A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND THE COMMITTEE ON GRADUATE STUDIES OF STANFORD UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Shyam Raghunathan

October 2010
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Krishna Saraswat, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Yoshio Nishi, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tejas Krishnamohan

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
This page is intentionally left blank
Abstract

Flash memory is the most widely used non-volatile information-storage device today. NAND flash memories are ubiquitous in their use as portable storage media in cellphones, cameras, music players, and other portable electronic devices. In addition, NAND flash memory has recently seen rapid adoption as solid-state drives (SSD) in place of hard-disk drives (HDD) in modern personal computers and data servers. In addition to greater speed and better endurance against mechanical damage, SSDs also consume much lesser power than HDDs.

The NAND flash memory device, consisting of a floating-gate transistor cell, is the most aggressively scaled electronic device, as evidenced by ever-increasing memory capacities. In this work, we will examine possible problems arising from continued scaling of these structures, and discuss novel solutions to overcome them.

Firstly, we investigate scaling of the conventional poly-silicon floating-gate, aimed at reducing cell-to-cell interference. We experimentally delineate a new reliability concern for the first time, with programming current through ultra-thin
Abstract

Poly-silicon floating-gates becoming increasingly ballistic. We also experimentally demonstrate doping-related issues in the poly-silicon floating-gate.

We then apply a novel metal-based floating-gate cell for the first time, designed to overcome the problems discussed above. We explore factors that influence the choice of metal, and demonstrate excellent functionality in ultra-thin metal floating-gate cells scaled down to 3 nm TiN floating-gate thickness, thus greatly reducing cell-to-cell interference.

Finally, in order to facilitate continued scaling of the control dielectric, we explore replacement of the conventional silicon oxide-nitride dielectric with high-k dielectric materials. We integrate poly-silicon floating-gate cells with Al₂O₃ high-k control dielectric, and show that the presence of a silicon nitride inter-layer improves the interface between the floating-gate and the control dielectric. We also integrate metal floating-gate cells with Al₂O₃ control dielectric, and these cells exhibit very good electrical characteristics. Further, we establish that a deeper work-function control gate is helpful in reducing gate-injection.

Combining ultra-thin metal floating-gate, high-k control dielectric and deep work-function control gate, we enable the planar floating-gate cell as a scalable candidate.
Acknowledgements

This work would not have been possible without the help and guidance of many individuals. First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Krishna Saraswat. Right from the day that I arrived at Stanford, his constant support and guidance has been invaluable to my progress. He is undoubtedly one of the kindest people I have ever met, and his calm and cheerful demeanor is very inspiring. I would like to specially thank him for his support when I wanted to change my topic of work. His support in encouraging his students to explore a range of topics, is critical to the successful and enjoyable experience that his students have. I have been very fortunate to have had the best advisor anyone could hope for, and it has been an honor to be part of his illustrious research group.

I would like to express my sincere gratitude to my co-advisor, Prof. Yoshio Nishi. I have received many valuable suggestions and feedback from him during the course of my time at Stanford. I am very grateful to have been acquainted with him.

I am indebted to Prof. Tejas Krishnamohan for his guidance and support all through my time here. Even as a student in the group, he was an inspiration to me. He is truly one of the smartest people I have ever interacted with, and his guidance has
Acknowledgements

been extremely critical to my progress. I am especially thankful for the long discussions that we often had late in the night, in spite of his incredibly busy schedule. I am extremely fortunate to have had the opportunity to be mentored by him.

I would like to thank Prof. Pierre Khuri-Yakub for agreeing to chair my orals. I am very grateful to Dr. Ann Marshall for her help with TEM imaging. I would like to express my gratitude to Dr. Jim McVittie and Dr. Eric Peroziello for all their help in the lab. I am very thankful to Prof. Ted Kamins for the many useful discussions and feedback.

All my experimental work was performed at the Stanford Nanofabrication Facility. I would like to thank the staff for enabling a world-class research facility.

I am very grateful to Dr. Pranav Kalavade, Dr. Krishna Parat, Dr. Kiran Pangal, Dr. Prashant Damle and Dr. Sanjay Rangan for their guidance and support during my internship at Intel.

I have been fortunate to have had the support of amazing friends and colleagues over the last few years at Stanford: Abhijit, Albert, Ali, Ammar, Aneesh, Anshul, Arunanshu, Arvind, Bob, Caner, ChiOn, Crystal, David, Debbie, Donguk, Donghyun, Donkoun, Duygu, Ed, Elmer, Emel, EricP, Eunji, Filip, Gail, Gaurav, Girish, Gunhan, Hai, Hemanth, Hiro, Hoon, Hoyoel, Hyun-Yong, Irene, J, Jae, Jason, Jeannie, Jenny, Jia, Jim, Jin-hong, John, Jungyup, Kavya, Kiran, Kishore, Krishna, Kyeongran, Kyung-hoaee, Lan, Mahnaz, Manar, Manoj, Marika, Mary, Masaharu, Masato, Maurice, Mihir, Munehiro, Nancy, Nevran, Nishant, Onur, Paul, Pawan, Peter, Pranav, Prasanthi, Prashant, Raghav, Raja, Rajesh, Ray, Rishi, Robin, Rohit, Rostam, Roozbeh, Sangbum, Sanjay, Sarves, Shankar, Serene, Shreekar, Simon,
Acknowledgements

Suman, Sumanth, Suyog, Swaroop, Szu-lin, Shen, Ted, Uli, Wooshik, Yasuhiro, Yeul, Yuan, Yuniarto, Ze.

No words would ever do justice to express my deepest thanks and gratitude to my most wonderful family – my parents and my sister. Their unconditional love, sacrifice, support and encouragement have allowed me to pursue my ambitions.
Acknowledgements

This page is intentionally left blank
# Table of Contents

Abstract ................................................................................................................................. v

Acknowledgements ............................................................................................................. vii

Table of Contents ............................................................................................................... xi

List of Tables ...................................................................................................................... xiv

List of Figures ..................................................................................................................... xv

Chapter 1 – Introduction ..................................................................................................... 1
  1.1 Motivation ...................................................................................................................... 1
  1.2 Thesis Organization .................................................................................................... 5

References ............................................................................................................................. 8

Chapter 2 – Flash Memory: Operation and Scaling .......................................................... 9
  2.1 Introduction ................................................................................................................... 9
  2.2 What is Flash Memory? ............................................................................................... 9
  2.3 The Floating-gate Cell ............................................................................................... 12
     2.3.1 Gate coupling ratio (GCR) .................................................................................. 15
  2.4 Read Operation .......................................................................................................... 16
  2.5 Multi-level Cell (MLC) ............................................................................................. 17
  2.6 Types of Flash Memory ............................................................................................. 18
  2.7 Mechanisms of Charge Injection ............................................................................. 21
     2.7.1 Channel hot electron (CHE) injection ................................................................. 21
     2.7.2 Fowler-Nordheim tunneling ............................................................................... 23
Table of Contents

2.8  Reliability of Flash Memories ................................................................. 27
    2.8.1  Endurance (Cycling) ...................................................................... 27
    2.8.2  Retention ......................................................................................... 28

2.9  Flash Memory Scaling ............................................................................. 30
    2.9.1  Challenges in scaling NOR flash ....................................................... 31
    2.9.2  Challenges in scaling NAND flash ..................................................... 33

References ....................................................................................................... 37

Chapter 3 – Scaled Poly-Silicon Floating-Gate NAND Flash Memory ............ 41

3.1  Introduction ............................................................................................... 41

3.2  Motivation .................................................................................................. 42
    3.2.1  Cell-to-cell interference ................................................................... 42
    3.2.2  Control gate wrap-around ................................................................. 45

3.3  Floating-gate thickness scaling ................................................................. 49

3.4  Experimental Test Structure ..................................................................... 50

3.5  Fabrication ................................................................................................. 52

3.6  Measurement Setup ................................................................................... 55

3.7  Program/Erase characteristics .................................................................. 56

3.8  Retention .................................................................................................... 62

References ....................................................................................................... 64

Chapter 4 – Scaled Poly-Silicon Floating-Gate: Problems to Ultimate Limit ...... 67

4.1  Introduction ............................................................................................... 67

4.2  The Ballistic Problem ............................................................................... 68
    4.2.1  Ballistic transport ........................................................................... 68
    4.2.2  Ballistic transport during programming ............................................. 69
    4.2.3  Test Structure ................................................................................... 71
    4.2.4  Measurement ..................................................................................... 73
    4.2.5  Results ............................................................................................... 77
    4.2.6  Verification ......................................................................................... 81
    4.2.7  Mean Free Path ................................................................................ 84
    4.2.8  Impact (Endurance) ......................................................................... 85
    4.2.9  Summary ........................................................................................... 89
4.3 Poly-silicon depletion ................................................................. 89
4.3.1 Poly depletion effect ............................................................ 89
4.3.2 Poly depletion in NAND flash memory ................................. 91
4.3.3 Experiment ........................................................................... 93
4.3.4 Summary ............................................................................... 98

References .................................................................................. 99

Chapter 5 – Metal Floating-Gate ....................................................... 101
5.1 Introduction ........................................................................... 101
5.2 Metal Floating-Gate Device ...................................................... 102
5.3 Ballistic Component in Metal Floating-Gates .............................. 105
5.4 Absence of Poly-depletion effects ............................................ 110
5.5 Metal floating-gate work-function ............................................ 112
5.6 Choice of metal : TiN, TaN ...................................................... 115
5.7 Ultra-thin 3 nm TiN floating-gate device .................................... 116
5.8 Summary .................................................................................. 118

References .................................................................................. 119

Chapter 6 – High-k dielectric IPD ..................................................... 121
6.1 Introduction ........................................................................... 121
6.2 Motivation ............................................................................... 122
6.3 Choice of high-k dielectric for IPD application .......................... 125
6.4 Poly-silicon FG with Al₂O₃ IPD ................................................ 128
6.5 Metal floating-gate device with Al₂O₃ IPD ................................. 135
6.6 Control-Gate ....................................................................... 139
6.7 Summary ............................................................................... 139

References .................................................................................. 142

Chapter 7 – Conclusions and Future Directions ................................. 145
7.1 The Planar Floating-Gate Device ............................................. 145
7.2 Future Directions ................................................................... 148

References .................................................................................. 142
List of Tables

Table 2.1: Major differences between NOR and NAND flash memories

Table 5.1: Work-functions and Fermi energies of some elements [5.11, 5.12]

Table 6.1: Common high-k dielectrics [6.14]
List of Figures

Figure 1.1 Applications of flash memory.............................................................. 1
Figure 1.2: Demand for NAND flash memory (Source: Gartner May 2009, Semiconductor Forecast Database (SEQS-WW-DB) [1.2]) .............................................. 2
Figure 1.3: Comparison of SSD and HDD (Source: Micron) .................................. 3
Figure 1.4: World-wide data center power savings with SSD compared to HDD. (Source: Samsung) ........................................................................................................ 3
Figure 1.5: Scaling of NAND flash memory. The pace is 2X per year as opposed to 2X every 1.5 years according to Moore’s Law. (Source: Samsung)...................... 4
Figure 2.1 : Non-volatile MOS memory qualitative comparison in the flexibility-cost plane (From [2.1]) ........................................................................................................ 10
Figure 2.2: Semiconductor memory market for flash, DRAM and SRAM (Source: [2.1], [2.10]) ................................................................................................................. 11
Figure 2.3: Cross-section schematic of a floating-gate transistor .............................. 12
Figure 2.4: Band diagrams for the neutral state (left) and charged state (right) of the floating-gate transistor. (Source: [2.1]) ................................................................. 13
List of Figures

Figure 2.5: Schematic showing the capacitive coupling of the floating-gate (Source [2.12]) ................................................................. 14
Figure 2.6: I–V curves of an FG device when there is no charge stored in the FG (curve A) and when a negative charge Q is stored in the FG (curve B) [2.12, 2.15]... 17
Figure 2.7: Two different threshold voltage states lead to 1 bit per cell storage while four different threshold voltage states lead to 2 bits per cell ......................... 18
Figure 2.8: NOR and NAND flash arrays with layout comparison (Source: Intel [2.11]) ................................................................. 19
Figure 2.9: NOR flash programming using CHE injection [Source: Intel] ............... 22
Figure 2.10: A schematic energy band diagram describing the three processes involved in electron injection [2.22] ................................................................. 23
Figure 2.11: F-N tunneling of electrons from the poly-Si FG into the Si substrate through the triangular energy barrier posed by the tunox. (Source: [2.12]) ................. 24
Figure 2.12: F-N tunneling current as a function of electric field (Source: [2.12], [2.27]) ........................................................................ 25
Figure 2.13: F-N programming in NAND flash (Source [2.11]) ......................... 26
Figure 2.14: Threshold voltage window closure caused by traps in the tunnel oxide as a function of program/erase cycles on a single cell, showing endurance characteristics [2.1]. ......................................................................................... 28
Figure 2.15: Band diagram showing charge-retention state and leakage path through dielectrics ................................................................................................. 29
Figure 2.16: NOR flash memory scaling (Source: Intel [2.11]) ......................... 30
Figure 2.17: NAND flash Bit Density Scaling Trend (Source: [2.34]) ................. 31
Figure 2.18: NOR flash ITRS Scaling Projection [2.36]........................................33
Figure 2.19: ITRS Scaling projections for floating-gate NAND flash. [2.36]..........36
Figure 3.2: Floating-gate interference causes undesirable shift in threshold voltage of
one cell due to adjacent cells (Source: [3.7]).........................................................43
Figure 3.3: This figure depicts the shifting of threshold voltage of Cell 1 when Cell 2
gets programmed due to the parasitic coupling between the adjacent cells. (Source:
[3.6]).........................................................................................................................44
Figure 3.4: The undesirable threshold voltage shift (interference) is getting worse with
scaling. (Source: [3.8], [3.9])..................................................................................44
Figure 3.5: Control-gate is wrapped around the FG shielding interference between
adjacent floating-gates.................................................................46
Figure 3.6: Image showing CG wrapping around the FG [3.10].................................46
Figure 3.7: Changing the height of the CG wrap-around has a significant effect on the
word-line direction interference [3.10].................................................................47
Figure 3.8: Interference along BL direction and WL direction [3.11].........................47
Figure 3.9: CG no longer has space to wrap around FG thereby leading to a serious
interference problem and GCR decrease.................................................................48
Figure 3.10: Image showing that the distance between cells is extremely less (Source:
[3.13]).......................................................................................................................48
Figure 3.11: Reduction of FG height greatly reduces FG interference.....................49
Figure 3.12: Shows that for future nodes, required FG height is a few nm. (Source:
[3.16]).......................................................................................................................50
Figure 3.13: Schematic of experimental test structure.............................................51
List of Figures

