Characterization and Modeling of Non-Volatile SONOS Semiconductor Memories with Gridded Capacitors

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Christopher John Barthol, B.S., M.S.

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2015

Dissertation Committee:

Dr. Marvin H. White, Advisor

Dr. Paul R. Berger

Dr. Steven B. Biblyk

Dr. Steven A. Ringel
© Copyright by

Christopher John Barthol

2015
Abstract

The demand for high-capacity, low-power memory is increasing rapidly as modern portable electronic devices boost performance while decreasing in size. Due to its many advantages over traditional floating gate memory, Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) type semiconductor memories and their various derivatives are the next step in the evolution of Non-Volatile Semiconductor Memory (NVSM) technology. As the SONOS cell size is decreased the gate stack has reached its limitations with silicon oxide used as the defacto oxide for both the blocking and tunneling layers. High K materials, which improve the electrical characteristics while maintaining a thick blocking or cap layer, have started to replace silicon dioxide. The change from a silicon dioxide that has seen over 60 years of development to new materials requires a rapid turnaround in the fabrication of devices as well as quick characterization of the gate stack.

We have fabricated gridded Metal-Aluminum Oxide-Nitride-Silicon Oxide-Silicon (MANOS) capacitors. The gridded capacitor structure allow carrier types other than what is originally present in the substrate. This structure is easier to fabricate while still allowing all characterization tests, such as speed write/erase and endurance tests, to be performed. This technique does not require ohmic contacts to the grid lines reducing the fabrication process. Two wafer sets have been created. One has an aluminum oxide block layer while the other has a hafnium/aluminum oxide blend.
The capacitors have been fabricated on <100> p-type silicon wafers with a resistivity of 20 Ω−cm. A grid structure with line widths of 5µm and line spacing ranging from 30µm to 300µm is defined and doped with phosphorous. The gate stacks consist of a high quality 2.4nm SiO$_2$ tunnel oxide layer thermally grown in a triple-wall oxidation furnace, a 7.7nm silicon-rich nitride deposited by LPCVD, and the two sets of clocking oxides deposited by atomic layer deposition (ALD).

Various fundamental and dynamic electric characterization techniques such as Capacitance-Voltage, Linear Voltage Ramp, Speed Write/Erase, and Retention measurements have been completed on the gridded capacitor structure. The nuances of each test is described in detail. A flatband voltage tracking system has been created to aid in these measurements. We also discuss a set of simulation programs for charge trap non-volatile memories. The programs show excellent agreement with both full transistors and the gridded capacitor structure.

We also present a method to extract carrier mobility with the two terminal gridded capacitor structure where the silicon substrate is implanted with a grid structure of the opposite carrier type. We can extract the carrier mobility as a function of vertical electric field with a combination of Capacitance-Voltage(C-V) and Conductance-Voltage(G-V) measurements. In addition, the structure eliminates the need for source/drain contacts, which lends itself well to material systems where low resistance, source/drain contacts are difficult to implement and reduce device performance.
This is dedicated to the love of my life Emily.
Acknowledgments

I would not have been able to make it this far and be so successful without a countless number of people, but I'll try to mention as many as possible. First of all, I would like to extend my sincere gratitude and thanks to my advisor, Dr. Marvin H. White, for his guidance, support, and encouragement with my research. He has provided me with great opportunities throughout the years. Dr. White welcomed me into the research world with a Research Experience for Undergraduates (REU) over the summer of 2008, let me stay and work in the lab while I finished my undergraduate degrees, and then took me under his wing as a Masters Student and now finally as a PhD Candidate at The Ohio State University. He has spent countless hours helping me understand semiconductor physics and deriving equations. I would be truly lost without his knowledge and guidance.

I would also like to thank Mr. Ray Filozof, Dr. Floyd Miller for their direction and help in the fabrication of my devices at Lehigh, and Dr. John A. Carlin, Derek A. Ditmer, Dr. Peter J. Janney III, Aimee L. Price, and Paul D. Steffen for their help in fabrication at The Ohio State University. Also a big thank you to Dr. Patrick Shea of Northrup Grumman Corporation for the deposition of High-K ALD films.

It has been a pleasure working in the Sherman Fairchild Center and The Ohio State University and I wish to thank all of the students that I have been fortunate enough to work with over the years. Thanks to Akwete Borkui-Doku, Gan Wang, Xiaochen...
(DeeDee) Zhang, Eric Winokur, Nathan C. Eichenlaub, Luckshitha S. Liyanage, Josh Goldman, Stephen Abate, Matthew Galligan, Marcela Meza, and Jordan Thimot.

Finally, I would like to thank my friends and family. First, to Emily, thank you for being there with me from the start of this great adventure and all the times to come. To all my friends, thank you for making my time at Lehigh and The Ohio State University extremely special. Most importantly, to my parents who have supplied time, love, and support throughout my entire life, you have always been there for me and words cannot express my gratitude.
Vita

December 30, 1985 . . . . . . . . . . . . . . . . . . . . . . . . . Born - Reeders, PA, USA

2008 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .B.S. Electrical Engineering, Lehigh University
2008 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .B.S. Engineering Physics, Lehigh University
2010 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .M.S. Electrical Engineering, Lehigh University
2010-present . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .Graduate Research Assistant, The Ohio State University.

Publications

Research Publications

C.J.Barthol and M. H. White “Characterization of NVSM Gate Stacks with Gridded Capacitors”. *International Semiconductor Device Research Symposium*, University of Maryland, MD, 2011


Fields of Study

Major Field: Electrical and Computer Engineering
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>1. Introduction to Non-Volatile Semiconductor Memories</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Non Volatile Semiconductor Memories</td>
<td>3</td>
</tr>
<tr>
<td>1.1.1 Floating Gate</td>
<td>3</td>
</tr>
<tr>
<td>1.1.2 SONOS and Derivatives</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Ferromagnetic (FeRAM)</td>
<td>13</td>
</tr>
<tr>
<td>1.3 Magnetic Tunnel Junction (MJT)</td>
<td>18</td>
</tr>
<tr>
<td>1.3.1 Field Switch</td>
<td>20</td>
</tr>
<tr>
<td>1.3.2 Spin Torque Transfer</td>
<td>22</td>
</tr>
<tr>
<td>1.4 Phase Change (PCRAM)</td>
<td>25</td>
</tr>
<tr>
<td>1.5 Experiment Memory Devices</td>
<td>29</td>
</tr>
<tr>
<td>1.5.1 Nano-Electro-Mechanical Memory (NEMM)</td>
<td>31</td>
</tr>
<tr>
<td>1.5.2 Mott Memory</td>
<td>34</td>
</tr>
<tr>
<td>1.5.3 Redox</td>
<td>37</td>
</tr>
<tr>
<td>1.5.4 Organic Memory</td>
<td>39</td>
</tr>
<tr>
<td>1.6 Memory Comparisons</td>
<td>41</td>
</tr>
<tr>
<td>1.7 Summary</td>
<td>46</td>
</tr>
</tbody>
</table>
2. Gridded Capacitors Fabrication ........................................ 48
   2.1 Tunnel Oxide Formation and Characterization .................. 53
      2.1.1 Deal-Grove and Massoud Model .......................... 53
      2.1.2 2.1.2 X-Ray Photoelectron Spectroscopy ................. 57
      2.1.3 Triple Wall Furnace ....................................... 68
      2.1.4 Tunnel Oxide Results ...................................... 72
   2.2 Nitride Charge Trapping Formation and Characterization ....... 77
   2.3 Blocking Oxide Formation and Characterization ................ 80
   2.4 High-K Oxides .................................................. 81
      2.4.1 Aluminum Oxide (Al₂O₃) Blocking Oxide ................ 89
      2.4.2 Aluminum Oxide (Al₂O₃)/ Hafnium Oxide (HfO₂) Blocking Oxide .................................................. 93
   2.5 Post Processing ................................................ 98
   2.6 TEM Characterization of Samples ................................ 98
   2.7 Summary ....................................................... 119

3. Electrical Characterization Techniques .............................. 121
   3.1 Capacitance-Voltage Measurements .............................. 122
      3.1.1 Capacitance-Voltage Theory .............................. 122
      3.1.2 Gridded Capacitor Capacitance-Voltage Results .......... 124
   3.2 Linear Voltage Ramp ............................................ 130
      3.2.1 Linear Voltage Ramp Theory .............................. 130
      3.2.2 Linear Voltage Ramp Measurements on Gridded Capacitors 135
   3.3 Mobility ........................................................ 137
   3.4 EOT Calculations ............................................... 145
   3.5 Summary ....................................................... 148

4. Speed Write/Erase .................................................... 149
   4.1 Write/Erase Speed Theory ...................................... 150
   4.2 Flatband Tracking Circuit ..................................... 158
   4.3 Write/Erase Speed Measurements ................................ 168
   4.4 Simulation of Non Volatile Charge Trap Memories ............ 178
   4.5 Summary ....................................................... 191

5. Retention .......................................................... 193
   5.1 Retention Theory ............................................... 196
   5.2 Retention Results .............................................. 199
6. Other Material Considerations ........................................ 202
   6.1 Gridded Capacitors in III-V Materials ....................... 202
   6.2 Fabrication of Devices ........................................ 202
   6.3 III-V Gridded Capacitor Results ............................ 216
   6.4 Summary .................................................. 222

7. Conclusions .................................................. 223
   7.1 Conclusions from This Dissertation .......................... 223
   7.2 Recommendations for Future Research ....................... 225

Appendices .................................................. 227
   A. MANOS Transistor Fabrication ................................. 227
   B. Gridded Capacitor Fabrication ............................... 232
   C. C-V Measurement ........................................... 234
   D. LVR Measurement ............................................ 236
   E. Speed Write/Erase Measurement .............................. 238
   F. Matlab Write/Erase Simulation Programs .................... 240
      F.1 Erase Program ........................................... 240
      F.2 Write Program .......................................... 244
      F.3 Jbtb Program ............................................ 247
      F.4 Jtat Program ............................................ 248
      F.5 Simpson Erase Program ................................ 249
      F.6 Simpson Write Program ................................ 250
      F.7 Modified Simpson Write Program ......................... 251
      F.8 Calculate Write Electric Field Program ................. 252
      F.9 Calculate Erase Electric Field Program ................. 253

Bibliography ............................................... 255
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A comparison of major memory systems. A variety of specifications is provided on the best devices which have provided data. Baseline technologies are listed as DRAM, SRAM and NAND with prototype technologies listed as FeRAM, MRAM/STT, and PCM</td>
</tr>
<tr>
<td>1.2</td>
<td>A list of which technologies have the best perform for each listed feature</td>
</tr>
<tr>
<td>2.1</td>
<td>Experimental and theoretical thickness calculations for given oxidation times. We originally calculated an initial starting oxide of 5.4Å, after switching argon tanks the tests wafers were calculated to have a starting oxide thickness of 16.4Å. The final column shows the thicknesses measured on our actual wafers</td>
</tr>
<tr>
<td>2.2</td>
<td>Static dielectric constant, experimental band gap, and conduction band offset on Si of some high-K materials. Aluminum oxide has nearly the same band gap but a much higher K value</td>
</tr>
<tr>
<td>2.3</td>
<td>TEM thickness measurements of the gridded capacitor with a 2.4nm tunnel oxide and aluminum oxide blocking layer and the gridded capacitor with a 4.0nm tunnel oxide and the aluminum/hafnium blocking oxide. Each measurement is done ten times and averaged</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Image Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>NAND flash memory market for the years 2007 to 2014. There is a 600% increase in the amount of memory over the last 4 years.[34]</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>A cross section view of a typical floating gate non-volatile memory structure. The control gate is separated from the floating gate by an ONO stack.</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Tunneling Electron Microscope (TEM) image of a floating gate structure. The Control Gate (CG) is separated from the Floating Gate (FG) by a narrow ONO stack. Tungsten silicon(WSi) is used as a contact material.[37]</td>
<td>5</td>
</tr>
<tr>
<td>1.4</td>
<td>Diagram of floating gate transistor (a) Write operation by CHE injection over the drain, (b) Erase by FN tunneling over the source</td>
<td>6</td>
</tr>
<tr>
<td>1.5</td>
<td>History of charge trapping devices. Flash memories started with metal gates, then shifted to poly-silicon gates, and back to metal gates.[106]</td>
<td>8</td>
</tr>
<tr>
<td>1.6</td>
<td>TEM of a charge trapping non-volatile memory. Aluminum oxide is used as the blocking oxide, silicon nitride as the charge trap material, and silicon oxide as the tunneling oxide.[49]</td>
<td>10</td>
</tr>
<tr>
<td>1.7</td>
<td>(a) Write operation for a charge trapping NVSM (b) Erase operation for a charge trapping NVSM</td>
<td>12</td>
</tr>
<tr>
<td>1.8</td>
<td>First schematic of a FeRAM. The ferroelectric material was deposited directly on top of the channel.[110]</td>
<td>13</td>
</tr>
<tr>
<td>1.9</td>
<td>Schematic hysteresis loop for a polycrystalline Fe Film. The higher polarization state is a binary &quot;1&quot; while the lower state is a binary &quot;0&quot;.[29]</td>
<td>14</td>
</tr>
</tbody>
</table>
1.10 Schematic of a 1T1C Ferromagnetic Memory Cell. The access transistor is on the left side of the image with the source and drain in red. The ferroelectric film is in blue with the dark blue line as the connector between the transistor and the capacitor. [27] 15

1.11 Device structure for a metal-insulator-ferroelectric-insulator-semiconductor field effect memory. The ferroelectric film is separated from the channel by a buffer layer, eliminating the need for a separate transistor. [29] 17

1.12 Magnetic Tunnel Junction (MJT) states a) Parallel State with both polarizations in the same direction b) Anti-parallel state with the polarization in opposite directions. [2] 19

1.13 External field switch MRAM. The hard axis allows access to the cell while the easy axis controls the memory device by which direction the current flows in. [2] 21

1.14 Programing of an STT MRAM a) Anti-parallel to parallel. Electrons flow from bottom to top. Electrons with the same spin pass through the pinned layer and interact with the free layer to switch states. b) Parallel to Anti-Parallel. Electrons flow from top to bottom. Electrons with the same spin of the free and pinned layer pass through the device. Opposite spin electrons are blocked by the pinned layer and start to build up. After a critical point the electrons of opposite spin exert a force on the free layer to switch polarization states. [42] 23

1.15 TEM sample of a MJT. Magnesium Oxide is used as a barrier layer between the two CoFeB layers. The advantage of such devices is that extremely small scale devices are available. [42] 24

1.16 Phase change memory a) Schematic showing the two electrodes, heater, and phase change material. A thin conductor heats the phase change material. b) Programming a PCRAM. A RESET pulse heats and cools the material quickly causing an amorphous region, while the SET pulse allows the material to form a crystalline structure. A read pulse does not change the material structure. [8] 26

1.17 PCRAM Endurance. The on and off resistance values remain constant over 10E8 cycles. After 10E8 cycles the values start to defer but the difference between on and off states are still very far apart. [47] 28
1.18 Fully Integrated 1Gb PCRAM Chip. The chip consists of individual partitions of 64Mb strung together to form a 1Gb chip. [50] . . . . . . 30

1.19 8Gb PCRAM features at the 20nm node. The PCRAM has quick Set pulses in a small area. [50] . . . . . . . . . . . . . . . . . . . . . . . . 31

1.20 Schematic design of a MEMS based nonvolatile memory device using a cantilever. The actuating electrode pulls the metal cantilever into direct contact of the floating gate to exchange charge. [51] . . . . . . . 32

1.21 Cross section of a vertical silicon cantilever switch at a) Neutral position, the cantilever is electronical isolated from either side. b) ON state, the cantilever is pulled to GND or VPi to exchange charge. [65] 33

1.22 FIB images of vertical NEMM a) Top down, sowing the cantilever in the air gap. b) Cross section of NEMM depicting a 23nm air gap. [65] 34

1.23 100X magnification top view image of CeRAM test structure [107] . . 35

1.24 Resistance values for ON and OFF state over 100 cycles for a CeRAM structure. There is a large variation of on off resistances that are separated by a factor of 10 resistance. [107] . . . . . . . . . . . . . . . . . . . . . . . . 36

1.25 Program/Erase cycle of a electrochemical metalization (ECM) memory. A) Undisturbed state with no dendrites. B) Ag+ diffusion across the insulator toward the Pt electrode. C) Ag+ dendrite forms. D) Dendrite connection between to contacts. E) Ag+ moves away from Pt electrode. [98] . . . . . . . . . . . . . . . . . . . . . . . . 38

1.26 Operational diagram of a molecular memory [77] . . . . . . . . . . . 39

1.27 Measured logic diagram of a molecular memory device. A write pulse first sets the device and subsequent pulsese have a corresponding high output. When an erase pulse is applied the next pulse shows no output, at which point it is rewritten and has an output with additional pulses. [77] . . . . . . . . . . . . . . . . . . . . . . . . 40
1.28 A graphical view of our Figure of Merit. The light blue line is a Universal memory, which all other memory systems are striving towards. NAND is remarkably lower due to the endurance characteristics of NAND devices.

2.1 Write and erase curves on a non gridded capacitor. The capacitor has a gate stack of aluminium oxide/silicon nitride/silicon oxide 8/8/3nm. For the write or program state the generation of carriers from light is slow and the device does not begin to write until after 1E-3 seconds.[99]

2.2 Large scale view of the gridded capacitor mask set. The blue circles are the metal gates of each individual capacitor. Tan lines indicate the grid lines with a spacing of 30um on the left and 300um on the right. Capacitor sizes increase to compensate with the larger grid sizes.

2.3 View of one grid spacing section of the gridded capacitor mask set. Blue indicates the gate metal, with tan lines indicating the grid lines. Each grid section is labeled for easy detection.

2.4 Values of the pre-exponential constants and activation energies for different crystal orientations in silicon.[61]

2.5 Pre-exponential constant and activation energies for different crystal orientation and dry oxidation for silicon.[59]

2.6 Oxidation thickness versus Time with varying temperatures.[59]

2.7 XPS Diagram. A focused beam of X-rays hits the sample which then emits electrons. The electrons are captured and analyzed resulting in a graph of the number of electrons at each energy level.[13]

2.8 Scienta ESCA-300 used at Lehigh University Material Science Department. The detector is towards the back of the photo.

2.9 XPS spectra on a hydrogen terminated silicon surface. The graph depicts the number of counts (y-axis) versus the electron binding energy (x-axis).
2.10 XPS spectra on a hydrogen terminated silicon wafer. Major peaks are automatically label by the CasaXPS graphing program. The labeled peaks correspond to the presence of Oxygen, Silicon, and Carbon atoms in the sample.

2.11 XPS Spectra of the silicon peak on a hydrogen terminated silicon wafer. 64

2.12 XPS Spectra of the silicon peak on a hydrogen terminated silicon wafer with labels. The labels are shifted left of the experimental results. This is due to a charge difference between the detector and the sample. This imbalance can be fixed by flowing electrons over the sample(dangerous) or through the CasaXPS program.[56] 65

2.13 XPS spectra of the carbon peak on a hydrogen terminated silicon wafer. The peak is much less defined indicating few carbon atoms on the sample. Carbon is expected as the wafers were transferred outside of a cleanroom environment. 67

2.14 XPS spectra of the oxygen peak on a hydrogen terminated silicon wafer. Like the carbon peak, the oxygen peak is ragged indicating few oxygen atoms. Again some oxygen is expected as the wafers were transferred outside of a cleanroom environment. 67

2.15 Schematic diagram of the triple wall oxidation furnace. The gasses enter a pre combustion furnace [A] to remove some contaminates. A cold trap [C] and cryopump [B] freeze out water molecules. Nitrogen [D], flows through the outer tubes [G] to stop containments from migrating into the oxidation furnace [E]. A glass tube called a white elephant [F] connects the wafers to the oxidation tube for a warm up process. Gases are expelled and monitored out of the back of the tube [H, I, J][113] 68

2.16 Photo showing the operator section of the triple wall furnace. The white elephant [1] connects to the inner furnace tube [2]. The cold trap [4] and furance controls [3] are also shown[113]. 70

2.17 XPS spectra of the oxygen peak on hydrogen terminated silicon wafers placed in the white elephant for one hour. The peak shows no difference from previous hydrogen terminated wafers. 73
2.18 XPS spectra of the silicon peak on hydrogen terminated silicon wafers placed in the white elephant for one hour. The peak shows no difference between from previous hydrogen terminated wafers. ........................................ 73

2.19 Original argon only test wafer showed a starting oxide of 5.4 angstroms. After running oxidation tests with a new argon cylinder the initial oxide jumped to 16.4 angstroms. ................................................................. 75

2.20 XPS spectra of the silicon peaks on a silicon wafer which has undergone oxidation. Area calculations are done through the CasaXPS software program (shaded region) for thickness measurements. This wafer shows the peaks for an oxide thickness of 2.4nm. ............................................. 77

2.21 Control panel of the LPCVD furnaces at Lehigh University. The tubes are directly behind the panel and the loading bay is at the back of the photo. The white tube visible on the right side of the photo is the POCl$_3$ diffusion furnace. ................................................................. 78

2.22 Full scan XPS spectra of a silicon wafer after a silicon nitride deposition. 78

2.23 XPS Spectra of the silicon peak on a silicon wafer after silicon nitride deposition. The double silicon peak is on the right with the nitride corresponding nitride bonding peak on the left. ................................. 79

2.24 XPS spectra of the carbon peak on a silicon wafer after silicon nitride deposition. The peak is low and rough depicting the lack of many carbon atoms. ................................................................. 80

2.25 XPS spectra of the nitrogen peak on a silicon wafer after silicon nitride deposition. ................................................................. 81

2.26 Schematic of a basic n channel MOSFET. The darker red shows the depletion region, while the orange region shows the gate oxide. Gate, Drain, and Source contacts are shown. In this transistor the source is tied to the substrate. ................................................................. 82

2.27 Capacitance equivalent of the gate stack. Each layer of the gate stack has its own associated capacitance. Referencing the capacitance to one individual layer we can provide an equivalent capacitance of the entire gate stack. ................................................................. 83
2.28 Static dielectric constants vs. band gap. The correlation between High K value and lower Band Gap is clearly seen. Other good candidates are hafnium oxide and zirconium dioxide. [79] ............... 86

2.29 Schematic of required band offsets. For good retention in non-volatile charge trap memories an offset of greater than 1V.[80] ............... 87

2.30 Band gap and band alignment for various materials on silicon. Both silicon oxide and aluminum oxide have offsets greater than 1V. A few higher K materials have no offset with silicon.[108] ............... 88

2.31 Atomic Layer Deposition (ALD) process for hafnium or zirconium oxide. The precursor is sent through until saturated adsorption occurs. The chamber is purged to prevent deposition and then water is passed through the chamber. The water reacts with the precursor to form a single layer of the desired oxide. .................. 90

2.32 Full scan XPS spectra of aluminum oxide deposited by ALD on silicon wafers. The deposition is over 100A so only oxygen, carbon, and aluminum are present. .................. 91

2.33 XPS spectra of the carbon peak of aluminum oxide on silicon wafers. A larger carbon peak is observed than on all other samples. This could be due to the length of time between deposition and measurement. . 91

2.34 XPS spectra of the oxygen peak on aluminum oxide deposited on silicon wafers. ................................. 92

2.35 XPS spectra of the aluminum peak on aluminum oxide deposited on silicon wafers. The peak is large and the two spins states have merged into a single visible peak. .................. 93

2.36 Full scan XPS spectra on aluminum/hafnium oxide layer deposition on a silicon wafer. The layer is thicker than XPS results can measure and so no silicon is visible in the scan. .................. 94

2.37 XPS spectra of the hafnium peak on the aluminum/hafnium oxide layer deposition on silicon wafers. The double peak due to spin orbital decoupling from the hafnium atom are clearly visible ................. 95
2.38 XPS spectra of the aluminum peak on the aluminum/hafnium oxide layer deposition on silicon wafers. Similar to the aluminum only samples no double peak is observed. ........................................... 95

2.39 XPS spectra of the oxygen peak on the aluminum/hafnium oxide layer deposition on silicon wafers. ...................................................... 96

2.40 XPS spectra of the oxygen peak on the aluminum/hafnium oxide layer deposition on silicon wafers. In addition to carbon, potassium is also present on the sample. ...................................................... 96

2.41 Schematic of a typical TEM machine. An electron gun produces electrons which are focused through a sample. The scattered electrons are then focused onto a recording system for viewing.[23] .................. 99

2.42 The Tecnai F20 system used for TEM measurements at The Ohio State University. ................................................................. 101

2.43 A closer image of the viewing screen on the Tecnai F20 system. A phosphorous screen is used for direct imaging while a computer system also captures and records the image produced. ...................... 102

2.44 Helios 600 Focused Ion Beam system used to cut and thin the sample. 103

2.45 TEM image showing the trench dug to find and extract the test sample. The device source and drain contacts are the two large dark rectangles. A protective metal has been deposited over the device. ............... 104

2.46 A closer view the test device to be extracted. The source and drain contacts and vias are clearly visible. The gate metal is a small rectangle between the two white vias. ............................................ 105

2.47 The extracted sample attached to a post. The sample has been slightly thinned. ................................................................. 107

2.48 The start of the shaving and thinning process. The device is still invisible. The debris from extraction is shown on the right side of the image. ................................................................. 108
2.49 The device thinned to be transparent to electrons. The sample is thinned at a slight angle so the top maintains its shaped. The bottom section as been overly thinned and some holes have appeared. The device to be imaged is in the center of the sample and slightly glows due to electron transmission.

