



**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

(660223-03-5539)

Date: _____

SIGNATURE OF SUPERVISOR

NAME OF SUPERVISOR

Date: _____

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

	PAGE
THESIS DECLARATION	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF SYMBOLS	xiii
LIST OF ABBREVIATIONS	xvi
ABSTRAK	xviii
ABSTRACT	xix
Chapter 1 Introduction	1
1.1 The Flash Memory in Brief	1
1.2 Problem Statement	8
1.3 Motivation of the Study	11
1.4 Research Objectives	12
1.5 Research Scope	12
1.6 Thesis Outline	13
Chapter 2 Review of Flash Memory Technology	15
2.1 Introduction	15
2.2 Basics of Flash Memory Devices	15
2.2.1 Operating Principles	16
2.2.2 Device Structure	17
2.2.3 Flash Memory Operations	19
2.3 Flash Memory Characteristics	21
2.3.1 Transient Characteristics	22
2.3.2 Endurance Characteristics	23
2.3.3 Retention Characteristics	24
2.4 Electron Tunneling Mechanism in Floating Gate Flash Device	25
2.4.1 Fowler-Nordheim Tunneling	26
2.4.2 Channel Hot Electron Injection	29
2.4.3 Direct Tunneling	31

2.4.4	Trap-Assisted Tunneling	33
2.5	Flash Memory Scaling	35
2.5.1	Scaling Issues	36
2.5.2	Program/Erase and Data Retention Trade-Off	38
2.5.3	The Effect of SILC on Data Retention	39
2.5.4	A Review on the Proposed Solutions	40
2.6	Tunnel Barrier Engineering	44
2.6.1	Crested Tunnel Barrier	44
2.6.2	VARIOT Tunnel Barrier	47
2.7	Chapter Summary	50
 Chapter 3 Single Layer Tunnel Barrier Floating Gate Flash		 51
3.1	Introduction	51
3.2	Floating Gate Flash Device Requirement	52
3.3	Floating Gate Flash Capacitor Model	55
3.3.1	The Concept of Threshold Voltage in Floating Gate Flash Device	56
3.3.2	The Floating Gate Capacitor Model	57
3.3.3	The Floating Gate Transient Characteristics	59
3.4	Trap Generation and Electrical Breakdown in Tunnel Oxide	61
3.4.1	Oxide Breakdown Mechanism – The General Model	62
3.4.2	Stress Induced Leakage Current (SILC)	64
3.4.3	Charge Trapping and Trap Generation	67
3.4.4	Techniques for Stressing and Measuring the Charge Trapping	69
3.5	Oxide Nitridation for Low-Field Characteristics Improvement	71
3.6	Experimental Details and Fabrication Process Flow	74
3.7	Device Characterizations	77
3.7.1	Current-Voltage Characterization	78
3.7.2	Capacitance-Voltage Characterization	80
3.8	Device Simulation	83
3.9	Experimental Results and Discussion	84
3.9.1	Stress Induce Leakage Current, Soft Breakdown and Hard Breakdown Regions	84
3.9.2	Current-Voltage Characteristics at High Field	88
3.9.3	Programming Time, τ_{prog}	92
3.9.4	Current-Voltage Characteristics at Low Field	93
3.9.5	Stress-Induced Leakage Current (SILC)	95
3.9.6	Device Retention Time, τ_{ret}	99
3.9.7	Oxide Trap Generation	99
3.10	Conclusion	105
 Chapter 4 Multi-Layer Tunnel Barrier for NAND Flash		 111
4.1	Introduction	111
4.2	Electron Tunneling Through Multiple Barrier	113
4.2.1	Tunneling Through Two-Layer Barrier	113

