ is the electric field. Electron escape frequency, $f_0$, can be obtained from the localization energy $E_{\text{loc}}$ as:
\[ f_0 = \frac{E_{loc}}{h} \]
where \( E_{loc} = \frac{3\hbar^2}{2m_\alpha} \left( \frac{\pi}{a} \right)^2 \)
with \( a \approx \sqrt{\frac{\sigma}{\pi}} \) \( \ldots (2.6.23) \)

where \( \sigma \) is the trap capture cross section area equal to \( 3 \times 10^{15} \text{ cm}^2 \).

The continuity model does not account for the occupancy factor of the trap which leads to unreasonable direct tunneling transparency in equation 2.6.21 - 22, \( J_{SiO2-SiN} \), for very thin tunnel oxide. To overcome this shortcoming of the continuity model, De Salvo et al. [9] proposed a two-step trap-assisted conduction mechanism. In the two-step trap-assisted model, the rate of change of electron concentration in the traps can obtained as follows:

\[ \frac{\partial(f_t)}{\partial t} = J_N (1 - f_t) - J_o f_t \]
\( \ldots (2.6.24) \)

where \( f_t \) is the probability of occupation of traps at the nitride/oxide interface, \( J_N \) is the supplying current and the \( J_o \) is the emptying current. Under steady state condition when \( \frac{\partial(f_t)}{\partial t} = 0 \), the occupancy probability \( (f_t) \), \( J_{NO} \) nitride-oxide current, tunnel oxide field \( F_{ox} \), and nitride field \( F_N \) can be found as follows:

\[ f_t = \frac{J_N}{J_N + J_o} \]
\( \ldots (2.6.25) \)

\[ J_{NO} = \frac{J_N J_o}{J_N + J_o} \]
\( \ldots (2.6.26) \)

\[ \varepsilon_{ox} F_{ox} - \varepsilon_N F_N = qN_o f_t \]
\( \ldots (2.6.27) \)
where equation 2.6.27 is obtained from Gauss’s Law. In summary, the following key points can be identified:

(i) Convolution of single-layer dielectric current conduction mechanism is used to formulate current conduction for multi-layer dielectric structures.

(ii) The current continuity model takes into account the interface nitride/oxide traps.

(iii) Two-step trap-assisted model is proposed to account for the occupancy factor of the discrete traps.
CHAPTER 3

PROPOSED MODEL

3.1 Motivation for Research

Tan et al. [7] investigated hafnium aluminum oxide (HfAlO) alloy as a possible replacement of HfO₂ high-κ material in Silicon Oxide High-κ Oxide Silicon (SOHOS) FLASH devices. Specifically, following two approaches were recommended to obtain better performance:

(i) Device 1: Replace the nitride storage layer in SONOS with HfAlO high-κ dielectric material.

(ii) Device 2: Replace the blocking oxide in SONOS with HfAlO high-κ dielectric material alloy.

HfAlO combines the advantages of better charge storage at low programming voltages, faster programming, and less over-erase of HfO₂ and good charge-retention of aluminum oxide (Al₂O₃). The second approach of replacing the blocking oxide by HfAlO alloy leads to increased voltage drop on the tunnel oxide due to higher dielectric constant of HfAlO. This increase on the tunnel oxide leads to higher current injection and hence efficient program/erase processes. As it was explained in Chapter 2, injection currents have exponential dependence on the electric field.

The model proposed in this thesis is used to obtain the threshold voltage shift vs. program-time curves for the gate voltages used by Tan et al. [7] to conduct experiments
on the two proposed devices denoted as device 1 and device 2. Moreover, threshold voltage shift vs. program time curve for an intermediate gate voltage of 8V is obtained and presented for device 2. The approach in developing the model is to relate the density of deep level traps in the nitride for device 2 as a function of the voltage drop on the tunnel oxide and nitride storage layer. The formula presented for the density of trap can be used to get an estimate for the maximum threshold voltage shift for all types of SONOS type devices with high-κ blocking oxide. The proposed model also provides an indication for the location of the nitride traps where majority of the charges are stored.

3.2 Device Architecture and Fabrication

Devices that form the center of this study were fabricated by Tan et al. [7]. A schematic representation of the architecture of these devices is presented in Figure 3.2.1. Tunnel oxide of thickness 25 Å was thermally grown at 800 °C over (100) p-type silicon substrate. Silicon Nitride was deposited using low pressure chemical vapor deposition (LPCVD) with the thickness of 60 Å for the SONOS type devices with high-κ storage layer and with the thickness of 50 Å for the SONOS type devices with high-κ blocking oxide. Pure HfO₂ and Al₂O₃ were deposited by atomic layer deposition (ALD) for SONOS type FLASH with high-κ storage layer. The thickness of the storage layer was kept at 60 Å for all SONOS type devices with high-κ storage layer. For SONOS type devices with high-κ storage layer, HfAlO (HfO₂ with 10% Al₂O₃ concentration) films were deposited by metal organic chemical vapor deposition using a single cocktail source. The blocking oxide with thickness 55 – 75 Å was deposited using LPCVD. Pure HfO₂, Al₂O₃, or HfAlO alloys (with 52% or 85% Al₂O₃ concentration) were deposited

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
using ALD with the thickness of 75 Å for the SONOS type devices with high-κ blocking oxide. HfN/TaN control gates were deposited using physical vapor deposition for all the fabricated devices. Additionally, all transistors underwent source/drain implantation followed by activation anneal at 950 °C for 30 s. Transistors have the dimensions of $W/L$ of 100 μm/20 μm ($W$ is the gate width and $L$ is the gate length) [7].

Figure 3.2.1: (a) SONOS/SOHOS device architecture with either Si$_3$N$_4$/HfAlO/HfO$_2$/Al$_2$O$_3$ storage layer. (b) SONOS device architecture with high-κ blocking oxide of either SiO$_2$/Al$_2$O$_3$/HfO$_2$/HfO$_2$/0.46(Al$_2$O$_3$)$_{0.52}$(HfO$_2$)$_{0.15}$(Al$_2$O$_3$)$_{0.85}$ [7]  

3.3 Model  

The model presented in this thesis is based upon the “current continuity” model described briefly in Chapter 2 Section 6. Lenzlinger et al. [10] stated that the FLASH device does not work as memory if the device is operated under steady state-conditions. The condition for steady-state is that the total current into the nitride is equal to the current leaving the nitride i.e. $J_{IN} = J_{OUT}$. Our model incorporates the information
regarding the trap cross section and trap density to build upon the existing current continuity model to predict SONOS programming characteristics.