2.50 The sample is placed in a container for transport. The small metal piece in the membrane container has three posts, and one of the posts has the sample attached to it.

2.51 TEM image of a SONOS transistor provided by Northrop Grumman Corp. The two metals for the source and drain contacts are clearly visible. The gate region is above of these two rectangles.

2.52 TEM image of a SONOS transistor provided by Northrop Grumman Corp. The gate metal is towards the top of the image. The gate stack is not yet visible.

2.53 TEM image of a SONOS transistor provided by Northrop Grumman Corp. The silicon substrate and poly silicon gate are labeled. The gate is barely visible between these two materials.

2.54 Maximum resolution TEM image of a SONOS transistor provided by Northrop Grumman Corp. The gate stack has a thickness of roughly 150A. The tunnel oxide is barely visible and the blocking oxide is roughly 20A.

2.55 TEM image of the gridded capacitor with an aluminum oxide blocking oxide. The silicon substrate is on the bottom of the image with the metal gate on the top. The silicon oxide tunnel oxide is the dark line across the image. The nitride layer is above this with the aluminum oxide layer above the nitride.

2.56 TEM image of the gridded capacitor with an aluminum/hafnium oxide blocking oxide. The silicon substrate is on the bottom of the image with the metal gate on the top. The silicon oxide tunnel oxide is the black line. The nitride layer is darker in this image with the aluminum/hafnium above this region. At the top of the blocking oxide is a white line which could be a halfnium only layer.
3.1 Capacitance vs Voltage measurement on a gridded capacitor with a capacitor radius of 100um, a grid spacing of 50um, and a measurement frequency of 100kHz. The accumulation capacitance is the same as the inversion capacitance. ................................. 125

3.2 Capacitance vs Voltage measurements on gridded capacitors. With a 300um grid spacing the inversion capacitance is lowered, while with a 30um grid spacing the accumulation capacitance is lowered. A grid spacing of 50-100um is ideal. ................................................................. 126

3.3 Capacitance vs Voltage measurement on a gridded capacitor with aluminum oxide as the blocking oxide. A dip is observed in the inversion region due to a trap state away from the silicon/silicon oxide interface. 128

3.4 Capacitance vs Voltage measurement on a gridded capacitor with aluminum/hafnium oxide as the blocking oxide. The same dip is in the inversion region is observed as is likely from a trap state near the silicon oxide/silicon nitride interface. ................................. 129

3.5 Linear Voltage Ramp Setup on a transistor. A function generator is tied to the source/drain/substrate while an electrometer measures the current out of the gate. The results are graphed in LabVIEW. For a gridded capacitor sample the source/drain and substrate are already tied together. ................................. 131

3.6 Example of an LVR measurement. At high voltages electron tunneling increases creating spikes. ................................................................. 133

3.7 LVR measurements on a gridded capacitor with 2.4/7.7/15.0nm silicon oxide/silicon nitride/aluminum oxide stack and a grid width of 30um and a capacitor radius of 40um. ................................. 136

3.8 LVR measurements on a gridded capacitor with 2.4/7.7/15.0nm silicon oxide/silicon nitride/aluminum oxide stack and a grid width of 30um and a capacitor radius of 70um. ................................. 136

3.9 Transmission line equivalent circuit for a transistor [8]. Using the minority and major carrier currents $i_1$ and $i_2$, the total capacitance and conductance per unit length can be obtained. From the peak in the conductance vs. frequency curve a channel mobility can be determined. 138
3.10 A transistor from [116] exhibits a peak at 10.5MHz 

3.11 Schematic and TEM of the gridded capacitor under test. The grid spacing is not visible in the TEM image. 

3.12 Gm vs Frequency curve on the gridded capacitor. The device exhibits a peak at 3.46MHz. 

3.13 Mobility measurements vs Grid Spacing. Too many grids reduces the mobility, while too few grids has an increased mobility. The optimal grid spacing lies between 50-100um. 

3.14 EOT calculations on the gridded capacitor with a gate stack containing an aluminum oxide blocking layer. The extracted K value approaches 7.8 for capacitance radii greater than 150um. 

3.15 EOT calculations on the gridded capacitor with a gate stack containing an aluminum/hafnium oxide blocking layer. The K value is smaller for smaller capacitances. The K value of the layer is roughly 14. 

4.1 Energy band diagrams and oxide electric field requirements for Fowler-Nordheim, Direct Tunneling, Modified Fowler-Nordheim, and Trap Assisted Tunneling. 

4.2 Charge transport mechanism for the write and erase operations in charge trap non-volatile memories. 

4.3 Simplified energy band diagram of a SONOS charge trap non-volatile memory. The model simplifies the structure by using a sheet of charge located in the nitride (Qn) a distance X from the tunnel oxide. 

4.4 A basic C-V measurement on a non-gridded capacitor. The inversion capacitance does not equal the accumulation capacitance due to the lack of carriers in the substrate. 

4.5 The write/erase pulse application process. A positive (write) pulse shifts the curve to the right and a negative (erase) pulse shifts it towards the left. Notice that the shift is parallel.
4.6 Old capacitance vs voltage measurement setup. The pulses are applied to the device, then disconnected and connected to the C-V analyzer for measurement. This process leaves a large time difference between pulse application and measurement.  

4.7 Front panel of the Boonton Capacitance Meter. The meter has two ports for connecting capacitance. The external voltage bias port is in the back of the meter.  

4.8 Boonton circuit to change applied voltages. The front has a the device under test connect as well as a reference capacitance. A diff amplifier and an integrator change the applied voltage until the difference between capacitances is zero.  

4.9 Full schematic of the flatband tracking circuit. Connections to the meter, voltages, and the computer are along the outside.  

4.10 Schematic of the ADG451SR switch. A voltage is applied to open the correct port to the device. Write, erase, and read signals are applied through the switch.  

4.11 The LabVIEW erase program control panel. The user can change pulse width, number of data points, and the path of the output. The test is graphed in real time.  

4.12 Final flatband tracking circuit board. The connections around the board go to the probe stand or LabVIEW. The connections in the top left are for applying write or erase voltages.  

4.13 Low voltage speed write/erase curves on 24/77/150A silicon/nitride/aluminum gate stack. Write and erase curves are shown for applied voltages of 6, 7, and 8V.  

4.14 Low voltage speed write/erase curves on 24/77/150A silicon/nitride/hafnium-aluminum gate stack. Write and erase curves are shown for 6, 7, and 8 applied voltages.  

4.15 Write curves on a gridded capacitor with the aluminum oxide blocking oxide. Results are shown for 9, 10, 11, 12, 13, and 14 volts.
4.16 Write curves on a gridded capacitor with the aluminum/hafnium oxide blocking oxide. Results are shown for 9,10,11, and 12 volts . . . . . . . . 173

4.17 Erase curves on a gridded capacitor with the aluminum oxide blocking oxide. Results are shown for -9,-10,-11,-12,-13, and -14 volts. . . . . . . . . 173

4.18 Erase curves on a gridded capacitor with the aluminum/hafnium oxide blocking oxide. Results are shown for -9,-10,-11, and -12 volts . . . . 174

4.19 Flatband shift for aluminum oxide gridded capacitors on the write state. 176

4.20 Flatband shift for hafnium/aluminum oxide gridded capacitors on the write state. ................................................................. 177

4.21 Flatband shift for aluminum oxide gridded capacitors on the erase state. 177

4.22 Flatband shift for hafnium/aluminum oxide gridded capacitors on the erase state. ................................................................. 178

4.23 Write curve and simulations on a Northrop Grumman charge trap non-volatile memory. It is important to know the thicknesses of the gate stack for accurate simulations since many gate stack combinations can produce the same curve. To speed up the device, the nitride layer can be reduce. Reducing the voltage will slow the device down. . . . . . . . 186

4.24 Erase curve and simulations on a Northrop Grumman charge trap non-volatile memory. The original program did not include electron back tunneling. Once it is added and exact match is shown. . . . . . . . . 187

4.25 Experimental and simulation write and erase curves performed at 8V on a gridded capacitor with an aluminum oxide blocking oxide. The gate stack is 2.4/7.7/15nm silicon oxide/silicon nitride/aluminum oxide. The simulation curves are an excellent match to the experimental curves. ................................................................. 189
4.26 Experimental and simulation erase curves performed at 7 and 8V on a gridded capacitor with an aluminum oxide blocking oxide. The gate stack is 2.4/7.7/15nm silicon oxide/silicon nitride/aluminum oxide. Curves are included for the 7V measurement with different threshold voltages. The original measurement of 0.2V was incorrect given poor agreement between the curves. With a 1V starting threshold voltage there is excellent agreement between the curves. 190

4.27 Experimental and simulation write curves performed at 7, 8, and 9V on a gridded capacitor with an aluminum oxide blocking oxide. The gate stack is 2.4/7.7/15nm silicon oxide/silicon nitride/aluminum oxide. The 9V curve is slightly off showing an issue with the starting threshold voltage value. 191

5.1 Diagram of Donor and Acceptor traps. A donor trap has a positive state which switches to a neutral state with the addition of an electron or the release of a hole. A acceptor trap has a neutral state which switches to a negative state with the addition of an electron or the release of a hole. 194

5.2 The amphoteric trap model has only one trap. This trap has three states, positive, neutral, or negative. 195

5.3 Band diagram depicting the different trap levels. Acceptor traps lie in the upper portion of the energy gap, while donor traps lie in the lower portion of the energy gap. 195

5.4 Energy band diagram of a MONOS charge trap memory at zero bias with electrons in the trapping material. Electrons can tunnel out from traps into the conduction band of the substrate or can rearrange themselves in the nitride through Frenkle-Poole emission. Holes can back tunnel into the nitride material from the valence band of the substrate. 198

5.5 Gate stack diagram and TEM of the device under test. We used a 2.4/7.7/15nm MANOS device. 200

5.6 Retention curves on a gridded capacitor with aluminum oxide blocking layer. The erase state shows excellent retention characteristics. The positive state starts to lose charge after a second. This is due to the poor silicon nitride process producing traps near the interface. The thickness of the device is 2.4/7.7/15nm. 201

xxv
6.1 Cross section of the III-V charge trap non volatile memory stack. Aluminum oxide has been used for both the blocking and tunneling oxide. 203

6.2 Cross section of the III-V transistor stack. An aluminium/hafnium oxide layer has been used as the insulating later. 203

6.3 SRIM ion distribution analysis on a GaAs active layer implanted with carbon at 50keV. 205

6.4 Ion doses used to implant InGaAs and GaAs wafers. 206

6.5 Silicon oxide sacrificial protective layer and the silicon nitride masking layer used while implantation is completed. 207

6.6 An SF6 plasma is used to create the grid lines in the silicon nitride masking layer. 208

6.7 Ion implantation is completed through the silicon oxide protective layer. 209

6.8 The final wafer with an implanted grid depicted in orange. 210

6.9 Halfnium ALD measurements. The Picosun system deposited 0.44 Angstroms per cycle. 211

6.10 An image of the Picosun ALD machine while open at The Ohio State University Nanotech Lab. 213

6.11 An image of the metal deposition machine at The Ohio State University Nanotech Lab. 214

6.12 Final gate stacks cross section on the III-V transistor gate stacks. 215

6.13 Final gate stacks cross section on the III-V charge trap non volatile memory stack. 216

6.14 C-V measurement on a 50um capacitor radius and 30um grid spacing at 1MHz and 100kHz. A slight C-V curve is seen. 219

xxvi
6.15 C-V measurement were scanned from high to low (blue) and low to high (red). There is a distinct shift in the curve which leads to voltage shift of 3.2V.
Chapter 1: Introduction to Non-Volatile Semiconductor Memories

Semiconductor memory has become a staple of modern electronic systems. It can be found in nearly every electronic device on sale today. Computers to automobiles and everything in-between contain at least some form of semiconductor memory. There are two main categories of semiconductor memory: volatile and nonvolatile. Volatile memory need to be constantly refreshed. The information will be lost if power to the circuit is lost. Memory components such as Dynamic Random Access Memories (DRAMs) are examples of such volatile memory systems. A basic DRAM cell has an access transistor with an attached capacitor. The capacitor stores the charge/information. However, capacitors are leaky and therefore the memory needs to be refreshed, typically every 50ms. Nonvolatile memory systems can hold information for much longer periods of time. The industry standard for information storage is 10 years.

The demand for electronic devices has greatly increased over the last few years. This increase, in turn, has led to a greater demand for nonvolatile memory storage. Nonvolatile memory has a long list of applications. It may be used as data storage in small, hand-held flash drives or cell phones, as hard drives in laptops, netbooks, or tables, or as memory storage on military and commercial satellites. As the list of
applications continues to grow, so does the demand. In Figure 1.1 the NAND flash market is shown[34]. In 2014, over 40% of the mark is tied to solid state drives and mobile phones.

Figure 1.1: NAND flash memory market for the years 2007 to 2014. There is a 600% increase in the amount of memory over the last 4 years.[34]

The great demand for nonvolatile memory has pushed the industry to create new memory structures with low-power, high density, and long retention to meet the electronic industry needs. Nonvolatile memory comes in many different shapes and forms which will be discussed in this chapter. We start with a discussion of the most prolific memory, the charge trap memory. These memory devices come in two major flavors, the floating gate, and the charge trap variety. Next, we look into the other major memory types such as Ferroelectric (FeRAM), Magnetic Tunnel Junction (MJT), and
Phase Change (PC) memories. Then we examine new experimental memory devices which are still under investigation.

1.1 Non Volatile Semiconductor Memories

The floating gate structure is the most abundant non-volatile semiconductor memory device available on the market to date. It also happens to be the first non-volatile memory device which was originally proposed by Kahng and Sze in 1967[40].

1.1.1 Floating Gate

The structure of a floating gate non-volatile memory device contains a poly-silicon gate that is isolated from the capacitor coupled control gate and the silicon substrate by oxide barriers. Since the poly-silicon gate has no external connection it is referred to as floating, hence the name for these devices. The oxide between the floating gate and the substrate is called the tunneling oxide and is usually composed of silicon oxide (SiO$_2$). The oxide between the floating gate and the control gate is called the blocking oxide. Today's modern floating gate structures implement a triple dielectric blocking oxide consisting of silicon oxide, silicon nitride, and silicon oxide (ONO) stack shown in Figure 1.2.
Figure 1.2: A cross section view of a typical floating gate non-volatile memory structure. The control gate is separated from the floating gate by an ONO stack.

The multi-dielectric blocking oxide configuration allows an improved capacitive couple ratio between the control gate and the floating gate. The increased capacitive coupling ratio facilitates the use of lower programming and erase voltages as well as allows for continued scaling of the device.

Being electrically isolated, the floating gate acts as a potential well for charge. Floating gate devices store charge in the poly-silicon floating gate as free carriers with a continuous spatial distribution in the conduction band. The configuration has a serious drawback. Since the carriers are free in the conduction band any damage
Figure 1.3: Tunneling Electron Microscope (TEM) image of a floating gate structure. The Control Gate (CG) is separated from the Floating Gate (FG) by a narrow ONO stack. Tungsten silicon(WSi) is used as a contact material.[37]

which allows a path for charge to leave the material will drain all the charge from the floating gate and lose (discharge) the data which is stored.

Charge, which is injected into the floating gate, is maintained due to its isolation, and also allows for the modulation of the apparent threshold voltage as seen from the control gate. When electrons are stored in the floating gate, the device is in a negatively charged state. This is associated with the logical 0, while the neutral or positive state is associated with the logical 1 state.

A floating gate non-volatile memory device stores information by holding charge that is injected from the substrate into the floating gate. Programming of the device is accomplished through a process known as channel hot electron (CHE) injection.
CHE injection occurs when electrons are accelerated to a high enough energy level to surmount the Si-SiO\textsubscript{2} energy barrier, which is roughly 3.2eV by a lateral field in the channel. During CHE injection, the electrons in the channel must be accelerated to an energy level in excess of the barrier height, which is accomplished by a large positive bias being applied to the drain of the device. Next, a scattering event must take place for the electrons to be redirected normally to the Si-SiO\textsubscript{2} surface. Finally, the electrons must reach the oxide interface without its energy dropping below the barrier height of the Si-SiO\textsubscript{2} interface. The few electrons that meet these criteria and make it into the floating gate are referred to as lucky electrons, and are responsible for the shift of the threshold voltage in the device. Figure 1.4 (a) below shows the write process in a floating gate non-volatile memory device.

Figure 1.4: Diagram of floating gate transistor (a) Write operation by CHE injection over the drain, (b) Erase by FN tunneling over the source
Shown in Figure 1.4(b) is the erase procedure of the floating gate non-volatile memory device which employs Fowler-Nordheim (FN) tunneling. FN tunneling was first described by Fowler and Nordheim in 1928[18], and was eventually named after them. In 1969, Lenzlinger and Snow observed the same phenomenon for the Si-SiO$_2$ system[52]. Under a large electric field, the energy bands of the tunnel oxide become distorted. When this occurs there is a small but finite probability that an electron in the conduction of the floating gate will tunnel through the SiO$_2$ tunneling oxide and appear in the silicon substrate. This process increases exponentially with an increase in applied electric field. The process becomes readily observable when the field across the tunnel oxide is on the order of 10MV/cm. For the floating gate device pictures in Figure 1.4b, a negative voltage is applied on the control gate while a positive voltage is applied on the source of the device. The potential difference between the control gate and the source provides the necessary voltage across the tunnel oxide for FN tunneling to take place between the floating gate and the source of the device. It should be noted that a floating gate device uses electrons to both write (CHE injection to the floating gate) and erase (FN tunneling from the floating gate to the source) the device.

Year after year, semiconductor memories are scaled down to smaller dimensions to achieve higher storage density. Floating gate devices greatest challenge is to maintain a gate capacitive coupling ratio, which is typically around 0.6, with a reduction of the tunnel oxide thickness while minimizing word line disturbance [99, 7]. A coupling ratio of 0.6 means control voltages need to be in the 18-20 volt range to establish appropriate electric fields across a 8-9nm tunnel oxides. Even with these challenges industry has been able to reduce the NAND floating gate memory down to 22nm[17]
1.1.2 SONOS and Derivatives

The leading alternative to the floating gate non-volatile memory device is the Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) charge trapping non-volatile memory. Charge trapping NVSMs were first introduced in 1967, which happens to be the same year the floating gate device was introduced. The device was introduced by Wegener et al. and they consisted of a Metal-Nitride-Oxide-Silicon (MNOS) structure [104]. Figure 1.5 traces the development of charge trapping NVSMs over the last half century.

The first charge trapping devices employed a metal (aluminum) gate and an n-type substrate. Shortly after their introduction the technology shifted to a p-type substrate and a poly-silicon gate. Both of these early types of devices used large silicon nitride trapping regions and so they did not require a blocking oxide.
Around the early 1980's the SONOS (Silicon-Oxide-Nitride-Oxide-Silicon) NVSM was introduced. The devices used a blocking oxide between the poly gate and nitride charge trapping layer. This design allowed for a great decrease in the nitride thickness as well as great improvement in retention. In subsequent years, much research and development took place to perfect the SONOS structure [53, 9, 54, 88, 19].

In 1998, Boaz Eitan developed and patented multiple bit storage [16]. These devices had a thicker nitride layer to store a larger amount of charge to aid in storage. A larger tunnel oxide was also introduced to prevent leakage of charge out of the nitride layer. The device was given the name NROM (Nitride-based Read Only Memory). The larger tunnel oxide made it more difficult to write and erase the device, but prevented charge loss for long memory storage. By 2005, the modern charge trapping NVSM had been developed. These devices included the use of a high-k dielectric as a blocking oxide, in addition to the reintroduction of metal gates.

The structure of charge trapping NVSM is very similar topologically to the floating gate structure previously discussed, but with one large difference. In a charge trapping device, the poly-silicon floating gate is replaced by a charge trapping material. This material stores charge in spatially isolated deep level traps. The setup has a great advantage over floating gate devices; if the device is damaged in such a way that there is a path for charge to escape the material, only charge local to that defect will leak out. The rest of the memory remains intact and the information stored would be retained. The main charge trap material used today is silicon nitride with a thickness of 4-8nm.

There have been quite a few studies of the nitride material [35], its traps [45], and ways to improve performance and reliability [84], [109]. Recently, many new charge
Figure 1.6: TEM of a charge trapping non-volatile memory. Aluminum oxide is used as the blocking oxide, silicon nitride as the charge trap material, and silicon oxide as the tunneling oxide.[49]

trapping materials have been investigated to see if they can be used in the charge trapping non-volatile memories. These materials include: \( \text{Y}_2\text{O}_3 \) [69], [70], various Hf compositions [94, 95, 114], \( \text{Ta}_2\text{O}_5 \) [101], [102], \( \text{La}_2\text{O}_3 \) [43], \( \text{DyO}_2 \) [30], \( \text{Yb}_2\text{O}_3 \) [68], \( \text{ZrO}_2 \) [115]. Just as with the floating gate device, the material which holds the charge is isolated from the gate and substrate by oxides.

The oxide between the nitride layer and the substrate of the device is called the tunnel oxide, since it is the layer which electrons and holes tunnel through to get to the nitride layer. The tunnel oxide is usually composed of silicon oxide (SiO\(_2\)) and has a thickness around 2-4nm.
The oxide between the charge storage layer and the gate electrode is called the blocking oxide. This layer prevents holes and electrons from tunneling to or from the charge storage material. Due to this tunneling restriction, the blocking oxide must be large (usually greater than 5nm) and possess a sufficient barrier height to both electrons and holes. In the past, the preferred material was SiO$_2$, but a shift has begun to use high-K dielectrics as the blocking oxide for reasons which will be discussed in the later section.

The tunnel oxide, charge storage layer, and the blocking oxide are commonly referred to as the gate stack, as they are the layers which reside directly underneath the gate. Modern devices use poly-silicon, or a metal such as TaN [101] or Al as the gate electrode. A full charge trapping NVSM implementing a Poly Si/Al$_2$O$_3$/SiN/Si$_2$ structure is shown via Transmission Electron Microscopy (TEM) in Figure 1.6.

Charge trap devices use tunneling mechanisms for both write and erase operations. Both of these operations are visible in Figure 1.7. During the write operation, a positive bias is applied to the gate while the source and drain is grounded. Electrons accumulate at the Si-SiO$_2$ surface and when a large enough field appears across the tunnel oxide, the electron tunnel into the charge trapping material via FN tunneling and fall into the deep level traps. This operation is shown in Figure 1.7a.
To erase the device, a negative bias is applied to the gate while the source and drain are grounded. The applied bias allows holes to be present at the Si-SiO$_2$ interface and can tunnel through the oxide and neutralize the electrons in the charge storage material. It should be noted that the electrons do not tunnel out of the charge trapping material at low level fields as the electrons are located in deep level traps. Since this is not practical, the holes are needed to perform the erase operation. At higher voltages electron back-tunneling out of the nitride is possible. It should be noted that the barrier height of SiO$_2$ is larger for holes (4.1 eV) than for electrons (3.2eV). For holes to be able to tunnel through the oxide barrier, the thickness of the oxide must be small. As a result, devices employ a tunnel oxide thickness of 2nm. Such a thin tunnel oxide, however, leads to poor data retention due to direct tunneling can occur from the substrate when the device is off. To mitigate this problem, groups

Figure 1.7: (a) Write operation for a charge trapping NVSM (b) Erase operation for a charge trapping NVSM
have used various techniques to increase the tunnel oxide while still allowing enough tunnel current to take place during write/erase operations. These techniques include: replacing the silicon oxide with silicon nitride [89, 91], or adding in a high-K dielectric with smaller SiO\textsubscript{2} layer [10, 100].

