FABRICATION AND CHARACTERIZATION OF SINGLE AND MULTILAYER TUNNEL DIELECTRICS FOR ADVANCED FLOATING GATE FLASH MEMORY

By

RAMZAN MAT AYUB
(0740110161)

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

SCHOOL OF MICROELECTRONIC ENGINEERING
UNIVERSITI MALAYSIA PERLIS

2014

THESIS DECLARATION FORM

UNIVERSITI MALAYSIA PERLIS

DECLARATION OF THESIS

Author's full name : Ramzan b Mat Ayub

Date of birth : 23 February 1966

Title : Fabrication and characterization of single and multi-layer tunnel dielectrics for advanced floating gate flash memory

Academic Session : 2012/2013

I hereby declare that this thesis becomes the property of Universiti Malaysia Perlis (UniMAP) and to be placed at the library of UniMAP. This thesis is classified as:

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*

☐ **RESTRICTED** (Contains restricted information as specified by the organization where research was done)*

☐ **OPEN ACCESS** I agree that my thesis is to be made immediately available as hard copy or on-line open access (full text)

I, the author, give permission to the UniMAP to reproduce this thesis in whole or in part for the purpose of research or academic exchange only (except during a period of _____ years, if so requested above).

Certified by:

SIGNATURE	**SIGNATURE OF SUPERVISOR**
(660223-03-5539)	**NAME OF SUPERVISOR**
Date: _________________	Date: _________________

***NOTES**: If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentially or restriction.

Acknowledgement

In the name of Allah, Most Gracious, Most Merciful.

Praise be to Allah, for the strength, knowledge and perseverance that He has been bestowed upon me, not only to complete this research, but indeed throughout my life.

The accomplishment of the research as described in this thesis would have not been possible without the help and support from numerous people and entities that I would like very much, to acknowledge and extend my deepest appreciation.

First of all, I would like to express my exceptional thanks to my main supervisor, Prof Dr Uda Hashim for the scientific guidance, support and encouragement in so many ways and forms. The special thanks also go to my second supervisor, Dr Nazri Abdul Halif for the support, guidance and insights into both technical and non-technical matters.

Huge thanks to Dr Mohd Khairuddin Md Arshad for the help in MATLAB coding, as well as other numerous tips and guides on general thesis preparation.

My deep appreciation also goes to my research collaborators and the lab's technical staff; Mr. Mohd Rosydi Zakaria, Ms Zarimawaty Zailan, Mr. Azman Hassan, Miss Norhafizah, Mr. Haffiz Abdul Razak, Mr. Bahari, Mr. Jasni Ismail, Mr. Hasrul, Miss Nursyamira and many others.

I sincerely acknowledge the Ministry of Higher Education (MoHE) and Universiti Malaysia Perlis (UniMAP) for providing the scholarship under SLAB/SLAI program. My special appreciation to the Vice Chancellor of UniMAP, Brig. General Dato' Prof Dr Kamarudin Hussin and the Deputy Vice Chancellor for Academic and Internalization, Dato' Prof Dr Zul Azhar Zahid Jamal, and the Centre for Graduate Study (CGS) for their support.

I would like also to acknowledge The Ministry of Science, Technology and Innovation (MOSTI) through the financial support provided under the ScienceFund Research Grant (Grant No: 9005-0035), titled: ”Advanced Flash Memory Development for 32 nm Technology Node and Beyond” which made this research work possible.

Last but not least, my very special appreciation goes to my family who always beside me with their unconditional love and support. My wife Nor Azliza, sons and daughters: Abdul Muizz, Nurul Iman, Irfan, Sufya, Mohammad Rafiq and Nur Adelia. Without their support, backing and understanding, this thesis could not be materialized. This thesis is exclusively dedicated to our new family members, Nur Mawadda, who was born on October 28th, 2013.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