Figure 3.14: TEM image of the gate with ultra-thin 7 nm poly-silicon floating-gate. A very smooth layer is obtained at such a low thickness. .............................................................. 54

Figure 3.15: TEM image showing ultra-thin 7 nm poly-silicon floating-gate .......... 55

Figure 3.16: Programming characteristics of 75 nm thick n-type poly-silicon floating-gate ................................................................................................................................................ 57

Figure 3.17: Erasing characteristics of 75 nm n-type poly-silicon floating-gate ...... 58

Figure 3.18: Programming characteristics for various thicknesses of n-type poly-silicon FG devices ............................................................................................................................................ 61

Figure 3.19: Erasing characteristics for various thicknesses of n-type poly-silicon FG devices ............................................................................................................................................ 61

Figure 3.20: Retention for n-type poly-Si FG devices of various thicknesses .......... 62

Figure 4.1: Schematic band diagram of ballistic transport and emission via multiple tunneling through interconnected nanocrystallites in porous silicon under a high electric field. [4.7] ............................................................................................................. 69

Figure 4.2: Band diagram describing ballistic transport during programming. ........ 70

Figure 4.3: Test Structure used to study ballistic transport ...................................... 71

Figure 4.4: Band diagram during measurement at low FG bias. Substrate is grounded and CG is fixed at 20 V. ............................................................................................................................................. 74

Figure 4.5: Shows the IPD leakage current at the start of measurement.................. 74

Figure 4.6: Band diagram of measurement at high FG bias .................................... 75

Figure 4.7: Shows schematics of expected currents at mid-to-high floating-gate bias. 75

Figure 4.8: Schematic of expected currents during the ballistic current measurement technique described. ........................................................................................................................................... 77
List of Figures

Figure 4.9: Ballistic measurement data: 75 nm n-type poly-silicon floating-gate device ....... 78

Figure 4.10: Ballistic component measurement data in 30 nm n+ poly-silicon floating-gate device ............................................................................................................. 79

Figure 4.11: Ballistic component measurement data for 10 nm n+ poly-silicon floating-gate sample ........................................................................................................... 80

Figure 4.12: Shows two floating-gate structures with the same IPD area (region between FG and CG) but different active areas (tunnel oxide region) ...................... 82

Figure 4.13: Ballistic component measurement data on 11 nm poly-silicon floating-gate sample for structures with different active areas ................................................... 83

Figure 4.14: Magnified version of figure 4.13 .................................................................. 83

Figure 4.15: Inelastic mean free path extraction by slope of ballistic component versus floating-gate thickness .................................................................................. 85

Figure 4.16: Cycling data (+- 14 V, 100 ms) on 75 nm, 10 nm and 7 nm poly-silicon floating-gate devices. The 7 nm floating-gate devices show more degradation .......... 86

Figure 4.17: Cycling data (+17 V, 100 µs and -14 V, 100 ms) on 75 nm and 10 nm poly-silicon floating-gate devices. The 10 nm floating-gate devices show more degradation ........................................................................................................... 87

Figure 4.18: Capacitance-Voltage curves of cycled devices indicating more degradation in the 10 nm poly-silicon floating-gate device compared to the 75 nm floating-gate device. ........................................................................................................... 88

Figure 4.19: The poly-depletion effect. Due to the depletion in the poly-silicon gate, the effective oxide thickness increases and the effective capacitance decreases. Source: [4.13] .......................................................................................................................... 90
List of Figures

Figure 4.21: As a result of the poly-depletion, there is a decreased field across the tunnel oxide which makes the erase slower. Source: [4.15]

Figure 4.22: Shows the capacitance-voltage plots for poly-silicon floating-gate devices. The 7 nm device shows a characteristic poly-depletion dip in the capacitance.

Figure 4.23: The poly-depletion effect is much more pronounced for larger emulated wraps.

Figure 4.24: Medici simulation potential contours for depleted and undepleted FG. Shows effect of emulated wrap in amplifying poly-depletion.

Figure 4.25: Shows slower programming in 7 nm poly-silicon floating-gate device due to poly-depletion.

Figure 5.1: Floating-gate capacitor structure used to study program/erase and reliability performance of metal floating-gate devices.

Figure 5.2: TEM of a metal (TiPt) floating-gate device gate-stack.

Figure 5.3: Programming characteristics of TiPt floating-gate device.

Figure 5.4: Erasing characteristics of TiPt floating-gate device.

Figure 5.5: Structure with floating-gate contact used to measure the ballistic component of tunnel oxide current.

Figure 5.6: Ballistic component measurement data on 7 nm TiPt metal floating-gate sample. The metal has much lower ballistic component compared to poly-silicon of the same thickness.
Figure 5.7: Percentage of ballistic current plotted versus floating-gate thickness for a variety of materials. In general, it can be observed that metals have lower ballistic component compared to poly-silicon due to increased inelastic electron scattering. 107

Figure 5.8: Shows the definition of work-function and Fermi energy in a metal. Also shows the density of states (DOS) versus energy. 108

Figure 5.9: Capacitance-Voltage plot of 3 nm TiN FG capacitor showing no characteristic poly-depletion dip in the capacitance. 111

Figure 5.10: Flat-band voltage versus programming time for 3 nm TiN FG sample. There is no slow-down in the beginning as observed with depleted poly-silicon samples due to decreased GCR. 111

Figure 5.11: Band diagrams showing that deeper work-function FG leads to slower erase. Top figure shows shallower work-function FG device and bottom figure shows deeper work-function FG device. Deeper work-function FG has larger barrier to tunneling out from FG to substrate. 113

Figure 5.12: Band diagrams showing that deeper work-function FG leads to better retention. Top figure shows shallower work-function FG device and bottom figure shows deeper work-function FG device. Deeper work-function FG device stores charge in a deeper potential well. 114

Figure 5.13: Shows deeper work-function (Pt) FG device erase slower than shallower work-function (Ti+Pt) FG device. 115

Figure 5.14: Programming characteristics of 3 nm TiN FG device showing excellent performance. 117
List of Figures

Figure 5.15: Erasing characteristics of 3 nm TiN FG device showing excellent performance .................................................................................................................. 117
Figure 5.16: Retention data for 3 nm TiN floating-gate device ............................... 118
Figure 6.1: Image of control-gate wrapping around floating-gates [6.11] ............ 123
Figure 6.2: CG no longer has space to wrap around FG thereby leading to a serious interference problem and GCR decrease ........................................................................ 123
Figure 6.3: Image shows that the distance between cells is very less (Source: [6.12]) ......................................................................................................................................................... 124
Figure 6.4: Shows erase-saturation occurring as a result of gate-injection. A larger conduction barrier height of the IPD would reduce injection from the control-gate. 126
Figure 6.5: Common high-k dielectrics. In general, larger the dielectric constant, smaller the band-gap. [6.14] .................................................................................................................. 127
Figure 6.6: TEM image of the gate-stack of a poly-silicon FG Al₂O₃ high-k IPD device................................................................................................................................................ 128
Figure 6.7: Program characteristics of 60 nm poly-silicon floating-gate devices with different IPD layers. ........................................................................................................ 130
Figure 6.8: Erase characteristics of 60 nm poly-silicon floating-gate devices with different IPD layers. ........................................................................................................ 130
Figure 6.9: Cycling data on 60 nm poly-silicon FG Al₂O₃ high-k IPD samples with and without nitride inter-layer. The samples were cycled to +/- 3 V flat-band voltage. The sample with nitride inter-layer performs better .................................................................................. 131
Figure 6.10: Retention characteristics for 60 nm poly-silicon FG Al₂O₃ high-k IPD device with and without nitride interlayer .................................................................................. 132
Figure 6.11: Program characteristics of poly-silicon FG devices of various thicknesses with nitride + Al₂O₃ high-k IPD. ................................................................. 133

Figure 6.12: Erase characteristics of poly-silicon FG devices of various thicknesses with nitride + Al₂O₃ high-k IPD. ................................................................. 133

Figure 6.13: Cycling data of 60 nm and 7 nm poly-silicon FG devices with nitride + Al₂O₃ high-k IPD. The 7 nm poly-silicon FG sample performs worse, likely due to ballistic carrier damage. ................................................................. 134

Figure 6.14: TEM image of a TiN floating-gate device with Al₂O₃ high-k IPD ....... 135

Figure 6.15: Program characteristics of TaN FG compared to poly-silicon FG, both with high-k IPD. ................................................................................................. 136

Figure 6.16: Erase of TaN FG compared to poly-silicon FG, both with high-k IPD. 137

Figure 6.17: Retention of TaN FG device with high-k IPD .................................. 137

Figure 6.18: Endurance data for TaN FG device with Al₂O₃ high-k IPD .......... 138

Figure 6.19: Band diagrams depicting deeper work-function metal with lesser gate-injection. ................................................................................................................. 140

Figure 6.20: Pt CG (deeper work-function) showing improved erase saturation compared to Al CG (shallower work-function) .............................................. 141

Figure 7.1: The planar (no wrap) floating-gate device ........................................... 146

Figure 7.2: Non-volatile memory potential solutions (Source: ITRS 2009) .......... 149
## Common Abbreviations Used

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FG</td>
<td>Floating-gate</td>
</tr>
<tr>
<td>CG</td>
<td>Control-gate</td>
</tr>
<tr>
<td>IPD</td>
<td>Inter-poly dielectric</td>
</tr>
<tr>
<td>ONO</td>
<td>oxide-nitride-oxide</td>
</tr>
<tr>
<td>CHE</td>
<td>Channel hot electron</td>
</tr>
<tr>
<td>F-N</td>
<td>Fowler-Nordheim</td>
</tr>
<tr>
<td>GCR</td>
<td>Gate-coupling ratio</td>
</tr>
<tr>
<td>SSD</td>
<td>Solid-state drive</td>
</tr>
<tr>
<td>tunox</td>
<td>Tunnel oxide</td>
</tr>
<tr>
<td>MLC</td>
<td>Multi-level cell</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

Flash memory is the most widely used non-volatile memory device today [1.1]. Flash memories are ubiquitous in their use as portable storage media in cellular phones, media players, cameras and other portable electronic devices (figure 1.1).

![Flash Memory Applications](image)

Figure 1.1 Applications of flash memory
The demand for these is projected to dramatically increase (figure 1.2) over the coming decade [1.2].

NAND-type flash memory, optimized for higher density, has overtaken dynamic random access memory (DRAM) as the technology driver today. A key new application of NAND flash memory is the solid-state drive (SSD). The solid-state drive is used as a replacement for hard-disk drives (HDD) in modern computers. Solid-state drives hold many advantages over HDDs, as shown in figure 1.3. In addition to providing better performance, SSDs also consume much lesser power than HDDs. This leads to massive savings in power/cost in data servers, shown in figure 1.4.
In order to keep up with the demand for increased memory capacities, flash memory has been continuously scaled (figure 1.5) to smaller and smaller dimensions.
Chapter 1 - Introduction

The scaling of NAND flash memory has been faster than Moore’s Law [1.3], referred to as Hwang’s Law [1.4] but has started to slow down due to difficulties in scaling.

Figure 1.5: Scaling of NAND flash memory. The pace is 2X per year as opposed to 2X every 1.5 years according to Moore’s Law. (Source: Samsung)

Further scaling of NAND flash memory faces serious roadblocks. Some of the biggest challenges include increased parasitic floating-gate (FG) interferences and decreased gate-coupling ratio (due to lack of space for control-gate (CG) wrap around the floating-gate) [1.5]. These arise due to the ultra-high density-scaling requirements of NAND flash memory. This thesis deals with these serious challenges in the continued scaling of NAND flash memory technology and demonstrates novel solutions to overcome them.
1.2 Thesis Organization

This thesis is organized into seven chapters, including this introduction chapter.

In chapter 2, brief overviews of flash memory history, operation and scaling are presented. The ‘floating-gate transistor’ cell is introduced with its operation principle. The types of flash memory are described and the main charge injection mechanisms are discussed. Then, reliability considerations of flash memory are briefly described, followed by the scaling trends and the main scaling challenges of flash memory.

Chapter 3 studies poly-silicon floating-gate devices with vertical scaling of the floating-gate. In the first section, the cell-to-cell interference challenge in scaling NAND flash memory is discussed, which motivates thickness scaling of the floating-gate. Subsequently the test structure used for this study, and the fabrication of scaled poly-silicon floating-gate devices, are described. This is followed by detailed analysis for the correlation in program/erase/retention characteristics between devices with poly-silicon floating-gate thicknesses ranging from 75 nm to ultra-thin 7 nm.

In chapter 4, important challenges to the scaling of the poly-silicon floating-gate thickness are discussed. A new reliability concern, arising from the programming current increasingly going ballistically through the ultra-thin floating-gate, is described. A new test structure is developed to experimentally investigate this phenomenon. Experimental results are presented which quantitatively describe the magnitude of the effect. The implication of this effect is shown by means of endurance
data, which show thinner floating-gate devices degrading faster. This reliability concern poses a major scaling challenge. Following the discussion of this problem, other issues of doping that limit ultra-thin poly-silicon floating-gates are discussed.

In chapter 5, the use of a metal floating-gate as a replacement for poly-silicon is explored. The metal floating-gate is experimentally demonstrated to provide excellent electrical performance. The metal floating-gate can also provide solution to the ballistic carrier and poly-depletion issues discussed in chapter 4. The factors underlying the choice of metal are discussed in detail. From the various options available, TiN (or TaN) is chosen as the preferred floating-gate. The TiN floating-gate is scaled down to ultra-thin 3 nm, with excellent electrical performance at even such thicknesses. Utilization of an ultra-thin metallic floating-gate greatly reduces cell-to-cell interference.

In chapter 6, the replacement of the conventional oxide-nitride control dielectric with high-k dielectric is explored. The motivation to use high-k control dielectric is to improve the gate-coupling ratio (GCR), which is lowered as a result of the inability to wrap the control-gate around the floating-gate. The requirements of the high-k material for control dielectric application are enumerated. Al₂O₃ is chosen as the high-k control dielectric material for experimental demonstration. Poly-silicon floating-gate devices are integrated with Al₂O₃ control dielectric, and their electrical characteristics are presented. The usefulness of a silicon nitride interlayer is proven, and metal (TaN) floating-gate is integrated with Al₂O₃ control dielectric. Finally, the role of the control-gate work-function is discussed.
In chapter 7, the conclusions and future directions are discussed, in which the planar floating-gate device is proposed with ultra-thin metal floating-gate, high-k control dielectric and deep work-function control-gate. Finally, directions for future research are discussed.
Chapter 1 - Introduction

References


[1.4] Hwang’s Law was named after Hwang Chang-gyu, former head of Samsung Electronics' semiconductor business.