### 1.2 Ferromagnetic (FeRAM)

The ferromagnetic memory device has celebrated 40 years since its introduction. The first mention of using ferroelectrics as a memory device took place in 1974 by Shu-Yau Wu[110]. Wu described a simple memory device where a ferroelectric Bi\textsubscript{4}Y\textsubscript{13}O\textsubscript{2} was deposited between the channel and a gate of the MOSFET. The first design is shown in Figure 1.8

FeRAM works through a change in polarization. A hysteresis loop shows this polarization in Figure 1.9. The relation between the polarization of the ferroelectric film and the applied electric field is given by the P-E curve depicted. Ps is the saturation polarization while Pr is the remnant polarization, and Ex is the coercive
Figure 1.9: Schematic hysteresis loop for a polycrystalline Fe Film. The higher polarization state is a binary ”1” while the lower state is a binary ”0”. [29]

field. To program these materials an electric field is applied to the device in a certain direction. Once the applied field is removed the material is still polarized at the point Pr. When you polarize in one direction, one could say you have a binary 1 while the other direction would be a binary 0. The coercive field $E_c$ depicts how large of an electric field is needed to switch the polarization state.

The majority of FeRAMs today are used in the 1T1C configuration. This configuration contains one transistor, the access transistor, and a ferromagnetic capacitor. A cross section of such a device is shown in Figure 1.10.
Figure 1.10: Schematic of a 1T1C Ferromagnetic Memory Cell. The access transistor is on the left side of the image with the source and drain in red. The ferroelectric film is in blue with the dark blue line as the connector between the transistor and the capacitor. [27]
The device shown uses a Pb(Zr,Ti)O$_3$ better known as PZT ferroelectric material. To program the device the word line and bit line are pulled high, while the plug is grounded. This configuration is such that a high voltage is present on the bottom electrode while the top electrode is grounded. The voltage configuration allows a binary 1 to be written. To write a 0 the word line and plug are pulled high while the bit is grounded. The "0" state is such that a high voltage is present on the top electrode while the bottom electrode is grounded. Therefore each polarization state is selected by the electric field direction.

To read the device, the word line and plug are selected as high, while the bit line is floating. The amount of current drawn from the capacitor is then sense to see if the state stored is a 1 or a 0. A severe drawback of ferroelectric capacitors is that the read process is destructive. Since the voltage on the bit line is floating an unknown electric field is placed across the ferroelectric material which can alter its polarization state. For this reason the memory needs to be refreshed after a read is performed. Another issue is scalability. As the dimensions of the capacitor are reduced the polarization effects is also reduced. It has been suggested that scaling is the limiting factor and will limit arrays to 128Mb[90]. Despite these problems 1Mb ferroelectric random access memories have been created [75].

To combat the destructive read, ferroelectric materials have been incorporated into field effect transistors in the gate region as shown in Fig 1.11.
When a positive voltage is applied to the gate, polarization of the ferroelectric is directed upward. This causes an inversion layer to form. With carriers at the surface, a connection is made between the source and drain and with a small applied voltage the transistor is on and a current flows. This configuration is known as the binary 1 state. When a negative pulse is applied the field is removed. The inversion layer is no longer present and when a small voltage is applied no current flows and the transistor remains off. With this configuration the device can be read many
times with no alteration to the polarization field in the ferroelectric. However, other problems remain with ferroelectrics.

A main concern with ferroelectrics and the describe configuration is the loss of polarization over time. Once the polarization field is loss, so is the stored information. The loss of polarization is on the order of days[29]. Two main depolarization sources have been identified.

The first depolarization source is from the substrate. Whenever a ferroelectric film is placed on a semiconductor, a depolarization field is present which opposes the remnant polarization [58]. This effect is reduced by having a metal-ferroelectric-metal configuration. However, such a device cannot be integrated into todays semiconductor manufacturing process. The second retention lose is due to charge trapping at the interface as well as trapped charges in the insulator. The excess charge and traps at the interface and dielectric can destabilize and reduce the ferroelectric polarization. There has been a growing investigation of the material systems that could be used with ferroelectrics to overcome the issues raised here [36, 71, 78]. Many believe these issues cannot be solved and that FeRAMs will be limited to DRAM applications.

1.3 Magnetic Tunnel Junction (MJT)

Magnetic Tunnel Junction (MJT) memories consist of a soft magnet, or free layer, which stores the information, a tunneling layer, and a pinned reference magnet. The pinned layer remains in the same magnetic state permanently while the free layer is able to switch magnetic states. The two states are known as the parallel and anti-parallel state which are shown in Figure 1.12. A large majority of MJTs use magnesium oxide as the tunneling layer.
Figure 1.12: Magnetic Tunnel Junction (MJT) states a) Parallel State with both polarizations in the same direction b) Anti-parallel state with the polarization in opposite directions. [2]

MJTs are currently being investigated due to their extremely low switching power. The magnets undergo a process which is characterized as collective switch behavior. In collective switching the whole magnet behaves as if it has one large spin [83]. This means that the switching behavior is not dependent on the number of electrons in the magnet. The magnet switch spins at the same energy level regardless of the size of the magnet or how many electrons are present. This is radically different than conventional charge base memories. In charge based memories the switching behavior from a 1 to a 0 state is dependent on the number of electrons that are present. Hence conventional charge based memories require more energy to switch between states than MJTs.

Magnetic tunnel junctions are resistive based memories. The difference between a 1 state and a 0 state is deduced by whether the current through the MJT is high or low. The resistance is dependent on the relative orientation of the magnetization.
in the free and fixed layer. The resistance will be altered depending on whether the states are parallel or anti-parallel. This resistance change is referred to as the tunnel magneto-resistance (TMR) ratio. The ratio is taken between the high resistive state and the low resistive state. A higher TMR ratio is required to distinguish between a 1 and the 0 state. The TMR effect was first discovered by Julliere in 1975[39].

There are two basic flavors of magnetic tunnel junction memories. The first one is known as the field switch MJT while the second is the Spin Torque Transfer (STT). Both technologies use the same configuration as shown in Figure 1.12, however the difference between the two devices lies in how they switch the free layer.

1.3.1 Field Switch

In field switching MJTs a current is passed close to the free layer. The current in the wire has an associated magnetic field which exerts itself on the magnetic material in the free layer. A schematic of this step is shown in Figure 1.13.
With the field switch setup shown in Figure 1.13 two currents are required, one along the hard axis and the other along the easy axis. These two currents are perpendicular. The axes can be thought of as an extra set of word and bit lines to select each cell. First a current and magnetic field is generated along the hard access. Then the easy axis line has a current flow in either direction. The direction of the current determines the generated magnetic field and thus how the free layer is switched into a parallel or anti-parallel state.

The field switch MJT is no longer a major player in memory development. This is due to the extra lines that are required to program the cell. Along with space and scaling concerns, the extra axis lines waste a lot of energy with the dual current
flow. A majority of these issues have been mitigated with the development of the Spin Torque Transfer (STT) MRAM.

### 1.3.2 Spin Torque Transfer

Instead of using a magnetic field to switch the spin, the magnets can be switched directly with a current through the device. This phenomenon is known as spin torque transfer (STT) which was predicted by Slonczewski [92] and Berger [3]. For spin torque transfer memories the extra lines needed to switch the magnetic field are eliminated. The current is passed directly through the MJT. This is best shown in Figure 1.14.
Figure 1.14: Programming of an STT MRAM a) Anti-parallel to parallel. Electrons flow from bottom to top. Electrons with the same spin pass through the pinned layer and interact with the free layer to switch states. b) Parallel to Anti-Parallel. Electrons flow from top to bottom. Electrons with the same spin of the free and pinned layer pass through the device. Opposite spin electrons are blocked by the pinned layer and start to build up. After a critical point the electrons of opposite spin exert a force on the free layer to switch polarization states. [42]

The STT MRAM works through an interaction between magnetization and a spin polarized current, which is attributed to angular momentum exchange between the spins of local magnetic moment and free electrons passing through the MJT [2]. To switch the STT MRAM to a parallel state from an anti-parallel state, as shown in Figure 1.14b, the current must pass through the pinned layer first. In essence the pinned layer only allows electrons with the same spin as the layer to pass, and flow through to the free layer. Once a critical current density through the free layer is reached the whole free layer magnetization switches to be the same as the pinned layer. To go from a parallel state to an anti-parallel state the current must flow through the free layer first. In this configuration, only electrons with the same spin
Figure 1.15: TEM sample of a MJT. Magnesium Oxide is used as a barrier layer between the two CoFeB layers. The advantage of such devices is that extremely small scale devices are available. [42]

as the pinned layer flow through the device. The other electrons of opposite spin are reflected at the barrier. This current of opposite spin electrons exerts spin torque transfer on the magnetization of the free layer. Again, when the current density reaches a certain threshold the whole layer will switch to the anti-parallel state.

Spin torque transfer memories have been demonstrated down to a feature size of 65nm [17], however they still have a variety of challenges to overcome. One of the main challenges of MJTs is the deposition of the magnetic and barrier layers. Figure 1.15 shows a TEM sample of the MJT layers.

As seen in Figure 1.15 the deposition of the MJT layers are uneven and can vary over large and small areas. This is a major challenge in the scaling of devices as
well as for large scale manufacturing of the devices. Without improvements to the processing systems, reducing the feature size to the 4F2 theoretical limits remains out of reach. The yield of such devices would be greatly reduced due to the lack of adequate processing of the materials over a large area. Another major issues that MJTs need to overcome to become more prevalent is read disturb of the individual cells. To read a STT cell the current which runs through an access transistor as well as the STT cell is sensed. The level of resistance dictates whether the cell is in a 1 or 0 state. However, one reads the cell the same way one writes a cell. Therefore there are limits to the amount of current one can pass through the MJT to read the state of the cell. An issue arises when one wants to scale the memory up in size. The current needs to be large enough to be detected by a sense amplifier. The low read currents can be hard to measure in a large array, and increasing the read current is not an option as it may program a cell.

1.4 Phase Change (PCRAM)

One of the most compelling alternatives to conventional charge trap memories is the phase change (PC) memory. Phase change memory is a resistive base memory system. The 1 and 0 states are defined through high and low resistance values when a current passes through the material. Phase change devices were first proposed by S. Ovshinsky, who in the 1960s reported the observation of a reversible memory switching in chalcogenide materials [67]. Chalcogenides are semi-conducting glasses made by elements of the VI group such as sulfide, selenium, and tellurium. Chalcogenides were first used in rewritable optical media due to their ability to undergo quick crystalline-amorphous phase changes [111, 46]. Dewald first proposed their use
Figure 1.16: Phase change memory a) Schematic showing the two electrodes, heater, and phase change material. A thin conductor heats the phase change material. b) Programming a PCRAM. A RESET pulse heats and cools the material quickly causing an amorphous region, while the SET pulse allows the material to form a crystalline structure. A read pulse does not change the material structure.[8]

as a memory device in 1962[66]. Today they are considered one of the most promising candidates to become the next mainstream nonvolatile memory device, due to their large cycling endurance [62, 73], fast program and access times as well as scalability [44, 72].

A basic phase change memory cell is shown in Figure 1.16. The configuration is known as a 1T1R cell design where there is one access transistor and a resistor. The resistor is the chalcogenide layer (using Ge$_2$Sb$_2$Te$_5$ or GST) which is placed between two electrodes. One of the electrodes has a heating element. This is a narrow conducting region that leads to the phase change material. A current is forced through the device which first goes through the narrow heater region. The narrow
conductor heats the phase change material which then changes from a crystalline or amorphous state to a malleable state. Depending on the current pulse and cooling the material can be reset in a crystalline or amorphous state. The crystalline state has a low resistance while the amorphous state has a higher resistance. Programming of a phase change memory device is accomplished by driving a 50-100 ns current pulse through the cell [47]. The amorphous state is obtained by driving the device with a current pulse for 100 ns. This is known as the RESET pulse as shown in Figure 1.16. The pulse heats the phase change material to temperatures in excess of 620 degrees Celsius, which is then quickly cooled down. The quick cooling of the GST material doesn’t allow the atoms to rearrange bonds to set up in a crystalline structure and the material remains amorphous. To obtain the SET condition and return the material to a crystalline state, a lower current pulse is used over a longer time. The current pulse is established to heat the material up to approximately 550 degrees Celsius. The lower temperature, which is held longer, allows for the GST material to rearrange its bonds into a crystalline state. To read the resistance of the cell a low current is passed through the device. The current is low enough, around 100uA, that it in no way alters the state of the phase change material. The transition can take place repeatable for many cycles with little to no degradation of the device. Figure 1.17 shows the endurance for a phase change memory.

The memory shows little degradation over $10^8$ cycles, and newer work shows endurance levels much higher than shown here [47]. Since the phase change is done with localized heating of the material there is little chance of read disturb or retention loss. One of the main issues with phase change memories is heat cross talk. As scaling of devices continues to lower and lower dimensions, the heating elements get closer and
Figure 1.17: PCRAM Endurance. The on and off resistance values remain constant over 10E8 cycles. After 10E8 cycles the values start to defer but the difference between on and off states are still very far apart. [47]
closer together. Heating one element can propagate heat to a nearby element and alter the state of adjacent cells. In a large memory that is used frequently heat dissipation could become a large issue. Despite the issues mentioned here, manufactures have produced large arrays at small dimensions.

In 2009 G. Servalli from Numonyx of Milan Italy debuted a 1Gbit PCRAM at the 45nm node \[87\]. This memory ran at a supply voltage of only 1.8V. However the effective cell size was 5.5F2. In 2011, S.H. Lee et al \[50\], showed a 1Gbit PCRAM that had an effective cell size of 4F2. This memory is shown in Figure 1.18. Even more amazing, M.J. Kang et al \[41\] debuted an 8Gb phase change memory at the 20nm node. The memory features are shown in Table 1.19.

Development on PCRAMs continue today. Current research initiatives have allowed some phase change memory systems to surpass conventional nonvolatile memory systems in many design areas. As shown by the latest work much of the thermal cross talk problems have been solved or mitigated, however, it remains unseen how the memory performs under large batch updates to many cells over a short period of time. PCRAMs remain the most promising alternative to todays conventional charge trap memories.

1.5 Experiment Memory Devices

In the previous section the most developed and well research memory systems were discussed. However, many new memory devices are under current investigation. A few of these memory systems include mechanical switches, Mott, and Redox memories. All previously discussed memories use inorganic materials that are aligned well with
Figure 1.18: Fully Integrated 1Gb PCRAM Chip. The chip consists of individual partitions of 64Mb strung together to form a 1Gb chip. [50]
current semiconductor processes, however there is a growing body of research into organic semiconductors.

1.5.1 Nano-Electro-Mechanical Memory (NEMM)

Nano-Electro-Mechanical Memory is any memory system that employs moving mechanical parts. These devices are mostly used as switches. They are under investigation due to their low power consumption and nanosecond switching speeds [51]. However, their best feature may be their abbreviated name NEMmory.

There are many different types of mechanical switches being investigated. One of these is the cantilever design proposed by Sang Wook Lee et al.[51] shown in Figure 1.20.
This design is similar to floating gate devices of today. The device uses a floating gate to store charge and hence information. A source and drain is used to read the amount of stored charge in the floating gate. However, this design differs in how the charge is placed in the floating gate. To program or erase the device an actuating electrode is activated to pull the cantilever down into contact with the floating gate. The cantilever has a supplied voltage to erase or program the floating gate. Since the cantilever comes into contact with the floating gate, charge transfer and hence switching is almost instantaneous.

The main advantage of such a device is its switching speed. This particle design was able to switch in around 130ns. The memory has several disadvantages. First and foremost is the endurance of such a device is extremely poor. The cantilever only lasted for approximately 500 cycles. Also, the addition of a cantilever takes up space and limits scaling. Other groups have looked into designs to reduce the amount of space used. One of these designs is the vertical NEMM.
E.J. Ng, et al [65] has proposed taking the horizontal cantilever switch and flipping it upright. This configuration is shown in Figure 1.21.

The switch is a vertical pylon between the source and drain. The fabrication process consists of all standard CMOS processes. The device characteristics, however, are poor. As shown in Figure 1.21b, when a voltage is applied between the two electrodes, it cause the pillar to deflect to one side. The voltage required to switch the device is 17V. Once the voltage is removed the pillar stays attached to the side due to van der Waals forces.

NEMM memory systems are intriguing, yet they are extremely far from conventional nonvolatile memory devices. Having mechanical parts in any system is prone
to mechanical failure of the device. It is also hard to see how well such systems would scale down to current technology limits.

### 1.5.2 Mott Memory

Mott memory is a new definition of memory recently added by the ITRS [17]. It is a resistive based memory device. In the Mott memory device, charge injection induces a transition from strongly correlated to weakly correlated electrons which result in an insulator-metal transition (IMT) or Mott transition [17]. These memories are also sometimes referred to as correlated electron random access memories (CeRAMs). They are called Mott memories because of the insulator to metal transistor described by Mott and later Hubbard [63, 33]
In 2011, C.R. McWilliams et. al. created a CeRAM resistive memory based on a Mott-like charge transfer metal-insulator transition in NiO[107]. The memory was a simple capacitor on silicon substrates. The silicon substrate had 500nm of thermally grown oxide, a Pt bottom electrode, NiO layer, and a Pt top electrode. The test structures are shown in Figure 1.23.

The switching mechanism is still being explored, but it is argued that the mechanism is activated by a critical electron population described by the Mott-Hubbard model [67,68]. To program a device, a large voltage is applied till the current saturates. This changes the material into a metal state. To reset the device the opposite voltage is applied to the capacitor. This can be repeated to go back and forth between a high resistance and low resistance. The endurance of such a device is shown in Figure 1.24.
Figure 1.24: Resistance values for ON and OFF state over 100 cycles for a CeRAM structure. There is a large variation of on off resistances that are separated by a factor of 10 resistance.[107]
A critical issue for Mott memory is the sensitivity to changes in parameters. A small change in parameters such as charge density, strain, disorder, and local chemical composition drastically changes the ON and OFF resistances. This prevents the formation of large memory arrays. However, this is a very new memory device with a lot of potential growth.

1.5.3 Redox

Redox based memories are also resistance memories. A change in resistance of a metal-insulator-metal structure is caused by ion (cation or anion) migration combined with redox processes involving the electrode material or the insulator material, or both[17]. Redox memories are broken into three categories. The first is electrochemical mettallization (ECM). This memory is shown in Figure 1.25.

For this memory device Ag and Pt electrodes are separated by an insulator. As the voltage is increased Ag ions propagate through the insulator and form on the Pt electrode. The Ag ions keep collecting on the far electrode and start to form Ag dendrites. Soon, the Ag dendrites reach back to the Ag electrode and a low resistance path is created between the two electrodes. To reverse the effect and opposite voltage is applied to the electrodes. Currently the switching capabilities are slow as the dendrites have to physically move atoms to form a physical connection between the electrodes. Endurance issues are a critical setback for these memories as the physical movement of the atoms degrades the devices performance.

The second type of redox memory is called the valence change mechanism (VCM). This memory effect results in a change of stoichiometry on the insulator due to oxygen migration. The effect of migration is controlled through voltage pulses where the
Figure 1.25: Program/Erase cycle of an electrochemical metalization (ECM) memory. 
A) Undisturbed state with no dendrites. B) Ag⁺ diffusion across the insulator toward the Pt electrode. C) Ag⁺ dendrite forms. D) Dendrite connection between contacts. E) Ag⁺ moves away from Pt electrode. [98]
polarity of the voltage dictates the direction of change. The third type of Redox is called the thermo-chemical mechanism (TCM). This memory is similar to VCM however the stoichiometric change is caused by a temperature change. Redox memories are still very new and much about how they operate is not known. Still, recent results are encouraging [103].

1.5.4 Organic Memory

Organic memory refers to any polymer or organic resistive memory. The memory element of such devices consists of an organic material sandwiched between two metal electrodes. The organic film is relatively thick, and the operational voltages are high usually greater than 20 volts. These first organic memory devices are now referred to as macromolecular memory. To reduce the size and operating voltages, the move to molecular memory has increased. A molecular memory is shown in Figure 1.26.
Figure 1.27: Measured logic diagram of a molecular memory device. A write pulse first sets the device and subsequent pulses have a corresponding high output. When an erase pulse is applied the next pulse shows no output, at which point it is rewritten and has an output with additional pulses. [77]

In this particular molecular memory 2-amino-4-ethynylphenyl-4-ethynylphenyl-5-nitro-1-benzenethiolate is strung between two Ag electrodes. By applying a high voltage to one of the electrodes, the conductance through the material becomes high. When a negative voltage is placed on the electrode, the conductance becomes low.

By using only a molecule between the electrodes the vertical size of the memory is reduced. Keep in mind that many strands of the organic material are strung between the electrodes, not just one as is picture in Figure 1.26.

Figure 1.27 shows the measured logic diagram picture above. Since the storage mechanism of the molecular memory is through conductance the output was dropped across a resistor which was sent to a comparator, inverted, and gated with the read
pulse. The upper trace shows the input wave form to the memory cell. The positive pulses are write pulses while the negative ones are erases. The output is shown on the lower trace. The first pulse writes the cell, while the two following pulses read the cell. This diagram shows the cell can be written and rewritten. The group has stated this can be extended to the hour timescale with little degradation [77]. Even with these advancements organic memory has many issues.

A major problem is retention time in organic devices. In the device shown in Figure 1.26, the retention time was on the scale of approximately 15 minutes. This is hardly a nonvolatile memory. However, much research is taking place to investigate the mechanism behind the conductance change. In the future, it is likely there will be organic memory devices that can hold information for much longer periods of time.

1.6 Memory Comparisons

We have discussed many different memory types; however how they compete against each other is difficult to see. Table 1.1 shows a comparison of all major memory technologies.

These technologies include SRAM, DRAM, NAND, FeRAM, MJT, and PCM. The other memory types discussed in this paper have not be included due to their under development status. The table shows the latest results as of 2011. As can be seen there is a lot of differences between each of the memories. Scaling takes place very well in NAND and PC memories, yet poorly in FeRAM. Obviously, DRAM and SRAM have no retention times to speak about. All the write voltages are very low (1-2 volts), however NAND write voltages are much higher at 15V. This is due to the tunnel mechanism discussed in section 2. Each memory device currently has its
Table 1.1: A comparison of major memory systems. A variety of specifications is provided on the best devices which have provided data. Baseline technologies are listed as DRAM, SRAM and NAND with prototype technologies listed as FeRAM, MRAM/STT, and PCM

<table>
<thead>
<tr>
<th></th>
<th>Baseline Technologies</th>
<th>Prototype Tech.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DRAM</td>
<td>SRAM</td>
</tr>
<tr>
<td>Feature Size</td>
<td>36nm</td>
<td>45nm</td>
</tr>
<tr>
<td>Cell Area</td>
<td>6F2</td>
<td>140F2</td>
</tr>
<tr>
<td>Read Time</td>
<td>&lt;10ns</td>
<td>.2ns</td>
</tr>
<tr>
<td>W/E Time</td>
<td>&lt;10ns</td>
<td>.2ns</td>
</tr>
<tr>
<td>Retention Time</td>
<td>64ms</td>
<td>-</td>
</tr>
<tr>
<td>Write Cycles</td>
<td>&gt;1E16</td>
<td>&gt;1E16</td>
</tr>
<tr>
<td>Write Voltage</td>
<td>2.5V</td>
<td>1V</td>
</tr>
<tr>
<td>Read Voltage</td>
<td>1.7V</td>
<td>1V</td>
</tr>
<tr>
<td>Write Energy</td>
<td>4.00E-15</td>
<td>5.00E-16</td>
</tr>
</tbody>
</table>
Table 1.2: A list of which technologies have the best performance for each listed feature.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Current Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Size</td>
<td>22nm (NAND)</td>
</tr>
<tr>
<td>Cell Area</td>
<td>4F² (NAND)</td>
</tr>
<tr>
<td>Read Time</td>
<td>.2ns (SRAM) 2ns (DRAM)</td>
</tr>
<tr>
<td>W/E Time</td>
<td>.2ns (SRAM) 2ns (DRAM)</td>
</tr>
<tr>
<td>Write Cycles</td>
<td>10¹⁶ (DRAM)</td>
</tr>
<tr>
<td>Write Voltage</td>
<td>1V (SRAM)</td>
</tr>
<tr>
<td>Read Voltage</td>
<td>1V (SRAM)</td>
</tr>
<tr>
<td>Write Energy</td>
<td>~1E-16 J/Bit (NAND)</td>
</tr>
<tr>
<td>Retention Time</td>
<td>&gt;10 Years (Non-Volatile)</td>
</tr>
</tbody>
</table>

place as the best of one category. Among the current industrial leaders DRAMs are extremely fast, SRAM have low voltages, NANDs have the best retention. As the new technologies progress, a memory device called the universal memory is under investigation. Such a memory would be used in every part of a circuit. It would be a best of all worlds memory device. A universal memory device should have the characteristics listed in Table 1.2.