Similar to the continuity model of Lundström et al. [12], nitride current conduction due to Poole-Frenkel emission is ignored due to the facts that large activation energy is required for the electrons to free themselves from the traps and the existence of unfavorable electric field conditions. The fact that nitride has deep level charge storage traps has been widely reported [6, 29 – 30]. Low electric field “ohmic” current component of the nitride current is also ignored simply because large gate voltages are used for programming. Moreover, the effects of gate current are ignored because it is assumed that the blocking oxides are thick enough to prevent tunneling of electrons from the nitride into the gate.

Feeding current for the model are based on the electron limited oxide tunnel currents namely, FNT, DT and MFNT. All the traps are assumed to be neutral. Also, the assumption that a trap can only be occupied by one electron is made. Based on this assumption, the model is used to find the density of occupied traps as a function of conduction band potential energy difference between the silicon substrate, near silicon/tunnel oxide interface, and the silicon nitride, near nitride/blocking oxide interface, under applied gate bias. The model also predicts the location of the traps, interface or nitride bulk, where the charges are stored.
3.3.1 Electrostatics

The electrostatic calculation for the SONOS device is rather simple and is based on the well known MOS theory [2, 18]. "Flat Band" condition is defined under which there is no voltage drop on any of the materials in the SONOS structure as depicted in Figure 3.3.1.1. For p-Silicon substrate, flat band voltage, \( V_{FB} \), can be calculated by energy balance of the band diagram as follows:

\[
V_{FB} = - \left( \Phi_M - (\Phi_B + E_g / 2 + \Phi_F) \right) / q
\]  

\[ \ldots (3.3.1.1) \]

where \( q \) is the charge of electron, \( \Phi_M \) is the potential energy barrier seen by the electrons in the conduction band of the control gate, \( \Phi_B \) is the Si/SiO\(_2\) electron barrier, and \( E_g \) is the

41

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
silicon band gap, $\Phi_F$ is the potential related to the Fermi level $= \left( \frac{kT}{q} \ln \left( \frac{N_A}{n_i} \right) \right)$ where $k$ is the Boltzmann constant, $N_A$ is the p-Si substrate doping, and $n_i$ is the intrinsic carrier concentration. When a positive gate voltage is applied to the structure, an inversion layer is formed under the tunnel oxide. The probability of electrons tunneling from the silicon substrate is obtained using the WKB approximation as discussed in Chapter 2 Section 3.2.

![Schematic band diagram depicting positive gate voltage band bending](image)

Figure 3.3.1.2: A schematic band diagram depicting positive gate voltage band bending for SONOS structure

Electrons tunnel from the inversion layer at the Si/SiO$_2$ interface to the storage layer for the voltages used to program the FLASH memory. Total voltage drop on the oxide can be obtained from the energy balance equation as follows:
where \(V_g\) is the applied gate voltage, \(V_{OX}\) is the voltage drop on the dielectric layers, and \(\psi_b(F_i)\) is the silicon substrate band bending due the electric field on the tunnel oxide \(F_i\). These quantities are represented pictorially in Figure 3.3.1.2. Combining equations 3.3.1.1 & 3.3.1.2, equation 3.3.1.2 can be rewritten as [26]:

\[
V_g = V_{OX} + \psi_b(F_i) + V_{FB} \quad \text{.... (3.3.1.3)}
\]

Electric field (\(F_{OX}\)) can be calculated using the above equation as:

\[
F_{OX} = \frac{V_g - (\psi_b(F_i) + V_{FB})}{t_{eq}} \quad \text{.... (3.3.1.4)}
\]

In equation 3.3.1.4, \(t_{eq}\) is the equivalent oxide thickness which is given by

\[
t_{eq} = t_{BO} \frac{\varepsilon_{OX}}{\varepsilon_{BO}} + t_{ST} \frac{\varepsilon_{OX}}{\varepsilon_{ST}} + t_{TO} \frac{\varepsilon_{OX}}{\varepsilon_{TO}} \quad \text{.... (3.3.1.5)}
\]

where subscript \(t_{BO}\), \(t_{ST}\), \(t_{TO}\), and \(t_{OX}\) represents the thickness of the blocking oxide, storage layer, tunnel oxide, and silicon dioxide, respectively whereas \(\varepsilon_{BO}\), \(\varepsilon_{ST}\), \(\varepsilon_{TO}\), and
\( \varepsilon_{\text{OX}} \) represents the relative dielectric constants for the blocking oxide, storage layer, tunnel oxide, and silicon dioxide, respectively.

The field \((F_{\text{OX}})\) is calculated once during the initial electrostatic calculations of the structure under applied bias. The same field is used to calculate the tunnel current densities as it is assumed that the traps are isolated and that the traps do not interact with each other. Furthermore, a restriction of single occupancy for traps, i.e., only one electron can occupy a trap, is placed. In other words, electron tunneling and storage is allowed only when empty traps in the storage layer are available and, the electric field for empty traps is favorable for occupancy. Spatial energy band diagrams for device 1, under flat band condition, with (a) silicon nitride, (b) hafnium aluminum oxide, and (c) aluminum oxide storage layer are shown in Figure 3.3.1.3.

![Spatial energy band diagrams](image)

Figure 3.3.1.3: Spatial energy band diagrams depicting flat band condition for SONOS device with (a) Si\(_3\)N\(_4\), (b) HfAlO, and (c) Al\(_2\)O\(_3\) storage layer [7]
3.3.2 Current Tunneling

Electron tunneling from the p-Si substrate into the storage layer forms the current used in programming the FLASH cell. As explained in Chapter 2, tunneling depends on the distribution of the occupied states in the injected material, and on the shape, height and width of the potential energy barrier [2, 3, 5, 13]. Probability of carrier tunneling is calculated using the WKB approximation. The spatial dependence of the oxide barrier seen by the electrons in the conduction band of the substrate heavily depends upon the voltage drop on the tunnel oxide. Voltage drop on the tunnel oxide, $V_{TO}$, can be found using the oxide electric field, $F_{OX}$, and the thickness of the tunnel oxide, $t_{TO}$, as:

$$V_{TO} = F_{OX}t_{TO} \quad \ldots \quad (3.3.2.1)$$

Following are the three cases of tunneling which depends upon the voltage drop on the tunnel oxide:

(i) $V_{TO} \geq \Phi_B$: Fowler Nordheim Tunneling

(ii) $(\Phi_B - \Phi_S) < V_{TO} < \Phi_B$: Direct Tunneling

(iii) $[\Phi_B - (\Phi_S + \Phi_i)] < V_{TO} < (\Phi_B - \Phi_S)$: Modified Fowler Nordheim Tunneling

where $\Phi_i$ is the voltage drop on the nitride.