Chapter 2

Flash Memory: Operation and Scaling

2.1 Introduction

In this chapter, a brief overview of flash memory operation and scaling are presented, including its short history, the ‘floating-gate transistor’ cell and its operation. The types of flash memory and the main charge injection mechanisms are discussed. Then, reliability considerations of flash memory are briefly described, with scaling trends and main scaling challenges.

2.2 What is Flash Memory?

Flash memory is a type of non-volatile memory [2.1]. A non-volatile memory device is one that can retain stored information in the absence of power (figure 2.1).
Other examples of non-volatile memory include read-only-memory (ROM), hard disk drives, floppy disks, optical discs etc.

![Non-volatile MOS memory qualitative comparison in the flexibility-cost plane (From [2.1])](image)

The first category of non-volatile MOS memories consists of ROMs, in which the data is permanently written during manufacturing. While these are very inexpensive, they completely lack flexibility as the data cannot be altered by users. The first floating-gate memory device was introduced by Kahng and Sze [2.2] in 1967. In 1970, Frohman-Bentchkowsky [2.3] developed a poly-silicon floating-gate transistor. In this device, hot electrons are injected into the floating-gate and removed by ultraviolet (UV) photoemission. This device was called the erasable programmable read-only memory (EPROM) or UV-EPROM. While this device provided more flexibility, it needed to be removed from the system to erase it by exposure to UV light.

Following this, there was a lot of effort to develop electrically erasable programmable ROMs (EEPROMs) [2.4, 2.5]. The EEPROM cell, consisting of two
transistors and a tunnel oxide was developed. While this had byte-rewrite and erasure capability, it was a larger cell since it consisted of two transistors.

Subsequently, the flash-EEPROM was developed by Toshiba [2.6] in 1984. A single-transistor cell was achieved by means of hot carrier programming and tunnel erase. According to Toshiba, the name "flash" was suggested because the erasure process of the memory contents reminded them of the flash of a camera.

The first commercial flash memory chip was produced by Intel in 1988 [2.7, 2.8]. This was a NOR-type flash memory chip where there is direct access to each cell. The first NAND-type flash memory, optimized for higher density, was developed by Toshiba in 1987 [2.9]. The growth of flash memory was sluggish initially due to reliability problems but took off dramatically after 2000 due to improvements in manufacturing process and reliability technology (figure 2.2).

![Figure 2.2: Semiconductor memory market for flash, DRAM and SRAM (Source: [2.1], [2.10])](image)
2.3 The Floating-gate Cell

A simple floating-gate transistor cell is shown in Figure 2.3.

![Figure 2.3: Cross-section schematic of a floating-gate transistor](image)

A floating-gate transistor is similar to a regular MOS transistor, except for the additional electrode called the floating-gate and an additional dielectric layer between the floating-gate and the main gate (or control-gate). The floating-gate is electrically isolated from surrounding electrodes on all sides by dielectrics. The dielectric layer between the floating-gate and the control-gate is called the control-dielectric or inter-poly dielectric (IPD) because both gates are usually made of poly-silicon. The inter-poly dielectric is typically composed of a triple layer of oxide-nitride-oxide (ONO) [2.11]. The dielectric closest to the substrate is called the tunnel dielectric (or tunox). The name originates from the fact that the erase operation and in some cases, the program operation, occurs through this oxide using quantum mechanical tunneling.
The source and drain are typically n+ type and the substrate is p-type (NMOS transistor).

The floating-gate serves as the charge storage node. This electrode is capacitively coupled to the control-gate and other nodes. During the programming operation, charge is injected into the floating-gate from the substrate using some mechanism, usually hot-electron injection or tunneling. The charge is sensed as a shift in the threshold voltage (the turn-on voltage) of the transistor. The charge is removed from the floating-gate by tunneling through the tunnel oxide back into the substrate. Figure 2.4 shows the band diagrams for the charged and neutral states of the floating-gate transistor.

Figure 2.4: Band diagrams for the neutral state (left) and charged state (right) of the floating-gate transistor. (Source: [2.1])

In the neutral state, there are no electrons on the floating-gate. Electrons are pushed from the substrate onto the floating-gate during a programming operation. In
the charged state, there are electrons on the floating-gate which raises the potential of the floating-gate. The potential well formed by the dielectrics on either side of the floating-gate, enables it to retain charge.

Figure 2.5 is a simple schematic of the floating-gate transistor where the various capacitive couplings to the floating-gate are shown. The floating-gate is capacitively coupled not only to the control-gate, but also to the source, drain and body of the transistor. Following [2.12], consider the case when there is no charge on the floating-gate ($Q = 0$),

$$Q = 0 = C_{FC} (V_{FG} - V_{CG}) + C_{S} (V_{FG} - V_{S}) + C_{D} (V_{FG} - V_{D}) + C_{B} (V_{FG} - V_{B})$$

Where

$C_{FC}$ is the capacitance between floating-gate and control-gate

$C_{S}$ is the capacitance between floating-gate and source

$C_{D}$ is the capacitance between floating-gate and drain
C_B is the capacitance between floating-gate and body

V_{FG} is the potential of the floating-gate

V_{CG} is the potential of the control gate

V_D is the potential of the drain

V_S is the potential of the source

V_B is the potential of the body

If \( C_{\text{Tot}} = C_{FC} + C_{D} + C_{S} + C_{B} \) = Total FG capacitance,

Then

\[
V_{FG} = GCR \cdot V_{CG} + DCR \cdot V_D + SCR \cdot V_S + BCR \cdot V_B
\]

Where

\( GCR = \text{Gate coupling ratio} = \frac{C_{FC}}{C_{\text{Tot}}} \)

\( DCR = \text{Drain coupling ratio} = \frac{C_{D}}{C_{\text{Tot}}} \)

\( SCR = \text{Source coupling ratio} = \frac{C_{S}}{C_{\text{Tot}}} \)

\( BCR = \text{Body coupling ratio} = \frac{C_{B}}{C_{\text{Tot}}} \)

### 2.3.1 Gate coupling ratio (GCR)

The Gate coupling ratio (GCR) is a very important parameter in the design of flash memories. It is a measure of how strongly the control gate is coupled to the floating-gate, and as a consequence, it determines the electric fields across the tunnel oxide and the control oxide. The field across the tunnel oxide is critical to the speed of operation of the device. A device with a lower GCR would operate at lower tunnel...
oxide fields compared to a device with higher GCR, at the same voltages. Therefore, the device with lower GCR needs higher voltages to operate at the same speed (the same tunnel oxide field). This also means that the field in the IPD would be larger, which is undesirable.

The GCR has been increased in flash memories by wrapping the control-gate around the floating-gate ([2.13]). However, due to increasing density of cells in NAND-type arrays, there is no longer enough space to wrap the control-gate around the floating-gate, thereby leading to a large decrease in GCR. This problem and the solution will be discussed in the following chapters.

An important point to note is that, in today’s high density arrays, the floating-gate is also capacitively coupled to adjacent floating-gates and other electrodes of adjacent cells causing interference in their operation. This has not been considered in the simple mathematical approach described above. This parasitic capacitive coupling between neighboring cells has started to become a significant fraction of the overall coupling and this is one of the biggest challenges facing high-density NAND flash memory technology today.

### 2.4 Read Operation

The state of the floating-gate transistor is ascertained from the threshold voltage of the FG MOS transistor. From elementary MOS theory [2.14], it is known
that a sheet of charge in the gate stack produces a shift in the threshold voltage. Therefore, following [2.12], we have

$$\Delta V_t = \text{shift in threshold voltage} = -\frac{Q}{C_{FC}}$$

Where

$Q$ is the charge on the floating-gate

$C_{FC}$ is the capacitance between CG and FG

And $\Delta V_t$ is shift in threshold voltage as measured at the CG

The shift in threshold voltage can be obtained by measuring the drain current ($I_D$) – control gate voltage ($V_{CG}$) transfer characteristics, as shown in Figure 2.6.

![I-V curves of an FG device](image)

Figure 2.6: I–V curves of an FG device when there is no charge stored in the FG (curve A) and when a negative charge $Q$ is stored in the FG (curve B) [2.12, 2.15].

### 2.5 Multi-level Cell (MLC)

Figure 2.6 shows two different levels of charge states, which are used to store ‘1’ and ‘0’ and hence, a single bit per cell. However, it is also possible to store many
more levels of charge in a cell. For example, a cell with 4 levels of charge states can be used to store ‘11’, ‘10’, ‘01’ and ‘00’, thereby storing 2 bits per cell (figure 2.7). Similarly, a cell with 8 levels of charge states can store 3 bits per cell. In general, a cell with ‘$2^n$’ charge states can store ‘n’ bits per cell [2.11, 2.13].

![Figure 2.7: Two different threshold voltage states lead to 1 bit per cell storage while four different threshold voltage states lead to 2 bits per cell.](image)

While multi-level cells are very important and useful in increasing density of storage in flash memories, they also make design much more restrictive and stringent due to the smaller window available between adjacent threshold voltage states [2.13].

### 2.6 Types of Flash Memory

There are two main types of flash memory in use today, called NOR flash memory [2.6] and NAND flash memory [2.9]. While both are based on the floating-gate transistor cell, their principal difference is in the way the cells are connected together in an array. The words ‘NOR’ and ‘NAND’, refer to the NOR-gate type and NAND-gate type connections of the cells in an array. Figure 2.8 shows NOR and NAND-type flash memory arrays.
From the figure, we can see that there is direct access to each cell in NOR-type flash memory. The control-gate electrode is connected to the word-line and the drain is connected to the bit-line. The sources are grounded. The NAND-type cell cross-section is similar except for the contacts. The contact corresponding to each cell is eliminated in the interest of higher packing density. However, the direct cell access present in NOR-type memory is not available in NAND-type memory. Hence, the latency for random access in NAND flash is much longer (1-10 microseconds) compared to NOR flash (10s of nanoseconds) due to the larger resistance of the line discharging the bit-line capacitance [2.11].

Figure 2.8: NOR and NAND flash arrays with layout comparison (Source: Intel [2.11])
The differences between NOR and NAND-type flash memories [2.11] are shown in Table 2.1.

<table>
<thead>
<tr>
<th>NOR Flash Memory</th>
<th>NAND Flash Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell size = 10(\lambda^2)</td>
<td>Cell size = 5(\lambda^2)</td>
</tr>
<tr>
<td>Direct access to cell</td>
<td>Indirect access to cell</td>
</tr>
<tr>
<td>Fast random access</td>
<td>Slow random access. Fast page access</td>
</tr>
<tr>
<td>Slower Program/Erase</td>
<td>Faster Program/erase</td>
</tr>
<tr>
<td>Channel hot electron programming</td>
<td>Fowler-Nordheim tunneling programming</td>
</tr>
<tr>
<td>Byte addressable</td>
<td>I/O type interface</td>
</tr>
<tr>
<td>Lower density</td>
<td>Higher density</td>
</tr>
<tr>
<td>Used for code-storage applications, which require execution-in-place (XIP)</td>
<td>Used for data-storage applications, like media players, which require fast program/erase and fast page access</td>
</tr>
</tbody>
</table>

Table 2.1: Major differences between NOR and NAND flash memories

Therefore, NAND flash memories are well-suited for data-storage applications like media players, cameras and solid-state drives, while NOR flash memories are better suited for storing and executing operating system code in cellular phones. Since NAND flash memories are experiencing a dramatic increase in demand [2.16] for very high density storage, they are the main technological driver. Hence, the problems and solutions discussed in this thesis are focused on NAND flash memory.
The mechanisms of charge injection are different in NOR and NAND flash memories, and this is critical to the operation of the floating-gate transistor. The two most commonly used mechanisms of charge injection are described in the next section.

### 2.7 Mechanisms of Charge Injection

In this section, two main types of charge injection mechanisms are discussed. Channel hot electron (CHE) injection is used to program NOR flash memory, while Fowler-Nordheim (F-N) tunneling is used to program NAND flash memory. Both NOR and NAND flash memories are erased using F-N tunneling. These mechanisms are briefly described below.

#### 2.7.1 Channel hot electron (CHE) injection

In CHE injection ([2.18], [2.19], [2.20], [2.21], [2.22]), the basic idea is to provide enough energy to the channel electrons to overcome the tunnel oxide – silicon energy barrier. A large drain voltage produces a lateral electric field (between source and drain). This field increases the energy of the electrons, thus making them ‘hot’. The drain voltage needs to be higher than the oxide-silicon barrier (3.2 eV). These hot electrons are injected from the channel into the floating-gate, due to the perpendicular field applied from the gate. This injection primarily happens near the drain, because
that is the region where the electrons have gained maximum energy, or are the ‘hottest’. A NOR flash programming operation using CHE injection is depicted in Fig 2.9.

![Diagram of NOR flash programming using CHE injection](https://example.com/nor-diagram)

Figure 2.9: NOR flash programming using CHE injection [Source: Intel]

Hot-electron injection is commonly described using the ‘lucky-electron’ model [2.21]. This is based on the premise that the electron needs to be ‘lucky’ to be injected because it should have sufficient energy higher than the barrier, which means that it should not suffer too many scattering events that reduce its energy while it moves from the source to the drain. In addition, the momentum of the electron should be directed towards the interface, possibly by a scattering event. Since the probability of such conditions is very low, the electron which does satisfy these conditions is ‘lucky’. This also implies that hot electron injection is an inefficient method of programming since only a tiny fraction of channel electrons end up getting injected. The band diagram is depicted in Fig. 2.10 along with the slight lowering of the barrier due to the Schottky barrier (image charge) lowering effect [2.22].
More advanced quasi-thermal equilibrium models exist for describing CHE injection ([2.12], [2.17], [2.23], [2.24]).

## 2.7.2 Fowler-Nordheim tunneling

Fowler–Nordheim (F-N) tunneling [2.25] refers to quantum-mechanical tunneling through a thin potential barrier, induced by an electric field. Figure 2.11 shows F-N tunneling from the poly-silicon FG into the substrate. The oxide band bends under the application of an electric field, and presents a triangle-shaped barrier to the FG. The larger the field, the thinner the tunneling barrier is, so the current is larger. The probability of tunneling also depends on the distribution of carriers in the injecting material and the height of the barrier. Unlike the ‘lucky electron’ hot carrier injection discussed earlier, F-N tunneling injection is 100% efficient.
Using a free-electron gas model for the metal and the Wentzel–Kramers–Brillouin (WKB) approximation for the tunneling probability [2.25, 2.26, 2.37], the following expression for current density is obtained:

\[
\frac{J}{E^2} = A \exp(-B/E),
\]

Where

\( J \) = Current density

\( E \) = Electric field across the dielectric
Chapter 2 – Flash Memory: Operation and Scaling

\[ A = \frac{e^3 m}{16 \pi^2 \hbar m_{ox} \Phi_0} \]

\[ B = \frac{4}{3} \left( \frac{2 m_{ox}}{e \hbar} \right)^{1/2} \Phi_0^{3/2} \]

e = electronic charge

\( 2\pi \hbar = \) Planck’s constant

\( \Phi_0 = \) Barrier height

\( m_{ox} = \) electron mass in oxide

\( m = \) free electron mass

Figure 2.12 shows F-N tunneling current as a function of electric field. The current is exponentially dependent on the field. This is important, because at low fields when we just want to retain charge, the tunneling current is very low, and at high fields the current is very high which enables fast charge injection.