Table 1.1 takes the best of each major technology. For instance NAND has the best scaling at 4F², while DRAMs have the quickest speed. A universal memory should have all of these characteristics. Currently there is no universal memory but with many new technologies gaining acceptance in the industry is there any new memory that might meet this requirement? How far is each of the technologies from a universal memory?
To answer these questions we can devise a simple figure of merit (FOM). A figure of merit is used to characterize the performance of a device relative to its alternatives. Looking at Table 1.2, we can see that the only quantities that we wish to be larger are the write cycles and retention time, the rest of the features should be as small as possible. Therefore we can arrive at the following figure of merit.

\[
\text{Figure of Merit} = \frac{(\text{Write Cycles})(\text{Retention Time})}{\text{Feature Size}(\text{Read } t)(\text{Write } t)(\text{Write } V)(\text{Read } V)\text{Energy}}
\]  
(1.1)

By using this figure of merit we wish to obtain the largest value. This value will have arbitrary units, and is just used for a comparison of memory technologies. Computing the figure of merit for a universal memory we arrive at Equation 1.2.

\[
\text{Figure of Merit(Universal)} = \frac{(10^{16})(10)}{4(22^2)(.2)(.2)(1)(1)10^{-16}} = 1.42 \times 10^{22}
\]  
(1.2)

Computing the rest of the memory devices figure of merit and plotting them we arrive at Figure 1.28.

Figure 1.28 shows each of the main memory devices figures of merit as calculated by Equation 1.2. For comparison DRAM and SRAMs are included; however they have little chance of becoming a universal memory since they lack any meaningful retention time.

The first point of interest is how poorly NAND memories compare to the other memory devices. This can be attributed to one factor, endurance. The endurance given in Table 1.2 is much lower than competing technologies. This is not because NAND devices have poor endurance. NAND manufacturers give NAND endurance
cycles that they know all their memory devices will meet. Many NAND memories can be cycled millions of time. However, the quantity of NAND devices being produced means there are always a few bad devices and therefore they must list the maximum amount of cycles that they can guarantee for all devices produced. The other memories listed have not been scaled to the levels that NAND memories have. Therefore, the endurance listed for those devices is the best performer, not the worst. Increasing the endurance to cycles to one million shows that NAND is at the same level as other technologies (NAND +Ret column).

The next point is that the best performer on this figure of merit scale is FeRAMs. This is mainly due to their long endurance cycles and low voltages. However, FeRAMs are currently facing enormous scaling issues. Reducing the size of the structure reduces the level polarization in the material and it currently seems impossible to
scale below 180nm. FeRAM also needs to be refreshed after a read cycle. This is not included in the figure of merit; however it is a major hurdle for FeRAM acceptance in the marketplace.

This leaves three main technologies that could advance to a universal memory. NAND devices need to improve their endurance cycles and lower their voltages. Improvement can be done with a detailed look at the gate stack. MJTs are still relatively new and improvements in these devices are already taking shape as mentioned. Phase change memories are a jack of all trades memory. They are good in nearly every category listed in Table 1.2; however they are not the best in any of them. Currently many manufactures are turning to phase change since they are overly good in most categories.

Currently, there is no memory device that can be declared a universal memory. Three memory structures, NAND, MJT, and PC, are close but each has their own challenges ahead. Still there is great hope that one of these technologies or one of the emerging memory technologies can be declared a universal memory.

1.7 Summary

This section provides an overview of the nonvolatile memory systems in development today. It provides the latest information on all major nonvolatile memory systems. These devices include charge trap (CT), ferroelectric memories (FeRAM), magnetic tunnel junctions (MJT), and phase change (PC) memories. In each case a description of the basic operating principles are given. Also provided are the current challenges and limitations to each technology. This section also provides a look at
new upcoming memory systems. These include organic memory, Mott memory, nano-eleco-mechanical, and redox. To aid in the understanding of the various challenges of each memory type, a table and discussion of the major memories is given. Also provided is a rough figure of merit to determine which memory configuration is closest to obtaining the elusive universal memory.
The standard procedure to test a new gate stack design for a non-volatile memory requires fabrication of a complete transistor device. This fabrication process is long, complicated, and most importantly expensive. A standard transistor fabrication process is given in Appendix A.

An easier, quicker, and less costly alternative is required to sufficiently study new gate stack designs. One alternative is to fabricate capacitors. These traditional NVSM capacitor test structures have a uniformly doped substrate which provides a fixed amount of free carriers. For an n-type substrate the free carriers are electrons while for p-type substrate the free carriers are holes. Unfortunately, these devices only have one carrier type present in the substrate limiting their usefulness for full non-volatile memory tests. The write operation of a charge trapping NVSM requires the use of electrons, while the erase operation requires the use of holes. To test both operations, one needs to either create multiple stacks with different substrate dopings, or use another carrier generation method such as light or temperature. The first method requires more test area and is not an accurate representation of a final device, as a final device will have only one substrate doping. The second carrier generation method is also not an accurate representation of a final device.
Figure 2.1: Write and erase curves on a non gridded capacitor. The capacitor has a gate stack of aluminium oxide/silicon nitride/silicon oxide 8/8/3nm. For the write or program state the generation of carriers from light is slow and the device does not begin to write until after $1E-3$ seconds.\[99\]
To overcome the traditional capacitors shortcomings, we have developed the grid-ded capacitor structure, further mentioned as a gridded capacitor or GC. The gridded capacitor requires grid lines that are doped with the opposite carrier type than the substrate and implanted before gate stack deposition. These minority carrier grid lines traverse the entire wafer, which permits the implanted grid lines to establish the same DC and AC potential between the lines and the substrate due to the high leakage and capacitance of the grid. The long grid lines have a low resistance to the substrate due to DC leakage shunted with a large junction capacitance. This allows for a representation of source/drain contacts imbedded in the substrate without the actual need of make contacts. The structure lends itself well to material systems where low resistance source/drain contacts reduce device performance and are difficult and time consuming to fabricate. For example, compound materials [31], carbon nanotubes [38], and graphene [32].

The gridded capacitor test structure requires only two masks, one mask for the definition of doped grid lines, and another mask for the definition of the metal gates. A detailed view of the overlapped masks on the whole wafer is shown in Figure 2.2, while Figure 2.3 shows a detailed view of one of the grid testing areas. The current mask set contains grid spacings of 30, 40, 50, 70, 100, 200, and 300um. The capacitor radii vary from 40-250um depending on the grid size; however, there is always a 250um radius capacitor on each grid. Grid and capacitor size varies to allow for selecting the optimal structure for the labs equipment and the required test and will be fully explained later in the chapter.
Figure 2.2: Large scale view of the gridded capacitor mask set. The blue circles are the metal gates of each individual capacitor. Tan lines indicate the grid lines with a spacing of 30\,\mu m on the left and 300\,\mu m on the right. Capacitor sizes increase to compensate with the larger grid sizes.
Figure 2.3: View of one grid spacing section of the gridded capacitor mask set. Blue indicates the gate metal, with tan lines indicating the grid lines. Each grid section is labeled for easy detection.
2.1 Tunnel Oxide Formation and Characterization

2.1.1 Deal-Grove and Massoud Model

The standard silicon oxide growth model is the Deal-Grove model. When the Grove-Deal model was first developed in the mid 1960’s oxides were grown larger than 100nm[14]. It has been successfully used for years, but it is not applicable for thicknesses under 10nm. As the oxide thicknesses were steadily decreased, it became apparent that a new model would be needed to calculate accurate oxidation times. This task was taken up by Massoud and others[61]. The group modified the Grove-Deal model to more accurately describe and predict thin oxidation.

Massoud and others found that the SiO$_2$ growth rate for thin oxides could be expressed by the following equation[61].

$$\frac{dx_o}{dt} = \frac{B}{2x_o + A} + C_1 \exp \left(-\frac{x_o}{L_1}\right) + C_2 \exp \left(-\frac{x_o}{L_2}\right)$$

(2.1)

The first term on the right hand side of Equation 2.1 is the linear-parabolic term where B and B/A are the parabolic and linear rate constants, respectively, as defined by Deal and Grove. These values are significantly altered in the Massoud model [60]. The rate constants B and B/A can be written in another form using an Arrhenius equation. An Arrhenius equation is a simple formula for the temperature dependence of the rate constant, and therefore, the rate of a chemical reaction. The Arrhenius equation gives the dependence of the rate constant $k$ of chemical reactions on the temperature $T$ and activation energy $E\alpha$ [14] as shown in Equation 2.2

$$k = A \exp \left(-\frac{E\alpha}{RT}\right)$$

(2.2)
Figure 2.4: Values of the pre-exponential constants and activation energies for different crystal orientations in silicon.[61]

A is the pre-exponential factor and R is the gas constant. Returning to the rate constants A and B/A, which in Arrhenius form are shown in 2.3 and 2.4

\[
B = C_B \exp \left( -\frac{E_B}{kT} \right) \tag{2.3}
\]

\[
\frac{B}{A} = C_{B/A} \exp \left( -\frac{E_{B/A}}{kT} \right) \tag{2.4}
\]

The values for the pre-exponential constants and activation energies for different crystal orientations are provided in Table 2.4

The two exponential terms in Equation 2.1 represent the rate enhancement in the thin regime. They can be defined in terms of pre-exponential constants C1 and C2 with characteristic lengths L1 and L2 [60]
Another formulation of Equation 2.1, where the two terms that represent the rate enhancement in the thin regime are decaying exponentially with time, can be expressed as:

\[
\frac{dx_o}{dt} = \frac{B + K_1 \exp\left[-\frac{t}{\tau_1}\right] + K_2 \exp\left[-\frac{t}{\tau_2}\right]}{2x_o + A} \quad (2.5)
\]

The four parameters \(K_1, K_2, \tau_1, \tau_2\) can be fitted into an Arrhenius equation as shown:

\[
K_1 = K_1^0 \exp\left[-\frac{E_{K_1}}{kT}\right] \quad (2.6)
\]

\[
K_2 = K_2^0 \exp\left[-\frac{E_{K_2}}{kT}\right] \quad (2.7)
\]

\[
\tau_1 = \tau_1^0 \exp\left[-\frac{E_{\tau_1}}{kT}\right] \quad (2.8)
\]

\[
\tau_2 = \tau_2^0 \exp\left[-\frac{E_{\tau_2}}{kT}\right] \quad (2.9)
\]

The pre-exponential constant and activation energies in the above equations for different crystal orientations and dry oxidation in the temperature range from 800-1000°C are given in Table 2.5.

To predict the oxide growth we wish to have an analytical expression in which we solve for the oxide thickness \(x_o\). For this purpose we rewrite Equation 2.5

\[
[2x_o + A]dx_o = [B + K_1 \exp\left[-\frac{t}{\tau_1}\right] + K_2 \exp\left[-\frac{t}{\tau_2}\right]] dt \quad (2.10)
\]
We can now integrate from time 0, where the native oxide thickness is $x_i$, to an oxidation time $t$ where the oxide thickness is $x_o$ which results in the following equations:

\[
x_o^2 + Ax_o = Bt + M[1 - \exp\left(-\frac{t}{\tau_1}\right)] + M_2[1 - \exp\left(-\frac{t}{\tau_2}\right)] + M_o
\]  

(2.11)

\[M_o = (x_i^2 + Ax_o)M_1 = K_1\tau_1M_2 = K_2\tau_2\]  

(2.12)

At which point we solve the quadratic equation for $x_o$

\[
x_o = -\frac{A}{2} + \sqrt{\frac{A^2}{2} + Bt + M[1 - \exp\left(-\frac{t}{\tau_1}\right)] + M_2[1 - \exp\left(-\frac{t}{\tau_2}\right)] + M_o}
\]  

(2.13)

Figure 2.5: Pre-exponential constant and activation energies for different crystal orientation and dry oxidation for silicon.[59]

<table>
<thead>
<tr>
<th>Crystal Orientation</th>
<th>(100)</th>
<th>(111)</th>
<th>(110)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_1^0$ [nm$^2$/min]</td>
<td>$2.49 \times 10^{11}$</td>
<td>$2.70 \times 10^9$</td>
<td>$4.07 \times 10^8$</td>
</tr>
<tr>
<td>$E_{K1}$ [eV]</td>
<td>2.18</td>
<td>1.74</td>
<td>1.54</td>
</tr>
<tr>
<td>$K_2^0$ [nm$^2$/min]</td>
<td>$3.72 \times 10^{11}$</td>
<td>$1.33 \times 10^9$</td>
<td>$1.20 \times 10^8$</td>
</tr>
<tr>
<td>$E_{K2}$ [eV]</td>
<td>2.28</td>
<td>1.76</td>
<td>1.56</td>
</tr>
<tr>
<td>$\tau_1^0$ [min]</td>
<td>$4.14 \times 10^{-6}$</td>
<td>$1.72 \times 10^{-6}$</td>
<td>$5.38 \times 10^{-9}$</td>
</tr>
<tr>
<td>$E_{\tau1}$ [eV]</td>
<td>1.38</td>
<td>1.45</td>
<td>2.02</td>
</tr>
<tr>
<td>$\tau_2^0$ [min]</td>
<td>$2.71 \times 10^{-7}$</td>
<td>$1.56 \times 10^{-7}$</td>
<td>$1.63 \times 10^{-8}$</td>
</tr>
<tr>
<td>$E_{\tau2}$ [eV]</td>
<td>1.38</td>
<td>1.90</td>
<td>2.12</td>
</tr>
</tbody>
</table>
One can now solve for any number of variables with a few given initial conditions. Normally, one would have a desired thickness in mind and calculate how long an oxidation needs to take place. One could then find an initial oxide by knowing the final thickness, oxidation time, and temperature of the furnace. A plot of oxide thickness versus time for various temperatures is given in Figure 2.6 with an initial oxide thickness of zero.

### 2.1.2 2.1.2 X-Ray Photoelectron Spectroscopy

X-Ray photoelectron spectroscopy (XPS) is a quantitative spectroscopic technique that measures the elemental composition, empirical formula, chemical state, and electronics state of the atoms that exist within a material. XPS spectra are obtained by
Figure 2.7: XPS Diagram. A focused beam of X-rays hits the sample which then emits electrons. The electrons are captured and analyzed resulting in a graph of the number of electrons at each energy level.[13]

Irradiating a material with a beam of X-rays while simultaneously measuring the kinetic energy and number of electrons that escape from the top 100 Angstroms of the material. A diagram showing the technique is shown in Figure 2.7.

For an XPS measurement a one inch by one inch sample is prepared. X-ray photoelectron spectroscopy only measures the top 100Å (maximum) of the sample. The sample is then placed in an ultra-high vacuum (< 10^{-8} torr). This usually requires samples to be solid although a few XPS systems allow volatile liquids and operate at lower pressures. Once in the chamber and at ultra-high vacuum a focused beam of X-rays is fired at the sample. The X-rays excite atoms and electrons are freed from the surface. The freed electrons are collected and analyzed by a detector. A graph of electron counts versus energy reveals different peaks. These peaks correspond to
the atoms chemical and electronic state and elemental composition in the surface. By measuring the electron counts at different x-ray applied angles, one can perform a rudimentary depth analysis. This is a non-destructive measurement, however it can only provide resolution of the first 100Å. For greater depths one needs to perform a destruction Auger Electron Spectroscopy.

Our XPS measurements were performed in the Material Science Department of Lehigh University on a Scienta ESCA-300 picture in Figure 2.8. The Scienta ESCA 300 is generally regarded as one of the best XPS instruments in operation today. A combination of design factors permit data to be obtained from samples quickly without sacrificing energy resolution. A few basic facts about the ESCA-300 are provided:
• X-ray source: Monochromatic Al K X-rays generated from a rotating anode which can operate at power levels of up to 7.5 kW. The monochromator consists of seven crystals mounted on three Rowland circles. The x-ray spot is a line >5mm long and >1mm wide.

• Energy analyzer: High resolution 300 mm mean radius hemispherical electrostatic analyzer designed for high throughput with small solid angle acceptance for enhanced angle resolved XPS capabilities.

• Detector: Multi channel micro-channel plate-CCD camera combination.

• Charge compensation: Non-line-of sight monochromatic low energy electron flood gun

• Sample stage and manipulation: 5 axis computer controlled precision manipulator. Sample heating to >600°C and sample cooling to -100°C.

• Analyzed area: The nominal analyzed area is rectangular with approximate dimensions of 4 mm x 0.2 mm for a total area of about 0.8 mm². The size of the analyzed area can be adjusted electronically and mechanically. The minimum practical area for routine sample analysis is about 0.5 mm x 0.1 mm.

• In situ sample preparation: The sample chamber and/or attached chambers provide the ability to heat specimens to >1000°C without the use of electron bombardment, expose surfaces to various vacuum compatible reactant gases, deposit thin films from a precision Knudsen cell and monitor the thickness with a crystal monitor, monitor the vacuum with an RGA, fracture brittle samples in situ, sputter clean surfaces, and scrape surfaces in UHV. The sample entry
chamber, which can have base pressures in the low $10^{-8}$ torr range, can be used to carry out experiments at temperatures $>400^\circ$C at pressures ranging from $10^{-8}$ to 700 torr.

- Sample types: Virtually any material with outgassing rates that allow the pumps to maintain pressures of $10^{-7}$ torr or lower can be analyzed.

To provide a baseline for all further silicon oxide measurements, a fresh silicon wafer went through an RCA cleaning process followed by a Hydrofluoric (HF) dip. The HF dip removes any native oxide and terminates the silicon surface with hydrogen. The hydrogen bond can only be broken over a period of time (hours) or at elevated temperatures. This allows the wafer to be minimally exposed to air without the growth of an oxide layer. The treated wafer was brought immediately to be loaded into the Scienta ESCA-300.
Figure 2.9 and 2.10 show the results of the XPS scan on a hydrogen terminated silicon wafer. The peaks are represent electron shell orbital energies. The simple names of the orbitals are s, p, d, and f, which correspond to names sharp, principal, diffuse, and fundamental respectively. The electron filling of the orbitals are built up in the following sequence: 1s2s2p3s3p4s3d4p5s. The S orbital has one state, which can contain two electrons with opposite spin. The P orbital has three states hence it can contain 6 electrons. Since the energy of the x-ray wavelength is known, the electron binding energy of each emitted electron can be determined by using the following equation based on the work of Ernest Rutherford.

\[ E_{\text{binding}} = E_{\text{photon}} - (E_{\text{kinetic}} + \phi) \]  

(2.14)

\( E_{\text{binding}} \) is the binding energy of the electron, \( E_{\text{photon}} \) is the energy of the x-ray photons being used, \( E_{\text{kinetic}} \) is the kinetic energy of the electron as measured by the instrument, and \( \phi \) is the work function of the spectrometer (not the material). Therefore, mapping the energy peaks, the amount of electrons ejected or counts per second (CPS) from the material permit us to measure the electron binding energies and allow us to determine the elements present in the material.
Figure 2.10: XPS spectra on a hydrogen terminated silicon wafer. Major peaks are automatically label by the CasaXPS graphing program. The labeled peaks correspond to the presence of Oxygen, Silicon, and Carbon atoms in the sample.

In Figure 2.10, the CPS peaks match with Oxygen, Silicon, and Carbon atoms. The silicon peaks are far larger than the oxygen peaks indicating a surface relatively clean of oxygen. The oxygen 1s state is easily visible as well as the KLL transition. The KLL peak represent the energy of the electrons ejected from the atoms due to the filling of the O 1s state (K shell) by an electron from the L shell coupled with the ejection of an electron from the L shell. This transition is less likely than an electron ejected from the O 1s K shell and hence has a lower peak. Heavier atoms will have more energy peaks due to more states as well as acceptable transitions.
Figure 2.11: XPS Spectra of the silicon peak on a hydrogen terminated silicon wafer.

One can also zoom into certain sections of the binding energy to obtain a clearer picture. An important peak occurs at a binding energy of 99eV, as shown in Figure 2.11. Figure 2.11 is again taken on a hydrogen terminated surface and therefore has very few SiO\(_2\) atoms. The two peaks are due to spin orbital splitting. Also Figure 2.11 shows a small hump around a binding energy of 102-105eV. This bump is caused by the formation of SiO\(_2\). Since Figure 2.11 is for a hydrogen terminated surface the counts from SiO\(_2\) atoms is nearly non-existent, further strengthening our result of a clean SiO\(_2\) free starting surface.

Figure 2.11, but with known binding energies is shown in Figure 2.12. The known binding energy values do not exactly match with the results. This is due to a potential difference between the sample and the detector. The potential difference will shift the
Figure 2.12: XPS Spectra of the silicon peak on a hydrogen terminated silicon wafer with labels. The labels are shifted left of the experimental results. This is due to a charge difference between the detector and the sample. This imbalance can be fixed by flowing electrons over the sample (dangerous) or through the CasaXPS program.[56]
entire XPS spectra left or right and is easily distinguishable from elemental shifts. The potential difference can be corrected two ways. The first way is to use an electron gun and flood the surface of the sample with electrons. Adjusting the amount of electrons on the samples surface will alter the overall potential difference. This method is available on the ESCA 300 but it is time consuming and it has the potential to destroy samples due to the additional charge. For our samples we use the second way to account for the potential different which is to measure the shift at multiple points in the spectra and manually shift the curve in a computer program.

For completion we can also examine the carbon and oxygen peaks, which are shown in Figure 2.13 and 2.14 respectively. Anytime a wafer comes out of processing and contacts air, some amount of carbon will deposit on the wafer. The test wafers were brought out of the fabrication lab to another lab and hence have some carbon on them. Fortunately, the amount of carbon is very low. Examining the carbon peak will also help determine how clean our lab environment is and the fabrication process. Any extra carbon, and a larger peak, will mean something is wrong with our processing equipment. Both Figures indicate a low level of carbon and oxygen present on the wafer.
Figure 2.13: XPS spectra of the carbon peak on a hydrogen terminated silicon wafer. The peak is much less defined indicating few carbon atoms on the sample. Carbon is expected as the wafers were transferred outside of a cleanroom environment.

Figure 2.14: XPS spectra of the oxygen peak on a hydrogen terminated silicon wafer. Like the carbon peak, the oxygen peak is ragged indicating few oxygen atoms. Again some oxygen is expected as the wafers were transferred outside of a cleanroom environment.
2.1.3 Triple Wall Furnace

The first layer of the gate stack is the silicon oxide, also referred to as the tunneling oxide. This oxide must have good uniformity, a small defect density, high dielectric breakdown strength, high charge-to-breakdown, small interface trap density, and the ability to endure hot electron injection and ionizing radiation without creating interface trap charges and fixed oxide charges. Conventional single wall oxidation furnaces introduce defects or impurity traps in the oxide by allowing the diffusion of heavy metal ions and mobile alkali ions through the quartz tube wall[85]. Also, a considerable amount of moisture is incorporated into the oxide by the diffusion of water through the single quartz wall. To eliminate these adverse conditions, we used a special triple-wall oxidation furnace[113].
The triple wall oxidation system consists of four major parts: a triple-wall oxidation furnace tube, a pre-combustion furnace, a cold trap assemble, and a hydrometer assembly. Figure 2.15 shows a schematic diagram of the system while Figure 2.16 shows an actual picture of the furnace system. The custom-designed, triple-wall oxidation furnace tube (E) is made of three concentric quartz tubes whose ends are sealed together to leave gaps between each quartz tube. Through the gap between the outer wall and the middle wall, both N\textsubscript{2} gas and O\textsubscript{2} gas are flowing to getter the heavy metal ions such as Fe, Cr, Ni, etc. and mobile alkali ions such as Na, K, etc., that come from the heating elements of the oxidation furnace. The inner gap, located between the middle wall and the inner wall, supports a N\textsubscript{2} gas evaporated from a liquid N\textsubscript{2} source. This operation prevents the diffusion of moisture from the outer gap through the inner wall into the oxidation chamber. The inner tube is used for the oxidation and its atmosphere is controlled very tightly to minimize the introduction of any impurities into the oxidizing gas. This is accomplished with the use of the highest purity O\textsubscript{2} gas which flows through a pre-combustion furnace and a cold trap assembly. To further clean the oxidation chamber, TCA is flowed for 30 min at 1000\textdegree C a day before the oxidation. The temperature is controlled by a front panel readout (3).