A general expression for quantum mechanical tunneling current density, $J$, can be obtained by using an independent electron approximation and an elastic tunneling process. The current density, $J$, for a parabolic dispersion relation with transversal energy component, $E_t$, and effective electron mass, $m_t$ is given by [27]:

45
where $E$ is measured from the Si conduction band edge at the Si/SiO$_2$ interface and represents the kinetic energy of the tunneling electron, and $T_t$ is the tunneling transmission probability calculated using the WKB approximation and is given by [27]:

$$T_t(E, k_t) = \exp \left[ -2 \int_0^{x_1} k_{ox} (E, k_t, x) \, dx \right] \quad \text{....(3.3.2.3)}$$

where $k_{ox}$ is the imaginary part of the complex electron wave vector of the tunneling electron, $x_1$ is the distance from the Si/SiO$_2$ interface to the classical turning point, and $k_t$ is the transversal component of the tunneling electron wave vector. Equations 3.3.2.2 & 3.3.2.3 can be used to obtain the current density equations for FNT, DT, and MFNT for different cases of classical turning point, $x_1$, as follows [8, 27]:

$$J_{FNT} = \frac{q^3 m_o}{8nA(q\phi_B)m_{ox}} F_{ox} \exp \left( -\frac{8\pi \sqrt{2m_{ox}(q\phi_B)^3}}{3qhF_{ox}} \right) \quad \text{....(3.3.2.4)}$$

$$J_{DT} = \frac{AF_{ox}^2}{\left(1 - \sqrt{1 - V_{TO}/\phi_B} \right)^2} \exp \left( -\frac{B}{F_{ox}} \left( 1 - \left( 1 - \frac{V_{TO}}{\phi_B} \right)^{3/2} \right) \right) \quad \text{....(3.3.2.5)}$$
where $q$ is the charge of a single electron, $\Phi_B$ is the energy barrier between Si-SiO$_2$ interface, $h$ is the Planck’s constant, $F_{ox}$ is the electric field on the oxide, $m_o$ is the mass of free electron, $m_{ox}$ is the effective mass of electron in SiO$_2$, $V_{TO}$ is the voltage drop on the tunnel oxide, $m_{ST}$ is the effective mass of electron in the storage layer, $\varepsilon_{ST}$ and $\varepsilon_{ox}$ are the dielectric constants of the storage layer and the oxide, respectively, and $A$ and $B$ are constants given as $A = \frac{q^3 m_s}{8 \pi \hbar (q \phi_B) m_{ox}}$ and $B = \frac{8 \pi \sqrt{2 m_{ox} (q \phi_B)^2}}{3 q h}$, respectively. FNT, DT, and MFNT mechanisms are shown in Figures 3.3.2.1 and 3.3.2.2.

Figure 3.3.2.2: Schematic energy band diagrams depicting FNT for classical turning point $x_1$ [8]
3.3.3 Threshold Voltage Shift

Current density expressions given by equations (3.3.2.4 - 3.3.2.6) from Section 3.3.2 are integrated over time to find the amount of charges stored per unit area during a programming step. The differential accumulated charge density, $\Delta Q_I$, in a differential time, $\Delta t$, is given by:

$$\Delta Q_I = \frac{J_{\text{IN}} N_{Tt}(t) \sigma_T}{q} \Delta t$$  \hspace{1cm} (3.3.3.1)

where $J_{\text{IN}}$ is the either FNT, or DT, or MFNT, $N_{Tt}(t)$ is the net density of available traps per unit area, $\sigma_T$ is the area of the trap, and $q$ is the charge of a single electron.
Number of charges stored is calculated for each time step and then added into the total stored charges from the previous time steps of the program cycle. In other words, 
\[ Q_{\text{tot}}(t) = Q_{\text{tot}}(t - \Delta t) + \Delta Q_t. \]
Simultaneously, the number of charges stored, \( Q_{\text{tot}}(t) \), for the programming time is subtracted from the available density of traps. The balanced density of available traps per unit area is used as \( N_{\text{Ti}}(t + \Delta t) \) for the next time step. This process is repeated until, either all the traps are occupied or the programming time runs out. The threshold voltage shift, \( \Delta V_{\text{TH}} \) defined as the difference between the instantaneous threshold voltage and the initial threshold voltage, due to stored charges is calculated for each time step as follows [19]:

\[
\Delta V_{\text{TH}}(t) = \frac{qQ_{\text{tot}}(t)}{\varepsilon_{\text{ox}}} \left( t_{BO} \frac{\varepsilon_{\text{ox}}}{\varepsilon_{BO}} + t_{ST} \frac{\varepsilon_{\text{ox}}}{\varepsilon_{ST}} \right) \quad .... (3.3.2)
\]

where \( Q_{\text{tot}}(t) \) is the total charge including the charges from all the previous time step and current time step summed up together. Values for \( \Delta V_{\text{TH}}(t) \) can be plotted against programming time for different gate voltages to predict/confirm FLASH programming behavior.

3.3.4 Relative Dielectric Constant

Note that the current density, electrostatic, and the threshold voltage shift depend on the dielectric constant of the tunnel oxide, the storage layer, and the blocking oxide. Values for the relative dielectric constants of various materials used in our research are reported in Table 3.3.4.1. Dielectric constant values for Silicon Nitride, Hafnium Oxide,
and Aluminum Oxide have been widely reported [7–8, 18]. The dielectric constant used for Si₃N₄ (7.5) is close to the reported values. To calculate the dielectric constants for Hafnium Aluminum Oxide alloys, following weighted sum formula is used:

\[ \varepsilon_{HfAlO} = x\left(\varepsilon_{HfO_2}\right) + y\left(\varepsilon_{Al_2O_3}\right) \]  \hspace{1cm} (3.3.4.1)

where x and y denotes the percentage of HfO₂ and Al₂O₃ in the HfAlO alloy, respectively.

Table 3.3.4.1: Dielectric constants for the material used in our study

<table>
<thead>
<tr>
<th>Material</th>
<th>Relative dielectric constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>11.8</td>
</tr>
<tr>
<td>SiO₂</td>
<td>3.9</td>
</tr>
<tr>
<td>Si₃N₄</td>
<td>7.5</td>
</tr>
<tr>
<td>Al₂O₃</td>
<td>9</td>
</tr>
<tr>
<td>HfO₂</td>
<td>25</td>
</tr>
<tr>
<td>HfAlO₁₀% Aluminum Oxide</td>
<td>23.4</td>
</tr>
<tr>
<td>HfAlO₅₂% Aluminum Oxide</td>
<td>16.68</td>
</tr>
<tr>
<td>HfAlO₈₅% Aluminum Oxide</td>
<td>11.4</td>
</tr>
</tbody>
</table>
CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, the procedure for obtaining the programming characteristics for SONOS type devices proposed by Tan et al. [7] is presented. Results obtained for several different SONOS type structures with various storage and blocking dielectrics are presented and compared with the experimental results provided by Tan et al [7]. Based on the results, trap density in silicon nitride is modeled as a function of energy. All results are discussed in the light of available physical insight.