![Figure 2.12: F-N tunneling current as a function of electric field (Source: [2.12], [2.27])]
The classic expression for F-N current density shown above does not consider some issues: the temperature dependence of the phenomenon, the quantum effects at the silicon interface, the influence of band bending at the Si/SiO$_2$ interface, and the voltage drop in silicon, the fact that the correct statistics for electrons are not Maxwellian but Fermi–Dirac, and the collision-broadening barrier lowering [2.12, 2.28].

In the case of poor quality oxides [2.29] that contain large numbers of bulk and interface traps, trap-assisted tunneling [2.30] increases tunneling current density to much greater values than expected due to the effective decrease in barrier.

Figure 2.13 shows a NAND flash memory device programmed by F-N tunneling.

![Figure 2.13: F-N programming in NAND flash (Source [2.11])](image)
2.8 Reliability of Flash Memories

So far, the basic structure and operation of the floating-gate device has been described along with the two main types of flash memory – NOR and NAND. These descriptions have established the mechanisms through which charge is injected and removed from the floating-gate and the means of reading the state of the device. In this section, a very important issue in flash memory technology is discussed – the reliability [2.31].

2.8.1 Endurance (Cycling)

The flash memory device is required to retain its properties on being subjected to repeated program/erase cycles. When thin dielectrics are repeatedly stressed at high fields, interface and bulk traps develop in the dielectric [2.29]. Charge is trapped and released from these traps, and this modifies the fields across the dielectric. This tends to modify the program/erase characteristics over time, as damage is induced in the dielectric. Figure 2.14 shows the threshold voltage window closing with cycling due to traps induced in the tunnel oxide.
Flash memories are typically expected to last for 100,000 cycles [2.13] without major degradation. Modern multi-level cells have lower endurance benchmarks due to extremely stringent threshold voltage windows.

### 2.8.2 Retention

The ‘retention’ [2.32] is critical in any non-volatile memory device. The ability to retain charge without supplied power is the definition of non-volatile memory. ‘Retention’ is a benchmark used to quantify the extent of time for which the stored charge is retained in a device. Figure 2.15 shows the band diagram of a flash memory device in the charge-retention state.
The charge-loss in the retention state is determined by tunneling leakage under weak fields through adjoining dielectrics. This charge-loss would be greatly amplified if the dielectrics contained defects, since that would enhance trap-assisted tunneling [2.30].

A typical retention benchmark for flash memories is 10 years [2.13], i.e. the floating-gate device should not lose more than a small fraction of its carriers for 10 years. In practice, it is not possible to quantify retention by doing measurements for such a long period of time. Hence, retention tests are accelerated at higher temperatures. Typically, a 24-hour benchmark at 150 °C is utilized. Loss of charge with time at higher temperatures is measured to make a comparison of retention properties of the devices under study. Mean times to failure and defect density are other metrics used for quantifying reliability of flash memory [2.33].
Chapter 2 – Flash Memory: Operation and Scaling

2.9 Flash Memory Scaling

Flash memories, like most other electronic devices, have been continually scaled since their introduction, to obtain increasing densities.

Figure 2.16: NOR flash memory scaling (Source: Intel [2.11])

NOR-type flash memories were the first to be introduced. Figure 2.16 shows the scaling of NOR flash memory for two decades. Until recently, non-volatile memory half-pitches have lagged those for DRAM or CMOS logic devices in the same year [2.13]. Rapid progress in NAND flash technology has not only reversed this trend, but also surpassed the half-pitches of DRAM and CMOS logic devices. NAND flash memories have seen tremendous demand in this decade, and their bit density scaling trend is shown in Figure 2.17.
Both NOR and NAND flash memories are now facing major roadblocks in continued scaling. The major scaling difficulties of NOR and NAND flash memories are summarized in following sections.

### 2.9.1 Challenges in scaling NOR flash

a) Drain-voltage scaling [2.11, 2.13, 2.35]: NOR flash memory devices are programmed by channel hot-electron injection. Since the silicon – oxide barrier height is 3.2 eV, the drain voltage has to be at least greater than 3.2 V for reasonable efficiency. Therefore, there is a major challenge in scaling the drain voltage in NOR flash devices.
b) Tunnel oxide scaling [2.13, 2.35]: The scaling of tunnel oxide is limited by concerns for reliability issues. Since NOR flash memories provide direct cell access, the reliability issues are more stringent, unlike NAND flash memories which can use error code correction and data re-mapping strategies. The tunnel oxide thickness for NOR flash devices is essentially stuck at 8-9 nm and not scaled anymore.

c) Channel length scaling [2.13, 2.35]: There are a number of factors contributing to difficulty in scaling channel length in NOR flash devices. NOR flash devices use abrupt drain junctions to generate high lateral fields, which in turn are required to efficiently generate hot electrons. However, these abrupt junctions also increase short-channel effects and leakage current. Use of halo implants helps with leakage but reduces junction breakdown voltage. All this is exacerbated by the fact that the tunnel oxide does not scale anymore, because a thinner tunnel oxide reduces short-channel effects. Hence, the channel length faces potentially game-ending scaling issues and consequently, so does the cell area.

d) Control Dielectric scaling [2.35]: This is also limited by reliability concerns, similar to the tunnel oxide.

The ITRS scaling projection [2.36] for floating-gate NOR flash is shown in Figure 2.18.
2.9.2 Challenges in scaling NAND flash

a) Cell-to-cell interference [2.13, 2.35]: The scaling of high-density NAND flash memory is limited by parasitic interference between adjacent cells since they are extremely close to each other. The floating-gate, being a capacitive-coupled electrode, has started to have significant coupling with the floating-
gates of adjacent cells and other electrodes of neighboring cells. This causes an undesirable shift in the state of one cell due to neighboring cells. Especially with the advent of multi-level cells, the window between states is not much, which is the biggest challenge in the scaling of NAND flash memory.

b) Maintaining GCR [2.13]: A GCR > 0.6 is necessary for satisfactory operation of these devices, and this is typically achieved by wrapping the control-gate around the floating-gate, thereby increasing the area and consequently the capacitive coupling between the two gates. However, with decreasing spacing between adjacent cells, there is no longer any space for wrapping the control-gate in between floating-gates. This leads to significant reduction in the GCR and needs to be compensated by some method.

c) Tunnel oxide scaling [2.13, 2.35]: Like NOR flash, tunnel oxide scaling in NAND flash is also limited by reliability concerns. However, due to the use of error correction codes (ECC), NAND flash is a little more tolerant to tunnel oxide scaling than NOR flash. Still, the tunnel oxide thickness is stuck at ~6 nm and needs breakthroughs to continue scaling beyond that.

d) Control dielectric scaling [2.36]: The control dielectric thickness is also limited by reliability concerns, similar to the tunnel oxide. Reliability concerns are also very serious if the control dielectric is modified to use high-k dielectrics, which may be necessary to improve GCR.

e) Lithography concern [2.35]: Lithography limitations are also an important problem in further scaling of NAND flash. Extreme ultraviolet (EUV) lithography is the main candidate for next generation lithography.
f) Few electron phenomena [2.35]: The ultimate intrinsic limits of NAND flash memories are likely to be due to statistical fluctuations induced by too few electrons stored.

Figure 2.19 shows the ITRS scaling projections for floating-gate NAND flash memory.
### NAND Flash Technology Nodes and Scaling Projections

<table>
<thead>
<tr>
<th>Year of Production</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAF 90nm (comsat)</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>8MPU/65C Metal 1 (ML) 90nm (comsat)</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
</tr>
<tr>
<td>64T NAND Flash 90nm (comsat)</td>
<td>39</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
</tr>
</tbody>
</table>

### NAND Flash Technology Nodes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
<td>6T</td>
</tr>
<tr>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
<td>6F</td>
</tr>
<tr>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
<td>10F</td>
</tr>
</tbody>
</table>

**Figure 2.19:** ITRS Scaling projections for floating-gate NAND flash. [2.36]
References


[2.36] ITRS Table PIDS5, 2009.

Chapter 3

Scaled Poly-Silicon Floating-Gate NAND Flash Memory

3.1 Introduction

This chapter studies poly-silicon floating-gate devices with vertical scaling, i.e., thickness scaling of the floating-gate. In the first section, the cell-to-cell interference challenge in scaling NAND flash memory is discussed, which motivates thickness scaling of the floating-gate. Subsequently the test structure used for this study, and the fabrication of scaled poly-silicon floating-gate devices are described. This is followed by the results of the study comparing and discussing in detail the program/erase and retention characteristics of devices with poly-silicon floating-gates of various thicknesses ranging from 75 nm to ultra-thin 7 nm.
3.2 Motivation

3.2.1 Cell-to-cell interference

NAND flash memory is optimized for use with high-density applications [3.1, 3.2, 3.3, 3.4, 3.5]. As the density gets higher and higher, the floating-gate cells get closer and closer to each other. The floating-gate, being a capacitive-coupled electrode, has started to have significant coupling with the floating-gates of adjacent cells and other electrodes of neighboring cells [3.2, 3.4, 3.5, 3.12]. Figure 3.1 shows an SEM image of a NAND flash array showing the parasitic coupling between cells.

![Figure 3.1: SEM of NAND flash array (Source: [3.7])](image)

This causes an undesirable shift in the state of one cell due to neighboring cells (figure 3.2).
The undesirable shift in state of a cell due to capacitive coupling with adjacent cells in the array (in the bit-line direction), is depicted in figure 3.3. With multi-level cells (MLC), the design window available for the threshold voltage states is much reduced due to the need to accommodate a large number of states in a single cell. Figure 3.4 depicts the interference problem getting worse with technology node scaling and shows the relative magnitudes of the interference along different directions. This is one of the biggest challenges in the further scaling of NAND flash memory.
Figure 3.3: This figure depicts the shifting of threshold voltage of Cell 1 when Cell 2 gets programmed due to the parasitic coupling between the adjacent cells. (Source: [3.6])

Figure 3.4: The undesirable threshold voltage shift (interference) is getting worse with scaling. (Source: [3.8], [3.9])
3.2.2 Control gate wrap-around

In floating-gate NAND flash devices, the control-gate is wrapped around the floating-gate [3.5] as shown in figure 3.5. A TEM image of the control-gate wrap-around is shown in figure 3.6. The CG wrap-around serves two very important purposes:

(i) The conductive CG electrode shields the interference field lines between adjacent floating-gates. Changing the height of the CG shield has a significant effect on the interference in the word-line direction, as shown in figure 3.7. Figure 3.8 shows the FG-FG coupling in the bit-line and word-line directions.

(ii) The CG wrap-around increases the area of overlap between the CG and FG, and therefore the capacitive coupling between them. This capacitive coupling is related to the GCR, which needs to be 0.6-0.7 for satisfactory operation. The CG wrap-around is very important for achieving a desirable GCR value.

As the density of NAND flash memory increases, there is no more space [3.14] for the control gate to go in between floating-gates of adjacent cells (figures 3.9, 3.10). As a result, the interference problem is greatly enhanced and the GCR also faces a significant drop. In this chapter, a solution for the interference problem is demonstrated. The GCR issue is handled in chapter 6.
Figure 3.5: Control gate is wrapped around the FG shielding interference between adjacent floating gates.

Figure 3.6: Image showing CG wrapping around the FG [3.10]
Figure 3.7: Changing the height of the CG wrap-around has a significant effect on the word-line direction interference [3.10]

Figure 3.8: Interference along BL direction and WL direction [3.11]
Figure 3.9: CG no longer has space to wrap around FG thereby leading to a serious interference problem and GCR decrease.

Figure 3.10: Image showing that the distance between cells is extremely less (Source: [3.13])
3.3 **Floating-gate thickness scaling**

The parasitic interference between adjacent cells can be lowered by reducing the thickness of the floating-gate [3.15, 3.16, 3.17]. Figure 3.11 shows that by reducing the thickness of the floating-gate, we reduce the area of overlap between the FG and interfering electrodes, thereby reducing the capacitive coupling and the interference between them. Figure 3.12 indicates that we need floating-gate thicknesses to be on the order of a few nm, to meet interference requirements for future nodes. Conventional NAND flash memory devices use poly-silicon floating-gates of thickness ~45 nm ([3.10]). Here, we systematically study the scaling of the poly-silicon floating gate by fabricating devices with floating-gate thicknesses going from 75 nm to ultra-thin 7 nm and comparing their electrical characteristics.

![Control gate](image)

**Figure 3.11:** Reduction of FG height greatly reduces FG interference.
3.4 Experimental Test Structure

The purpose of the experiment is to fabricate and study floating-gate devices with a range of floating-gate thicknesses varying from thick 75 nm floating-gate to as thin as possible. The devices need to be compared in terms of program/erase characteristics, retention and endurance. The floating-gate capacitor shown in Figure 3.13 is chosen as the test structure.
A capacitor is chosen instead of a transistor because the capacitor enables the study of program/erase characteristics, retention and endurance with reduced complexity of fabrication compared to the transistor. The state of the device is sensed by measuring the flat-band voltage from a capacitance-voltage sweep.

An important thing to note from Figure 3.13 is the emulated control gate wrap-around. The area between the control gate and the floating gate is fixed while the area between the floating-gate and channel has been made smaller, by modifying the active area size. By doing this, the increased area between the CG and FG mimics the control-gate wrapping around the floating-gate in NAND flash memory devices. This
enables study of the devices with a range of GCR values since the GCR value is a function of the ratio of the IPD capacitance and the tunox capacitance.

Since the structure is a capacitor, minority carriers need to be generated and do not have a source, like in a transistor. This means that the generation of minority carriers will take time, and hence the erase operation can not be performed at speeds faster than 1 ms in the case of n-type substrate. There is a similar restriction on programming speed in the case of p-type substrate. This is because the substrate goes into deep depletion when the appropriate bias is applied, implying that most of the voltage applied drops across the substrate and not the tunnel oxide field. By giving sufficient time for the minority carriers to be generated, the substrate goes from deep depletion to inversion, making the tunnel oxide field as high as expected.