ψ	Psi
π	Pi
λ	Lambda
λ_D	Extrinsic Debye length
α	Capacitance coupling ratio
ε_0	Permittivity of free space, 8.85x10^{-14} F/cm
ε_{Si}	Relative permittivity of Si, 11.9ε_0
ε_{SiO2}	Relative permittivity of SiO_2, 3.9ε_0
ε_{N1}	Relative permittivity of Si_3N_4 layer, 7.8ε_0
ε_{O1}	Relative permittivity of bottom SiO_2 layer, 3.9ε_0
ε_{O2}	Relative permittivity of top SiO_2 layer, 3.9ε_0
k	Boltzmann constant, 1.38x10^{-23} eV/°K
kT	Thermal energy at room temperature, 4.046 x 10^{-21}J
h	Planck constant, 6.625x10^{-31} J-s
$\hbar$	Planck constant over 2π, $\frac{6.625x10-31\ J-s}{2\pi}$
τ_c	Trap capture time
τ_e	Trap emission time
τ_{prog}	Programming time of the memory cell
τ_{ret}	Retention time of the memory cell
γ	Maximum charge loss from the floating gate
υ	Traps escape frequency
ϕ_B	The barrier height at the conductor and insulator interface
ϕ_{BN1}	The barrier height of Si_3N_4 layer
ϕ_{BO1}	The barrier height of bottom SiO_2 layer
ϕ_{BO2}	The barrier height of top SiO_2 layer
ϕ_F	Fermi potential of the semiconductor at interface
ϕ_{ms}	Work function difference between the gate metal and bulk material, -0.95V
eV	Electron volt
q	Charge of electron, 1.60x10^{-19} C
k_H	Dielectric constant of high-k material
k_L	Dielectric constant of low-k material
m*	Mass of free electron, 9.1x10^{-31} kg
m_{ox}	Electron effective mass in SiO_2, 0.45m*
m_{N1}	Electron effective mass in Si_3N_4 layer, 0.3m*
m_{O1}	Electron effective mass in bottom SiO_2 layer, 0.45m*
m_{O2}	Electron effective mass in top SiO_2 layer, 0.4m*
n_T	Concentration of trapped electron
t_H	Thickness of high-k material

t_L	Thickness of low-k material
A_{inj}	Area of injection current
C_{dif}	Differential capacitance
C_{FB}	Flatband capacitance
C_{FD}	Capacitance between FG and source
C_{FG}	Floating gate capacitance
C_{FS}	Capacitance between FG and source
C_{ox}	Gate oxide capacitance
C_T	Total capacitance
$\mathrm{C_T}$	Capture cross section
C_{TUN}	Capacitance between FG and tunnel
E_c	Conduction band of the material
E_{inj}	The electric field at the injecting interface
E_{OT}	Effective oxide thickness
E_{ox}	Electric field across oxide
E_v	Valens band of the material
F	Minimum feature size of certain semiconductor technology
I_D	Drain current
I_{prog}	Programming current
J	Tunneling current density
J_{FN}	F-N tunneling current density
J_{ret}	Retention current density
N_{BULK}	Bulk doping
N_{EFF}	Effective oxide charge concentration
N_T	Trap concentration
P	Tunneling probability
Q_D	Charge in the silicon depletion layer
Q_{EFF}	Effective oxide charge
Q_{FG}	Total charge in the floating gate
Q_I	Fixed charge at the silicon/insulator interface
Q_{ox}	Equivalent fixed oxide charge
Q_T	Total charge stored in the gate oxide
t_{ox}	Oxide thickness
T	Temperature
TC	Transmission coefficient
T_{N1}	Thickness of Si_3N_4 layer
T_{O1}	Thickness of bottom oxide
T_{O2}	Thickness of top oxide
V_B	Body (bulk) voltage
V_{CG}	Control gate voltage
V_D	Drain voltage
V_G	Gate voltage

V_{FB}	Flatband voltage
V_{FG}	Floating gate voltage
V_S	Source voltage
ΔV_{TH}	Threshold voltage shift
V_{TH}	Threshold voltage
V_{TO}	Threshold voltage of FG-Oxide-Substrate
V_{pp}	Programming voltage of memory cell
V_{read}	Read voltage
V_{ret}	Retention voltage