4.2 Procedure

To model the programming characteristics of the SONOS device, the electrostatics are calculated first for the flat band condition and for the case when a gate voltage is applied to program the device using equations 3.3.1.1 & 3.3.1.3 from Section 3.3, respectively. Along with the information of applied gate voltage, silicon substrate band banding, flat band voltage, and equivalent oxide thickness, the electrostatic condition for the device under applied bias can be used to determine the electric field over the Oxide-Storage Layer-Oxide (OSO) structure from equations 3.3.1.4 & 3.3.1.5. The tunnel oxide voltage drop is a function of the OSO stack electric field and is given by equation 3.3.2.1.
Tunnel oxide voltage drop is needed to determine the classical turning point, $x_1$, in the tunneling transmission probability, $T_t$, obtained from the WKB approximation (equation 3.3.2.3). Hence, the voltage drop on the tunnel oxide determines the shape of the tunnel oxide barrier seen by the electrons in the conduction band of the silicon substrate (inversion channel at the Si/SiO$_2$ interface) and the type of quantum mechanical tunneling, i.e., Fowler Nordheim Tunneling or Direct Tunneling.

To distinguish between the Direct Tunneling and Modified Fowler Nordheim Tunneling, the voltage drop on the storage layer is also required. The storage layer voltage drop, $\Phi_h$, can be obtained in a similar manner to the voltage drop on the tunnel oxide as

$$\phi_i = F_{ox}t_{ST} \frac{\varepsilon_{ox}}{\varepsilon_{ST}}$$

where $t_{ST}$ is the thickness of the storage layer, and $\varepsilon_{ox}$ and $\varepsilon_{ST}$ are the relative dielectric constants for the tunnel oxide and the storage layer, respectively. The conditions for MFNT and DT are described in Chapter 3 Section 3.

Equation 3.3.3.1 gives the accumulated charge, $\Delta Q_h$, in the storage layer for a chosen time step ($10^{-7} \text{ s} - 10^{-5} \text{ s}$) during programming cycle. Accumulated charge depends on the tunneling current density, $J$, and the current density equations can be found from equations 3.3.2.4 - 3.3.2.6 for FNT, DT, or MFNT, respectively. The shift from the nominal threshold voltage, $\Delta V_{TH}$, of the device is a function of the total accumulated charge, $Q_{tot}(t)$, and is described in equation 3.3.3.2. The shift in the threshold voltage, obtained from our model, is plotted against the programming time along with the experimental results of Tan et al. [7]. The density of traps, $N_{TH}$, in equation 3.3.3.1 is used as a fitting parameter in the model. The value for the density of traps is obtained using the final/saturation value for the threshold voltage shift from Tan et al. experimental results.
and the equation 3.3.3.2. $\Delta V_{TH}$ saturates once all the available traps, corresponding to the programming gate voltage, in the storage layer are occupied by electrons.

4.3 Effect of Storage Layer on Programming Efficiency of SONOS FLASH

The architecture of the structures used to study the effect of replacing the traditional silicon nitride storage layer in SONOS FLASH, with high-κ dielectrics, on the programming efficiency, is shown in Figure 4.3.1. The thickness of the tunnel oxide, storage layer, and blocking oxide are kept constant at 2.5 nm, 6 nm, and 5.5 nm, respectively.

![Figure 4.3.1: A schematic representation for the architecture of SONOS type devices with various storage layers, namely Si$_3$N$_4$, Al$_2$O$_3$, and HfAlO [7]](image)

Density of traps in the storage layer is used as a fitting parameter in the model. The number of traps for each material is obtained from the saturated/final value for the threshold voltage in the experimental data. For the SONOS device with Si$_3$N$_4$ as the

53
storage layer, the value for $\Phi_B$ (Si/SiO$_2$ conduction band potential barrier) and $\Phi_s$ (Si$_3$N$_4$/SiO$_2$ conduction band potential barrier) are taken to be 3.0 eV and 1.2 eV, respectively. The values used for the potential barriers in our model fits well with the reported values of 3.1 eV and 1.1 eV for $\Phi_B$ and $\Phi_s$, respectively [7]. The density of available traps, $N_{Tt}$, in Si$_3$N$_4$ for charge storage ($1.75 \times 10^{16}$ m$^{-2}$) compares well with the value of interface trap density of $10^{16}$ m$^{-2}$ reported in the literature [9]. Large trap capture cross section area of $10^{-17}$ m$^2$ is used and the value compares well with the reported one by Yang et al. [22].

For SONOS with Al$_2$O$_3$ storage layer, $\Phi_B$ is taken to be equal to 2.75 eV and $\Phi_s$ (Al$_2$O$_3$/SiO$_2$ electron potential barrier) to be equal to 0.4 eV. The reported value for $\Phi_B$ and $\Phi_s$ for Al$_2$O$_3$ case, by Tan et al. [7], are 3.1 eV and 0.4 eV, respectively. Trap density of $1.1 \times 10^{16}$ m$^{-2}$ and trap cross section area of $10^{-16}$ m$^2$ used for Al$_2$O$_3$ storage layer are well below the maximum trap density $3.2 \times 10^{17}$ m$^{-2}$ and $9.3 \times 10^{-19}$ m$^2$ reported by Ortiz et al. [20]. Large trap cross section area is used to allow for good charge retention behavior shown by Al$_2$O$_3$ [7, 12].

For the HfAlO storage layer SONOS devices, 3.1 eV and 1.47 eV values, reported by Tan et al. [7], are used for $\Phi_B$ and $\Phi_s$ (HfAlO/SiO$_2$ electron potential barrier), respectively. Trap density of $5.97 \times 10^{16}$ m$^2$ is used and it compares well with the reported value of $10^{17}$ m$^{-2}$ for HfO$_2$ (HfAlO has 90% HfO$_2$) [6]. Small capture cross section of $10^{-18}$ m$^2$ for electrons is used similar to small cross section area for holes reported by Yousif et al. [21] for hafnium oxide case.

Effective mass of electron in the oxide, $m_{ox}$, is kept constant for all devices at $0.28m_o$, $m_o$ is the mass of electron in free space, and is close to the value $0.29m_o$ reported by
Depas et al. [27]. The constants used in our model such as, Si/SiO$_2$ conduction band barrier ($\Phi_B$), storage layer/silicon dioxide conduction band barrier ($\Phi_S$), effective mass of electron in oxide ($m_{ox}$), effective mass of electron in storage layer ($m_{st}$), fitted density of traps ($N_{7T}$), and trap cross section area ($\sigma$), used in the model are listed in Table 4.3.1.