3.5 Fabrication

The starting substrate is an (100) 1-5 Ω cm n-type silicon substrate. The device isolation is performed by growing a 0.5 micron thick thermal oxide and opening up active areas with a wet buffered oxide etch. Following that, the tunnel oxide is grown in dry oxygen at 800 °C to obtain ~7 nm oxide. This thickness is limited by reliability concerns. Following this, the critical step of poly-silicon floating-gate deposition is performed.
Chapter 3 – Scaled Poly-Silicon Floating-Gate NAND Flash Memory

The smoothness and quality of the poly-silicon floating-gate are very important to the functionality of the floating-gate device. A rougher poly-silicon floating-gate would result in a leakier oxide on top on it [3.18]. To obtain smoother layers, the floating-gate can be deposited in amorphous form and subsequently crystallized. Therefore, the deposition of amorphous silicon is performed at 580 °C in-situ doped with phosphorus as high as possible. Annealing at high temperatures can cause significant grain growth and roughen the surface. Hence, the annealing step is postponed until after the IPD stack is deposited so that the roughness of the floating-gate under the stack does not get affected much.

Following the deposition of the amorphous silicon, a thin 20 nm capping layer of low temperature silicon oxide (LTO) is deposited at 300 °C to protect the floating-gate surface during subsequent processing. The floating-gate is patterned and dry etched in an SF₆/O₂ plasma. The capping LTO is removed and the IPD stack deposition begins.

A well-characterized LTO of reproducible quality is deposited at 450 °C and annealed at 750 °C in dry oxygen to enhance its quality. Following this, silicon nitride is deposited at 785 °C. Following that, another step of LTO deposition is performed, followed by another anneal at 750 °C in dry oxygen and an anneal at 950°C in nitrogen. The 950 °C nitrogen anneal also serves to activate the dopants in the floating-gate in addition to improving the quality of the IPD stack. All these higher temperature steps also help crystallize the amorphous silicon into poly-silicon. Since the crystallization occurs with the IPD stack on top of the floating-gate, it is less likely to be roughened during the process.
It is very important to be careful while cleaning the floating-gate surface, especially when it is ultra-thin. Heavily doped poly-silicon tends to get etched faster than undoped single-crystal silicon by buffered oxide etches and other reagents.

Following the 950 °C nitrogen anneal, the control gate is deposited. In the case of these capacitors, it was usually aluminum. Then, the aluminum is patterned into the control gate. Finally, a forming-gas anneal is performed at 350-400 °C to improve the quality of the interfaces.

Figure 3.14 and 3.15 show high resolution TEM images of the gate stack of an ultra-thin 7 nm poly-silicon floating-gate device. As the image reveals, the poly-silicon floating-gate is obtained to be very smooth at such a low thickness. This was the thinnest reported poly-silicon floating-gate device at the time [3.19].
Figure 3.15: TEM image showing ultra-thin 7 nm poly-silicon floating-gate

3.6 Measurement Setup

Measurement was performed on a Cascade prober. To obtain program/erase characteristics, it is necessary to apply a pulse and measure the state of the cell, and repeat this until the state of the cell saturates. The state of the cell is measured by performing a capacitance-voltage sweep and obtaining the flat-band voltage. Program/erase characteristics were obtained by developing a Labview program which controlled a pulse generator, an LCR meter and a switching unit. The program directed
the pulse generator to send a pulse, and then directed the switching unit to switch connections to the LCR meter. The LCR meter performed a capacitance-voltage sweep to obtain the flat-band voltage. The program then directed the switching unit to switch back to the pulse generator, and the process of pulse-switch-measure-switch is repeated until the flat-band voltage saturates. Similar Labview routines were developed to measure retention and endurance.

### 3.7 Program/Erase characteristics

The program/erase characteristics of the floating-gate devices are discussed in this section. The programming characteristics of 75 nm thick n-type poly-silicon floating-gate devices are shown in figure 3.16. As mentioned before, the measurement is performed by first applying a pulse, say, at 10 V (control-gate), followed by the measurement of the flat-band voltage followed by a pulse at 10.5 V, then a measurement and so on.
The flat-band voltage keeps increasing with programming voltage until it saturates. The voltage at which it saturates is called the saturation threshold voltage or $V_{T,SAT}$. It can be noted from the figure that the saturation voltage is quite large, $\sim 8$ V. Figure 3.17 shows the erasing characteristics of 75 nm thick n-type poly-silicon floating-gate. Here, again we note that the saturation voltage is quite large, $\sim -8$ V. Therefore, the window of operation is large.
The next important thing to note from figures 3.16 and 3.17 is that the slope of the program/erase characteristics is exactly 1 until close to saturation. This is indicative of a high quality device where the dielectrics are not leaky or trappy. If the dielectrics are excessively leaky or trappy, the slope would be lesser than 1.

The reason for why the slope is 1 is described here. Following the development in [3.20], this is based on the following two assumptions: a) the tunox current $i_g$ increases with increasing floating-gate potential $V_{FG}$ (this is true for F-N tunneling current in the normal range of operation); b) the floating gate potential $V_{FG}$, which controls electron injection, depends on control gate potential $V_{CG}$ and threshold...
voltage/flat-band voltage $V_T$ only through their difference $V_{CG} - V_T$ (also true, shown below)

Ignoring source, drain and body couplings,

$$V_{FG} = GCR \cdot V_{CG} - \left( \frac{Q}{C_{TOT}} \right)$$

$$= GCR \cdot V_{CG} - (C_{IPD} \cdot V_T / C_{TOT}) \text{ because } Q = C_{IPD} \cdot V_T$$

$$= GCR \cdot (V_{CG} - V_T) \text{ because } GCR = C_{IPD} / C_{TOT}$$

Where

$Q$ = Charge on FG

$C_{TOT}$ = Total FG capacitance

$C_{IPD}$ = IPD capacitance

Based on the two assumptions, we have $i_G = f \left( V_{CG} - V_T \right)$

Also, $i_G = -dQ/dt = C_{IPD} \cdot dV_T/dt$

Leading to $dV_T/dt = f \left( V_{CG} - V_T \right) / C_{IPD}$

If the starting value of $V_{CG}$ is low, $f \left( V_{CG} - V_T \right)$ would be low, thereby the change in $V_T$ would be low. The next $V_{CG}$ is higher by one step, say 0.5 or 1 V. Since the $V_T$ only increases by a small amount, the new $V_{CG} - V_T$ is higher than the last time, therefore $f \left( V_{CG} - V_T \right)$ is higher which implies that the change in $V_T$ is higher. If the change in $V_T$ is still not as high as one $V_{CG}$ step, then the value of $V_{CG} - V_T$ increases further during the next step, and consequently does the change in $V_T$. So the
change in $V_T$ keeps increasing until it reaches the point when the change in $V_T$ is equal to the $V_{CG}$ step. At this point, $V_{CG} - V_T$ is constant, and continues to remain constant. A similar argument can be made for a high starting value of $V_{CG}$. The nature of the function $f(V_{CG}-V_T)$ determines the speed at which the sequence converges.

Therefore, this explains the nature of the presented program/erase characteristics. Figures 3.16 and 3.17 also show program/erase characteristics for different pulse widths. The longest pulse width starts programming at the smallest voltage, and the curves for different pulse widths are shifted from each other by approximately the same shift amount for each order of magnitude in pulse-width. The same phenomenon is also seen in erase characteristics.

Figures 3.18 and 3.19 present the comparison of programming and erasing characteristics for n-type poly-silicon FG of thicknesses 75 nm, 11 nm and 7 nm. It is noted that all three have excellent saturation voltages and window, and the right slope of 1. (These characteristics are for 100 ms pulse widths.) In general, the program/erase characteristics are well-matched to one another when scaled from 75 nm thick n+ poly silicon to ultra-thin 7 nm.
Figure 3.18: Programming characteristics for various thicknesses of n-type poly-silicon FG devices

Figure 3.19: Erasing characteristics for various thicknesses of n-type poly-silicon FG devices
### 3.8 Retention

The retention of these n+ poly-silicon floating-gate devices is tested at a benchmark of 24 hours at 150 °C. The measurement is performed by programming a number of devices to various flat-band voltages, performing the 24 hour 150 °C bake and then measuring the threshold voltages of all the programmed cells. The retention is plotted in figure 3.20 as loss in the flat-band voltage versus the starting (pre-bake) flat-band voltage. From the figure, the retention is very good, as only less than 0.2 V is lost for 5 V starting flat-band voltage, and the retention appears well-matched between the poly-silicon FG devices of various thicknesses.

![Figure 3.20: Retention for n-type poly-Si FG devices of various thicknesses](image)
So far, the poly-silicon devices seem to perform well with scaling in floating-gate thickness. However, there are new problems that arise when scaling to ultra-thin poly-silicon FG. The endurance of these devices will be discussed in the next chapter when problems in scaling of poly-silicon FG devices are demonstrated.
Chapter 3 – Scaled Poly-Silicon Floating-Gate NAND Flash Memory

References


Chapter 3 – Scaled Poly-Silicon Floating-Gate NAND Flash Memory


Chapter 3 – Scaled Poly-Silicon Floating-Gate NAND Flash Memory

Chapter 4

Scaled Poly-Silicon Floating-Gate: Problems to Ultimate Limit

4.1 Introduction

A new reliability concern, arising from the programming current increasingly going ballistically through the ultra-thin floating-gate, is described here. A new test structure is built to experimentally investigate this effect. Experimental results quantitatively show the magnitude of the effect. The implication of this effect is observed by means of endurance data, which show thinner floating-gate devices degrading faster. This reliability concern poses a major scaling challenge. Following the discussion of this problem, other issues due to doping in ultra-thin poly-silicon floating-gates are also discussed.
4.2 The Ballistic Problem

4.2.1 Ballistic transport

Ballistic transport refers to carriers traversing a region without scattering [4.1-4.6]. In general, carriers get scattered by phonons (lattice vibrations), dopant atoms, other carriers etc. when traversing through a region. When the mean free path (average distance between two scattering events) of the carrier is larger than the distance of the region being traversed, there is a significant chance that the carrier will not face any scattering event.

There are two kinds of scattering events: elastic and inelastic. Elastic scattering refers to scattering events which do not involve change in energy of the carrier, but only a change in momentum. Inelastic scattering refers to change in energy of the carrier as a result of the scattering event.

Ballistic transport in Si/SiO$_2$ stacks has been observed by Murota and co-workers [4.7] using porous silicon. Figure 4.1 shows the schematic band diagram of their study. Electrons are injected through Si/SiO$_2$/Si/SiO$_2$... layers of interconnected nanocrystallites in porous silicon at a high electric field and the energy distribution of the arriving electrons is measured. The energy distribution is shown on the right-side of the figure. From the energy distribution, it can be clearly seen that the electrons have not suffered many inelastic (energy-reducing) scattering events, and the energy distribution is well-above expected levels for electrons in thermal equilibrium. In fact, at 100K, the electrons seem to have experienced no inelastic scattering events at all.
4.2.2 Ballistic transport during programming

Figure 4.2 shows a band diagram describing ballistic transport during programming.
During programming, the electrons are injected into the floating-gate from the substrate through F-N tunneling. Normally, these electrons get scattered and reach thermal equilibrium in the floating-gate. Depending on the field across the IPD, a small fraction of these scattered electrons will tunnel through the IPD to the control-gate.

However, if the floating-gate is made thinner and thinner, some of the injected electrons will not have sufficient scattering in the floating-gate. These ‘ballistic’ electrons will traverse the IPD at high energy, posing a major reliability risk for the IPD [4.8, 4.9]. Note that the electron does not have to be completely ballistic (zero inelastic scattering events) in the FG for it to traverse the IPD at high energy. ‘Quasi-
ballistic’ electrons, which only lose a small amount of energy in the FG, may still have enough energy to traverse the IPD energy barrier at sufficiently high energy.

The upcoming sections will explore the magnitude of this effect experimentally. This issue is expected to become exponentially worse as the FG is thinned down.

### 4.2.3 Test Structure

The test structure used for studying this issue is shown in Figure 4.3.
This structure starts out with similar structure as described in figure 3.13. The substrate, tunnel oxide, floating-gate, IPD and control-gate are described before. The control-gate used here is n+ poly-silicon (~80 nm). A range of different active areas is used as before. Following the patterning of the control-gate, two additional processing steps are performed with the aim of introducing a contact to the floating-gate. The idea behind introducing a floating-gate contact is that electrons scattered in the floating-gate will be collected away by this contact, while the ballistic (and quasi-ballistic) electrons will traverse across the IPD, and be collected at the control-gate.

To achieve the floating-gate contact, back-end low temperature silicon oxide (LTO) is deposited on top of the patterned control-gate. Contact holes are etched in the LTO to terminate on the control-gate and floating-gate. This is a very difficult step in the fabrication of these structures, because the floating-gate is only a few nm thick and it is easy to etch through it.

The contact hole etch was performed as follows: A well-characterized and reproducible CHF$_3$/O$_2$ plasma dry etch was utilized to etch 80-90% of the LTO. Following this dry etch, the photoresist was stripped and the remaining part was wet-etched by a solution of 50:1 HF (with added surfactant). Buffered oxide etch (BOE) was found to etch highly doped n-type poly-silicon relatively fast, hence HF was utilized for the wet etch. Since HF has a problem with penetrating into smaller holes, a surfactant was added to the HF. This combination of etches worked to achieve the contact holes to the floating-gate and control-gate as desired.

Following the contact hole etch, aluminum was deposited as the metallization layer and patterned to provide contacts to the floating-gate and control-gate.
4.2.4 Measurement

The measurement of ballistic current using the structure shown in figure 4.3 is described here. There are three terminals in the structure – the control-gate, the floating-gate and the substrate. The substrate is grounded. The control-gate is fixed at a sufficiently large programming bias, say, 20 V. The measurement is performed by sweeping the floating-gate voltage from 0 in the positive (programming) direction. The currents from all three terminals are measured.

Figure 4.4 shows the band diagram of the structure at the beginning of measurement. As shown in the figure, the field across the tunnel oxide is very small at the beginning, because the substrate is grounded and the floating-gate is at low bias. The field across the IPD is very large at the beginning, as the control gate is fixed at 20 V, with the floating-gate at low bias. Figure 4.5 shows a schematic plot of the IPD leakage current versus the floating-gate bias. As the figure shows, the IPD leakage current first starts at a high magnitude due to the large IPD field. As the floating-gate bias increases, the field across the IPD reduces, and this IPD leakage current falls exponentially. The IPD leakage current flows between the floating-gate and the control-gate. At low FG bias, there is negligible tunnel oxide current because of the small field across the tunnel oxide.
Chapter 4 – Scaled Poly-Silicon Floating-Gate: Problems to Ultimate Limit

Figure 4.4: Band diagram during measurement at low FG bias. Substrate is grounded and CG is fixed at 20 V.