A specially designed white elephant (F,1) provides the necessary tight fit to the inner wall of the triple wall oxidation furnace tube at the wafer loading end (2). The white elephant also has an inert gas inlet, an inert gas outlet, an exhaust gas outlet, and an opening for the push rod. Argon gas is used for the loading, unloading, and cooling of wafers in the white elephant after oxidation. Argon is also used as the carrier gas since nitrogen can ‘nitrodize’ the surface. The exhaust gas outlet is
Figure 2.16: Photo showing the operator section of the triple wall furnace. The white elephant [1] connects to the inner furnace tube [2]. The cold trap [4] and furnace controls [3] are also shown[113].
connected to the humidity sensor. The O$_2$ and Ar gases used in the oxidation process and introduced into the pre-combustion furnace (A, 4) are very pure (both 99.99%) and very dry; however, they still contain hydrocarbons which can be dissociated into water at the oxidation temperature. The purpose of the pre-combustion furnace is to burn the hydrocarbons in O$_2$ and Ar gases before oxidation and eliminate the moisture generating sources in the oxidation chamber. Generated moistures are condensed in the down-line cold trap assembly. Pre-combustion takes place at a compromised temperature of 600°C. The pre-combustion furnace tube is packed with quartz pieces to increase the gas traveling time in the combustion zone and promotes a longer reaction time.

A cold trap assembly (B and C) is needed to eliminate the moisture in the oxidizing gases after they pass through the pre-combustion furnace. It consists of a Neslab immersion cooler which can maintain -60°C without the refill of coolant, a custom designed quartz cooling coil, and a dewar containing a cooling medium. Methanol is used as a cooling medium. The temperature of the cold bath is monitored by a copper-constantan thermocouple and a digital thermometer with automatic cold junction compensation. The humidity level in the O$_2$ gas is measured by a Shaw dew point meter (H, I, J). It can detect less than 1ppm level of humidity. The meter has a capacitance moisture sensor made of a metal core, coated with a hygroscopic dielectric and finally covered with a porous gold film. The moisture sensor is operated at room temperature. The moisture level is monitored in the exhaust gas[113].
2.1.4 Tunnel Oxide Results

Prior to the final tunnel oxide layer oxidation, we must characterize the furnace. The triple wall oxidation furnace is built to only provide a growth of SiO$_2$ when oxygen is turned on and flowing through the tube. However, oxygen may find its way into the tube from a couple of sources such as the argon bottles and while loading the wafers. This additional oxygen may promote a small SiO$_2$ growth which must be quantified to provide accurate tunnel thickness results. From the XPS data on a hydrogen terminated silicon wafer surface previously shown in Figures 2.9, 2.11, 2.13, and 2.14, we can characterize the furnace up to the moment when the oxygen is turned on.

First, wafers were placed in the white elephant for one hour. This procedure was done to make sure that no leaks were present in the white elephant and that the heat flowing out of the tube was low enough to not provide a oxide growth. XPS spectra of these wafers showed results indistinguishable from the hydrogen terminated wafers. Next, a group of wafers were loaded into the furnace at $900^\circ$C with only argon following in the tube and allowed to rest for a half hour. This experiment simulates the loading and warmup time the wafers undergo before oxidation. Selected XPS spectra of the oxygen and silicon peaks are given in Figure 2.17 and 2.18.

Figure 2.17 shows the XPS spectra peak of oxygen on a hydrogen terminated wafer that has been in the furnace at $900^\circ$C for a half hour. It is clearly seen that the oxygen peak is more well defined than a purely hydrogen terminated wafer. This indicates additional oxygen has made its way into the tube. Figure 2.18 shows the XPS spectra peak of silicon on a hydrogen terminated wafer that has been in the furnace at $900^\circ$C for a half hour. This figure shows a more well defined hump in the
Figure 2.17: XPS spectra of the oxygen peak on hydrogen terminated silicon wafers placed in the white elephant for one hour. The peak shows no difference from previous hydrogen terminated wafers.

Figure 2.18: XPS spectra of the silicon peak on hydrogen terminated silicon wafers placed in the white elephant for one hour. The peak shows no difference between from previous hydrogen terminated wafers.
102-105eV binding energy range, which indicates a larger presence of SiO$_2$ atoms. By taking a ratio of the oxygen peak area to the silicon peak area and apply equation 2.15 we are able to determine the thickness of the SiO$_2$ layer.

\[
\text{Thickness}(A) = 30\ln\left(\frac{R}{45} + 1\right) \tag{2.15}
\]

where \(R\) is given by:

\[
R = \frac{\text{Oxygen Area}}{\text{Silicon Area}} \tag{2.16}
\]

This empirical relation was originally developed through a comparison of XPS and ellipsometry data[28]. For graphing as well as area calculations we access the XPS data through a program CasaXPS[56]. CasaXPS, Computer Aided Surface Analysis for X-ray Photoelectron Spectroscopy, is a computer program that allows powerful data processing of XPS spectra. The program can handle angle resolved samples, XPS depth profile, as well as Auger Electron Spectroscopy analysis. CasaXPS allows us to easily and accurately obtain the oxygen and silicon peaks areas.

Obtaining the peak areas from Figure 2.18 and using equations 2.15 and 2.16 the SiO$_2$ thickness is 5.41 Angstroms. By placing the wafers in a 900C furnace with only argon, an inert gas, flowing we already have a thin oxide formation. This leads to the conclusion that either there is a leak in the tube allowing oxygen to enter or the oxygen is from the argon flow. We can use the Massoud model discussed in the previous section to determine the oxidation times and use 5.41Å as an initial oxide value.

The theoretical oxidation times are given in Table 2.1 and graphed in Figure 2.19. Next, a test run was completed to verify the Massoud model oxidation times, this
Table 2.1: Experimental and theoretical thickness calculations for given oxidation times. We originally calculated an initial starting oxide of 5.4 Å, after switching argon tanks the tests wafers were calculated to have a starting oxide thickness of 16.4 Å. The final column shows the thicknesses measured on our actual wafers.

<table>
<thead>
<tr>
<th>Oxidation Time (min)</th>
<th>Experimental Data (Å)</th>
<th>Theory with 5.4 Å (Å)</th>
<th>Theory with 16.4 Å (Å)</th>
<th>Final (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:32</td>
<td>20.9</td>
<td>9.98</td>
<td>20.94</td>
<td>23.96</td>
</tr>
<tr>
<td>1:10</td>
<td>26.07</td>
<td>14.96</td>
<td>25.88</td>
<td>28.87</td>
</tr>
<tr>
<td>1:53</td>
<td>31.29</td>
<td>20.11</td>
<td>30.99</td>
<td>31.63</td>
</tr>
<tr>
<td>2:38</td>
<td>34.33</td>
<td>25.04</td>
<td>35.87</td>
<td>39.19</td>
</tr>
<tr>
<td>3:28</td>
<td>40.277</td>
<td>30.05</td>
<td>40.84</td>
<td>41.85</td>
</tr>
<tr>
<td>4:22</td>
<td>45.408</td>
<td>35.03</td>
<td>45.78</td>
<td>47.41</td>
</tr>
<tr>
<td>5:21</td>
<td>49.828</td>
<td>40.07</td>
<td>50.77</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.19: Original argon only test wafer showed a starting oxide of 5.4 angstroms. After running oxidation tests with a new argon cylinder the initial oxide jumped to 16.4 angstroms.
information is also given in Table 2.1 and graphed in Figure 2.19. The experimental
curve has a large shift up in oxidation thickness with what was calculated. The theo-
retical curve uses a starting oxidation thickness of 5.4Å, but using an initial oxidation
of 16.4Å gives a matching result between theoretical and experimental curves. After
a careful investigation, we discovered that after the original thickness and XPS mea-
surements a new argon cylinder was installed. Our argon cylinders are only 99.99%
pure. The oxygen content of the second cylinder was higher than the original resulting
in a large amount of oxygen being fed into the tube on warm-up and hence a larger
initial oxide growth. Discovering this issue we re-characterized the tube by placing
a fresh wafer in the tube for half an hour at 900°C with argon flowing. The XPS
results from this test indicate an initial oxide of 16Å which is in agreement with our
theoretical results. Through furnace characterization the following results are used

- The argon cylinder is a source of oxygen resulting in an initial oxide growth
- Each argon cylinder needs to be characterized through a furnace run
- Use one cylinder for the entire run

After the furnace and argon cylinder characterization the final run was made. The
final wafers were to range by 5Å from 20Å to 45Å. The results of the final run are
given in Table 2.1. The XPS measurements on the test wafers in the final run show
good agreement with the theoretical results, however they differ in being roughly
2Å thicker. This discrepancy is attributed to the short oxidation times. The oxide
cylinder is turned off at the required time, however there is still oxide in the tube
which continues the oxidation.
Figure 2.20: XPS spectra of the silicon peaks on a silicon wafer which has undergone oxidation. Area calculations are done through the CasaXPS software program (shaded region) for thickness measurements. This wafer shows the peaks for an oxide thickness of 2.4nm.

2.2 Nitride Charge Trapping Formation and Characterization

The nitride layer was deposited using a Low Pressure Chemical Vapor Deposition (LPCVD) furnace picture in Figure 2.21. The furnace operates at 680°C at a pressure of 0.3 Torr. Silicon nitride is deposited by a reaction between SiH$_2$Cl$_2$ and NH$_3$ at a ratio of 10:1000 scm respectively.

A fifteen minute deposition results in a 70Å thick oxygen rich silicon nitride layer. A test run was completed under the conditions mentioned above which resulted in a 75.79Å thick Si$_4$N$_4$ layer. This is well within the target range of 60-80Å. During the final production run a thickness of 77Å was measured.
Figure 2.21: Control panel of the LPCVD furnaces at Lehigh University. The tubes are directly behind the panel and the loading bay is at the back of the photo. The white tube visible on the right side of the photo is the POCl$_3$ diffusion furnace.

Figure 2.22: Full scan XPS spectra of a silicon wafer after a silicon nitride deposition.
Figure 2.23: XPS Spectra of the silicon peak on a silicon wafer after silicon nitride deposition. The double silicon peak is on the right with the nitride corresponding nitride bonding peak on the left.

The silicon nitride layer measurements were also completed through XPS spectra. Figure 2.22 shows the full XPS spectrum with labels while Figure 2.23 focuses on the silicon peak. The samples were grown directly on a clean hydrogen terminated silicon wafer and not onto a SiO$_2$ layer.

In Figure 2.23 there is a large oxygen peak. This is due to the Si$_3$N$_4$ being oxygen rich. The oxygen peak is shown in Figure 2.23. The bonding structure of silicon nitride is similar to silicon oxide. Therefore, it is proposed that the same XPS technique and formula shown in equations 2.15 and 2.16 can be used to calculate the thicknesses. Due to the atomic differences between silicon nitride and silicon oxide, this formula could be off as much as 5%. This is an acceptable offset as the silicon
nitride layer is not as important as the tunneling silicon oxide layer. We will revisit these assumptions in the TEM section of this chapter.

Figure 2.24 shows the XPS spectra of carbon on the silicon nitride coated wafers. Again, we show that the carbon peak is low and not well defined indicating few carbon atoms present in the material. This peak comes from any contact with air, which takes place between deposition and measurement. Lastly for completion, Figure 2.25 shows the nitrogen peak on the sample. This figure merely indicates that the reaction took place and there is nitrogen in the sample.

2.3 Blocking Oxide Formation and Characterization

For the third layer of the gate stack, we have used a high-K oxide such as aluminum oxide or hafnium oxide. The reasons for using high-K oxides over a standard silicon
oxide layer are provided in section 2.4. Wafers with two different high-K oxides were fabricated. One set of wafers received an aluminum oxide layer for the blocking oxide while the other set received a aluminum/hafnium oxide layer. Each of these layers are examined in section 2.4.1 and 2.4.2 respectively.

2.4 High-K Oxides

Silicon Oxide has been the go-to dielectric to prevent electrons from tunneling out of the nitride layer to either the substrate or the gate. Continued scaling of NVSM devices have reduced the thickness of SiO$_2$ layer less than 5nm thick. These thin films have allowed for an increase in tunneling through the SiO$_2$ layer. Today’s devices require a thicker film that also improves performance. To completely understand the
A FET device, pictured in Figure 2.26 has a source, drain, and a gate. The gate is separated from the rest of the device by an insulator. The insulator is normally been composed of SiO$_2$ due to its superior quality as a native oxide. As the device dimensions have been reduced this is no longer a suitable insulator and needs to be replaced.

When the source is grounded and a positive voltage is placed on the gate and drain, a channel will form between the source and drain N+ regions. This channel will allow electrons to flow and form a current to form. If the gate insulator is too thin, the electrons can tunnel through the insulator and flow into the gate. This has the negative effect of reducing the current between the source and drain. In essence, a FET is a capacitance-operated device. In every model, the source-drain current depends on the gate capacitance $C_{ox}$. The capacitance $C_{ox}$ is defined as:
Figure 2.27: Capacitance equivalent of the gate stack. Each layer of the gate stack has its own associated capacitance. Referencing the capacitance to one individual layer we can provide an equivalent capacitance of the entire gate stack.

\[ C_{ox} = \frac{\varepsilon_0 K_{ox} A}{t_{ox}} \] (2.17)

In Equation 2.17 \( \varepsilon_0 \) is the permittivity of free space, \( K_{ox} \) is the relative permittivity of the material, \( A \) is the area, and \( t_{ox} \) is the thickness of the insulator. All of the above definitions are either constants, or controlled by the technology process except \( t \), the thickness of the material. Therefore, the solution to the reduced dimensions and tunneling problem is to replace the thin SiO\(_2\) insulator with a thicker layer of a material which has a higher dielectric constant \( K \). The device will still have the small capacitance and hence source-drain current, with mitigated tunneling due to the thicker layer. These new insulators are referred to as high-K oxides.

In much the same way, we can look at a charge trapping NVSM gate stack as a series of capacitances as shown in figure 2.27.
Here $C_I$ is the blocking oxide capacitance, $C_N$ is the capacitance of the nitride layer, and $C_O$ is the tunnel oxide capacitance. The three capacitances can then be merged to form one total capacitance $C_T$.

\[
\frac{1}{C} = \frac{1}{C_O} + \frac{1}{C_N} + \frac{1}{C_I} \quad (2.18)
\]

Using Equation 2.17 we can rewrite 2.18

\[
\frac{1}{C} = \frac{1}{K_O \epsilon_o X_O} + \frac{1}{K_N \epsilon_o X_N} + \frac{1}{K_I \epsilon_o X_I} \quad (2.19)
\]

$X_I$ is the physical thickness of the blocking oxide, $X_N$ is the physical thickness of the nitride layer, $X_O$ is the physical thickness of the tunnel oxide, $\epsilon_o$ is again the permittivity of free space, $K_I$ is the dielectric constant of the blocking oxide, $K_N$ is the dielectric constant of the nitride layer, and $K_O$ is the dielectric constant of the SiO$_2$ tunnel oxide. In this equation, the area $A$ has already been factored out.

We can also rewrite the total capacitance and relate it to the SiO$_2$ tunnel oxide layer.

\[
\frac{1}{C} = \frac{1}{K_O \epsilon_o EOT} \quad (2.20)
\]

We rewrite the total thickness as the Equivalent (Electrical) Oxide Thickness (EOT). The EOT is not a physical thickness but a measure of what the equivalent thickness would be if the layer was replaced by an SiO$_2$ layer. Inserting Equation 2.20 into 2.19 we have a full expression for a device’s EOT.

\[
EOT = X_O + \frac{K_O X_N}{K_N} + \frac{K_O X_I}{K_I} \quad (2.21)
\]
Equation 2.21 assumes that the tunnel oxide is still composed of SiO$_2$. To improve device performance we want the smallest EOT so that we can use smaller voltages, but a large thicknesses to prevent tunneling and improve retention and device performance. To aid in this goal we replace the blocking oxide which has traditionally been silicon oxide with a High-K oxide that has a larger thickness.

Fortunately, there are a lot of materials that have a K value higher than silicon oxides value of 3.9. There are a lot of materials that have high K value which we need to test for device compatibility. To limit the number of materials to study a few rules have been devised.

1. The material must have a high enough K that it will be used for a reasonable number of years of scaling.

The SiO$_2$ layer that we are to replace has been used in production for over 50 years. We would like to use a material which we will be able to use for the next 50 years of scaling. This requires that the dielectric constant be much higher than 3.9. Table 2.2, as well as Figure 2.28, gives a few examples of some high-K materials as well as their K values.

2. The material must act as an insulator, by having band offsets with Si of over 1eV to minimize carrier injection into its bands.

As seen in Table 2.2 and Figure 2.28, the materials with the largest K values have small band gaps. The material used to replace SiO$_2$ must act as an insulator, so the potential barrier at each band must be over 1eV in order to prevent conduction by Schottky emission of electrons or holes into the oxide bands as shown in Figure 2.29[80]
Table 2.2: Static dielectric constant, experimental band gap, and conduction band offset on Si of some high-K materials. Aluminum oxide has nearly the same band gap but a much higher K value.[79]

<table>
<thead>
<tr>
<th>Material</th>
<th>$K$</th>
<th>Gap (eV)</th>
<th>CB offset (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SiO$_2$</td>
<td>3.9</td>
<td>9</td>
<td>3.2</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 (not ALD)</td>
</tr>
<tr>
<td>Ta$_2$O$_5$</td>
<td>22</td>
<td>4.4</td>
<td>0.35</td>
</tr>
<tr>
<td>TiO$_2$</td>
<td>80</td>
<td>3.5</td>
<td>0</td>
</tr>
<tr>
<td>SrTiO$_3$</td>
<td>2000</td>
<td>3.2</td>
<td>0</td>
</tr>
<tr>
<td>ZrO$_2$</td>
<td>25</td>
<td>5.8</td>
<td>1.5</td>
</tr>
<tr>
<td>HfO$_2$</td>
<td>25</td>
<td>5.8</td>
<td>1.4</td>
</tr>
<tr>
<td>Hf$_2$SiO$_4$</td>
<td>11</td>
<td>6.5</td>
<td>1.8</td>
</tr>
<tr>
<td>La$_2$O$_3$</td>
<td>30</td>
<td>6</td>
<td>2.3</td>
</tr>
<tr>
<td>Y$_2$O$_3$</td>
<td>15</td>
<td>6</td>
<td>2.3</td>
</tr>
<tr>
<td>α-LaAlO$_3$</td>
<td>30</td>
<td>5.6</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Figure 2.28: Static dielectric constants vs. band gap. The correlation between High K value and lower Band Gap is clearly seen. Other good candidates are hafnium oxide and zirconium dioxide. [79]
As shown in Figure 2.30, SiO$_2$ has a band gap of 9eV, so it has high barriers to electrons and holes. SrTiO$_3$, which was shown to have a very large K value, has a band gap of around 3.3eV. This material would need to be perfectly aligned with the Si band gap to achieve the 1eV barrier condition. However, this is not the case, as SrTiO$_3$ does not provide a barrier to electrons. In practice, the conduction band offset is usually smaller than the valence band offset. This configuration limits the choice of oxides to those with band gaps of over 5eV [55]. In charge trapping NVSM, the barrier height restrictions are much more strenuous. One must trap the electrons or holes in the nitride layer. To accomplish this task, the tunnel oxide, as well as the blocking oxide, must have a barrier height greater than the silicon nitride. As shown in Figure 2.30, silicon nitride has a conduction band offset of 2.4eV with respect to silicon. The only oxides with a conduction band offset greater than 2.4eV are SiO$_2$ at 3.2eV and aluminum oxide at 2.8eV. There are a few more restrictions on which
Figure 2.30: Band gap and band alignment for various materials on silicon. Both silicon oxide and aluminum oxide have offsets greater than 1V. A few higher K materials have no offset with silicon.[108]
high-k material one should use (particularly if you also consider replacing the tunnel oxide), but details are beyond the scope of this thesis[80].

These restrictions have led researchers to narrow the options down to a few high-K materials. These materials include: Gd$_2$O$_3$/GdAlO [74], HfO$_2$[22, 6], LaAlO/LaHfO[11, 26], and Al$_2$O$_3$[86, 1, 5]. As noted before, only Al$_2$O$_3$ and SiO$_2$ have a conduction band higher than silicon nitride. Therefore, it is common to mix aluminum oxide or silicon oxide into another high-k oxide. For our devices, we are making one wafer set with an Al$_2$O$_3$ blocking oxide and one wafer set with an Al$_2$O$_3$/HfO$_2$ layered blocking oxide.

2.4.1 Aluminum Oxide (Al$_2$O$_3$) Blocking Oxide

Aluminum oxide is an excellent high-k oxide replacement for silicon oxide. The band gap of aluminum oxide is similar to the band gap of silicon oxide while having a K value of 9. For one set of test wafers we have used a 150A layer of aluminum oxide which was deposited by Atomic Layer Deposition (ALD).

ALD is an extremely powerful deposition technique. ALD can deposit extremely high quality thin films. A process flow of an ALD deposition of either ZrO$_2$ or HfO$_2$ is given in Figure 6.10

A wafer is placed under vacuum. A precursor gas flows into the chamber. The precursor gas is such that only a mono-layer sticks to the wafer surface. After the precursor gas, the chamber is vented and purged.

Next, a reacting gas flows into the chamber. This gas has a reaction with the monolayer of precursor gas on the wafer to form the necessary layer. The chamber is again vented and purged. Since only a mono-layer of the desired material is deposited
Figure 2.31: Atomic Layer Deposition (ALD) process for hafnium or zirconium oxide. The precursor is sent through until saturated adsorption occurs. The chamber is purged to prevent deposition and then water is passed through the chamber. The water reacts with the precursor to form a single layer of the desired oxide. On the wafer, the steps can be repeated to obtain the desired thickness. ALD has the benefit of depositing a high quality film, but has the drawback of being an extremely slow process.

Test wafers were returned with aluminum oxide ALD deposition on a hydrogen terminated silicon wafer. This allowed for XPS measurements to be taken to determine the quality of the oxide on the devices. XPS can only detect the composition of the top 100Å while the oxide thickness is 150Å. Therefore we can only determine the quality of the oxide and whether or not any contaminants (other atoms) are present in the film. We can not measure the thickness using XPS measurements.

The full XPS survey is shown in Figure 2.32. This shows that carbon, oxygen, and aluminum are present on the wafer. Figure 2.33 shows the carbon peak of the XPS spectra. Comparing this figure to the other carbon peaks of Figures 2.9 and 2.13 shows this peak to be much higher. This is due to the amount of time the wafer was
Figure 2.32: Full scan XPS spectra of aluminum oxide deposited by ALD on silicon wafers. The deposition is over 100Å so only oxygen, carbon, and aluminum are present.

Figure 2.33: XPS spectra of the carbon peak of aluminum oxide on silicon wafers. A larger carbon peak is observed than on all other samples. This could be due to the length of time between deposition and measurement.
exposed to air. The previous samples were only exposed to air for a maximum of one full day. Due to shipping and obtaining time to use the ESCA 300 the blocking oxide samples were exposed to air for roughly one week. The carbon is also present in the final wafers due to the shipping time. Although larger than the tunnel oxide carbon peak, this larger carbon peak is still comparatively small to the overall composition of the wafer.

For completion, the oxide and aluminum peaks are shown in Figures 2.34 and 2.35. These images show a smooth well defined peak for both. These figures along with full survey show that film is free of other unwanted atoms indicating a very clean high quality film.
2.4.2 Aluminum Oxide (Al$_2$O$_3$)/ Hafnium Oxide (HfO$_2$) Blocking Oxide

Although aluminum oxide has a higher K value (9) than silicon oxide, this is still too low a value to last more than a few years of scaling. However, as seen previously the higher the K value the lower the band gap. This leads to an interesting dilemma. How can we maintain a high band gap and obtain a high K value in the same film? To address this challenge we theorize that a combination blocking oxide of both aluminum oxide(large band gap) and hafnium oxide(high K 24) will obtain a high K value while still maintaining the large band gap.

The second set of wafers received the Al$_2$O$_3$/HfO$_2$ layer. This layer is also 150Å thick and was deposited by atomic layer deposition by Dr. Patrick Shea at Northrop Grumman Corporation. We used a Al$_2$O$_3$/HfO$_2$ layer with a 7:3 ratio. This required
seven layers of aluminum oxide to be deposited followed by three layers of hafnium oxide.

Test wafers with the Al₂O₃/HfO₂ blend directly deposited on a hydrogen terminate wafer were returned from Northrop Grumman Corporation for XPS testing. The full XPS survey results are shown in Figure 2.36. Again silicon is not present in the results due to the thickness of the film. The full result shows many peaks indicating a complex film.

Figure 2.37 shows the hafnium peak, Figure 2.38 shows the aluminum peak, and Figure 2.39 shows the oxygen peak. All of these peaks are well defined and consistent with the film that was deposited. The hafnium is double peaked due to spin orbital decoupling, much like the silicon peak.
Figure 2.37: XPS spectra of the hafnium peak on the aluminum/hafnium oxide layer deposition on silicon wafers. The double peak due to spin orbital decoupling from the hafnium atom are clearly visible.