Table 4.3.1: List of parameters used in the model to study SONOS programming characteristics with various storage layers.

<table>
<thead>
<tr>
<th>Storage Layer</th>
<th>$\Phi_B$ (eV)</th>
<th>$\Phi_S$ (eV)</th>
<th>$m_{ox}$</th>
<th>$m_{st}$</th>
<th>$N_{7T}$ (m$^{-2}$)</th>
<th>$\sigma$ (m$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HfAlO</td>
<td>3.1</td>
<td>1.47</td>
<td>0.28 $m_o$</td>
<td>0.28 $m_o$</td>
<td>5.97E+16</td>
<td>1.00E-18</td>
</tr>
<tr>
<td>Si$_3$N$_4$</td>
<td>1.0</td>
<td>1.2</td>
<td>0.28 $m_o$</td>
<td>0.28 $m_o$</td>
<td>1.75E+16</td>
<td>1.00E-17</td>
</tr>
<tr>
<td>Al$_2$O$_3$</td>
<td>2.75</td>
<td>0.4</td>
<td>0.28 $m_o$</td>
<td>0.28 $m_o$</td>
<td>1.10E+16</td>
<td>1.00E-16</td>
</tr>
</tbody>
</table>

The programming characteristics of the SONOS type devices with three different storage layers, namely hafnium aluminum oxide, silicon nitride, and aluminum oxide are shown in Figure 4.3.2. The data for the shift in threshold voltage as a function of program time, obtained from the model, is plotted among the experimental results provided by Tan et al. [7]. The agreement between the results from the model and the experimental data for Si$_3$N$_4$, Al$_2$O$_3$, and HfAlO are within ±20%, ±28%, and ±24%, respectively. Hafnium aluminum oxide alloy (mixture of 90% HfO$_2$ and 10% Al$_2$O$_3$) shows the fastest programming speed among the three storage layers. Relatively higher dielectric constant, $\kappa$, value for HfAlO of 23.4, leads to smaller voltage drop on the storage dielectric and hence increased electric field on the tunnel-oxide. Increased electric fields, translates to higher injection current density and hence better programming efficiency. Combined with large high energy trap density, the threshold voltage shift obtained by using HfO$_2$ is
greater than for the case of Si$_3$N$_4$ under same applied gate bias. Even though the dielectric constant for Al$_2$O$_3$ (9) is greater than that of Si$_3$N$_4$ (7.5), the threshold voltage shift is smaller for Al$_2$O$_3$ because the trap density is suspected to be low [12]. The discrepancy between the threshold voltage saturation time from the model and the experimental data can be attributed to the fact that the reverse/de-trapping current from Al$_2$O$_3$ is ignored. It can be concluded that Tan et al. [7] successfully showed that HfAlO can be a possible replacement for Si$_3$N$_4$ in SONOS devices, for lower programming voltages, alleviating the need for smaller tunnel oxide thickness to increase performance.

Figure 4.3.2: Programming characteristics ($\Delta V_{th}(t)$ vs. program time) of SONOS FLASH with storage layer (i) HfAlO, (ii) Si$_3$N$_4$, and (iii) Al$_2$O$_3$
4.4 Effect of High-κ Dielectric Blocking Oxide and Gate Voltage on the Programming Efficiency of the SONOS FLASH

The architecture of the structures used to study the effect of replacing the traditional silicon dioxide blocking dielectric in SONOS FLASH with high-κ dielectrics, on the programming efficiency, is shown in Figure 4.4.1. The thickness of the tunnel oxide, storage layer, and blocking oxide are kept constant at 2.5 nm, 5.0 nm, and 7.5 nm, respectively. Transient programming characteristics of the SONOS devices are studied at three gate biases, $V_{g} - V_{fb} = 6V$, 7V, and 9V.

![Figure 4.4.1: A schematic representation for the architecture of SONOS type devices with various blocking oxides, namely SiO₂, Al₂O₃, HfO₂, HfAlO (with 15% HfO₂ concentration), and HfAlO (with 48% HfO₂ concentration) [7]](image)

Similar to the case of SONOS devices with high-κ storage dielectrics, the density of traps in Si₃N₄, storage layer, for SONOS type devices with high-κ blocking oxides is used as a fitting parameter to generate the threshold voltage shift vs. program time curves. The
number of traps for each device structure is obtained from the saturated/final value of the threshold voltage reported in the experimental data by Tan et al. [7]. The density of available traps, \(N_{t1}\), in Si\(_3\)N\(_4\) ranges from \(4.2 \times 10^{15}\) m\(^{-2}\), for SiO\(_2\) blocking oxide case under gate bias, \(V_g - V_{fb}\), of 6V, to \(1.25 \times 10^{17}\) m\(^{-2}\), for HfO\(_2\) blocking oxide case under gate bias, \(V_g - V_{fb}\), of 9V. \(N_{t1}\) for 9V HfO\(_2\) case is one order magnitude greater than the reported value of interface trap density \(10^{16}\) m\(^{-2}\) for Si\(_3\)N\(_4\)/SiO\(_2\) [9]. This leads to the speculation that the bulk traps in the nitride at different energy levels, and not the interface states, play dominant role of charge storage.

The value for \(\Phi_B\) (Si/SiO\(_2\) conduction band potential barrier) and \(\Phi_s\) (Si\(_3\)N\(_4\)/SiO\(_2\) conduction band potential barrier) are taken to be equal to the reported value, by Tan et al. [7], of 3.1 eV and 1.1 eV, respectively. Large capture cross section area for nitride traps, reported by Yang et al. [22], with the value of \(5 \times 10^{-17}\) m\(^2\) is used to model the programming behavior. Effective mass of electron in oxide, \(m_{ox}\), ranges from 0.28\(m_o\) - 0.32\(m_o\) (\(m_o\) is the mass of electron in free space) and fits well the values of 0.29\(m_o\) - 0.36\(m_o\) reported by Depas et al. [27]. Model parameters such as, Si/SiO\(_2\) conduction band barrier \(\Phi_B\), storage layer/SiO\(_2\) conduction band barrier \(\Phi_s\), effective mass of electron in oxide \(m_{ox}\), effective mass of electron in storage layer \(m_s\), and trap cross section area \((\sigma)\) are listed in Table 4.4.1.