Figure 4.5: Shows the IPD leakage current at the start of measurement.
Figure 4.6: Band diagram of measurement at high FG bias.

Figure 4.7: Shows schematics of expected currents at mid-to-high floating-gate bias.
As the floating-gate bias is increased further (figure 4.6), the electric field across the tunnel oxide increases, while the electric field across the IPD decreases. At these low electric fields across the IPD, the IPD leakage is negligible. As the tunnel oxide field increases, the tunneling current starts flowing from the substrate across the tunnel oxide. Most of the tunneling carriers will be scattered and thermalized in the floating-gate, and these electrons will be picked up at the floating-gate contact. A fraction of the tunneling carriers will not be inelastically scattered in the floating-gate, and these ballistic (or quasi-ballistic) electrons will be picked up at the control-gate. As shown in figure 4.7, the tunnel oxide current from the substrate increases exponentially with floating-gate bias. A fraction of that tunnel oxide current goes ballistic from the substrate to the control-gate through the IPD, thus undetected by the floating-gate contact.

Figures 4.5 and 4.7 are combined in figure 4.8 for the full range of measurement. As the floating-gate bias increases, the IPD leakage decreases exponentially and dies down. Beyond a certain floating-gate bias, the tunnel oxide current begins from the substrate. The inelastically scattered electrons comprising the tunnel oxide current, flow into the floating-gate contact. The ballistic current, a fraction of the tunnel oxide current comprising the ballistic (and quasi-ballistic) electrons, starts flowing into the control gate across the IPD. Thus, the measurement of the control gate current would reveal a decreasing IPD leakage first, followed by an increasing ballistic current. This measurement is applied to quantify the magnitude to which carriers are ballistic in poly-silicon floating-gates.
4.2.5 Results

The measurement structure was fabricated with a range of poly-silicon floating-gate thicknesses, going from 7 to 75 nm. The above-described measurements were performed for these structures, and the results are summarized here. Figure 4.9 shows the results of the measurement performed on a 75 nm thick poly-silicon floating-gate structure.
The control gate current ($I_{CG}$) decreases as the floating-gate bias is increased, corresponding to the IPD leakage falling. As the floating-gate bias continues to increase, the tunnel oxide current starts increasing. However, no concurrent increase suggesting the ballistic component of tunnel oxide current, is observed in control gate current. Hence, there is no ballistic component in the 75 nm thick poly-silicon floating-gate device.
Figure 4.10: Ballistic component measurement data in 30 nm n+ poly-silicon floating-gate device

Figure 4.10 shows the ballistic current measurement data in a device with 30 nm poly-silicon floating-gate. Here again, the control-gate current initially decreases as the floating-gate bias is increased, eventually falling to very small levels. As the floating-gate bias continues to increase, the tunnel oxide current starts increasing. Now, a concurrent increase in the control-gate current is observed. As explained, this control-gate current is present as a result of tunnel oxide current carriers travelling ballistic through the floating-gate, collecting at the control-gate. From the data, it is
observed that this ballistic component is about 0.01% of the tunnel oxide current, for the 30 nm poly-silicon floating-gate sample.

Figure 4.11: Ballistic component measurement data for 10 nm n+ poly-silicon floating-gate sample

Figure 4.11 shows the ballistic component measurement data for a 10 nm poly-silicon floating-gate device. Similar to the 30 nm floating-gate sample data, the presence of the ballistic component is noted. The ballistic component is measured to be 1.6% of tunnel oxide current, which is quite significant. It is also noted that the ballistic component has shown an exponential increase from 0.01% at 30 nm floating-
gate to 1.6% at 10 nm floating-gate thickness, thereby underscoring the seriousness of the problem as the floating-gate thickness is scaled down further.

**4.2.6 Verification**

In order to verify that the measured ‘ballistic’ component has indeed originated from the substrate tunnel oxide current, and is not any kind of IPD leakage current, another experiment was performed. In this experiment, the measurement was performed on structures with different active areas but same control-gate area, as shown in figure 4.12. The figure shows two structures with the same gate stack but different active areas (the tunnel oxide region). This means that the tunnel oxide area through which the tunnel oxide current flows is different for the two structures, but the IPD area (the area between floating-gate and control-gate) through which the IPD leakage current flows is the same for the two structures.

Figure 4.13 shows the ballistic component measurement data for 11 nm poly-silicon floating-gate using structures with different active areas. The result shows the initial control-gate current, which is the IPD leakage current, remains the same for the structures with different active areas. This is expected because the area of the IPD region, which determines the IPD leakage current, remains the same. The tunnel oxide current shown, scales proportional to the active region areas for the different structures, as expected. The ballistic component of the control-gate current also scales proportional to the active region areas. This proves that the origin of the ballistic
component is indeed the substrate tunnel oxide current and not any IPD leakage. A magnified version of figure 4.13 is shown in figure 4.14.

![Diagram of floating-gate structures](image)

Figure 4.12: Shows two floating-gate structures with the same IPD area (region between FG and CG) but different active areas (tunnel oxide region)
Figure 4.13: Ballistic component measurement data on 11 nm poly-silicon floating-gate sample for structures with different active areas.

Figure 4.14: Magnified version of figure 4.13
4.2.7 Mean Free Path

The ballistic component of tunnel oxide current is measured as described for various n+ poly-silicon floating-gate thicknesses, and plotted in figure 4.15. The data fits well on a straight line (note the vertical axis is on an exponential scale). Therefore, a mean free path can be extracted by fitting to a simple exponential law equation.

\[
\text{Mean Free Path } \lambda \\
\text{Probability of ballistic electron } P(t_{FG}) \\
\text{Where } t_{FG} \text{ is FG thickness} \\
P(t_{FG}) = A \exp\left(-\frac{t_{FG}}{\lambda}\right)
\]

Applying this, the mean free path is obtained as $\sim 4$ nm in n+ poly-silicon. Note that this is an *inelastic* mean free path, because inelastic scattering events that reduce energy of the carriers, are those that determine ballistic transport of the electron through the floating-gate. (Elastic scattering events will have a second order influence, because a carrier subject to more elastic scattering takes longer to traverse the region, making it more likely to be inelastically scattered.)
4.2.8 Impact (Endurance)

The ballistic component of programming current has been demonstrated and quantified so far. In this section, the impact of these ballistic carriers is discussed. This is accomplished by performing endurance tests on floating-gate capacitors with various poly-silicon floating-gate thicknesses ranging from 75 nm to 7 nm. The endurance test involves repeated program/erase cycles. During programming, the ballistic high-energy carriers that traverse the IPD pose a reliability threat to the IPD.
Repeated programming events lead to an accumulation of damage, which is expected to be observed in these tests.

Figure 4.16 shows cycling data on 75 nm, 10 nm and 7 nm thick n+ poly-silicon floating-gate devices. These devices are cycled at 14 V programming bias and -14 V erasing bias for 100 ms, corresponding to flat-band voltages of ±5 V. The 75 nm poly-silicon floating-gate device does not show much degradation, while the thinner devices, especially the 7 nm poly-silicon floating-gate device shows much more degradation, pointing to ballistic carriers causing reliability issues.

Figure 4.16: Cycling data (+- 14 V, 100 ms) on 75 nm, 10 nm and 7 nm poly-silicon floating-gate devices. The 7 nm floating-gate devices show more degradation.
Figure 4.17 shows similar endurance testing performed on 75 nm and 10 nm poly-silicon floating-gate devices programmed at 17 V for 100 µs and erased at -14 V for 100 ms (corresponding ± 5 V flat-band voltages). Again, it is clearly observed that the thinner poly-silicon floating-gate device degrades more. It is also worth noting that programming at higher fields for shorter pulse times (for the same flat-band voltage) degrades the device more.

Figure 4.17: Cycling data (+17 V, 100 µs and -14 V, 100 ms) on 75 nm and 10 nm poly-silicon floating-gate devices. The 10 nm floating-gate devices show more degradation
Figure 4.18 shows the capacitance-voltage plots for devices subjected to the endurance tests in figure 4.17. The capacitance-voltage plot of the 10 nm poly-silicon floating-gate device shows a larger degradation in slope and more pronounced hump compared to the 75 nm poly-silicon floating-gate device, indicating poorer dielectric quality. The high energy ballistic electrons break bonds in the IPD, and create an increasing number of traps over time, which are responsible for the degradation in the capacitance-voltage curves.

Figure 4.18: Capacitance-Voltage curves of cycled devices indicating more degradation in the 10 nm poly-silicon floating-gate device compared to the 75 nm floating-gate device.
4.2.9 Summary

In this section, a new reliability failure mechanism was introduced. As the floating-gate is thinned down, the programming carriers start going increasingly ballistic, and these high-energy carriers pose a reliability concern when they traverse the IPD. A new measurement structure was built to measure this ballistic effect experimentally. The experiments quantified the magnitude of this and extracted the inelastic mean free path of electrons in poly-silicon. Finally, the impact of this was shown by means of endurance tests on devices with various poly-silicon floating-gate thicknesses, revealing that the thinnest floating-gate devices degraded the most.

4.3 Poly-silicon depletion

4.3.1 Poly depletion effect

This effect [4.13] is well known in CMOS technology. Consider figure 4.19. In MOS transistors, high electric field due to a combination of higher supply voltage and thinner gate oxide causes band bending even in a heavily doped poly-silicon gate. This causes depletion in the poly-silicon gate. The effect of depletion is to increase the effective oxide thickness and reduce the effective oxide capacitance.
Figure 4.19: The poly-depletion effect. Due to the depletion in the poly-silicon gate, the effective oxide thickness increases and the effective capacitance decreases. Source: [4.13]
4.3.2 Poly depletion in NAND flash memory

As observed in MOS transistor technology, poly-silicon depletion can also occur in NAND flash memory devices. Consider figure 4.20.

![Figure 4.20](image)

Figure 4.20: Shows a plot of poly-silicon floating-gate doping versus technology node. Also shows an increase in the IPD thickness variation due to poly-depletion with scaling. Source: [4.14]

The band diagram in the figure shows that there is an effective increase in the EOT of the tunnel oxide and the IPD due to poly-depletion, which widens the threshold voltage distribution. The variation gets worse with scaling, which is a serious concern, given the small margins available for design.

Spessot and co-workers [4.15] demonstrate a decrease in the erase efficiency as a result of depletion in the poly-silicon floating-gate (figure 4.21). From the band diagram, it is observed that the floating-gate goes into deep-depletion near the IPD side (due to lower doping on the IPD side in this particular experiment). The voltage
drop across the depletion region, leads to a smaller voltage drop and electric field across the tunnel oxide. In other words, the additional capacitance caused by the depletion reduces the effective GCR of the device, which leads to a smaller field across the tunnel oxide. As a result of the smaller tunnel oxide field, the erase is slower.

Figure 4.21: As a result of the poly-depletion, there is a decreased field across the tunnel oxide which makes the erase slower. Source: [4.15]
4.3.3 Experiment

The test structure used was a floating-gate capacitor like the one described in figure 3.13. Figure 4.22 shows the capacitance-voltage plots for 20 nm, 12 nm and 7 nm thick poly-silicon floating-gate devices.

![Capacitance-voltage plots for poly-silicon floating-gate devices.](image)

Figure 4.22: Shows the capacitance-voltage plots for poly-silicon floating-gate devices. The 7 nm device shows a characteristic poly-depletion dip in the capacitance.

It can be observed that the 7 nm device shows the characteristic poly-depletion dip in the capacitance. This occurs as a result of the ultra-thin poly-silicon floating-gate getting fully depleted. The ultra-thin 7 nm layer may also have lower doping density, contributing to easier depletion, because phosphorus tends to segregate into grain boundaries [4.16, 4.17] during slow cooling after the high temperature anneal.
The grains are smaller at the bottom of a poly-silicon layer, therefore a thinner poly-silicon layer tends to have a higher proportion of grain boundaries, which can lead to more segregation of phosphorus. (Note that this segregation could be avoided by rapid cooling.) The sheet resistances were – 130 Ω/sq for 70 nm poly-silicon versus ~10,000 Ω/sq for 7 nm poly-silicon, also suggesting lower doping density in the ultra-thin layer. (Note that a direct extraction of doping cannot be performed from the sheet resistance, because the ultra-thin layer may have decreased mobility due to grain boundary and surface scattering.)

It is possible to amplify the effect of poly-depletion, and observe it more easily by using an emulated wrap structure (larger area between control-gate and floating-gate) like the one in figure 3.13. The data in figure 4.22 is for an emulated wrap of ~2. Figure 4.23 shows the effect of wrap by displaying capacitance-voltage plots for various levels of wrap. It can be clearly observed that the effect is much more pronounced for larger wraps. This is explained by figure 4.24. This figure shows the Medici simulated potential contours for an emulated wrap structure. In this figure, the area between the floating-gate and control-gate is larger than the tunnel oxide area. (The substrate is STI isolated). The first plot is for an undepleted floating-gate device. In this device, the control-gate couples perfectly to the floating-gate over its entire area, and the floating-gate couples to the substrate, as is clear from the potential contours. The second plot is for a depleted floating-gate device. The additional capacitance due to the depleted floating-gate is shown. In this case, the potential contours reveal that the control gate has to couple to the substrate directly, in the area outside of the tunnel oxide stack region. This leads to a much larger decrease in the
effective GCR as a result of poly-depletion, compared to the case without wrap, thereby making the poly-depletion effect easier to observe.

![Graph showing normalized capacitance vs. gate voltage for 7 nm and 12 nm poly-silicon FG with different wraps.](image)

Figure 4.23: The poly-depletion effect is much more pronounced for larger emulated wraps
Figure 4.24: Medici simulation potential contours for depleted and undepleted FG. Shows effect of emulated wrap in amplifying poly-depletion.
The programming characteristics are measured versus programming time, and the results are shown in figure 4.25. This measurement is performed by applying a fixed bias for varying pulse widths going from shorter to longer, and the flat-band voltages are measured. It is observed that the 7 nm poly-silicon floating-gate device starts programming slower due to apparent decrease in GCR from poly-depletion. As time increases and more charge is added into it, the floating-gate inverts and the GCR returns to normal. This experiment clearly demonstrates the negative effect that poly-depletion can have on the electrical characteristics.