4.2.2	Tunneling Through Three-Layer Barrier	115
4.3	Effective Oxide Thickness (EOT)	117
4.4	Device Simulation	119
4.5	Experimental Details and Process Flow	121
4.6	Device Characterization	125
4.7	Result and Discussion	125
4.7.1	Simulation Results	126
4.7.1.1	The I-V Characteristics of Two-Layer Tunnel Barrier	126
4.7.1.2	The I-V Characteristics of Three-Layer Tunnel Barrier	134
4.7.1.3	Programming Time, τ_{prog} of Multi-Layer System	143
4.7.2	Experimental Results	146
4.7.2.1	I-V Characteristics: Experimental versus Simulations	147
4.7.2.2	The Effect of Individual Dielectric Layer on τ_{prog}	151
4.7.2.3	The Effect of Individual Dielectric Layer on τ_{ret}	153
4.7.2.4	The Correlation Between Trap Generation and SILC	154
4.8	Conclusion	159
Chapter 5	Conclusion and Future Work	160
5.1	Introduction	160
5.2	Conclusion	160
5.3	Future Work	163
References		165

LIST OF TABLES

NO.		PAGE
1.1	NOR and NAND Features Comparison	4
1.2	Floating Gate Flash Technology Roadmap	10
3.1	Floating Gate Device Requirement	54
3.2	Calculated τ_{prog} based on 18 nm Technology Node for Conventionally Grown Tunnel Oxide	92
3.3	Calculated τ_{prog} based on 18 nm Technology Node for Oxynitrides	93
3.4	Data retention time for conventional oxide	99
3.5	Data retention time for oxynitride	100
3.6	Important parameters in trap generation calculation	102
4.1	Energy Barrier Parameters	120
4.2	Simulation Matrices for 2-Layer ETB	120
4.3	Simulation Matrices for 3-Layer ETB	121
4.4	Fabrication Matrices for 2-Layer ETB	124
4.5	Fabrication Matrices for 3-Layer ETB	124
4.6	Calculated τ_{prog} based on 18 nm Technology Node for 2-Layer Tunnel Barrier (Simulation Data)	144
4.7	Calculated τ_{prog} based on 18 nm Technology Node for 3-Layer Tunnel Barrier (Simulation Data)	145
4.8	The Summary of τ_{prog} for ETB Configurations	152
4.9	The Summary of τ_{ret} for the respective ETB Configurations	153
5.1	Summary of τ_{prog} and τ_{ret} performances for Engineered Tunnel Barrier	162

LIST OF FIGURES

NO.		PAGE
2.1	NMOS Field Effect Transistor, showing the oxide charge Q_T stored in the gate oxide	16
2.2	Charge Trapping (CT) Flash Memory, which operates based on the charge storage in Si_3N_4 layer.	18
2.3	The Schematics of Floating Gate (FG) Flash Memory, showing four it's main components; Control Gate, Inter Poly Dielectric, Floating Gate and Tunnel Oxide.	19
2.4	Plot of I_D versus V_{CG} to illustrate the I-V curves behavior with no electrons in the floating gate (black line), and with electrons in the floating gate (red line). Red colored e- represents the presence of electrons. V_{read} is the voltage applied to the Control Gate to sense the memory cell contents.	20
2.5	The figures illustrate the relationship between the basic operation mechanisms with the electrons in the floating gate; (a) Electrons were pushed into the floating gate during WRITE process. (b) Electrons were pushed out of the floating gate during ERASE process. (c) Electrons were contained inside the floating gate during data RETENTION mechanism, with e- represents an electron.	21
2.6	Typical endurance characteristics of a floating gate memory cell showing the threshold voltage in the written and erased state as a function of the number of applied Write / Erase cycles. The threshold voltage shows a threshold voltage window opening during the first tens of cycles, followed by a window closure after $10^5 - 10^6$ cycles.	24
2.7	The energy band diagram showing 4 main electron injection / tunneling mechanisms through the energy barrier of single layer dielectric (tunnel barrier). CHE and F-N Tunneling are the main programming mechanisms while DT and TAT are the unwanted effects as a result of tunnel barrier scaling. E_c , E_v and $q_{\phi B}$ is the conduction band, valence band and tunnel barrier height respectively.	26
2.8	Energy band representation of Si-SiO ₂ -Poly Si system showing: (a) without the external bias, there is no electron tunneling through the energy barrier, (b) with strong external bias, electron tunneled through the energy barrier as shown by the red-dotted line. E_c and E_v are the silicon's conduction and valence bands respectively	28
2.9	Energy band representation of Si-SiO ₂ -Poly Si system showing hot-electron injection in the oxide. The oxide field is low but the electrons are heated by high lateral fields in the channel. Some of the electrons would acquire enough energy to overcome the energy barrier as represented by the red-dotted line. E_c and E_v are the silicon's conduction and valence bands respectively	31
2.10	Energy band representation of Si-SiO ₂ -Poly Si system showing direct tunneling of electrons in the oxide when the oxide thickness is less than 4 nm. E_c and E_v are the silicon's conduction and valence bands respectively	32
2.11	(a) Schematic of band diagram during retention for a flash memory with thicker tunnel oxide (6-10 nm) (b) Similar schematic during retention for flash memory with thinner tunnel oxide (< 6 nm)	39
2.12	The schematic representing the generated traps after P/E cycles facilitate the loss of electrons from the floating gate during low field condition. The red-dotted arrow is the graphical representation of the electron path during the escape, in the form of SILC	40
2.13	Conduction band edge diagrams for tunnel barriers without (solid lines) and with	46