Figure 2.38: XPS spectra of the aluminum peak on the aluminum/hafnium oxide layer deposition on silicon wafers. Similar to the aluminum only samples no double peak is observed.
Figure 2.39: XPS spectra of the oxygen peak on the aluminum/hafnium oxide layer deposition on silicon wafers.

Figure 2.40: XPS spectra of the oxygen peak on the aluminum/hafnium oxide layer deposition on silicon wafers. In addition to carbon, potassium is also present on the sample.
Figure 2.40 shows the carbon XPS survey on the Al₂O₃/HFO₂ films. The survey shows three peaks. The main peak is the carbon peak and is much higher than the Al₂O₃ film. The other two peaks are due to potassium in the film. The presence of potassium is not fully understand. Since the aluminum oxide film was potassium free, the contamination did not come from the ALD deposition chamber or from shipping, transportation, or handling. The likely cause of contamination is from the precursors and precursor tubes. ALD precursors are regularly switched so one ALD machine can deposit many different films. If the precursory and precursory tubes are not adequately purged gas and material can be left behind and deposited on the wafer during the next run. The contamination stems from an incomplete purge of the precursor tubes.
2.5 Post Processing

Once the gridded capacitors are returned from the ALD blocking oxide deposition the wafers undergo a Rapid Thermal Anneal (RTA) in nitrogen. This RTA step is performed at a temperature of 900°C for 60 seconds, with a 25 second temperature ramp up. The RTA is performed to densify and compact the ALD films which are deposited at a much lower temperature. Next aluminum is deposited by electron beam on both side of the wafer for the gate and substrate contacts. The front gate capacitors are defined and etched. A full and detailed list depicting the gridded capacitor process is presented in Appendix A.

2.6 TEM Characterization of Samples

Transmission electron microscopy is a microscopy technique where a beam of electrons are passed through an ultra thin sample. The electrons interact with the sample as they pass through forming an image that is magnified and focused onto a detector. Since electrons have a smaller de Broglie wavelength, TEMs are capable of producing higher resolution images than traditional light microscopes.

A basic Transmission electron microscopy setup cross-section is shown in Figure 2.41[23]. At the top is the electron gun which produces the electrons. The emission source is usually a tungsten filament or a lanthanum hexaboride source[15]. Electromagnets are used to manipulate the beam while a series of lenses focus the electron beam. A TEM machine typically consists of three stages of lenses, which are the condenser, object, and projector lenses. The condenser lenses are primarily responsible for beam formation. The object lenses then focus the beam that comes through the ultra thin sample. In Scanning Transmission electron microscopy there are also object
Figure 2.41: Schematic of a typical TEM machine. An electron gun produces electrons which are focused through a sample. The scattered electrons are then focused onto a recording system for viewing.[23]
lenses above the sample to make the incident electron beam convergent. Finally the
projector lenses expand the electron beam to form an image on a detector, such as
phosphorus screen, film, or a CCD.

Transmission electron microscopy (TEM) and scanning transmission electron mi-
croscopy (STEM) images were completed on an FEI Tecnai F20 S/TEM system shown
in Figure 2.43.

The Tecnai F20 system is a field emission 200kV S/TEM with an X-TWIN lens
and high brightness field emission electron gun (FEG). This lens allows a 30 degree
tilt with a low background double tilt holder, and allows +/-70 degrees of tilt with the
tomography holder, allowing for efficient collection of X-rays for elemental analysis
down to the sub-nanometer level. The EDS solid angle for collection is 0.3sr, this
results in a factor of 3-4 improvement in EDS collection efficiency when compared
to the more established S-TWIN configuration. The system has a GIF which can
be used between imaging and spectroscopy modes, in spectroscopy mode the energy
resolution is routinely 0.8eV. Scanning Transmission Electron Microscopy (STEM) is
a mode routinely used for analytical experiments, and the X-TWIN system has an
enhanced STEM performance by incorporating a 1.0 mm probe Cs value.

The system can be used to analyze elemental compositions down to the sub-nm
range. The GIF operation is embedded in the microscope user interface, so in TEM
mode it is used just like a CCD camera by the operators. TEM Bright-Field resolution
is 0.25nm point to point, with a line resolution of 0.12nm achievable. While the area
of interest is centered, the mode of the microscope can be changed to STEM mode
for analytical investigations, without losing the area of interest. STEM resolution is
0.18nm on the HAADF detector, and Spectrum Imaging can be performed where the
Figure 2.42: The Tecnai F20 system used for TEM measurements at The Ohio State University.
Figure 2.43: A closer image of the viewing screen on the Tecnai F20 system. A phosphorous screen is used for direct imaging while a computer system also captures and records the image produced.

STEM image, the EDS signal and the PEELS signal can all be obtained at the same time. This is a very powerful way to cross correlate the analytical techniques.[97]

Sample preparation for TEM measurements requires a ultra thin cross section of the device under study. We will show the process using a MOSFET device, and the same process was used to study the gridded capacitor gate stacks. First, we cut a wafer down to an inch by inch square that include the device in the middle. To obtain a cross section we must first dig the device out of the inch by inch square, which is done using a Helios 600 Focused Ion Beam.

Dual Beam FIBs are a relatively new type of instrumentation. They consist of a high-resolution SEM column with a fine-probe ion source (Focused Ion Beam). These instruments allow the preparation of samples from specific areas of a sample
as well as nano-machining. The Helios NanoLab 600 is equipped with an extremely high resolution Elstar electron column with a Field Emission Gun (FEG) electron source. It is capable of \(\pm 1\text{nm} @ 15\text{kV}\) and \(\pm 2.5\text{nm} @ 1\text{kV}\) electron beam resolution. The Ga+ ion source can image and machine down to 5nm resolution levels. Adding to the imaging and sputtering capabilities are a light element X-ray EDS detector and an electron back scatter diffraction orientation imaging camera. The OmniProbe AutoProbe 200 in-situ sample lift-out system allows the preparation of site specific TEM samples without the need for support films. This is necessary to fully utilize the high resolution capability of modern TEMs such as the Titan. The combined sputtering and imaging and analytical capability makes DualBeam FIBs extremely versatile and a key component for TEM studies.[96]
Figure 2.45: TEM image showing the trench dug to find and extract the test sample. The device source and drain contacts are the two large dark rectangles. A protective metal has been deposited over the device.

Once the sample is inside the Helios 600 and brought to vacuum a platinum organic metal is deposited on the surface. This deposition serves two purposes. First it outlines the area where the device is and secondly provides a protective barrier once the sample is extracted from the wafer. Deposition is performed by a localized CVD method. A gas is sourced onto the wafer surface. A low beam electron current is passed through and reacts with the gas particles to deposit the metal on the source. Next, the FIB digs two trenches on either side of the metal deposition area. Both the metal and trenches are shown on in Figure 2.45.
Figure 2.46: A closer view of the test device to be extracted. The source and drain contacts and vias are clearly visible. The gate metal is a small rectangle between the two white vias.

Figure 2.45 is a scanning electron microscope (SEM) image taken at a 52 degree angle on a Northrup Grumman SONOS device. At this angle we can clearly see the device we wish to extract. The two large dark areas are contact lines which extend down to the source and drain of the device. Figure 2.46 shows a closer view of the device. The bright areas at the center of the figure are the vias to the source and drain. Between the two vias the gate metal is visible. The gate stack we wish to see is not visible.
At the 52 degree angle, a FIB cut is made under the device to separate it from the substrate. After this cut the device is only connected to the wafer on either side. A fourth cut is made on one side, leaving only one side connecting the sample to the wafer.

Next, an in-situ probe is lowered down to the device. The probe is very carefully brought into contact with the top metal deposition. Another metal deposition is performed to attach the probe to the sample. Once the sample is connected to the probe the final FIB cut is made to release the sample from the wafer. The chamber is vented and the sample is removed and replaced with sample holder. The sample holder consists of three metal posts. The chamber is again pumped down to vacuum and the sample which was cut out is brought down to one of the metal posts. Through delicate positioning the sample makes contact with one of the posts. A deposition is performed to attach the sample to the post and a FIB cut is made to release the probe from the sample. The final product of the sample attached to the post is shown in Figure 2.47.

The final steps in the Helios 600 require the sample to be shaved till it is thin enough to pass electrons through it. Figure 2.48 shows the start of this process. Both sides of the sample are shaved. On the right top of the sample is where the probe was cut from the sample. The source and drain contacts and vias are visibly seen. On the bottom of the sample is a lot of debris expelled from the FIB cuts. A user must be careful to keep the device area clear of this debris. At first the FIB cuts are at high power and take large chunks from the surface. As the sample becomes thinner and thinner a lower power more precise cut is made to achieve an ultra thin sample.
Figure 2.47: The extracted sample attached to a post. The sample has been slightly thinned.
Figure 2.48: The start of the shaving and thinning process. The device is still invisible. The debris from extraction is shown on the right side of the image.
Figure 2.49: The device thinned to be transparent to electrons. The sample is thinned at a slight angle so the top maintains its shaped. The bottom section as been overly thinned and some holes have appeared. The device to be imaged is in the center of the sample and slightly glows due to electron transmission.

Figure 2.48 shows the sample once it is TEM transparent. This occurs when the sample is roughly 70A thick. The shaving process is done at a slight angle so that the bottom of the sample is thinner than the top. This promotes sample stabilization. The bottom of the sample is only a few angstroms thick and starts to curl. A hole is also clearly seen in Figure 2.49.

The two large rectangles are the contact lines for the source and drain. Below and at the center of these rectangles, another rectangular box is shown. This is the metal gate. The devices gate stack is still not visible at this resolution as it is between the
metal gate and substrate. This area is visibly brighter indicating that electrons are passing through the device and thus the sample is TEM transparent.

Next the chamber is vented and the sample holder is removed and placed in a membrane box shown in Figure 2.50. This box has a very thin membrane suspended in it onto which the sample holder rests. This membrane box allows the sample to be transported with no damage to ultra thin sample. At this point the sample is ready to be imaged with the Tecnai F20 system.

Once the sample is in the Tecnai F20 system we can start to image the device more closely. First, we align the sample into view as shown in Figure 2.51. We make sure the contrast is reasonable and move up in magnification as shown in Figure 2.52.
Figure 2.51: TEM image of a SONOS transistor provided by Northrop Grumman Corp. The two metals for the source and drain contacts are clearly visible. The gate region is above of these two rectangles.
Figure 2.52: TEM image of a SONOS transistor provided by Northrop Grumman Corp. The gate metal is towards the top of the image. The gate stack is not yet visible.
In both Figures, the two large rectangles are the contact lines to the source and drain. Previously, the vias down to the source and drain were visible. Figure 2.51 and 2.52 lacks these features since the sample was thinned down. Between the two contact lines and more towards the top is a smaller rectangle. This is the gate polysilicon. The polycrystalline gate structure is clearly visible in Figure 2.52. At the top of the image is the silicon substrate.

Figure 2.53: TEM image of a SONOS transistor provided by Northrop Grumman Corp. The silicon substrate and poly silicon gate are labeled. The gate is barely visible between these two materials.
Figure 2.54: Maximum resolution TEM image of a SONOS transistor provided by Northrop Grumman Corp. The gate stack has a thickness of roughly 150Å. The tunnel oxide is barely visible and the blocking oxide is roughly 20Å.

Further magnification on the gate region reveals the image in Figure 2.53. The image shows the entire gate region. The polysilicon crystal is visible as well as the substrate. Between these two regions lies the gate stack which we wish to image. At this level of magnification (38000x) the gate stack is barely visible and the individual layers are not discernable.
Figure 2.54 shows the gate stack region at maximum magnification (200,000x). The polysilicon gate is on the left hand side of the image, while the silicon substrate is on the right hand side. Between the two regions lies the gate stack region we wish to image. Starting on the left hand side a darker line is visible. This line is the silicon oxide blocking layer. Using a measuring tool on the Techani system the region has a thickness of roughly 20 Angstroms. The next region is the silicon nitride charge trap region which was measured at 120 Angstroms. Lastly a very thin dark line is visible between the silicon nitride layer and the silicon substrate. This layer is the silicon oxide tunnel oxide and was measured to be roughly 10 Angstroms thick.

We have also performed TEM measurements on the gridded capacitors samples to obtain an accurate representation of the physical layer thicknesses. Figure 2.55 shows the maximum magnification available on the aluminum oxide sample while Figure 2.56 shows the maximum magnification on the aluminum/hafnium oxide sample.

For the aluminum oxide sample in Figure 2.55 five distinct layers are visible. On the bottom of the image is the silicon substrate. A dark thick layer is visible halfway up the image. This layer is the silicon oxide tunnel layer. Above that is the silicon nitride layer. The layer is not well defined and appears rough on the upper portion. This is indicative of an uneven layer and the possibility of processing problems during the LPCVD deposition. The lighter portion of the TEM above the silicon nitride is the aluminum oxide layer. The lighter dots appearing in the picture are not part of the sample but rather random scattering events in the TEM.

With the hafnium oxide/aluminum oxide layer we see some interesting features. Again the bottom of the image is the silicon substrate. The darkest layer is the silicon oxide tunnel layer. Above this region is the silicon nitride layer. Again the
Figure 2.55: TEM image of the gridded capacitor with an aluminum oxide blocking oxide. The silicon substrate is on the bottom of the image with the metal gate on the top. The silicon oxide tunnel oxide is the dark line across the image. The nitride layer is above this with the aluminum oxide layer above the nitride.
Figure 2.56: TEM image of the gridded capacitor with an aluminum/hafnium oxide blocking oxide. The silicon substrate is on the bottom of the image with the metal gate on the top. The silicon oxide tunnel oxide is the black line. The nitride layer is darker in this image with the aluminum/hafnium above this region. At the top of the blocking oxide is a white line which could be a hafnium only layer.
Table 2.3: TEM thickness measurements of the gridded capacitor with a 2.4nm tunnel oxide and aluminium oxide blocking layer and the gridded capacitor with a 4.0nm tunnel oxide and the aluminum/hafnium blocking oxide. Each measurement is done ten times and averaged.

<table>
<thead>
<tr>
<th>Aluminum Oxide</th>
<th>Hafnium Oxide</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO2(nm)</td>
<td>SiN(nm)</td>
</tr>
<tr>
<td>2.678</td>
<td>4.106</td>
</tr>
<tr>
<td>2.5</td>
<td>3.411</td>
</tr>
<tr>
<td>2.321</td>
<td>4.3</td>
</tr>
<tr>
<td>2.143</td>
<td>4.646</td>
</tr>
<tr>
<td>2.857</td>
<td>4.999</td>
</tr>
</tbody>
</table>

Averages in Angstroms

| 23.94 | 43.969 | 174.753 | 40.407 | 58.893 | 160.925 | 32.87 |

The nitride layer appears uneven and rough on the upper part of the layer. Next is the hafnium/aluminum oxide layer. It appears as a solid band and there is no definition of individual layers. This means that during the post processing steps the aluminum oxide and hafnium oxide layer mixed together to form a single blended layer. At the very top of the stack is a thin white layer. This layer is most likely a hafnium layer. During the post processing steps the wafers undergo a high temperature anneal. The anneal is for 90 seconds at 900 degrees Celsius. This is sufficient time and a hot enough temperature for the hafnium atoms to migrate out the stack and form an individual layer.

Thickness measurements are provided in Table 2.3. For the aluminum oxide sample we original had a 24/7/150Å silicon oxide/silicon nitride/aluminum oxide sample
and for the hafnium/aluminum oxide sample we had originally deduced a 40/77/150 Å silicon oxide/silicon nitride/hafnium-aluminum oxide sample. Our TEM measurements exactly match our original measurements through XPS measurements. The silicon nitride TEM measurements are varying and much less than what was originally deposited. This could be due to two reasons.

1. The XPS measurement technique does not work with a silicon nitride layer.

2. The LPCVD process did not produce a high quality silicon nitride layer.

During the measurement we assumed that a silicon nitride atom took up roughly the same space as a silicon oxide atom (i.e., the ratio of silicon nitride atoms to silicon atoms is equal to the ratio of silicon oxide atoms to silicon atoms). This is not true and could throw off our measurements by as much as 10%, which was an acceptable difference. However, the difference from our XPS measurements to the TEM measurements are more than 10%, which brings us to the second point. It is probable that the LPCVD process did not produce a high quality layer. This is readily seen in the TEM pictures as the nitride layer is uneven. We conclude that the differences in layer thickness are the result of both the XPS measurements and an uneven LPCVD process.

2.7 Summary

This chapter has detailed the fabrication process for non-volatile charge trap gridded capacitors. The final capacitors have a variety of silicon oxide tunnel oxide thicknesses, a 7.7 nm silicon nitride charge trap material, and a 150 nm aluminum oxide or aluminum/hafnium oxide blocking layer with a metal gate. Thickness measurements
were taken via XPS for silicon oxide and silicon nitride regions. TEM images were also performed on the gridded capacitors showing the silicon oxide tunnel layer to be 2.4nm. The silicon nitride layer differed in thickness and was measured at between 4.4-6nm, while the blocking oxide had a thickness between 16-17.5nm. For the aluminum/hafnium oxide layer the hafnium has diffused out of the layer towards the metal gate creating a 3nm layer.
Chapter 3: Electrical Characterization Techniques

Electrical characterization of NVSM plays a crucial role in understanding device operation mechanisms, the extraction of device modeling parameters, evaluating the devices working performance, and ultimately achieving device design improvements. A set of characterization techniques has been developed to analyze charge trapping NVSM MOSFETs. We have extended these techniques such that we can use them to obtain the same information with the gridded capacitor structure as with a full NVSM MOSFET. We have classified characterization techniques into two categories: electrical characterization techniques and dynamic electrical characterization techniques.

Electrical characterization techniques are device physics orientated measurements which include C-V measurements and Linear Voltage Ramp measurements. They are used to study electron mobility, gate dielectric capacitance, threshold voltage shifts, etc. Dynamic electrical characterization techniques are device performance orientated measurements which include speed write/erase, retention, and endurance measurements. This chapter focuses on the fundamental electrical characterization techniques. In Chapter 4 we will focus on speed write/erase measurements, and Chapter 5 will include retention and endurance measurements.
3.1 Capacitance-Voltage Measurements

3.1.1 Capacitance-Voltage Theory

Analysis of the behavior of small-signal capacitance variation with a bias voltage on a MANOS capacitors provides a great understanding of the electrical behavior of the device. In this section, we present the theoretical background for calculating the effective dielectric thickness of a MANOS device based on its differential capacitance behavior. The difference capacitance is defined as:

\[ C_{\text{diff}} = \left| \frac{\delta Q_G}{\delta V_g} \right| \]  \hspace{1cm} (3.1)

The capacitance considered in this section is capacitance per unit area. The differential capacitance differs from the static capacitance since the charge on the gate capacitor varies nonlinearly with gate voltage. If the charge in the nitride layer remains constant, classical MOS capacitance models are still applicable with the simple modification of \( C_{\text{ox}} = C_{\text{eff}} \). Therefore, the differential capacitance \( C_{\text{diff}} \) can be viewed as the series combination of the fixed gate dielectric capacitor \( C_{\text{eff}} \) and the variable silicon capacitor \( C_{\text{Si}} \) as follows:

\[ \frac{1}{C_{\text{diff}}} = \frac{1}{C_{\text{eff}}} + \frac{1}{C_{\text{Si}}} \]  \hspace{1cm} (3.2)

\[ C_{\text{Si}} = \left| \frac{\delta Q_s}{\delta \psi_s} \right| \]  \hspace{1cm} (3.3)

The evaluation of \( C_{\text{Si}} \) requires representing \( Q_s \) as a function of \( \psi_s \). The potential \( \psi_s \) can be obtained from Poisson’s equation, which includes the contribution of electrons and holes at the Si-SiO\(_2\) interface. By using an integrating factor, Poisson’s equation
can be made an exact differential which permits integration. In addition, the use of boundary conditions at the Si-SiO$_2$ interface and at the edge of the depletion region permits obtaining the electric field $E_s$ at the Si-SiO$_2$ interface[25].

The charge in the silicon layer per unit area is obtained from Gauss’s law as:

$$Q_s = -K_{Si} \epsilon_o E_s = \pm \frac{\sqrt{2} K_{Si} \epsilon_o kT}{q L_D} \sqrt{\left[ \frac{q \psi_S}{kT} + e^{-\frac{q \psi_S}{kT}} - 1 \right] + \left( \frac{n_i}{N_a^-} \right)^2 \left[ e^{\frac{q \psi_S}{kT}} - \frac{q \psi_S}{kT} - 1 \right]}$$

(3.4)

where

$$L_D = \sqrt{\frac{K_{Si} \epsilon_o kT}{q^2 N_a^-}}$$

(3.5)

is the Debye Length for p-Si with p=N$_a$=N$_a^-$. The Debye length represents a distance over which the free carriers reduce the potential from the fixed impurity ions, and the reduced potential would decay as $\exp \left[-x/L_d\right]$. Next by substituting Equation 3.4 into Equation 3.3 we obtain:

$$C_{Si} = \frac{K_{Si} \epsilon_o}{\sqrt{2} L_D} \frac{1 - e^{-\frac{q \psi_S}{kT}} + \left( \frac{n_i}{N_a^-} \right)^2 e^{\frac{q \psi_S}{kT}}}{\sqrt{\left[ \frac{q \psi_S}{kT} + e^{-\frac{q \psi_S}{kT}} - 1 \right] + \left( \frac{n_i}{N_a^-} \right)^2 \left[ e^{\frac{q \psi_S}{kT}} - \frac{q \psi_S}{kT} - 1 \right]}}$$

(3.6)

Equation 3.6 is applicable in the accumulation, depletion, flat-band, and inversion region. In the accumulation and depletion region, when $\psi_s < 0$, Equation 3.6 can be further simplified by neglecting the term:

$$\left( \frac{n_i}{N_a^-} \right)^2 e^{\frac{q \psi_S}{kT}}$$

(3.7)

so that a simple expression for $C_{si}$ can be written as:

$$C_{Si} = \frac{K_{Si} \epsilon_o}{\sqrt{2} L_D} \frac{1 - e^{-\frac{q \psi_S}{kT}}}{\sqrt{\left[ \frac{q \psi_S}{kT} + e^{-\frac{q \psi_S}{kT}} - 1 \right]}}$$

(3.8)
For the flat-band condition ($\psi_s=0$), we can expand $e^{-\frac{q\psi_S}{kT}}$ to second order and obtain $C_{\text{diff}}$ as

$$C_{\text{diff}}(FB) = \frac{1}{X_{\text{eff}} K_{OX} \epsilon_o + \frac{L_D}{K_{Si} \epsilon_o}}$$

(3.9)

For the accumulation condition ($\psi_s < 0$), the term $e^{-\frac{q\psi_S}{kT}}$ becomes dominant so that $C_{\text{diff}}$ can be written as:

$$C_{\text{diff}}(\text{Accumulation}) = \frac{1}{X_{\text{eff}} K_{OX} \epsilon_o + \frac{L_D}{K_{Si} \epsilon_o} \frac{\sqrt{2}}{e^{\frac{q|\psi_S|}{kT}}}}$$

(3.10)

For the strong accumulation condition $|\psi_S| > \frac{6kT}{q}$, the second term in the denominator rapidly becomes smaller than $\frac{X_{\text{eff}} K_{OX} \epsilon_o}{K_{Si} \epsilon_o}$ so that:

$$C_{\text{diff}}(\text{StrongAccumulation}) = \frac{K_{OX} \epsilon_o}{X_{\text{eff}}} = C_{\text{eff}}$$

(3.11)

From the above analysis, we can see that $C_{\text{eff}}$ keeps increasing from the flatband to the accumulation region, and reaches its maximum during the strong accumulation condition. Therefore, for a p-type SONOS/MONOS NVSM device, $X_{\text{eff}}$ can be calculated from the maximum value on the strong accumulation region of a classical C-V curve [25]

### 3.1.2 Gridded Capacitor Capacitance-Voltage Results

Capacitor-Voltage (C-V) tests are extremely powerful semiconductor material and device characterization techniques. The measurement is enhanced with the gridded capacitor structure and the implanted grid. The grid lines are doped with the opposite carrier type than the substrate which allows both carrier types to be present. The
Figure 3.1: Capacitance vs Voltage measurement on a gridded capacitor with a capacitor radius of 100um, a grid spacing of 50um, and a measurement frequency of 100kHz. The accumulation capacitance is the same as the inversion capacitance. Additional carriers allow for both an accumulation and inversion layer to form across the entire surface quickly under appropriate bias conditions. A well designed gridded capacitor should have an adequate supply of both carriers. When a high frequency C-V measurement is taken the grid allows for the inversion capacitance to be the same as the accumulation capacitance. This is shown in Figure 3.1.