LIST OF ABBREVIATIONS

Al_2O_3	Aluminum Oxide
HF	Hydrogen Fluoride
HfAlO	Hafnium Aluminum Oxide
HfO_2	Hafnium Oxide
NH_3	Ammonia
NO	Nitric Oxide
NO_2	Nitrous Dioxide
Si_3N_4	Silicon Nitride
SiO_2	Silicon Dioxide
SiO_xN_y	Silicon Oxynitride
Ta_2O_5	Tantalum Oxide
TaN	Tantalum Nitride
A	Ampere
CG	Control Gate
CHE	Channel Hot Electron
CMOS	Complementary Metal Oxide Semiconductor
CP	Charge Pumping
CR	Coupling Ratio
CT	Charge Trapping
C-V	Capacitance Voltage
DC	Direct Current
DPN	Decouple Plasma Nitridation
DRAM	Dynamic Random Access Memory
DT	Direct Tunneling
EEPROM	Electrically Erasable Programmable Read Only Memory
EOT	Effective Oxide Thickness
EPROM	Electrically Programmable Read Only Memory
ETB	Engineered Tunnel Barrier
F	Minimum Feature Size of Specific Technology Node
FeRAM	Ferroelectric Random Access Memory
FG	Floating Gate
FM	Flash Memory
F-N	Fowler Nordheim
FOM	Figure of Merit
HBD	Hard Breakdown
HF C-V	High Frequency C-V
IC	Integrated Circuit
IPD	Inter Poly Dielectric
ITRS	International Technology Roadmap for Semiconductor
I-V	Current-Voltage
LOCOS	Local Oxidation

MATLAB	Matrix Laboratory
MLC	Multi-Level Cell
MOS	Metal Oxide Semiconductor
MV	Mega Volts
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MRAM	Magneto-resistive Random Access Memory
MTP	Multi Time Programmable
NC	Nano Crystal
NMOS	N-channel Metal Oxide Semiconductor
NVM	Non-Volatile Memory
ONO	Oxide Nitride Oxide
OTP	One Time Programmable
PBD	Post-Breakdown
PCM	Phase Change Memory
P/E	Program / Erase
PN	Plasma Nitridation
RAM	Random Access Memory
RTA	Rapid Thermal Annealing
RTN	Rapid Thermal Nitridation
SBD	Soft Breakdown
SEM	Scanning Electron Microscope
SHE	Substrate Hot Electron
SHH	Substrate Hot Hole
SMU	Source Measure Unit
SRAM	Static Random Access Memory
SILC	Stress Induced Leakage Current
SLC	Single Level Cell
SMU	Source Measure Unit
SONOS	Silicon Oxide Nitride Oxide Silicon
TAT	Trap Assisted Tunneling
TDDB	Time Dependent Dielectric Breakdown
TEM	Transmission Electron Microscope
TO	Tunnel Oxide
VARIOT	Variational Oxide Thickness
V	Volt
VM	Volatile Memory
WKB	Wentzel–Kramers–Brillouin
W/E	Write / Erase

Fabrikasi dan Pencirian Dielektrik Terowong Berlapisan Tunggal dan Berbilang Bagi Peranti Ingatan Kilat Berget Terapung Yang Termaju.