Figure 4.25: Shows slower programming in 7 nm poly-silicon floating-gate device due to poly-depletion.
4.3.4 Summary

The poly-depletion effect is proven to be a problem for scaled poly-silicon floating-gate NAND flash memory devices, in addition to the ballistic carrier issue. Other issues of poly-silicon floating-gate devices include poor compatibility with high-k dielectrics and difficulty in depositing smooth layers with ultra-small thicknesses. As a result of these problems, metals are explored as alternative materials for the scaled floating-gate.
References


Chapter 5

Metal Floating-Gate

5.1 Introduction

The scaling of poly-silicon floating-gate has been studied in detail so far, revealing serious challenges. In this chapter, the use of a metal floating-gate as a possible replacement for poly-silicon is explored. The metal floating-gate is experimentally demonstrated to provide excellent electrical performance. The metal floating-gate is also demonstrated to solve the ballistic carrier and poly-depletion issues discussed in chapter 4. The factors underlying the choice of metal are discussed in detail. From the various options available, TiN (or TaN) is chosen as the preferred floating-gate. The TiN floating-gate is scaled down to ultra-thin 3 nm, which shows excellent electrical performance at even such low thicknesses. Utilization of such an ultra-thin floating-gate would greatly reduce cell-to-cell interference.
5.2 Metal Floating-Gate Device

Cell-to-cell interference [5.1-5.4] is a major problem in ultra-scaled NAND flash memories. Scaling the thickness of the floating-gate [5.5-5.7] helps reduce the coupling between adjacent cells and reduce interference. The scaling of poly-silicon floating-gates was studied in detail in chapters 3 and 4. The serious issues of ballistic transport of injected carriers in the ultra-scaled floating-gate, and poly-depletion induced problems, were identified. In order to overcome these issues, the use of a metal-based floating-gate is explored.

The structure used for studying these metal floating-gate devices is the floating-gate capacitor structure as discussed before, and shown in figure 5.1.

![Figure 5.1: Floating-gate capacitor structure used to study program/erase and reliability performance of metal floating-gate devices.](image-url)
The fabrication of these devices is same as described for poly-silicon floating-gate devices, except for the replacement of the poly-silicon layer by a metal layer. The metal layer can be deposited in a number of ways depending on the choice of metal – evaporation, sputtering, chemical vapor deposition (CVD), atomic layer deposition (ALD) etc.

Figure 5.2 shows the TEM micrograph of a 7 nm metal (TiPt) floating-gate device gate-stack.

![TEM of a metal (TiPt) floating-gate device gate-stack](image-url)
The program/erase characteristics of this device are shown in figures 5.3 and 5.4. As shown, these devices have excellent window in excess of 20 V, demonstrating that metals are suitable for use as charge-storage electrodes in floating-gate memories.

Figure 5.3: Programming characteristics of TiPt floating-gate device

Figure 5.4: Erasing characteristics of TiPt floating-gate device
5.3 Ballistic Component in Metal Floating-Gates

In chapter 4, the problem of programming carriers going increasingly ballistic [5.8] as the poly-silicon floating-gate is scaled down, was studied. Here, the ballistic component in metal floating-gate devices is studied, as metal floating-gates are viewed as a solution to this issue due to much higher levels of inelastic electron scattering [5.9].

To study the ballistic component, the structure with floating-gate contact described earlier and shown in figure 5.5 is used. As described before, tunnel oxide carriers that get scattered and thermalized in the floating-gate, will be collected at the floating-gate contact, while the ballistic (and quasi-ballistic) carriers will be collected at the control-gate.

The measured data for the 7 nm TiPt floating-gate sample is shown in figure 5.6. The figure also shows a comparison with the 7 nm n+ poly-silicon floating-gate sample data. The nature of the plot is the same as described before – initial decrease in the IPD leakage current, followed by increase in the ballistic component concurrent with the onset of tunnel oxide current. From the figure, it is clear that the ballistic component is much lower (1000X) in the metal floating-gate device compared to the poly-silicon floating-gate device. It is also worth noting that the IPD leakage of the metal floating-gate device is lower than the n+ poly-silicon counterpart, due to the deeper work-function of the metal used here. Therefore, the metal floating-gate is found to be effective at reducing the ballistic carrier issue.
Figure 5.5: Structure with floating-gate contact used to measure the ballistic component of tunnel oxide current.

Figure 5.6: Ballistic component measurement data on 7 nm TiPt metal floating-gate sample. The metal has much lower ballistic component compared to poly-silicon of the same thickness.
The ballistic component in some other metals was also studied using the structure shown in figure 5.5. The results are summarized in figure 5.7.

![Figure 5.7: Percentage of ballistic current plotted versus floating-gate thickness for a variety of materials. In general, it can be observed that metals have lower ballistic component compared to poly-silicon due to increased inelastic electron scattering. ](image)

It is clearly observed that the ballistic component is smaller in metals compared to poly-silicon. Hence, use of a metal floating-gate would help to solve the ballistic carrier issue.

It is also observed that there are differences in the level of ballistic current between different metals. This is due to the differences in the level of inelastic scattering between different metals. This is illustrated using figure 5.8.
The level of scattering [5.10] in a material is directly related to the density of states of the material at the energy of interest. The larger the density of states, the higher is the level of scattering. Assuming a very simple parabolic model, the density of states increases with energy as shown. Differences in the band structures of different materials will also contribute to different density of states. Ignoring that for a simple first-order comparison, the density of states (and the level of scattering) can be compared by the energy of the electron above the conduction band of the material.

In metals, the Fermi-level is situated above the conduction band. The difference in energy between the conduction band and the Fermi-level is called the Fermi-energy. The difference in energy between the Fermi-level and the vacuum level is called the work-function. To compare two metals for a given incoming electron, the differences in both the Fermi energies and the work-functions have to be taken into account, as is clear from figure 5.8.
Taking Fe and Au for comparison, we can consider the work-functions and Fermi energies shown in table 5.1.

<table>
<thead>
<tr>
<th>Element</th>
<th>Work-function</th>
<th>Fermi energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li</td>
<td>2.93</td>
<td>4.74</td>
</tr>
<tr>
<td>Na</td>
<td>2.36</td>
<td>3.24</td>
</tr>
<tr>
<td>K</td>
<td>2.29</td>
<td>2.12</td>
</tr>
<tr>
<td>Rb</td>
<td>2.26</td>
<td>1.85</td>
</tr>
<tr>
<td>Cs</td>
<td>2.14</td>
<td>1.59</td>
</tr>
<tr>
<td>Cu</td>
<td>4.53-5.10</td>
<td>7.00</td>
</tr>
<tr>
<td>Ag</td>
<td>4.52-4.74</td>
<td>5.49</td>
</tr>
<tr>
<td>Au</td>
<td>5.1-5.47</td>
<td>5.53</td>
</tr>
<tr>
<td>Be</td>
<td>4.98</td>
<td>14.3</td>
</tr>
<tr>
<td>Mg</td>
<td>3.66</td>
<td>7.08</td>
</tr>
<tr>
<td>Ca</td>
<td>2.87</td>
<td>4.69</td>
</tr>
<tr>
<td>Sr</td>
<td>2.59</td>
<td>3.93</td>
</tr>
<tr>
<td>Ba</td>
<td>2.52-2.7</td>
<td>3.64</td>
</tr>
<tr>
<td>Nb</td>
<td>3.95-4.87</td>
<td>5.32</td>
</tr>
<tr>
<td>Fe</td>
<td>4.67-4.81</td>
<td>11.1</td>
</tr>
<tr>
<td>Mn</td>
<td>4.1</td>
<td>10.9</td>
</tr>
<tr>
<td>Zn</td>
<td>3.63-4.9</td>
<td>9.47</td>
</tr>
<tr>
<td>Cd</td>
<td>4.08</td>
<td>7.47</td>
</tr>
<tr>
<td>Hg</td>
<td>4.475</td>
<td>7.13</td>
</tr>
<tr>
<td>Al</td>
<td>4.06-4.26</td>
<td>11.7</td>
</tr>
<tr>
<td>Ga</td>
<td>4.32</td>
<td>10.4</td>
</tr>
<tr>
<td>In</td>
<td>4.09</td>
<td>8.63</td>
</tr>
<tr>
<td>Tl</td>
<td>3.84</td>
<td>8.15</td>
</tr>
<tr>
<td>Sn</td>
<td>4.42</td>
<td>10.2</td>
</tr>
<tr>
<td>Pb</td>
<td>4.25</td>
<td>9.47</td>
</tr>
</tbody>
</table>

Table 5.1: Work-functions and Fermi energies of some elements [5.11, 5.12]

The work-function of Fe is 4.7 compared to Au at 5.2. However, the Fermi energy of Fe is 11 compared to Au at 5.5. Therefore, based on the simple model discussed above, Fe is expected to scatter electrons more than Au. It should also be remembered that the work-function and other surface properties of metals are strongly
dependent on the surface morphology, orientation and other factors related to surface preparation.

5.4 Absence of Poly-depletion effects

In chapter 4, it was observed that problems occur as a result of depletion in the poly-silicon floating-gate. Since no doping is involved in metals, this effect is not expected to be observed, and is confirmed through the experimental results shown in figures 5.9 and 5.10.

Figure 5.9 shows the capacitance voltage plot of a 3 nm TiN floating-gate sample. There is no characteristic poly-depletion dip in the capacitance. Figure 5.10 shows the flat-band voltage versus programming time for a fixed bias. There is no initial slow period of programming as was observed in ultra-thin poly-silicon floating-gate samples. The slower period of programming was due to the decreased GCR of the device as the poly-silicon got depleted.

Hence, as expected, there are no doping-related issues observed in these metal floating-gate samples.
Figure 5.9: Capacitance-Voltage plot of 3 nm TiN FG capacitor showing no characteristic poly-depletion dip in the capacitance.

Figure 5.10: Flat-band voltage versus programming time for 3 nm TiN FG sample. There is no slowdown in the beginning as observed with depleted poly-silicon samples due to decreased GCR.
Chapter 5 – Metal Floating-Gate

5.5 Metal floating-gate work-function

The work-function of the metal floating-gate has a significant impact on the operation of the device. This is illustrated by figures 5.11 and 5.12.

Figure 5.11 shows band diagrams that explain why a deeper work-function FG device erases slower than a shallower work-function FG device. The deeper work-function FG device has a larger barrier height for F-N tunneling carriers out from the FG through the tunnel oxide onto the substrate.

Figure 5.12 shows band diagrams for a deeper work-function FG device having better retention compared to a shallower work-function FG device. This is because the deeper work-function FG device stores the charge in a deeper potential well (larger energy barriers).

Figure 5.13 shows experimental data confirming that a shallower work-function floating-gate device (Ti+Pt) erases faster than a deeper work-function floating-gate device (Pt).

Therefore, the metal work-function poses an important trade-off in the choice of metals for the floating-gate. A deeper work-function metal will provide better retention but will erase slower.
Figure 5.11: Band diagrams showing that deeper work-function FG leads to slower erase. Top figure shows shallower work-function FG device and bottom figure shows deeper work-function FG device. Deeper work-function FG has larger barrier to tunneling out from FG to substrate.
Figure 5.12: Band diagrams showing that deeper work-function FG leads to better retention. Top figure shows shallower work-function FG device and bottom figure shows deeper work-function FG device. Deeper work-function FG device stores charge in a deeper potential well.
5.6 Choice of metal: TiN, TaN

Since the work-function of the metal essentially determines a trade-off between erase-speed and retention, a mid-gap material such as TiN or TaN is picked as the floating-gate of choice. The following summarizes the reasons for picking TiN/TaN as the preferred floating-gate.

- Mid-gap material: Good trade-off between erase-speed and retention
- From figure 5.7, it is observed that TiN has a very small ballistic component.

Figure 5.13: Shows deeper work-function (Pt) FG device erase slower than shallower work-function (Ti+Pt) FG device
Chapter 5 – Metal Floating-Gate

- Ease of processing: TiN/TaN are already used in standard CMOS processing.
- Thermal stability: TiN/TaN have shown high thermal stability (upto 1000 °C) and compatibility with dielectrics. [5.13]

5.7 Ultra-thin 3 nm TiN floating-gate device

Based upon the reasons elaborated in section 5.6, TiN has been chosen as the floating-gate, and scaled down to ultra-thin 3 nm. The results are presented in this section.

Figures 5.14 and 5.15 show the programming and erasing characteristics of the 3 nm TiN floating-gate device. It can be observed that an excellent window of ~20 V is obtained with a perfect slope of 1 for the program/erase characteristics.

Figure 5.16 shows retention data for the 3 nm TiN floating-gate sample. As discussed before, the retention is characterized by programming devices to various levels and baking them at 150 °C for 24 hours. It can be observed that the retention is quite good as expected for a mid-gap work-function material.
Figure 5.14: Programming characteristics of 3 nm TiN FG device showing excellent performance

Figure 5.15: Erasing characteristics of 3 nm TiN FG device showing excellent performance
5.8 Summary

In this chapter, the use of a metal floating-gate as a replacement for poly-silicon was explored. The metal floating-gate was demonstrated to solve the ballistic carrier and poly-depletion issues discussed in chapter 4. The importance of the metal floating-gate work-function was discussed. From the various options available, TiN (or TaN) was chosen as the preferred floating-gate. The TiN floating-gate was scaled down to ultra-thin 3 nm, which demonstrated excellent electrical performance at even such low thicknesses. Utilization of such an ultra-thin floating-gate would greatly reduce cell-to-cell interference.
References


Chapter 5 – Metal Floating-Gate


Chapter 6

High-k dielectric IPD

6.1 Introduction

In this chapter, the replacement of the conventional oxide-nitride IPD with high-k IPD is explored. The motivation to use high-k IPD is to improve the GCR, which is lowered as a result of the inability to wrap the control-gate around the floating-gate. The requirements for the high-k material for use as IPD are enumerated. Al$_2$O$_3$ is chosen as the IPD material for experimental demonstration. Poly-silicon floating-gate devices are integrated with Al$_2$O$_3$ IPD, and their electrical characteristics are presented. The usefulness of a silicon nitride interlayer is shown. Metal (TaN) floating-gate devices are integrated with Al$_2$O$_3$ IPD, and the results are presented. Finally, the importance of the work-function of the control-gate is discussed.
Chapter 6 – High-k dielectric IPD

6.2 Motivation

As discussed in chapter 3, the control-gate is wrapped around the floating-gate in NAND flash devices [6.1]. An image of the control-gate wrap-around is shown in figure 6.1. The control-gate wrap-around has two important purposes:

(i) The conductive control-gate electrode shields the interference field lines between adjacent floating-gates. Changing the height of the control-gate shield has a significant effect on the interference [6.7 – 6.9].

(ii) The control-gate wrap-around increases the area of overlap between the control-gate and the floating-gate, thereby the capacitive coupling between them. This capacitive coupling is related to the GCR, which needs to be 0.6-0.7 for satisfactory operation. The control-gate wrap-around is very important for achieving a desirable GCR value.

As the density of NAND flash memory increases, there is no more space [6.2] for the control gate to go in between floating-gates (figures 6.2, 6.3). As a result, the interference problem is greatly enhanced, and the GCR also faces a significant drop. The interference problem was handled by scaling the height of the floating-gate [6.3 - 6.6] to a few nm.