Our mask set has varying grid spacing sizes of 30, 40, 50, 70, 100, 200, and 300µm. An optimal grid spacing needs to be determined. In inversion the minority carriers are supplied by the grid lines and spread out under the entire capacitor to form an inversion layer. If the grid lines are too far apart, the minority carriers will not be
Figure 3.2: Capacitance vs Voltage measurements on gridded capacitors. With a 300um grid spacing the inversion capacitance is lowered, while with a 30um grid spacing the accumulation capacitance is lowered. A grid spacing of 50-100um is ideal.

able to completely cover the entire capacitor area at high frequencies. This reduces the capacitance under surface inversion, but have no effect on the capacitors in the accumulation region. In accumulation, the grid lines become depleted at the surface while the non grid region has holes at the surface. When there are too many grid lines under the capacitor there is a larger depletion region and the capacitance in accumulation will be lower. The capacitor will be unaffected in the inversion region as more grid lines just allows the inversion region to form more easily.
Figure 3.2 shows two tests completed on two capacitors with a radius of 150\(\mu m\). The capacitor radius remained constant while one test was completed on a 300\(\mu m\) grid spacing and the other was completed on a 30\(\mu m\) grid spacing. For the capacitor on a grid spacing of 30\(\mu m\) the capacitance is reduced, while the capacitance is reduced in the inversion region with a capacitor on a grid spacing of 300\(\mu m\). Figure 3.1 shows a C-V measurement on a 100\(\mu m\) radius capacitor on a 50\(\mu m\) grid. Both the accumulation and inversion capacitances are the same. Grid spacings of 50\(\mu m\), 70\(\mu m\), and 100\(\mu m\) all show C-V measurements with the same capacitance in accumulation and inversion regions. Ideal gridded capacitors therefore have a grid spacing of between 50\(\mu m\) and 100\(\mu m\).

Additionally, it should be noted that some C-V measurements encountered a slight dip towards the inversion region as shown in Figure 3.3 and 3.4.
Figure 3.3: Capacitance vs Voltage measurement on a gridded capacitor with aluminum oxide as the blocking oxide. A dip is observed in the inversion region due to a trap state away from the silicon/silicon oxide interface.

Figure 3.3 is on a capacitor with grid spacing of 100µm, a capacitor radius of 250µm and an Al₂O₃ blocking oxide. Figure 3.4 is on a similar capacitor except the blocking oxide is the aluminum-hafnium blend. Since the dip is seen on both capacitors the dip is not caused by the blocking oxide. The dip is caused by a trap that is a distance away from the silicon/silicon oxide interface. We conclude that this is an interface trap at the silicon oxide/silicon nitride interface. As discussed in Chapter 2 the nitride layer had processing issues which have further manifest themselves into trap states.
Figure 3.4: Capacitance vs Voltage measurement on a gridded capacitor with aluminum/hafnium oxide as the blocking oxide. The same dip is in the inversion region is observed as is likely from a trap state near the silicon oxide/silicon nitride interface.
3.2 Linear Voltage Ramp

3.2.1 Linear Voltage Ramp Theory

The linear voltage ramp (LVR) measurement is a quasi-static C-V measurement. Instead of measuring the differential capacitance ($C_{\text{diff}}$) of the gate dielectric, we characterize the current flow through the gate when a linear bias is swept on the bulk.

$$\alpha = \frac{\delta V_{BG}}{\delta t}$$ (3.12)

The ramp rate, $\alpha$, is small and ranges from $60 \text{ mV/sec}$ to $300 \text{ mV/sec}$. The ramp rate is kept small to maintain quasi-thermal equilibrium in the device. This measurement can be used to observe the memory window of the device. Also, the flatband voltage shift observed can be related to the charge inject into the nitride layer.

The LVR setup used in our experiments is shown in Figure 3.5. In this setup, a slowly ramping voltage is applied directly to the substrate of the device, by a function generator (HP 33120A). The gate current is then measured by a Keithley 617 Digital Electrometer. Both the function generator and the electrometer are controlled and monitored by LabVIEW™.

The gate voltage applied is given by Equation 3.13

$$V_{BG} = -V_{GR} = V_o \pm \alpha t$$ (3.13)

If we neglect the parasitic components in the differential capacitance[112], the relationship between the LVR and C-V behaviors of a charge trapping NVSM device is given by Equation 3.14

130
Figure 3.5: Linear Voltage Ramp Setup on a transistor. A function generator is tied to the source/drain/substrate while an electrometer measures the current out of the gate. The results are graphed in LabVIEW. For a gridded capacitor sample the source/drain and substrate are already tied together.
\[ I_G = \frac{\delta Q_G}{\delta t} A_g = \frac{\delta Q_G \delta V_{GB}}{\delta t} A_g = \mp \alpha C_{diff} A_g \] (3.14)

At the strong accumulation region, \( C_{eff} \) can be calculated from the gate current since \( C_{diff} = C_{eff} \). Since we have a gridded capacitor structure, we can obtain the value of \( C_{eff} \) at the point when the device is biased in the inversion region. The current flow through the bulk can be formulated by means of calculating the derivative of \( Q_G \) with respect to \( t \).

\[ V_{GS} = V_{GB}(V_{SB} = 0) = V_{FB} + \phi_S - \frac{Q_S}{C_{eff}} - \frac{Q_{it}}{C_{eff}} \] (3.15)

\[ Q_G + Q_S + Q_{ss} + Q_{it} = 0 \] (3.16)

\( Q_S \) is the total semiconductor charge density, \( Q_{it} \) is the interface trap charge, and \( Q_{ss} \) is the equivalent charge at the Si-SiO\(_2\) interface due to the fixed charge in the blocking and tunneling oxides. By differentiating both sides of Equation 3.15

\[ \alpha = V'_{FB} + \phi'_S - \frac{Q'_S}{C_{eff}} - \frac{Q'_{it}}{C_{eff}} \] (3.17)

\[ Q'_s = -(I_G + Q'_N) \] (3.18)

\[ V_{FB} = \phi_{GS} - \frac{Q_{ss}}{C_{eff}} - \int_0^{x_N} \left( \frac{x_{OB}}{K_{OX}\varepsilon_o} + \frac{x_N - x}{K_N\varepsilon_o} \right) \rho_N(x)dx \]

\[ = \phi_{GS} - \frac{Q_{ss}}{C_{eff}} - \left( \frac{x_{OB}}{K_{OX}\varepsilon_o} + \frac{x_N - \bar{x}}{K_N\varepsilon_o} \right) Q_N \] (3.19)

If we substitute Equation 3.17 into Equation 3.18 and use the expression in 3.19 we have

\[ I_G = C_{eff} \left[ \alpha - \phi'_S - \frac{\delta (x_{OT})}{\delta t} \int_0^{x_N} \left( \frac{x_{OT}}{K_{OX}\varepsilon_o} + \frac{x}{K_N\varepsilon_o} \right) \rho_N(x,t)dx \right] \] (3.20)
Figure 3.6: Example of an LVR measurement. At high voltages electron tunneling increases creating spikes.

where the first term $C_{eff} \alpha$ represents the displacement current, and the last term relates to the changing rate of charge in the nitride layer.

Figure 3.6 shows an example LVR measurement. The value in strong accumulation or weak inversion is the effective gate dielectric capacitance $C_{eff}$ times the ramp rate $\alpha$. This corresponds to the case when the last two terms in Equation 3.20 are zero. In strong inversion there are two effects which can cause an increase in gate current. A rise in gate current can indicate that electrons in the nitride have begun to migrate to the gate electrode. Mathematically, this is seen when the later term in Equation 3.20 decreases leading to a larger gate current. Another issue occurs when the gate bias increases leading to an increase in the electric field across the blocking oxide.
The electric field lowers the energy barrier to electrons and tunneling begins to take place from the nitride region to the gate. The sharp increase in gate current towards the sides of the LVR curve are due to the tunneling of electrons out of the nitride to the gate.

The flatband voltage shift can also be observed and quantified in the LVR measurement. As the voltage increases from $V_{fb1}$ to $V_{fmax}$, electrons tunnel from the substrate to the nitride region through the tunnel oxide. After the gate bias reaches the maximum set voltage for the test, the electron current quickly reduces to zero.

As the gate voltage starts to decrease, electrons no longer tunnel to the nitride layer due to the internal electric field generated by the trap electrons. A component of the reverse current is caused by holes back tunneling from the nitride to the substrate. This component can be detected by a dual transistor structure. Therefore, for voltages between the two flatband voltages, the injected electrons and the back tunneling holes, can be determined by integrating the electron and hole currents ($I_n$ and $I_p$)[81]

\[
\delta Q_{n}^{inj} = \frac{1}{\alpha A_G} [(\int_{V_{FB1}}^{V_{Gmax}} I_N dV_G) - (\int_{V_{Gmax}}^{V_{FB2}} I_N dV_G)] \tag{3.21}
\]

\[
\delta Q_{p}^{inj} = \frac{1}{\alpha A_G} [(\int_{V_{FB1}}^{V_{Gmax}} I_P dV_G) - (\int_{V_{Gmax}}^{V_{FB2}} I_P dV_G)] \tag{3.22}
\]

Since the measured gate current can be viewed as the combination of hole and electron currents, the sum of equation 3.21 and 3.22 corresponds to the area $A_1 - A_2$ as shown in Figure 3.6. The charge transfer can be related to the flatband shift by the following equation:
\[ \Delta V_{FB} = \Delta Q^{inj}_n \left[ \frac{x_N - x_{inj}^n}{K_N \varepsilon_o} + \frac{x_{OB}}{K_{OX} \varepsilon_o} \right] + \Delta Q^{inj}_p \left[ \frac{x_N - x_{inj}^p}{K_N \varepsilon_o} + \frac{X_{OB}}{K_{OX} \varepsilon_o} \right] \] (3.23)

In the equation above, \( x_{inj}^n \) is the injected electron charge centroid and \( x_{inj}^p \) is the back-tunneled hole charge centroid. If we assume a uniform trap distribution in the nitride at high injection levels the charge centroid will approach half the nitride film thickness. Therefore, \( A_1 - A_2 \) can be approximated as the net injected electrons, which then relates the band shift as

\[ \Delta V_{FB} = \Delta Q^{inj}_n \left[ \frac{x_N}{2K_N \varepsilon_o} + \frac{x_{OB}}{K_{OX} \varepsilon_o} \right] = (A_1 - A_2) \left[ \frac{x_N}{2K_N \varepsilon_o} + \frac{x_{OB}}{K_{OB} \varepsilon_o} \right] \] (3.24)

This approximation allows us to evaluate the flat band shift \( \Delta V_{fb} \) from Figure 3.6. Also \( A_1 - A_2 \) can be obtained by integrating the gate current with respect to the gate bias. Therefore, by changing the upper limit of the gate ramp range, we can calculate \( A_1 - A_2 \) and the change in flatband voltage. If we were to graph the change in flatband voltage versus the area \( A_1 - A_2 \) we obtain a linear relationship with the slope equal to Equation 3.25

\[ \frac{x_N}{2K_N \varepsilon_o} + \frac{x_{OB}}{K_{OX} \varepsilon_o} \] (3.25)

3.2.2 Linear Voltage Ramp Measurements on Gridded Capacitors

Figure 3.7 and 3.8 show the effect of keeping a constant lower voltage \( V_{gmin} \) and increasing \( V_{gmax} \). Figure 3.7 is on the 24/77/150Å gate stack with a grid width of 30um and a capacitor radius of 40um. Figure 3.8 is also on same gate stack with a
Figure 3.7: LVR measurements on a gridded capacitor with 2.4/7.7/15.0nm silicon oxide/silicon nitride/aluminum oxide stack and a grid width of 30um and a capacitor radius of 40um.

Figure 3.8: LVR measurements on a gridded capacitor with 2.4/7.7/15.0nm silicon oxide/silicon nitride/aluminum oxide stack and a grid width of 30um and a capacitor radius of 70um.
grid of 30um, but has a capacitor radius of 70um. Due to the larger area, the current is greater in Figure 3.8. At the low voltages of -7 and +7 volts, we are able to obtain a window close to 4V.

### 3.3 Mobility

The gridded capacitor can be analyzed with transmission line theory [12, 88] applied to the equivalent circuit shown in Figure 3.9. We assume the source, drain, and substrate are all electrically tied together without any bias applied and the inversion charge \( Q_n \) is distributed uniformly under the gate electrode. For the gridded capacitor structure the minority carrier grid lines traverse the entire wafer, which permits the implanted grid lines to establish the same potential between the lines and the substrate due to the high leakage and capacitance of the grid. Figure 3.9 shows two currents \( i_1 \) and \( i_2 \) form when a small AC voltage is applied to the gate. The current \( i_1 \) is the minority carrier current and flows from source and drain through the channel and provides minority carriers while \( i_2 \) represents majority carriers from the substrate.

We can solve for \( i_1 \) and \( i_2 \) for a rectangular gate MOSFET using basic transmission line theories[76]. The details of this result are found in Chao and Wang[12].

\[
i_1 = -j\omega WL \frac{C_o C_I}{C_o + C_D + C_I} \frac{\tanh \lambda}{\lambda} \nu_o \tag{3.26}
\]

\[
i_2 = -j\omega WL \left[ \frac{C_o C_D}{C_o + C_D} - \frac{C_o C_D C_I}{(C_o + C_D)(C_o + C_D + C_I)} \frac{\tanh \lambda}{\lambda} \right] \nu_o \tag{3.27}
\]

and the propagation constant \( \lambda \) is given by

\[
\lambda = (j\omega \tau_{CG})^{1/2} \tag{3.28}
\]
Figure 3.9: Transmission line equivalent circuit for a transistor [8]. Using the minority and major carrier currents $i_1$ and $i_2$, the total capacitance and conductance per unit length can be obtained. From the peak in the conductance vs. frequency curve a channel mobility can be determined.

where the channel time constant $\tau_{CG}$ is defined as

$$\tau_{CG} = \frac{C'\rho I L^2}{4} = \frac{C' L^2}{4Q I \mu I}$$ (3.29)

$$C' = \frac{C_I(C_o + C_D)}{C_o + C_D + C_I} = \frac{C_{GC} C_o + C_D}{C_o}$$ (3.30)

The capacitance and conductance per unit area at the terminal calculated from $i_1$ alone are

$$C_1 = \frac{1}{WL} IM[\frac{i_1}{\omega v_o}] = C_{GC} Re[\frac{\tanh \lambda}{\lambda}]$$ (3.31)

and

$$\frac{G_1}{\omega} = \frac{1}{WL} Re[\frac{i_1}{\omega v_o}] = -C_{GC} Im[\frac{\tanh \lambda}{\lambda}]$$ (3.32)
where $C_{GC}$ is the gate-channel capacitance. The capacitance $C_1$ converges to $C_{GC}$ at low frequencies since $Re\left(\frac{\tanh\lambda}{\lambda}\right) \approx 1$ for a very small $\lambda$. Similarly, the total terminal capacitance and conductance per unit area resulting from $i_1$ and $i_2$ are

$$
C_T = \frac{1}{WL} Im\left[\frac{i_1 + i_2}{\omega v_o}\right] = C_{GC} \frac{C_o}{C_o + C_D} Re\left[\frac{\tanh\lambda}{\lambda}\right] + C_{HF}
$$

(3.33)

$$
\frac{G_T}{\omega} = \frac{1}{WL} Re\left[\frac{i_1 + i_2}{\omega v_o}\right] = -C_{GC} \frac{C_o}{C_o + C_D} Im\left[\frac{\tanh\lambda}{\lambda}\right]
$$

(3.34)

where $C_{HF}$ is the high-frequency capacitance given by

$$
C_{HF} = \frac{C_o C_D}{C_o + C_D}
$$

(3.35)

It can also be shown that at very low frequencies

$$
C_T = C_{LF} = \frac{C_o(C_I + C_D)}{C_o + C_I + C_D}
$$

(3.36)

$C_T$ degenerates to the ideal low-frequency capacitance. The function $Im\left(\frac{\tanh\lambda}{\lambda}\right)$ in Equation 3.32 exhibits a peak value of -0.4172 at $\omega \tau_{CG} = 2.54$. Therefore, both $\tau_{CG}$ and $C_{GC}$ can be obtained from the peaks of the $\frac{G_1}{\omega}$ versus $\omega$ plots. The channel mobility is then obtained from [76]

$$
\mu_I = C_{GC} \frac{C_o + C_D}{C_o} \frac{L^2}{4Q_I \tau_{CG}}
$$

(3.37)

where $Q_I$ is determined from $C_{GC}$. The ratio $\frac{C_o + C_D}{C_o}$ can be obtained from Equation 3.32 and Equation 3.34

$$
\frac{C_o + C_D}{C_o} = \frac{G_I}{G_T}
$$

(3.38)

where $G_I$ and $G_T$ are data points at 1MHz and are used to reduce the interface trap effects.
The mobility in the inversion layer, at a given gate bias, can be expressed as

$$\mu_I = \frac{\omega_p L^2 (1 + \frac{C_D}{C_D})}{4(2.54)(V_{GS} - V_{TH})} \approx \frac{\omega_p L^2}{4(2.54)(V_{GS} - V_{TH})}$$  \hspace{1cm} (3.39)

We have applied the G-V mobility extraction to a full SONOS device as a reference check. Using a poly-silicon – silicon oxide – silicon nitride – silicon oxide – p-type silicon substrate with thicknesses of 4.5nm/6.5nm/2.0nm, respectively, capacitance and conductance sweeps were completed. The device exhibited a peak at 10.5MHz which is shown in Figure 3.10. The calculated mobility is then $\mu_{inv} = \frac{322 \text{ cm}^2}{\text{Vs}}$. The gate area was 100$\mu$m by 100 $\mu$m and $V_{gs} - V_{TH} = 2V$ for the sample. The conductivity mobility for the transistor is defined as:
Figure 3.11: Schematic and TEM of the gridded capacitor under test. The grid spacing is not visible in the TEM image.

\[ \mu_{con} = \frac{I_D}{W C_{eff}(V_{GS} - V_{TH})V_D} \]  

(3.40)

This equation gives a calculated mobility of $318 \frac{cm^2}{Vs}$ at $V_{gs} - V_{TH} = 2V$ and $V_{DS} = 50mV$ for the same device. Thus, there is excellent agreement between the transmission-line extracted mobility and the conductance mobility.

In our second set of experiments we used the gridded capacitor structures. The gate stack in the gridded capacitor consisted of a high-quality 2.4nm SiO$_2$ tunnel oxide thermally grown in a triple-wall oxidation furnace, a 7.7nm silicon-rich nitride deposited by Low Pressure Chemical Vapor Deposition (LPCVD) and a 15nm Al$_2$O$_3$ blocking or cap oxide deposited with atomic layer deposition. A cross section and STEM image of the fabricated device is shown in Figure 3.11.
Figure 3.12 shows the G-ω measurement with a grid spacing of 100µm and a capacitor radius of 250µm. A peak is observed at 3.46MHz. Previously, Equation 3.39 was derived for rectangular devices, such as individual MOSFETs with L defined as the length of the channel. The grid structure is rectangular so no modification is necessary in the derivation, however the length of the channel needs to be modified. The simplest modification is to map the rectangular device to a circular structure with an equivalent source at the center and drain at the edge. This mapping leads to an equivalent L equal to the radius of the capacitor. Using this mapping and Equation 3.40 with a peak at 3.6MHz, \( V_{TH} = 4 \) V and \( L=250\mu m \) yields a mobility of \( 334 \frac{cm^2}{Vs} \).

In our experiments we noted the grid spacing and number of grids affected the mobility results. Our mask set has grid spacings of 30, 40, 50, 70, 100, 200, and 300 µm. The capacitor radii vary on each grid; however, there is always a 250 µm radius capacitor on each grid. In Figure 3.13 the mobility results obtained on a 250 µm radius capacitor are graphed vs. grid spacing. The mobility values are constant between the 50 µm and 100 µm grid spacing, which represents a range of optimal grid spacing. A larger than optimal grid spacing results in transit time limitations as the minority carriers cannot respond to the applied AC signal. A smaller than optimal grid spacing reflects the situation where the grid lines begin to dominate the capacitor area, which can be seen in Figure 3.13. For the larger grid spacing of 300 µm the inversion capacitance does not equal the accumulation capacitance as minority carrier response times influence the measurement. From Figures 3.12 and 3.2, we show the optimal grid spacing is between 50 µm and 100 µm. The optimal range likely continues a bit past 100 µm, but our mask does not allow for finer detail.
Figure 3.12: Gm vs Frequency curve on the gridded capacitor. The device exhibits a peak at 3.46MHz.
Figure 3.13: Mobility measurements vs Grid Spacing. Too many grids reduces the mobility, while too few grids has an increased mobility. The optimal grid spacing lies between 50-100um.
3.4 EOT Calculations

We have shown that the effective oxide thickness (EOT) is the sum of the all the layers in the gate stack related to the SiO$_2$ tunnel oxide. This equation is reproduced in Equation 3.41

$$EOT = X_o + \frac{K_o X_N}{K_N} + \frac{K_o X_I}{K_I}$$  \hspace{1cm} (3.41)

By using Equation 3.42 and solving for the EOT we obtain Equation 3.43

$$C_{inv} = \frac{K_o \epsilon_o A}{EOT}$$  \hspace{1cm} (3.42)

$$EOT = \frac{K_o \epsilon_o A}{C_{inv}}$$  \hspace{1cm} (3.43)

By equating the two EOTs, we can obtain an equation and solve for the $K_I$ value.

The final $K_I$ formula is shown in Equation 3.44

$$K_I = \frac{K_o X_I}{\frac{K_o \epsilon_o A}{C_{inv}} - X_o - \frac{K_o X_N}{K_N}}$$  \hspace{1cm} (3.44)

where $K_o$ is the dielectric constant of SiO$_2$, $X_I$ is the insulator thickness, $\epsilon_o$ is the permittivity of free space, $A$ is the capacitor area, $X_N$ is the silicon nitride thickness, $K_N$ is the dielectric constant of the silicon nitride, $X_o$ is the SiO$_2$ tunnel oxide thickness, and $C_{inv}$ is the inversion capacitance. From our previous XPS measurements, we have the values for the thicknesses. From the C-V measurements we can obtain a value for $C_{inv}$. The dielectric constants for the SiO$_2$ and silicon nitride have been found previously and their values are 3.9 and 6.5, respectively. We can now insert all the values and find what the dielectric constant is for both the Al$_2$O$_3$ blocking oxide. 145
and for the Al$_2$O$_3$/HfO$_2$ blocking oxide. Various $K_I$ values have been calculated and graphed below in Figure 3.14 for Al$_2$O$_3$ and Figure 3.15 for Al$_2$O$_3$/HfO$_2$.

In both figures, there is a strong dependence on capacitor area. As the area gets larger, the extracted $K_I$ value also becomes larger until, at roughly 100µm capacitor area, it levels off. This observation is attributed to the grid lines. As the capacitor area gets smaller, the grid width remains the same. Therefore, the smaller capacitor area has more of the doped grid lines underneath it. The grid lines are affecting the C-V measurement at small capacitance. As the capacitor area gets larger, the grid lines have less of an effect, and the extracted $K_I$ value flattens out to the true $K_I$ value. From these measurements, the $K_I$ value for Al$_2$O$_3$ is approximately 8, while
Figure 3.15: EOT calculations on the gridded capacitor with a gate stack containing an aluminum/hafnium oxide blocking layer. The K value is smaller for smaller capacitances. The K value of the layer is roughly 14.
the $K_1$ value for $\text{Al}_2\text{O}_3/\text{HfO}_2$ blocking oxide is roughly 14. Therefore, we conclude that the addition of $\text{HfO}_2$ into the blocking oxide has a positive effect on increasing the dielectric constant of the layer.

### 3.5 Summary

In this section we have provided the theory for Capacitance-Voltage and Linear Voltage Ramp measurements. We have completed both of these measurements on the gridded capacitor structure. The grid spacing is especially important for C-V measurements. We have shown that the optimal grid spacing is between 50µm and 100µm as this is when the accumulation capacitance is equal to the inversion capacitance. This region is ideal as the carrier concentrations from the grid lines and the substrate are roughly the same and one does not dominate the other.