Tan et al. [7] proposed the idea of replacing the blocking oxide in the ONO stack with a high-\(k\) material to increase programming efficiency and lowering operating voltages without decreasing the tunnel-oxide thickness. These devices have tremendous commercial potential and provide good reliability by minimizing defects associated with ultra-thin oxides because the need to scale tunnel oxide can be minimized. Results,
$\Delta V_{th}(t)$ vs. program time, obtained from our model and Tan et al. experiments are shown in Figures 4.4.2 - 4.4.6 for the cases of SiO$_2$, Al$_2$O$_3$, HfO$_2$, HfAlO (with 15% HfO$_2$ conc.), and HfAlO (with 48% HfO$_2$ conc.) blocking oxide, respectively.

Table 4.4.1: List of parameters used in the model to study SONOS programming characteristics with high-$\kappa$ blocking oxide.

<table>
<thead>
<tr>
<th>Blocking Oxide</th>
<th>$\Phi_B$ (eV)</th>
<th>$\Phi_S$ (eV)</th>
<th>$m_{ox}/m_0$</th>
<th>$m_{H}/m_0$</th>
<th>$\sigma$ ($m^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO$_2$</td>
<td>3.1</td>
<td>1.1</td>
<td>0.31</td>
<td>0.31</td>
<td>5.00E-17</td>
</tr>
<tr>
<td>Al$_2$O$_3$</td>
<td>3.1</td>
<td>1.1</td>
<td>0.28</td>
<td>0.28</td>
<td>5.00E-17</td>
</tr>
<tr>
<td>HfO$_2$</td>
<td>3.1</td>
<td>1.1</td>
<td>0.31</td>
<td>0.31</td>
<td>5.00E-17</td>
</tr>
<tr>
<td>HfAlO (15% HfO$_2$)</td>
<td>3.1</td>
<td>1.1</td>
<td>0.28</td>
<td>0.28</td>
<td>5.00E-17</td>
</tr>
<tr>
<td>HfAlO (48% HfO$_2$)</td>
<td>3.1</td>
<td>1.1</td>
<td>0.30</td>
<td>0.30</td>
<td>5.00E-17</td>
</tr>
</tbody>
</table>

Figure 4.4.2: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS device with SiO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V$, 7V, and 9V.
Figure 4.4.3: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with Al$_2$O$_3$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6\text{V}, 7\text{V},$ and $9\text{V}$

Figure 4.4.4: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6\text{V}, 7\text{V},$ and $9\text{V}$
Figure 4.4.5: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with HfAlO$_{15\%}$HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V$, 7V, and 9V

Figure 4.4.6: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type device with HfAlO$_{48\%}$HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V$, 7V, and 9V

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Higher value of the threshold voltage shift can be observed in each cases of blocking oxide from Figure 4.4.2 - 4.4.6 for higher programming gate voltages, ranging from 6V - 9V. Higher gate voltages leads to increased voltage drop on the OSO stack. The threshold voltage shift saturates at a higher value because as the voltage drop on the storage layer increases, the number of traps at higher energy levels, for charge storage, also increases. This holds true regardless of the blocking oxide material. Moreover, the results obtained from the model also indicate that the program time required to reach saturation of threshold voltage shift also decreases with higher programming voltages. This again can be attributed to higher current densities for higher gate biases which lead to faster occupation of available traps in the storage layer.

The effects of using high-K blocking oxide can also be observed from Figure 4.4.2 - 4.4.6. For low programming voltage of 6V, HfO₂ case shows the best programming efficiency. Considering from the electrostatic perspective, high-K blocking oxide reduces the voltage drop on the blocking oxide while simultaneously increasing the voltage drops on the tunnel oxide and the nitride. Increase in tunnel oxide voltage drop translates to higher electric field on the tunnel oxide and thus, increased tunnel density because tunneling current densities exhibit an exponential dependence on the tunnel oxide electric field. However, for higher programming voltages, \( V_g - V_{fb} \) greater than 7V, the shift in threshold voltage relatively decreases with higher concentration of HfO₂. This decrease in performance can be attributed to increased electron tunneling from the storage layer through the blocking oxide (\( \text{Si}_3\text{N}_4 \rightarrow \text{blocking oxide} \)). The band offset between the silicon nitride storage layer and the blocking oxide decreases proportionally as the concentration of HfO₂ is increased in the HfAlO alloy [7, 28]. Smaller band offsets can
lead to high leakage current at high program voltages because electrons see a thinner blocking oxide barrier. This charge loss due to leakage current results in less efficient programming for SONOS with high HfO₂ concentration blocking oxides. From Figure 4.4.2 - 4.4.6, it can be inferred that using high-κ blocking oxide can significantly increase the programming speed/efficiency of SONOS especially, for low programming voltages. The agreement between the results obtained from the model and the experimental results of Tan et al. for the SiO₂ blocking dielectric are within ±20% except for the case of 7V gate bias where at program time instance 10⁻⁴s the error is -75%. The correlation between the model and the experimental results of Tan et al. for the SONOS structures with Al₂O₃, HfO₂, HfAlO₁₅%HfO₂, and HfAlO₄₈%HfO₂ blocking oxide are within ±15%, ±9%, ±23.5%, and ±23.5%, respectively.

4.5 Dependence of Trap Density in Silicon Nitride on Energy under Applied Gate Bias

From the experimental data provided by Tan et al. [7], it can be observed that the shift in the threshold voltage saturates at higher voltages as the programming gate voltages are increased. It would be interesting to investigate the impact of higher tunnel oxide and silicon nitride storage layer voltage drops, due to high-κ blocking oxide, on the density of traps. Density of traps is used as fitting parameter in our model. Density of traps determines the total amount of charge that accumulates in the storage layer and hence, the final value of the threshold voltage shifts.
To understand the relation between the saturation of threshold voltage and the applied gate voltage, the density of traps is plotted against the potential energy difference between the conduction band edge at the Si/SiO$_2$ interface (tunnel oxide side) and the SiO$_2$/Si$_3$N$_4$ interface (control oxide side) for gate voltages of interest. Furthermore, it would be interesting to study the impact, if any; different blocking oxides have on the density of traps. The spatial energy band diagram for SONOS type device with high-$k$ blocking oxide under applied gate bias is shown in Figure 4.5.1. The potential energy difference between the conduction band edges, the Si/SiO$_2$ interface and the blocking SiO$_2$/Si$_3$N$_4$, denoted as $E_{\text{diff}}$ can be calculated as follows:

$$E_{\text{diff}} = (\phi_{st} + \phi_{s} + \phi_{ox}) - \phi_B$$ \hspace{1cm} \text{(4.5.1)}$$

where $\phi_B$ is the Si/SiO$_2$ conduction band electron energy barrier, $\phi_s$ is the Si$_3$N$_4$/SiO$_2$ conduction band electron energy barrier, $\phi_{ox}$ is the tunnel oxide conduction band energy

Figure 4.5.1: A schematic spatial energy band diagram depicting conduction band energy differences under applied bias
difference from the SiO$_2$/Si interface to the Si$_3$N$_4$/SiO$_2$ interface due to stored electron in the Si$_3$N$_4$, $\Phi_{st}$ is the silicon nitride conduction band energy difference from the (tunnel oxide) Si$_3$N$_4$/SiO$_2$ to the (control oxide) Si$_3$N$_4$/SiO$_2$ interface.