In order to improve GCR, the conventional oxide-nitride IPD needs to be replaced by a high-k dielectric. Due to the increased dielectric constant, there is greater coupling between the control gate and the floating-gate, and hence the GCR is improved. According to ITRS [6.10], target high-k EOT is 8-10 nm from 2012-2024.
Figure 6.1: Image of control-gate wrapping around floating-gates [6.11]

Figure 6.2: CG no longer has space to wrap around FG thereby leading to a serious interference problem and GCR decrease
Figure 6.3: Image shows that the distance between cells is very less (Source: [6.12])

Gate coupling ratio $GCR = \frac{C_{IPD}}{C_{TOT}}$.

Where $C_{IPD} = \text{Capacitance between CG and FG}$

And $C_{TOT} = \text{Total capacitance to FG}$

Since $C_{IPD} = k \times \varepsilon_r \times \text{Area} / t_{IPD}$

Where $k = \text{dielectric constant of IPD}$

$\varepsilon_r = \text{vacuum permittivity}$

Area = Area between CG and FG

$t_{IPD} = \text{thickness of IPD dielectric}$

we have $GCR \propto k$

Hence the use of high-$k$ IPD to improve GCR.
6.3 Choice of high-k dielectric for IPD application

Requirements of high-k IPD

- **Effectively block charge loss from the floating-gate:** This implies a high-quality dielectric without defects, and good conduction band barrier height.

- **Must not trap charge:** The high-k dielectric must not be trappy as this will lead to undesired level shifts, and retention issues related to de-trapping.

- **Large conduction-band barrier height:** Among other things, a large conduction band barrier height is important to prevent gate-injection. Gate injection is the undesirable injection of carriers from the control-gate into the floating-gate during the erase operation (figure 6.4). When gate injection is equal to the charge flowing out from the floating-gate into the substrate, the cell cannot be erased anymore. This condition is called erase-saturation. A large conduction band barrier height helps alleviate the erase-saturation issue because it reduces the amount of gate-injection.

- **Physically thin:** The high-k dielectric IPD cannot be physically thick because fringing field lines through the high-k dielectric will cause interference between adjacent cells. [6.13]
The parameters of common high-k dielectric materials are shown in figure 6.5 and table 6.1.

![Diagram](image)

**Figure 6.4:** Shows erase-saturation occurring as a result of gate-injection. A larger conduction barrier height of the IPD would reduce injection from the control-gate.

<table>
<thead>
<tr>
<th>Dielectric</th>
<th>Permittivity</th>
<th>Band Gap (eV)</th>
<th>$\Delta E_C$ to Si</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO$_2$</td>
<td>3.9</td>
<td>9</td>
<td>3.5</td>
</tr>
<tr>
<td>Si$_3$N$_4$</td>
<td>7</td>
<td>5.3</td>
<td>2.4</td>
</tr>
<tr>
<td>Al$_2$O$_3$</td>
<td>9</td>
<td>8.8</td>
<td>2.8</td>
</tr>
<tr>
<td>TiO$_2$</td>
<td>80</td>
<td>3.5</td>
<td>0</td>
</tr>
<tr>
<td>Ta$_2$O$_5$</td>
<td>26</td>
<td>4.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Y$_2$O$_3$</td>
<td>15</td>
<td>6</td>
<td>2.3</td>
</tr>
<tr>
<td>La$_2$O$_3$</td>
<td>30</td>
<td>6</td>
<td>2.3</td>
</tr>
<tr>
<td>HfO$_2$</td>
<td>25</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>ZrO$_2$</td>
<td>25</td>
<td>5.8</td>
<td>1.4</td>
</tr>
<tr>
<td>ZrSiO$_4$</td>
<td>15</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>HfSiO$_4$</td>
<td>15</td>
<td>6</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6.1: Common high-k dielectrics [6.14]
Figure 6.5: Common high-k dielectrics. In general, larger the dielectric constant, smaller the band-gap. [6.14]

Problems of high-k IPD

- Most high-k dielectrics have small band-gap and barrier height. This leads to gate-injection and charge loss issues.
- High-k materials good enough for logic may not be suitable as IPD since they may be trappy at the large thicknesses (8-10 nm EOT) used.
- High-k dielectrics tend to have earlier dielectric breakdown.
- High-k dielectrics are known to form a poor interface with poly-silicon.
- High-k dielectrics with smaller band-gap exhibit metal electrode Fermi level pinning changing the barrier height.

Given the above requirements and problems, Al$_2$O$_3$ is chosen as the test vehicle to demonstrate high-k IPD integration due to its reasonably large conduction band barrier height (2.8 eV) and extensive prior development.
6.4 Poly-silicon FG with Al₂O₃ IPD

Al₂O₃ high-k IPD is integrated with poly-silicon floating-gate devices. The test structure is a floating-gate capacitor (no wrap). The TEM image of the gate stack of one such device is shown in figure 6.6.

![TEM image of the gate-stack of a poly-silicon FG Al₂O₃ high-k IPD device](image-url)

Figure 6.6: TEM image of the gate-stack of a poly-silicon FG Al₂O₃ high-k IPD device
Figures 6.7 and 6.8 show the programming and erase characteristics of 60 nm thick poly-silicon floating-gate devices with different IPD splits. The different splits under consideration are Al₂O₃ IPD without any post-deposition anneal, Al₂O₃ IPD with 950 °C post-deposition anneal, and silicon nitride + Al₂O₃ IPD with 950 °C post-deposition anneal. Silicon nitride is considered as an inter-layer between the poly-silicon and Al₂O₃, in order to improve the interface quality between them.

From figures 6.7 and 6.8, it is observed that samples with 950 °C post-deposition anneal (PDA) have much better characteristics than those without. Both programming and erase characteristics exhibit slope of 1 for an increased range on the PDA samples compared to the non-PDA samples. The better quality is achieved with PDA due to the crystallization of Al₂O₃, which reduces bulk defects, compacts the layer, and is known to increase the dielectric constant [6.15].

Comparing the cases with and without nitride inter-layer, the erase characteristics look identical while the programming characteristics look slightly better for the nitride inter-layer case. However, it is observed that the programming characteristics have a slope lesser than 1 for much of the range in all cases, indicating poorer performance with the poly-silicon floating-gate.

Figures 6.9 and 6.10 show the cycling and retention characteristics, respectively, of the samples with and without nitride inter-layer. It is observed that the retention behavior is identical but the cycling behavior shows better characteristics for sample with the nitride inter-layer. High-k dielectrics are known to form a poor quality low-k interfacial layer with silicon. Presence of the nitride inter-layer appears to
suppress that, and improve the quality of the interface with the poly-silicon floating-gate.

Figure 6.7: Program characteristics of 60 nm poly-silicon floating-gate devices with different IPD layers.

Figure 6.8: Erase characteristics of 60 nm poly-silicon floating-gate devices with different IPD layers.
Figure 6.9: Cycling data on 60 nm poly-silicon FG Al₂O₃ high-k IPD samples with and without nitride inter-layer. The samples were cycled to +/− 3 V flat-band voltage. The sample with nitride inter-layer performs better.
The nitride + Al$_2$O$_3$ IPD, identified so far as the best option, is integrated with poly-silicon FG layers of various thicknesses. Figures 6.11 and 6.12 present the programming and erase characteristics of poly-silicon FG devices (7 to 60 nm) with nitride + Al$_2$O$_3$ IPD. It is observed that the programming and erase characteristics of thin poly-silicon floating-gate devices are matched to those of thick poly-silicon floating-gate devices. The slope is lesser than 1 for much of the programming region, as discussed before.

Figure 6.13 shows endurance characteristics for these devices. It is observed that the thick poly-silicon floating-gate device performs much better than the ultra-thin 7 nm poly-silicon floating-gate device. This is expected, because the ultra-thin 7 nm poly-silicon floating-gate device suffers from ballistic carrier damage.
Figure 6.11: Program characteristics of poly-silicon FG devices of various thicknesses with nitride + Al₂O₃ high-k IPD.

Figure 6.12: Erase characteristics of poly-silicon FG devices of various thicknesses with nitride + Al₂O₃ high-k IPD.
Figure 6.13: Cycling data of 60 nm and 7 nm poly-silicon FG devices with nitride+Al₂O₃ high-k IPD. The 7 nm poly-silicon FG sample performs worse, likely due to ballistic carrier damage.
6.5 Metal floating-gate device with Al$_2$O$_3$ IPD

TaN (or TiN) is the metal floating-gate of choice, and it is integrated with Al$_2$O$_3$ high-k IPD. TaN/TiN is chosen, because it is a mid-gap material and also based on other reasons discussed in chapter 5. A TEM image of one such device is shown in figure 6.14.

![TEM image of a TiN floating-gate device with Al$_2$O$_3$ high-k IPD](image)

Figure 6.14: TEM image of a TiN floating-gate device with Al$_2$O$_3$ high-k IPD
The programming and erase characteristics of a TaN metal floating-gate device with Al₂O₃ high-k IPD are presented in figures 6.15 and 6.16 along with poly-silicon floating-gate sample data for comparison.

It is observed that the metal has much better program characteristics with a slope of 1 for a wider range. It is also observed that the metal erases slower due to the deeper work-function compared to n+ poly-silicon floating-gate.

The retention and endurance data of these TaN samples is presented in figures 6.17 and 6.18 respectively. It is observed that the retention of these samples is very good as expected for a mid-gap floating-gate material like TaN. It is also observed that the cycling behavior is reasonably good. It can be improved further by optimizing the process more, better understanding the tunnel oxide–metal interface, and improving techniques to deposit dielectrics on metals.

Figure 6.15: Program characteristics of TaN FG compared to poly-silicon FG, both with high-k IPD.
Figure 6.16: Erase of TaN FG compared to poly-silicon FG, both with high-k IPD.

Figure 6.17: Retention of TaN FG device with high-k IPD
Figure 6.18: Endurance data for TaN FG device with Al₂O₃ high-k IPD
6.6 Control-Gate

The control-gate work-function is an important parameter because it determines the gate-injection, and consequently the erase-saturation. As shown in figure 6.19, the deeper the work-function of the control-gate, the lesser the gate-injection is. This is especially important, given the use of deeper work-function metal floating-gates which erase slower.

This is experimentally confirmed through the data shown in figure 6.20, for devices with aluminum control gate versus platinum control gate. The deeper work-function platinum control-gate device performs better with respect to erase-saturation, compared to the shallower work-function aluminum.

6.7 Summary

The replacement of the conventional oxide-nitride IPD in floating-gate NAND flash devices with high-k dielectric IPD was explored. The requirements of the high-k material for use as IPD were discussed. Both poly-silicon and metal floating-gate devices were integrated with Al_2O_3 IPD, and their electrical characteristics were presented. The role of the work-function of the control-gate was clarified.
Figure 6.19: Band diagrams depicting deeper work-function metal with lesser gate-injection.
Figure 6.20: Pt CG (deeper work-function) showing improved erase saturation compared to Al CG (shallower work-function)
References


Chapter 7
Conclusions and Future Directions

7.1 The Planar Floating-Gate Device

The planar floating-gate device with ultra-thin metal floating-gate, high-k IPD and deep work-function control gate, is proposed to extend scaling of NAND flash memories. The device is shown in figure 7.1. The development of the device came about as follows:

- Cell-to-cell interference was identified as a major barrier in the continued scaling of NAND flash memories. In order to overcome this interference issue, the floating-gate was proposed to be scaled.
- Poly-silicon floating-gate scaling was studied in detail, and these devices were scaled down to ultra-thin 7 nm poly-silicon floating-gate thickness, which was the thinnest reported at the time.
A new reliability failure mechanism was proposed and demonstrated for the first time. This mechanism is due to programming carriers going increasingly ballistic through the ultra-thin poly-silicon floating-gates. These high energy ballistic carriers pose a reliability risk to the IPD.

A new test structure was developed to study the ballistic carrier issue, and was used to extract the inelastic mean free path of electrons in poly-silicon.

Poly-silicon was demonstrated to have scaling issues due to the ballistic carrier issue, poly-depletion etc.
• Metal-based floating-gate was proposed to overcome challenges of the poly-silicon floating-gate, and demonstrated for the first time.

• Metal floating-gate was shown to successfully overcome ballistic and other issues, and demonstrated excellent electrical characteristics.

• The effect of metal floating-gate work-function was studied, and mid-gap materials, e.g. TiN/TaN, were chosen as the preferred floating-gate material.

• The TiN floating-gate was scaled down to ultra-thin 3 nm, exhibiting excellent performance, and reducing cell-to-cell interference greatly due to its very small thickness.

• High-k dielectric was proposed to replace conventional IPD in order to compensate loss of GCR caused by lack of CG wrap-around.

• The factors underlying choice of high-k dielectric were discussed and Al₂O₃ was picked.

• Both poly-silicon and metal floating-gate devices were integrated with Al₂O₃ high-k IPD.

• Use of a silicon nitride inter-layer improved the quality of the Al₂O₃ IPD on poly-silicon floating-gates.

• TaN floating-gate devices with high-k IPD were shown to have better programming characteristics than poly-silicon floating-gate devices with high-k IPD. The TaN floating-gate devices erased slower due to the deeper work-function. Overall, the TaN devices had very good electrical performance.
Chapter 7 – Conclusions and Future Directions

- A deeper work-function control-gate was shown to reduce gate-injection, and consequently erase-saturation.

- Combining all these layers, as shown in figure 7.1, we get the planar (no wrap) floating-gate device with ultra-thin TiN/TaN floating-gate, Al₂O₃ high-k IPD and deep work-function control gate. This structure would allow us to extend scaling of NAND flash memories.

### 7.2 Future Directions

- Tunnel oxide thickness scaling has essentially stopped. Work on barrier-engineered layers is an important area that has potential.

- The process steps involved in the planar floating-gate device need to be further optimized. For example, there needs to be better understanding for deposition of high quality high-k dielectrics on metals, and interaction of the metal-tunnel oxide interface.

- As the cell becomes smaller and smaller, it will be limited by the number of electrons stored. At that point, instead of scaling cell-area, it is better to integrate cells in 3D to increase density. Work towards this, like the bit cost scalable (BiCS) cell by Toshiba, is an important direction of research.

- At those ultra-small dimensions, as the flash cell becomes intrinsically limited, non-charge based cells like ferro-electric RAM (FeRAM),
magnetic RAM (MRAM), phase change RAM (PCRAM) and resistance change RAM (ReRAM) are promising alternatives.

Figure 7.2 shows potential solutions for non-volatile memory, as projected by ITRS.

Figure 7.2: Non-volatile memory potential solutions (Source: ITRS 2009)