	(dashed lines) applied electric field: (a) conventional barrier, (b) ideal crested barrier; and (c) tri-layer crested barrier with F-N tunneling (bold arrows) through sub-band. U and d are the energy barrier and dielectric thickness respectively.	
2.14	Tunneling current density, J (in A/m^2 , dashed lines) corresponding to the barrier schemes in Figs. 2-13(a) and 2-13(b), as a function of applied voltage V.	47
2.15	Band diagram illustrating the VARIOT concept at flatband and under applied voltage V. The low-k dielectric has a thickness t_L and dielectric constant k_L , and the high-k dielectric has a thickness t_H and dielectric constant k_H . (a) two-layer barrier (b) three-layer barrier	49
2.16	I-V curves showing current density across VARIOT (dark line) versus single layer SiO_2 (red line) stacks with the same EOT.	50
3.1	Schematic showing the FG device constraint using I-V curve	55
3.2	Floating Gate Capacitor Model	58
3.3	Sketch of wear-out/breakdown process showing trap generation which lead to destructive oxide breakdown	63
3.4	Current-Voltage characteristics of 10 nm SiO_2 showing different type of SILCs	66
3.5	Normalized leakage current versus stress time	67
3.6	(a) Bonding angle for Si – O – Si system showing the bridging angle θ . (b) Bond energy distribution of the bridging oxygen bond	69
3.7	Schematic of MOS capacitor used a test structure to represent the floating gate flash cell.	74
3.8	MOS capacitors fabrication process flow, showing the main fabrication steps such as thermal oxidation, deposition, lithography and etch	76
3.9	TEM pictures of MOS Capacitor test structure using 8 nm single-layers SiO_2 (a) 64,000x magnification (b) 225,000x magnification.	76
3.10	The summary of I-V and C-V electrical characterizations	79
3.11	The diagram showing the set up for I-V characterization using Keithley's Model 4200-SC	80
3.12	The C-V behavior for 2 conditions: (a) low frequency when the minority carriers in inversion contribute fully to the measured capacitance; (b) high frequency when the minority carriers do not contribute to the measured capacitance. The flatband voltage V_{FB} and threshold voltage V_{TH} also could be extracted from the curves	83
3.13	Current-Voltage curves of $2.5 \times 10^{-5} \text{ cm}^2$ MOS capacitors with 2 – 12 nm SiO_2 tunnel oxides	86
3.14	The average SBD (out of 30 capacitors) field for the respective SiO_2 thickness under V-Ramp test	87
3.15	Current-Voltage curve for MOS capacitor with 4 nm SiO_2 . Voltage is ramped until HBD	87
3.16	J versus E plot of $2.5 \times 10^{-5} \text{ cm}^2$ MOS capacitors with 2 – 12 nm SiO_2 tunnel oxides.	89
3.17	Measured versus calculated current density at pre-tunneling field for 4nm tunnel oxide	89
3.18	F-N Plot for MOS capacitors with tunnel oxides thickness of 2 and 4 nm. The black and red dotted lines represent linear extrapolation for 2 nm and 4 nm oxides respectively	90
3.19	J versus E plot for 4nm oxynitrides, compared with the corresponding conventional oxide	91
3.20	J versus E plot for oxynitride with growth conditions of (15% N_2 :85% O_2) as compared to corresponding conventional oxide.	91