ABSTRAK

Peranti get terapung adalah merupakan komponen utama kepada teknologi ingatan tidak-meruap sejak bermulanya era peranti ingatan kilat. Walaubagaimanapun, apabila peranti dikecilkan sehingga ke dimensi nanometer, get terapung kilat menghadapi satu laluan yang sukar. Pengecilan oksida penerowong mempunyai limit praktikal sekitar 8 nm disebabkan keperluan pengekalan data. Justeru, tujuan kajian ini ialah untuk mencirikan dan menilai prestasi oksida penerowong berlapisan tunggal dan berbilang, yang mana fokus utamanya adalah untuk mengecilkannya kurang dari 8 nm. Kajian ini dilakukan di dalam dua langkah. Pertamanya, ciri-ciri I-V peranti di selakukan menggunakan perisian MATLAB, berdasarkan model fizikal padat yang terkini. Kelajuan pengaturcaraan dan penahanan data kemudiannya di kira berdasarkan lenkung I-V yang diselakukan. Keduanya, pemuat MOS kemudiannya di fabrikasikan dan dicirikan untuk pengesahan keputusan penyelakuan. Prestasi oksida penerowong berlapisan tunggal telah ditunjukkan dengan jayanya. Prestasinya telah di nilaikan berasaskan dua aspek, iaitu kelajuan pengaturcaraan τ_{prog} dan penahanan data τ_{ret}. τ_{prog} untuk lapisan oksida dan oksinitrid berlapisan tunggal berketebalan 4 nm ialah masing-masingnya 110 μs dan 130 μs, tidak terlalu jauh dari kehendak teknologi iaitu selama100 μs. Walaubagaimanapun, prestasi τ_{ret} mereka adalah jauh lebih rendah dari yang diperlukan iaitu 10-tahun, yang mana kedua-duanya hanya mampu mencapai 3.1 dan 4.6 tahun masing-masing. Berdasarkan hal tersebut, boleh disimpulkan bahawa kedua-dua lapisan tunggal oksida dan oksinitrid berketebalan 4 nm telah gagal untuk memenuhi keperluan teknologi nod 18 nm. Walaubagaimanapun, telah dibuktikan bahawa oksida nitrid mampu untuk menambahkan prestasi τ_{ret} bagi lapisan tunggal SiO_2. Urutan dari itu, telah juga ditunjukkan bahawa ketebalan oksida berlapisan tunggal dan oksinitrid berketebalan masing-masingnya 8.25 dan 6.4 nm, adalah diperlukan untuk mencapai keperluan penahanan data selama 10 tahun. Juga telah berjaya ditunjukkan bahawa oksida nitrid berupaya untuk mengurangkan penghasilan perangkap secara berkesan, yang mana ini akan mengurangkan kebocoran peranti pada medan rendah, terutama di dalam bentuk SILC. Bagi kes dielektrik berbilang lapisan, telah ditunjukkan bahawa konfigurasi terbaik ialah yang mempunyai lapisan dasar SiO_2 paling tipis / Si_3N_4 paling tebal. Penyelakuan peranti menunjukkan bahawa untuk dielektrik berlapisan 2 dan 3, τ_{prog} adalah dalam julat 18 hingga 41 μs untuk lapisan berketebalan berkesan oksida (EOT) 4 dan 8 nm, manakala secara eksperimen nilainya adalah dalam julat 2 hingga 104 μs. Mengambilkira keperluan τ_{ret} walaubagaimanapun, hanya konfigurasi yang berketebalan berkesan oksida (EOT) 6 nm untuk kedua-dua dielektrik berlapisan 2 dan 3, serta 8 nm untuk dilektrik berlapisan-3yang telah berjaya memenuhi kehendak teknologi nod 18 nm.

Fabrication and Characterization of Single and Multi-Layer Tunnel Dielectrics for Advanced Floating Gate Flash Memory