The values for $E_{\text{diff}}$ and the density of traps, $N_{\text{Ti}}$, used to model the programming behavior of SONOS type devices with high-$\kappa$ blocking oxide proposed by Tan et al. [7] are listed in Table 4.5.1. Density of nitride traps is plotted against $E_{\text{diff}}$ in Figure 4.5.2 for all programming cases of the studied blocking oxides. A linear curve fitting on the data yields a correlation of 97.9%. The linear approximation for the density of traps, $N_{\text{Ti}}$ (m$^{-2}$), as a function of $E_{\text{diff}}$ (eV) is given as:

$$N_{\text{Ti}} = (3\times10^{16}) E_{\text{diff}}$$

(4.5.2)

Figure 4.5.2: Density of traps, $N_{\text{Ti}}$ (m$^{-2}$), vs. energy difference, $E_{\text{diff}}$ (eV), for all the investigated blocking oxide dielectrics and a linear approximation.
Table 4.5.1: A comparison of the density of traps, $N_{T1}$, obtained from the model with the densities obtained from the linear approximation given by equation 4.5.2. $N_{T1}$ is calculated as a function of $E_{diff}$ for all cases of blocking oxide including SiO$_2$, Al$_2$O$_3$, HfO$_2$, HfAlO$_{15\%}$HfO$_2$, and HfAlO$_{48\%}$HfO$_2$ and programming voltages of 6V, 7V, and 9V.

<table>
<thead>
<tr>
<th>Blocking Oxide</th>
<th>$V_g - V_{fb}$ (V)</th>
<th>$E_{diff}$ (eV)</th>
<th>$N_{T1\text{mod}}$ (m$^{-2}$) Model</th>
<th>$N_{T1\text{app}}$ (m$^{-2}$) Linear Approximation</th>
<th>Fitting Error $100\times(NT_{1\text{app}} - NT_{1\text{mod}})/N_{T1\text{mod}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO$_2$</td>
<td>6</td>
<td>0.1392</td>
<td>1.07E+15</td>
<td>4.176E+15</td>
<td>290.28 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.5439</td>
<td>5.34E+15</td>
<td>1.632E+16</td>
<td>205.56 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>1.353</td>
<td>3.67E+16</td>
<td>4.059E+16</td>
<td>10.60 %</td>
</tr>
<tr>
<td>Al$_2$O$_3$</td>
<td>6</td>
<td>1.228</td>
<td>3.87E+16</td>
<td>3.684E+16</td>
<td>-4.81 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.839</td>
<td>6.56E+16</td>
<td>5.517E+16</td>
<td>-15.90 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3.06</td>
<td>1.09E+17</td>
<td>9.18E+16</td>
<td>-15.78 %</td>
</tr>
<tr>
<td>HfO$_2$</td>
<td>6</td>
<td>2.299</td>
<td>6.87E+16</td>
<td>6.897E+16</td>
<td>0.39 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3.112</td>
<td>1.02E+17</td>
<td>9.336E+16</td>
<td>-8.47 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>4.739</td>
<td>1.45E+17</td>
<td>1.422E+17</td>
<td>-1.95 %</td>
</tr>
<tr>
<td>HfAlO (15%) HfO$_2$</td>
<td>6</td>
<td>1.516</td>
<td>4.59E+16</td>
<td>4.548E+16</td>
<td>-0.92 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2.181</td>
<td>7.06E+16</td>
<td>6.543E+16</td>
<td>-7.32 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3.512</td>
<td>1.14E+17</td>
<td>1.054E+17</td>
<td>-7.58 %</td>
</tr>
<tr>
<td>HfAlO (48%) HfO$_2$</td>
<td>6</td>
<td>1.933</td>
<td>5.20E+16</td>
<td>5.799E+16</td>
<td>11.52 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2.677</td>
<td>8.37E+16</td>
<td>8.031E+16</td>
<td>-4.05 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>4.165</td>
<td>1.36E+17</td>
<td>1.25E+17</td>
<td>-8.13 %</td>
</tr>
</tbody>
</table>

The linear fit in Figure 4.5.2 indicates that the blocking oxide affects the electrostatic of the SONOS type devices only. In other words, it can be concluded that the tunneled charges are stored in the bulk deep level nitride traps and the interface between the silicon nitride and high-$\kappa$ blocking oxide plays an insignificant role in providing trap centers for charge storage. Results, $\Delta V_{th}(t)$ vs. program time, obtained from the linear fit of equation 4.5.2 for density of traps in silicon nitride and the experiments, from Tan et al., for programming gate voltages, $V_g - V_{fb}$ of 6V, 7V, and 9V, are shown in Figures 4.5.3 - 4.5.6 for the cases of Al$_2$O$_3$, HfO$_2$, HfAlO (with 15% HfO$_2$ concentration), and
HfAlO (with 48% HfO₂ concentration) blocking oxide, respectively. SiO₂ blocking oxide case is ignored because the error in determining the density of traps by the linear approximation is large (Table 4.5.1). Reason for this large error could be attributed to the interface between SiO₂/Si₃N₄. Since both the nitride and oxide are silicon based materials it may be possible that the interaction between the two materials at lower programming voltage leads to interface traps being the dominant place for charge storage. A comparison between the final threshold voltage shifts obtained using the linear approximation equation and the shifts obtained from the experimental results is provided in Table 4.5.2. The maximum error in predicting the threshold voltage shift saturation using linear approximation is within ±16%.

![Graphic](image)

Figure 4.5.3: Programming transients (ΔV_{th}(t) vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with Al₂O₃ blocking oxide of thickness 7.5 nm for V_{g} - V_{th} = 6V, 7V, and 9V.