3.21	Low field characteristics for conventional tunnel oxide	94
3.22	Low field characteristics for oxynitride	95
3.23	J versus E plot for conventional oxide after 10, 100 and 1000 second constant voltage stress	96
3.24	The normalized pre-tunneling current increase versus stress time for conventional oxides at 1.5 MV/cm	97
3.25	The normalized pre-tunneling current increase versus stress time for oxynitrides, compared with 4 nm conventional oxides at 1.5 MV/cm	98
3.26	High Frequency CV (HF C-V) Curves for the conventional oxide with the thickness of 4 nm with the capacitor area of $2.5 \times 10^{-5} \text{ cm}^2$, with stressed at 7 MV/cm for 10, 100 and 1000 seconds	101
3.27	N_{EFF} as a function of stress time for 2, 4 and 6 nm oxides	102
3.28	N_{EFF} as a function of stress time for 2, 4 and 6 nm oxides and oxynitrides	103
3.29	N_{EFF} as a function of stress time for 4 nm oxynitrides grown at 850°C with various N_2/O_2 ratios.	104
3.30	N_{EFF} as a function of stress time for 4 nm oxynitrides grown at various thermal levels with 30% N_2/O_2 ratio.	104
3.31	The summary of Programming Time for both oxides and oxynitrides.	107
3.32	SILC and N_{EFF} as a function of film thickness, measured after 1000 seconds of voltage stress at 7 MV/cm	109
3.33	The summary of Retention Time for both Oxides and Oxynitrides	109
3.34	Retention time extrapolation for pure oxide	110
3.35	Retention time extrapolation for oxynitride	110
4.1	Schematic of 2-layer VARIOT tunnel barrier (a) Device cross-section	114
4.2	Schematic of 3-layer VARIOT tunnel barrier (a) Device cross-section	116
4.3	Schematic of 2-layer VARIOT MOS Capacitor	122
4.4	Schematic of 3-layer VARIOT MOS Capacitor	122
4.5	The VARIOT Capacitor Fabrication Process Flow	123
4.6	TEM pictures of VARIOT capacitor test structure with 6 nm SiO_2 / 4 nm Si_3N_4 2-Layer configuration. (a) 225,000x magnification, (b) 410,000x magnification.	125
4.7	Simulated J versus V_g for 4 nm EOT 2-Layer Tunnel Barrier, compared with the simulated 4 nm single layer SiO_2	128
4.8	Simulated J versus V_g for 4 nm EOT 2-Layer Tunnel Barrier at low field, compared with 4 nm single layer SiO_2	130
4.9	Simulated J versus V_g for 4 nm EOT 2-Layer Tunnel Barrier at high field, compared with 4 nm single layer SiO_2 .	130
4.10	Programming voltage, V_{pp} reduction for engineered tunnel barrier NAND flash	131
4.11	Simulated J versus V_g for 6 nm EOT 2-Layer Tunnel Barrier, compared with the simulated 6 nm single layer SiO_2 .	131
4.12	Simulated J versus V_g for 8 nm EOT 2-Layer Tunnel Barrier, compared with the simulated 8 nm single layer SiO_2	132
4.13	I-V characteristics comparison for 4, 6 and 8 nm EOT with fixed (1 nm) bottom oxide thickness	132
4.14	Programming Voltages comparison for 4, 6 and 8 nm EOT with fixed (1 nm) bottom oxide thickness.	133
4.15	The I-V curves of 4 nm EOT 3-layer tunnel barrier compared with single layer tunnel barrier with the same EOT	136
4.16	Programming voltages comparison for 4 nm EOT 3-layer barrier with fixed (1 nm)	137