ABSTRACT

The floating gate device has been the workhorse for the non-volatile memory technology since the beginning of flash memory era. However, as the device is scaled down towards the realms of nanometer dimension, floating gate flash faces a very steep scaling path. The tunnel oxide scaling has a practical limit of approximately 8 nm due to data retention requirement. Therefore, the purpose of this work is to characterize and to assess the performances of single and multi-layer tunnel oxide, which primary focus is to further scale it beyond 8 nm. This study was carried out in two steps. Firstly, device I-V characteristics were simulated using the MATLAB software, based on the most recent compact physical model. Programming speed and data retention were calculated based on the simulated I-V curves. Secondly, MOS capacitors were then fabricated and characterized to validate the simulation result. The performance of single layer tunnel oxide has been successfully demonstrated. Its performance has been mainly evaluated from two perspectives, namely the programming time τ_{prog}, and data retention τ_{ret}. The τ_{prog} for 4 nm single layer oxide and oxynitride were calculated to be 110 μs and 130 μs respectively, not too far off from 100 μs technological requirement. However, their τ_{ret} performance was well below 10-year requirement, with both dielectrics just been able to achieve 3.1 and 4.6 year respectively. In that sense, one can conclude that both 4 nm single layer oxide and oxynitride have failed to comply with the requirement of 18 nm technology node. However, it has been proved that nitrided oxide could improve the τ_{ret} of single layer SiO_2. Furthermore, it has also been demonstrated that the thickness of a single layer oxide and oxynitride of 8.25 and 6.4 nm respectively, would be required to achieve the 10-year data retention requirement. It has also been shown that nitrided oxide could serve as an effective way of suppressing trap generation which in turn would suppress low field device leakages, especially in the form of SILC. In the case of multi-layer dielectrics, it has been shown that the best configuration is the one with the thinnest bottom SiO_2 / thickest Si_3N_4. Device simulation shows that for 2 and 3-layer dielectrics, the τ_{prog} was in the range of 18 to 41 μs for the EOT of 4 to 8 nm, while experimentally it's in the range of 2 to 104 μs. Taking τ_{ret} requirement into consideration however reveals that only configurations with the EOT of 6 nm for both 2 and 3-layer dielectrics and 8 nm of 3-layer dielectric have successfully met the requirement for 18 nm technology nodes.

CHAPTER 1

INTRODUCTION

1.1 The Flash Memory In Brief

Semiconductor memory is an electronic data storage device that widely regarded as an essential element of today’s electronics industry. The device is normally used as computer memory and other integrated circuits (ICs) based product, with its construction is built around semiconductor processing technology.

In general, semiconductor memory exists in two different forms in ICs. The non-permanent type, normally called volatile memory (VM), which only retains its information as long as the power supply is connected. Examples of VM are the majority of RAMs (Random-Access Memory) such as SRAM (Static Random-Access Memory) and DRAM (Dynamic Random-Access Memory) (Bez, Camerlenghi, Modelli, & Visconti, 2003).

Another form of memory, which is the focus of this study, is called Non Volatile Memory (NVM). In this type of memory, the stored information is retained even after the power supply is removed. Examples of NVM are One Time Programmable (OTP) Memory, Electrically Erasable Programmable Read-Only Memory (EEPROM) and Flash Memory (FM).

NVM itself can be a One Time Programmable (OTP) or a Multi Time Programmable (MTP). In OTP memory, the information is programmed into the memory cell during the fabrication process (Bartolomeo et al., 2009). The main disadvantage with the OTP is it cannot be reprogrammed, which is a distracting factor for many forms of applications. MTP memory devices on the other hand, offer advantages in the way that its information can be stored and erased several times. The like of Electrically Programmable Read-Only Memory (EPROM), EEPROM and FM are all belong to this category (Brown,D. & Brewer,E. 1998e).

The history of FM started in 1967, when Kahng and Sze presented a novel concept of floating gate transistor, where electrons could be stored onto it (Kahng and Sze, 1967). Since then, the EPROM cell has been developed. This technology grew rapidly to become the most significant NVM technology in the 1980s. About the same period, the Flash EEPROM was introduced which add the electrically erasable feature to the existing EPROM (Mukherjee & Chang, 1985). Consequently, the first FM product was presented in 1988 (Kynett & Baker, 1988).

However, FM market did not take-off smoothly until the technology was proved to be reliable and manufacturable. Only by the late of 1990s, the demand for FM grew rapidly as the consumer products which require NVM for code and data storage, such as mobile application start to be of in high demand. Starting from year 2000, the FM can be considered as a really mature technology (Falan Yinug, 2007).

Since year 2000 onwards have witnessed the rapid growth of the FM due to mostly to ever increasing popularity of mobile and portable devices such as digital cameras, smartphones and computer tablets. This popularity of FM is due to its unique ability to erase the cells in blocks of data at a very fast rate (Falan Yinug, 2007).