67
Figure 4.5.4: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$.

Figure 4.5.5: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with HfAlO$_{15}\%$HfO$_2$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.5.6: Programming transients ($\Delta V_{th}(t)$ vs. program time) for SONOS type devices, obtained using linear approximation for density of traps in silicon nitride, with HfAlO$_{88}$HfO$_{2}$ blocking oxide of thickness 7.5 nm for $V_g - V_{fb} = 6V, 7V,$ and $9V$

The model can be used along with the linear approximation equation given by equation 4.5.2 to predict the programming transients for the SONOS devices with high-κ blocking oxides for the programming voltage, $V_g - V_{fb}$, of 8V. The energy difference, $E_{diff}$, required in the density of traps equation 4.5.2 can be found by following the procedure for electrostatics calculation described in Section 4.1. The programming behavior for SONOS type devices with high-κ blocking oxides for programming voltage of 8V is shown in Figure 4.5.7. The thickness of the blocking oxide is kept constant at 7.5 nm.
Figure 4.5.7: Programming transients ($\Delta V_{th}(t)$ vs. program time) for blocking oxides $\text{Al}_2\text{O}_3$, $\text{HfO}_2$, $\text{HfAlO}_{15}\%\text{HfO}_2$, and $\text{HfAlO}_{48}\%\text{HfO}_2$ at program voltage of $V_g - V_f = 8\text{V}$

Table 4.5.2: A comparison between the final value of threshold voltage shifts from the experimental results with the shifts obtained from the model with linear approximation fit for blocking oxides $\text{Al}_2\text{O}_3$, $\text{HfO}_2$, $\text{HfAlO}_{15}\%\text{HfO}_2$, and $\text{HfAlO}_{48}\%\text{HfO}_2$ and programming voltages of $6\text{V}$, $7\text{V}$, and $9\text{V}$.

<table>
<thead>
<tr>
<th>Blocking Oxide</th>
<th>$V_g - V_f$ (V)</th>
<th>Model $V_{th}$ (V)</th>
<th>Linear Fit $V_{th}$ (V)</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Al}_2\text{O}_3$</td>
<td>6</td>
<td>1.05</td>
<td>1</td>
<td>-4.762 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.78</td>
<td>1.5</td>
<td>-15.730 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2.95</td>
<td>2.49</td>
<td>-15.593 %</td>
</tr>
<tr>
<td>$\text{HfO}_2$</td>
<td>6</td>
<td>1.2</td>
<td>1.205</td>
<td>0.417 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.78</td>
<td>1.63</td>
<td>-8.427 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2.53</td>
<td>2.48</td>
<td>-1.976 %</td>
</tr>
<tr>
<td>$\text{HfAlO}_{15}%\text{HfO}_2$</td>
<td>6</td>
<td>1.1</td>
<td>1.09</td>
<td>-0.909 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.69</td>
<td>1.57</td>
<td>-7.101 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2.72</td>
<td>2.52</td>
<td>-7.353 %</td>
</tr>
<tr>
<td>$\text{HfAlO}_{48}%\text{HfO}_2$</td>
<td>6</td>
<td>1.05</td>
<td>1.17</td>
<td>11.429 %</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.69</td>
<td>1.62</td>
<td>-4.142 %</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2.75</td>
<td>2.52</td>
<td>-8.364 %</td>
</tr>
</tbody>
</table>
4.6 Summary

Following key points can be observed from the modeling of SONOS type devices proposed by Tan et al. [7]:

(i) A “trap-based” current continuity model is proposed and successfully used to predict the programming behavior of SONOS type FLASH with various storage layers. The correlation between the experimental data of Tan et al. [7] and the data obtained from the model is ±28% for the worst case.

(ii) Density of traps is used in the model, to generate the threshold voltage shift vs. program time plots for the SONOS type FLASH with various blocking oxides. The correlation between the model and the experimental results, of Tan et al. [7], for the SONOS structures with SiO$_2$, Al$_2$O$_3$, HfO$_2$, HfAlO (with 15% HfO$_2$ concentration), and HfAlO (with 48% HfO$_2$ concentration) blocking oxides are within ±20%, ±15%, ±9%, ±23.5%, and ±23.5%, respectively.

(iii) A linear fit is obtained between the fitted density of traps and the energy difference between the conduction band edges from the Si/SiO$_2$ (tunnel oxide) and the Si$_3$N$_4$/SiO$_2$ (control oxide). The programming transients (ΔV$_{th}$(t) vs. program time) curve have a maximum error of ±16% for predicting maximum threshold voltage shift.

(iv) It is observed that the bulk nitride traps play the dominant role in trapping tunnel charges and the high-κ blocking oxide affects only the tunneling efficiency/tunneling current.
CONCLUSION AND FUTURE WORK

A physics based model is developed to predict the programming transients ($\Delta V_{th}$ vs. programming time) of SONOS type devices. Specifically, the proposed model is based on the "current continuity" model and takes into account the information regarding trap density and trap capture cross section area. Density of trap in the storage layer is used as a fitting parameter in our model. Density of available trap for storing charges determines the final threshold voltage shift during programming cycle. Good agreement between the model and the experimental results of Tan et al. [7] is obtained for SONOS type devices with high-$\kappa$ dielectric storage layers and SONOS type devices with high-$\kappa$ dielectric blocking oxides.

For the SONOS type devices with various blocking oxide materials, it is concluded that the bulk, storage layer, traps play the dominant role in charge storage rather than the interface traps at the storage layer and the blocking oxide interface. A linear fit between the density of traps in the storage layer to the energy difference between the Si/SiO$_2$ (substrate/tunnel oxide) and the storage layer/blocking oxide conduction bands is also proposed. Results, obtained using the proposed density of trap formula, are within $\pm 16\%$ of the experimental results, from Tan et al. [7], in determining the final threshold voltage shift for SONOS type FLASH with high-$\kappa$ blocking oxide.
Following are some recommendations for future work:

(i) include the reverse current due to loss of stored electrons through the tunnel oxide and blocking oxide.

(ii) include the effects of interface states on programming characteristics.

(iii) study the effects of high programming gate voltage on high-κ storage layer SONOS devices.

(iv) develop the model for erase characteristics of the studied SONOS type devices.

(v) develop the model to predict the retention characteristics of the studied SONOS type devices.
BIBLIOGRAPHY


7. Yan Ny Tan, Chim, W.K., Wee Kiong Choi, Moon Sig Joo, and Byung Jin Cho, “Hafnium aluminum oxide as charge storage and blocking-oxide layers in


VITA

Graduate College
University of Nevada, Las Vegas

Varun Jain

Address:
2120 Ramrod Ave.
Henderson, NV 89014

Education:
Bachelor of Science, Electrical Engineering, 2004
Northern Arizona University, Flagstaff, Arizona

Honors:
Won the Outstanding Teaching Assistant Award, 2005-06 Academic Year
Member of Tau Beta Pi

Thesis Title:
A model for programming characteristics of SONOS type FLASH with high-κ dielectrics.

Thesis Examination Committee:
Chairperson, Dr. Rama Venkat, Ph. D.
Committee Member, Dr. Henry Selvaraj, Ph. D.
Committee Member, Dr. Paolo Ginobbi, Ph. D.
Graduate Faculty Representative, Dr. Ajit Roy, Ph. D.