	top oxide thickness	
4.17	The I-V curves showing the tunneling current density of ETB stacks surpassing that of single tunnel barrier	138
4.18	The I-V curves of stack with the thickest physical thickness showing the highest tunneling current density at 10 MV/cm of electric field.	138
4.19	Programming voltages comparison for 6 nm EOT 3-layer barrier with fixed (1 nm) top oxide thickness	139
4.20	I-V curves showing the tunneling current density of 8 nm ETB stacks at low fields	139
4.21	I-V curves of stack with the thickest physical thickness showing the highest tunneling current density at 10 MV/cm of electric field.	140
4.22	I-V curves showing stack with stacks the minimum bottom oxide and a maximum nitride (red-dotted line) thickness performs better than others.	140
4.23	Energy band diagrams (a) Single layer tunnel barrier (b) VARIOT 3-layer tunnel barrier. The red-dotted lines in both cases are the energy bands as a result of the applied voltages.	142
4.24	Band diagrams of (a) a conventional SiO ₂ tunnel barrier and of (b) a multi-layer tunnel barrier under low field. Without the applied bias, electrons could not tunnel through the barrier	143
4.25	Summary of Programming Time for 2 and 3-Layer ETB	146
4.26	Experimental versus Simulation Results for 4 nm EOT 2-Layer ETBs	148
4.27	Experimental versus Simulation Results for 4 nm EOT 2-Layer ETBs: Tunneling Current Density	148
4.28	I-V curves comparison for 4, 6 and 8 nm EOT 2-Layer ETB	149
4.29	I-V curves comparison for 4, 6 and 8 nm EOT 3-Layer ETB	149
4.30	2-Layer versus 3-layer ETBs.	150
4.31	Plot of Physical ETB Thickness versus Data Retention Time	154
4.32	J versus E plot for 2 and 3-layer ETBs with 4 nm EOT after 1000 s constant voltage stress	156
4.33	J versus E plot for 2-layer ETBs with 8 nm EOT after 1000 s constant voltage stress.	157
4.34	J versus E plot for 3-layer ETBs with 8 nm EOT after 1000 s constant voltage stress	157
4.35	The normalized pre-tunneling current versus stress time for several ETB's main configurations.	158
4.36	N_{EFF} as a function of stress time for several ETB's main configurations.	158

LIST OF SYMBOLS

ψ	Psi
π	Pi
λ	Lambda
λ_D	Extrinsic Debye length
α	Capacitance coupling ratio
ϵ_0	Permittivity of free space, 8.85×10^{-14} F/cm
ϵ_{Si}	Relative permittivity of Si, $11.9\epsilon_0$
ϵ_{SiO_2}	Relative permittivity of SiO ₂ , $3.9\epsilon_0$
ϵ_{N_1}	Relative permittivity of Si ₃ N ₄ layer, $7.8\epsilon_0$
ϵ_{O_1}	Relative permittivity of bottom SiO ₂ layer, $3.9\epsilon_0$
ϵ_{O_2}	Relative permittivity of top SiO ₂ layer, $3.9\epsilon_0$
k	Boltzmann constant, 1.38×10^{-23} eV/°K
kT	Thermal energy at room temperature, 4.046×10^{-21} J
h	Planck constant, 6.625×10^{-31} J-s
\hbar	Planck constant over 2π , $\frac{6.625 \times 10^{-31} \text{ J-s}}{2\pi}$
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
ν	Traps escape frequency
ϕ_B	The barrier height at the conductor and insulator interface
ϕ_{BN_1}	The barrier height of Si ₃ N ₄ layer
ϕ_{BO_1}	The barrier height of bottom SiO ₂ layer
ϕ_{BO_2}	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, 1.60×10^{-19} C
k_H	Dielectric constant of high-k material
k_L	Dielectric constant of low-k material
m^*	Mass of free electron, 9.1×10^{-31} kg
m_{ox}	Electron effective mass in SiO ₂ , $0.45m^*$
m_{N_1}	Electron effective mass in Si ₃ N ₄ layer, $0.3m^*$
m_{O_1}	Electron effective mass in bottom SiO ₂ layer, $0.45m^*$
m_{O_2}	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
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

© This item is protected by original copyright

LIST OF ABBREVIATIONS

Al ₂ O 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 ₃ N ₄	Silicon Nitride
SiO ₂	Silicon Dioxide
SiO _x N _y	Silicon Oxynitride
Ta ₂ O 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
TDDDB	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 dicecilkan sehingga ke dimensi nanometer, get terapung kilat menghadapi satu laluan yang sukar. Pengecilan oksida penerowong mempunyai limit praktikal sekitar 8 nm disebabkan keperluan pengekal 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 lengkung 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 selama 100 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 . Mengambil kira 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 dielektrik berlapisan-3 yang 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 nitrated 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 nitrated 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