Nowadays, the ubiquitous presence of the FM, especially of NAND cell architecture in almost all aspect of modern life especially, has led the flash memory to be considered as one of the integrated circuits technology driver towards 10 nm technology node with blistering speed, surpassing both logic and DRAM (Lu, 2012).

In semiconductor industry, cost and speed trade-off is always a serious deciding factor when designing a new product. As silicon real estate is becoming more expensive, the chip size emerges as the main cost contributing factor. For this reason, memory chip designers have developed several types of FM variant, namely the NOR, DINOR and NAND architectures to target for specific application. However, NAND and NOR architectures have emerged as the dominant FM variant, employed in contemporary electronic industry as the workhorse for wide spectrum of applications (Toshiba America, 2006).

The NOR architecture was optimized for speed. In NOR cell configuration, the individual memory cells are connected in parallel, which in turn requires one contact for every two memory cells, thus consuming significant chip area. This configuration enables the device to achieve random access, which result in shorter read times required for the random access of microprocessor instruction. Therefore, NOR is ideal for lower density, high-speed read applications in code storage and direct execution in portable electronic devices, such as smart phones and computer tablets.

NAND architecture on the other hand, was designed with a smaller chip size (about half of NOR) to enable a lower cost-per-bit of stored data. The reduced cell size was achieved by arranging an array of eight memory cells connected in series, thus saving an expensive silicon real estate for contact formation. NAND is ideal for the low-cost, high-density, high-speed program/erase applications such in the high-density data storage medium for consumer devices. The overall features comparison between NOR and NAND architectures is shown in Table 1.1.

Table 1.1: NOR and NAND Features Comparison (Micron Technology, Inc., 2013)

Serial NOR / Parallel NOR	**Single Level Cell (SLC) NAND / Multi Level Cell (MLC) NAND**
Low density, low pin count	High density, low pin count
Long life cycles	Less reliable and requires controller management
Reliability, high performance	Low performance
Reliable code and data storage	Mostly data-focused
Fast random access time	Fast writes and reads

Based on the way the devices store its information, FM device can be classified into two main classes. In the first class, the charge is stored on a conducting layer that is completely isolated from other structures by a dielectric film. This type of device is commonly referred to as a floating gate (FG) Flash. In the second class of FM, the charge is stored in discrete trapping centers of dielectric layer. These devices are therefore, commonly referred to as the charge-trapping (CT) device.

To date, FG Flash are the mainstream of FM and have followed Moore’s Law scaling through multiple technology generation, and mostly used in both NOR and NAND cells.

In a nutshell, the operational of FG Flash is based on the ability to bring electrons onto the floating gate and removing them again in order to change the threshold voltage of the memory cell. The pace at which these operations can be carried out is the most important FG Flash performance indicator and its normally termed as the programming speed. Nowadays, the programming operations for FG Flash are done by the methods of channel hot-electron (CHE) injection or Fowler-Norheim (F-N) tunneling.

The programming speed is proportional to the rate of electrons being injected onto the floating gate. The electron injection is carried out via ultra-thin dielectric layer, called the tunnel barrier, which transport the electrons under the influence of external electric field. Generally, the higher the electric field across the tunnel barrier, the higher the rate of electron injection through it.

If the applied voltage level is maintained and the thickness of tunnel barrier is reduced, the electric field will increase. As a result, higher rate of electrons would be injected onto the floating gate, achieving faster programming speed. This important concept underlies the device scaling philosophy, practiced by the NVM device technologists to improve the FG Flash speed performance.

However, as a result of a continuous and aggressive tunnel barrier scaling, especially when its thickness is reduced below 8 nm, several unwanted phenomenon such as Stress Induced Leakage Current (SILC) emerges (Wellekens & Houdt, 2008). The SILC would severely affect the FG Flash data retention capability, thus compromising the gain in the programming speed. A detail discussion on the tunnel barrier scaling is done in the next section.

Several approaches have been proposed as alternatives for the shortcoming encounters with further scaling of the tunnel barrier. Among the most widely