We also completed mobility measurements using the gridded capacitor structure. Using a combination of C-V and G-V measurements we can extract the carrier mobility as a function of vertical electric field. This gridded capacitor structure eliminates the need for source/drain contacts, which lends itself well to material systems where low resistance, source/drain contacts are difficult to implement or reduce device performance.
Chapter 4: Speed Write/Erase

To further evaluate, understand, and optimize SONOS NVSM devices we examine the speed/write mechanisms. Theoretical models have been developed over the years to analyze the speed/write operations. Libsch and White [55] proposed a model based on charge injection from the semiconductor bulk and gate electrodes with numerical simulations. French and White [21] obtained a simple closed-form solution valid for short programming time, with a relationship between the turn-on time, the write/erase slope, the dielectric thickness, initial stored charges in the nitride, and program voltages. Gritsenko et al [24] simulated the write/erase characteristics of SONOS devices with high-K dielectrics by solving five coupled differential equations numerically. They showed that the injection current through the top dielectric is reduced when a high-K material is used as the blocking oxide. Lee et al [48] compared the erase saturation level of SANOS and SONOS devices with a simple simulation model. They concluded that a lower gate injection current leads to a reduced erase saturation level with Al$_2$O$_3$ as the blocking oxide.

This chapter improves a speed simulation model by French and White [21] and Gan Wang and White [99] through speed and optimization improvements. We then apply this model to write/erase characteristics for conventional SONOS transistors.
as well as MANOS gridded capacitors. We highlight the use of gridded capacitors being able to perform the same characterization operations as transistor devices.

4.1 Write/Erase Speed Theory

Quantum mechanical tunneling is the only method of charge transport in the write/erase operation of SONOS/MANOS NVSM devices. Several tunneling mechanisms may take place in an NVSM device which include Fowler-Nordheim Tunneling (FN), Direct Tunneling (DT), Modified Fowler-Nordheim Tunneling (MFN), and Trap Assisted Tunneling (TAT). Which tunneling mechanisms takes place is determined by the electric field across the dielectrics and the energy band bending due to the applied field. Figure 4.1 shows the difference between each tunneling operation. In Figure 4.1 $\psi_1$ and $\psi_2$ are the barrier heights between the substrate to tunnel oxide and the tunnel oxide to nitride, $\psi_t$ is the trap energy level in the nitride referenced to the nitride conduction band, $X_1$ and $X_2$ are the thicknesses of the tunnel oxide and the nitride, and $K_1$ and $K_2$ are the dielectric constants for the two films, respectively.

Under a high electric field the band bending is strong enough to allow the carriers to tunnel through a triangle barrier. This condition of tunneling through a triangle barrier is known as Fowler-Nordheim Tunneling. A decrease in the electric fields allows the band bending to relax and carriers need to tunnel through the entire barrier layer, a condition known as Direct Tunneling. Further reduction in the electric field causes Modified Fowler Nordheim Tunneling. This mechanism occurs when carriers tunnel through a full film layer($I_1$) as well as a triangular barrier of a second film($I_2$). Direct tunneling can also occur through two full films ($I_1$ and $I_2$). The last tunneling mechanism is Trap Assisted Tunneling. Trap Assisted Tunneling occurs at low electric
Figure 4.1: Energy band diagrams and oxide electric field requirements for Fowler-Nordheim, Direct Tunneling, Modified Fowler-Nordheim, and Trap Assisted Tunneling. [116]
fields for thicker I$_2$ films when the tunneling path meets a trap level in the charge storage layer I$_2$.

In the NVSM tunneling derivation in [99] the tunneling currents are calculated assuming a WKB approximation [20] and Lundstrom’s[93,57] derivations for trap-assisted tunneling. The tunneling currents $J_{OT}$ and $J_{OB}$ are expressed using the WKB approximation[55,21,64] as

$$J_{OT}(E_{OT}); J_{OB}(E_{OT}) = C_T P_{OX} P_N$$ (4.1)

Where $C_T$ is the current constant, $P_{OX}$ is the tunneling probability through the tunneling oxide and $P_N$ is the tunneling probability through the nitride. The three parameters have different values for the different tunneling mechanisms.

For Fowler-Nordheim the parameters are

$$C_T = \frac{q^2 E_{OT}^2}{16\pi^2 \hbar \phi_1}$$ (4.2)

$$P_{ox} = \exp - \frac{4\sqrt{2qm_{ox}}}{2\hbar E_{OT}} \phi_1^{3/2}$$ (4.3)

$$P_N = 1$$ (4.4)

For Direct tunneling the parameters are

$$C_T = \frac{q^2 E_{OT}^2}{16\pi^2 \hbar} \left[ \phi_1^{1/2} - (\phi_1 - |E_{OT}|X_{OT})^{1/2} \right]^{-2}$$ (4.5)

$$P_{ox} = \exp - \frac{4\sqrt{2qm_{ox}}}{3\hbar E_{OT}} \left[ \phi_1^{3/2} - (\phi_1 - |E_{OT}|X_{OT})^{3/2} \right]$$ (4.6)
\[ P_N = 1 \] (4.7)

And for Modified Fowler-Nordheim tunneling the parameters are

\[ C_T = \frac{q^2 E_{OT}^2}{16\pi^2 h} [\phi_1^{1/2} - (\phi_1 - |E_{OT}|X_{OT})^{1/2} + \left(\frac{\phi_1}{\phi_1 - \phi_2}\right)^{1/2} (\phi_1 - \phi_2 - |E_{OT}|X_{OT})^{1/2}]^{-2} \] (4.8)

\[ P_{ox} = \exp - \frac{4\sqrt{2}q m^*_{ox}}{3h|E_{OT}|[\phi_1^{3/2} - (\phi_1 - |E_{OT}|X_{OT})^{3/2}]} \] (4.9)

\[ P_N = \exp - \frac{4\sqrt{2}q m^*_{ox}}{3h|E_{OT}|} \frac{\phi_1}{\phi_1 \phi_2} (\phi_1 - \phi_2 - |E_{OT}|X_{OT})^{3/2} \] (4.10)

For Trap Assistant tunneling the approach from Svennson and Lundstrom[93] is used.

\[ J_{TAT} = \frac{q N_t}{\tau_0 (q\gamma |E_{OT}|kT - 2k_n)} \exp \left[ -2(k_{ox} - \frac{k_n}{\gamma})X_{OT} - \frac{2k_n \phi_3}{\gamma |E_{OT}|} \right] \] (4.11)

Where

\[ k_{ox} = \frac{\sqrt{2m^*_{ox} q\phi_1}}{\hbar} \] (4.12)

\[ k_n = \frac{\sqrt{2m^* q\phi_t}}{\hbar} \] (4.13)

\[ \phi_3 = \phi_1 - \phi_2 - \phi_t \] (4.14)

\[ \gamma = \frac{\epsilon_{OT}}{eN} \] (4.15)

\( \tau_0 \) is a time constant on the order of \( 10^{-14} \) to \( 10^{-12} \) second, \( N_t \) is the density of nitride traps, \( \psi_t \) is the nitride trap energy level, \( \psi_1 \) is the barrier height of Si-SiO\(_2\), and \( \psi_2 \) is
the barrier height of SiO$_2$-Si$_3$N$_4$. The nitride effective mass $m^*_n$ can be related to the oxide effective mass $m^*_{ox}$[55] with Equation 4.16

$$
\frac{m^*_n}{m^*_{ox}} = \left( \frac{\epsilon_{OT}}{\epsilon_N} \right)^2 \frac{\phi_1}{\phi_1 - \phi_2}
$$

(4.16)

The tunneling model first discussed [99] is based on a NVSM devices with tunneling as the only charge transport mechanism for write/erase operations. The tunneling currents for both write and erase operations for a conventional p-type device are shown in Figure 4.2.

For the write operation a positive voltage is applied to the gate. During the write operation electrons tunnel through the tunnel oxide by Band-to-Band tunneling ($J_{OB,n}$) and/or Trap Assisted Tunneling ($J_{TAT,n}$). Hole back tunneling ($J_{BT,p}$) from the nitride to the channel is small and is therefore neglected in the model. Electron tunneling from the nitride through the blocking oxide to the gate electrode ($J_{OB,n}$) is possible at higher programing voltages and high electron current injection from
the channel. Under the conditions of small electron injecting and low programming voltages with relatively thick blocking oxides this current can be neglected.

In the erase operation a negative voltage is applied to the gate. During the erase operation holes tunnel across the tunneling oxide by Band-to-Band ($J_{OT_p}$) or Trap Assisted tunneling ($J_{TAT_p}$). Tunneling by electrons back to the channel ($J_{BT_n}$) is considered to be small and is neglected. To further simplify the model, interface traps and all other fixed charge in dielectrics are neglected, and stored charges are treated as a sheet charge in the nitride. This model simplification is shown in Figure 4.3.

We start by applying Gauss’s law to the interfaces of the ONO structure with a centroid model described by:
\[ E_{OT}(t_p)\epsilon_{OT} = E_{N1}(t_p)\epsilon_N \quad (4.17) \]

\[ E_{OB}(t_p)\epsilon_{OB} = E_{N2}(t_p)\epsilon_N \quad (4.18) \]

\[ E_{N2}(t_p) = E_{N1}(t_p) - \frac{Q_n(t_p)}{\epsilon_N} \quad (4.19) \]

where \( E_{OT}, E_{OB}, E_{N1}, E_{N2} \) are the electric fields in the tunnel oxide, blocking oxide, and inside and outside sections of the nitride respectively. Summing up the potentials across the device dielectrics allows for programming voltage to be written as shown in Equation 4.20

\[ V_p(t_p) = \phi_{GS} + \phi_S + E_{OT}(t_p)X_{OT} + E_{OB}(t_p)X_{OB} + E_{N1}(t_p)\bar{X} + E_{N2}(t_p)(X_N - \bar{X}) \quad (4.20) \]

where \( V_p(t_p) \) is the programming voltage and \( \bar{X} \) is the trap centroid in the nitride.

Substituting Equation 4.17, 4.18, 4.19 into Equation 4.20 leads to:

\[ V_P(t_p) = \phi_{GS} + \phi_S + E_{OT}(t_p)\epsilon_{OT}X_{OT} + \left( E_{OB}(t_p)\epsilon_{OB}X_{OB} + E_{N1}(t_p)\epsilon_NX_N - \bar{X}\right)Q_n(t_p) \quad (4.21) \]

where

\[ X_{eff} = X_{OT} + \frac{\epsilon_{OT}}{\epsilon_N}X_N + \frac{\epsilon_{OB}}{\epsilon_{OB}}X_{OB} \quad (4.22) \]

is the effective oxide thickness of the gate stack. Differentiation of Equation 4.21 gives:

\[ X_{eff} \frac{\delta E_{OT}(t_p)}{\delta t_p} = \left( \frac{X_{OB}}{\epsilon_{OB}} + \frac{X_N - \bar{X}}{\epsilon_N} \right) \frac{\delta Q_n}{\delta t_p} \quad (4.23) \]
Using current continuity, where the sum of the currents is equal to the change in charge we can write:

\[
\frac{\delta Q_N(t_p)}{\delta t_p} = -J_{OT}[|E_{OT}(t_p)|] - J_{TAT}[|E_{OT}(t_p)|] + J_{OB}[|E_{OB}(t_p)|] \tag{4.24}
\]

where \(J_{OT}\) is the band-to-band tunneling current from the substrate, \(J_{TAT}\) is the trap assisted current from the substrate, and \(J_{OB}\) is the band-to-band tunneling current from the gate electrode. We can related the electric field across the blocking oxide to the electric field across the tunnel oxide with Equation 4.25

\[
E_{OB}(t_p) = E_{OT}(t_p) \frac{\epsilon_{OT}}{\epsilon_{OB}} - \frac{Q_N(t_p)}{\epsilon_{OB}} \tag{4.25}
\]

By substituting Equation 4.24 into 4.23 gives Equation 4.26

\[
\frac{\delta E_{OT}(t_p)}{\delta t_p} = -AJ_{OT}[|E_{OT}(t_p)|] - J_{TAT}[|E_{OT}(t_p)|] + J_{OB}[|E_{OB}(t_p)|] \tag{4.26}
\]

where

\[
A = \left(\frac{X_{OB}}{\epsilon_{OB}} + \frac{X_N - \bar{X}}{\epsilon_N}\right)/X_{eff} \tag{4.27}
\]

Equation 4.26 can be transformed by separation of variable to yield the following:

\[
\int_{E_{OT}(0)}^{E_{OT}(t_p)} J_{OT}[|E_{OT}(t_p)|] - J_{TAT}[|E_{OT}(t_p)|] + J_{OB}[|E_{OB}(t_p)|]^{-1}dE_{OT} = -At_p \tag{4.28}
\]

We now use numerical integrations and iterations to solve Equation 4.28 which provides the tunnel oxide electric field \((E_{OT}(t_p))\) at a given programming time \(t_p\). The change of the threshold voltage can be derived as a change in the tunnel oxide electric field.
The inversion capacitance does not equal the accumulation capacitance due to the lack of carriers in the substrate.

\[
\delta V_{th}(t_p) = V_{th}(t_p) - V_{th}(0)
\]

\[
= - \left( \frac{X_{OB}}{\epsilon_{OB}} + \frac{X_N - \bar{X}}{\epsilon_N} \right) [Q_N(t_p) - Q_N(0)]
\]

\[
= -X_{eff} [E_{OT}(t_p) - E_{OT}(0)]
\]  \hspace{1cm} (4.29)

### 4.2 Flatband Tracking Circuit

A basic method of characterization of SONOS devices is to study the C-V characteristics of the device. The basic C-V theory was described in the previous chapter. As a refresher, these experiments apply a DC sweep with a superimposed small AC signal to the gate of the device and measures the gate capacitance. A basic C-V measurement is shown in Figure 4.4.
We can extract memory characteristics of a charge trap device by measuring the shift in the C-V curve. To quantify charge trap memories write/erase characteristics we can apply known programming pulses to the device and measure the changes in the C-V curve.

During the write/erase operation the device undergoes an injection of electrons or holes in the charge trap layer. This is a net gain or lose in the overall charge in the device. For this reason, when write/erase pulses are applied to the device the C-V curve shift is always parallel to the original curve. Since the curve is parallel, measuring the change in voltage at any point on the curve is equal to the change in the flatband voltage, or the change in the threshold voltage. Therefore, by applying pulses to the device and measuring the change in the voltage at any point on the curve we can obtain the write/erase characteristics of the device.

The present method of obtaining write/erase characteristics is slow and unreliable. This process is shown in Figure 4.6.
The first component of the system is the program/erase pulse applicator circuit. The circuit generates erase and write pulses which are then applied to the device under test on the probe station. One the pulses are applied the circuit is manually disconnected and the device is then attached to a 4192A CV Analyzer. The CV Analyzer applies the gate sweep of the device. The data is then transmitted to National Instruments LabVIEW\textsuperscript{TM} program. Once the data is collected the pulse circuit is applied to the device and a write measurement can be taken.

The whole process has a variety of negative aspects. First the method requires a lot of manually disconnecting and connecting of devices which slows the process down. Many pulses and measurements need to be completed on a single device and this current method can be extremely slow. Secondly, the method is unreliable as the C-V curve is not measured immediately after the applied pulse. During the time interval that the manual connections are made charge has already rearranged itself.
Figure 4.7: Front panel of the Boonton Capacitance Meter. The meter has two ports for connecting capacitance. The external voltage bias port is in the back of the meter.

and left the device. Therefore a new circuit is required to apply pulses and measure the C-V curve without any manual connections.

The redesigned circuit first starts with how the C-V sweep is performed. In reality, the sweep does not need to be completed. We only need to select a known capacitance on the C-V curve and measure the voltage at that capacitance before and after the applied pulses. To accomplish this we use a Boonton Capacitance meter shown in Figure 4.7

The Boonton capacitance meter has built in additional functions that make it ideal for our particular use case. As shown in Figure 4.7, the front of the meter has a panel that reads the capacitor values and has TEST and DIFF capacitor terminals that can read the individual capacitor values attached to those terminals. When a capacitor is connected to TEST and another is connected to the DIFF terminal, the
meter has the ability to read out the differential capacitance of the two attached capacitors.

The meter also has an analog voltage output that is directly proportional to difference between the attached capacitors. The output voltage depends is scaled since the maximum voltage output of the Boonton meter is 2V. The analog output of x pico-Farads in a 200pF scale which is given by the follow:

\[ V_{out} = \frac{x}{200}2V \]  

(4.30)

In addition to this features, the Boonton meter also has an external bias that is directly connected to the TEST terminal. This can be used to bias the gate terminal of CMOS devices. We can now measure a capacitance value while applying a DC bias externally. All of these mentioned properties are used in the new flatband tracking circuit.

From Figure 4.9 the initial bias applied to the CMOS capacitor at the TEST terminal is zero. When a fixed capacitor (reference) value is connected to the DIFF terminal, the meter reads the differential capacitance, as an example positive, at the front panel. We can first take a C-V sweep of the device and select a reference capacitance in the middle of the curve, preferably at the point of greatest slope. The Boonton produces a positive analog output, due to the differential capacitance, that is then applied to the negative terminal of the differential operational amplifier. The output of the amplifier is therefore negative and the integrator changes the voltage in the positive direction as long as the differential output is not zero. When the circuit becomes stable the differential output is zero, the integration has stopped, and the digital voltmeter reads the voltage corresponding to the reference capacitor value.
Figure 4.8: Boonton circuit to change applied voltages. The front has the device under test connect as well as a reference capacitance. A diff amplifier and an integrator change the applied voltage until the difference between capacitances is zero.
The full flatband tracking circuit is shown in Figure 4.9. The bottom left contains the boonton proportional of the circuit mentioned above. Various connectors to LabVIEW™, power supplies, and the probe stand are located around the perimeter. The additional circuitry is for applying the write and erase pulses.

The main chip associated with pulse application is the ADG451BR shown in Figure 4.10. The chip has four independent switches that closes once a voltage lower than 2V is applied to the corresponding input (IN) terminal. For example the D1 terminal is connected to the voltage source while the S1 terminal is connected to the gate of the charge trap memory. The voltage applied to the input terminal IN1 is provided by LabVIEW™ through a SCB-68 connector board. When the voltage signal on the IN1 terminal goes low the switch connects D1 with S1 allowing a positive
Figure 4.10: Schematic of the ADG451SR switch. A voltage is applied to open the correct port to the device. Write, erase, and read signals are applied through the switch.
voltage to be applied to the device. All of the S terminals are connected to the device while the other switches are used to apply a negative voltage or a read pulse. In the case of a read signal, no voltage is applied. Rather, we connect it to the boonton circuitry mentioned previously to measure the flat band voltage and log the change in the voltage.

The LabVIEW™ controllers all the inputs and timing needed. The front panel of the erase program is shown in Figure 4.11. Pulse widths, read delay, and how many read cycles to perform are entered into the front panel. Voltages are set on the individual power supplies. For each point the program takes an average of 5 read cycles. The information is the graphed in real time, allowing the user to stop the program earlier if something seems amiss. A separate write programs logs the data for the write state of the device.
Figure 4.12: Final flatband tracking circuit board. The connections around the board go to the probe stand or LabVIEW. The connections in the top left are for applying write or erase voltages.
4.3 Write/Erase Speed Measurements

In order to understand the device respond rate, we need to examine the influence of the write/erase pulse width on the threshold voltage shift for different write/erase pulse magnitudes. Therefore, the speed measurement result presents the threshold voltage shift versus the write/erase pulse width relationship. We use the flatband voltage tracking circuit, as well as LabVIEW™, to apply various pulse widths and pulse amplitudes to the gate stack and record the change in the flat band voltage. First, a C-V measurement is performed on the device. A reference capacitance is chosen where the C-V measurement has the steepest downward slope. This reference capacitor is placed onto the Boonton meter, as previously described, to seek out the flat band voltage.

Each data point is recorded after a pre-cycle pulse is applied to the device. The pre-cycle pulses have to ensure that the device under test (DUT) can be initialized to a given written or erased state. For all of our samples, we use a 100ms write pulse before an erase and a 100ms erase pulse before a write to initialize the device. After the pre-cycle pulse is applied, the write or erase pulse is applied and then the voltage is recorded. The write and erase pulses are varied from 1E-7 seconds to 1E-1 seconds. The speed write and speed erase curves are measured from two different programs. The two measurements should be taken alternately in case of a threshold voltage shift, which would be caused by the residual charge resulting from consistent write or erase operations. A detailed description of the test procedures can be found in Appendix E.

For a given write/erase bias condition, we can obtain some important information about the device under test from a pair of write/erase speed curves. First, we can
examine the initial state of the device. The starting point of the write curve can be regarded as the erased state, while the starting point of the erase curve can be regarded as the written state. The initial pulse times of 1E-7 seconds are too short to change the amount of charge in the nitride layer from the previous cycle. This allows us to check the accuracy of the speed measurement. It is abnormal if the write curve flatband voltage for the longest pulse time is not near the initial erase state, or if the erase curve flatband voltage for the longest pulse time is not near the initial write state. The discrepancy usually results from a residual charge effect. The pre-cycle pulse is not writing or erasing to the same state. This effect can be reduced by adjusting the pre-cycle pulse width to a longer time.

Additionally, we can examine the cross over point of the write and erase curves. This point gives an indication of the overall speed performance of the device for a given bias voltage. Generally, the shorter the cross-over point, the faster the device responds.

Finally, the curve slope gives a large indication of the device performance. The slope of the write/erase curve is the threshold voltage shifting rate as the write/erase pulse width increases. For a MANOS NVSM device, the slope of the write/erase curves decreases with an increase in the pulse width. Several mechanisms contribute to this effect: 1) as the electrons or holes are injected into the nitride layer, the injected charge establishes an electric field against the gate bias created field, which decreases the charge injection efficiency; 2) the amount of traps for electrons or holes is limited, so as more charge is injected there are less traps available to tunnel to reducing the tunnel current; 3) opposite charge which tunnels through the blocking oxide to the nitride layer neutralizes the charge injected from the substrate. Sometimes,
the write/erase curve flattens for longer pulse widths, which is the charge injection saturation effect.

In the operation of MANOS devices, charge injection is governed by the tunnel process, which is a strong function of the electric field [21]. Therefore, the magnitude of the applied programming voltage and the amount of the initial nitride charge are important parameters in determining the speed of a MANOS device. By comparing the write/erase curves using different write/erase voltages, we can choose appropriate bias conditions to optimize the speed performance of the device.

In Figures 4.13 and 4.14 we can see the results for a variety of low voltage speed write/erase tests for aluminum oxide and hafnium/aluminum oxide samples, respectively.

In these figures, we can see some of the major differences between the two samples. The cross over point on aluminum oxide gridded capacitor is around 80ms pulse width at 8V write and -8V erase, while on aluminum/hafnium oxide gridded capacitor the cross over point is at 10ms pulse width, with the same voltages. Therefore, we can deduce that the Al₂O₃/HfO₂ blocking oxide has increased the dielectric constant of the oxide, which then allows more voltage to be applied over the tunnel oxide. Because of this configuration, more band bending occurs, making faster speed write/erase curves. Also, the initial and final states for the Al₂O₃/HfO₂ blocking oxide curves are further apart. This indicates that more charge is allowed to cross into the nitride over the given pulses. Furthermore, it shows that the Al₂O₃/HfO₂ blocking is superior to the Al₂O₃ blocking oxide, even while operating at low voltages.

If we start to increase the voltages, which allows us to reduce the time it takes to write or erase the devices, we obtain the curves depicted in Figures 4.15-4.18. The
Figure 4.13: Low voltage speed write/erase curves on 24/77/150A silicon/nitride/aluminum gate stack. Write and erase curves are shown for applied voltages of 6, 7, and 8V.

Graphs are separated into write curves in Figure 4.15 and 4.16 and erase curves in Figure 4.17 and 4.18 for aluminum and aluminum/hafnium oxide gridded capacitors, respectively.

Focusing on the erase curves in Figure 4.17 and 4.18 first, we can see that the curves are very similar. The sample with the Al$_2$O$_3$/HfO$_2$ blocking oxide has higher initial erase states. This was seen previously on the low voltage tests. More importantly, the erase curves go down to -14V on aluminum oxide sample and only -12V on aluminum/hafnium oxide sample. This is due to the fact that hafnium/aluminum capacitors would break down above 12V write or erase. The aluminum only gridded capacitors were only limited to the voltages of the test circuit and did not show any signs of breakdown. With voltages above 12V on the hafnium/aluminum, the
Figure 4.14: Low voltage speed write/erase curves on 24/77/150A silicon/nitride/hafnium-aluminum gate stack. Write and erase curves are shown for 6, 7, and 8 applied voltages.

Figure 4.15: Write curves on a gridded capacitor with the aluminum oxide blocking oxide. Results are shown for 9, 10, 11, 12, 13, and 14 volts.
Figure 4.16: Write curves on a gridded capacitor with the aluminum/hafnium oxide blocking oxide. Results are shown for 9, 10, 11, and 12 volts.

Figure 4.17: Erase curves on a gridded capacitor with the aluminum oxide blocking oxide. Results are shown for -9, -10, -11, -12, -13, and -14 volts.
Al₂O₃/HfO₂ blocking oxide, the device would start to write or erase and then fail as the pulse width became large. Due to the higher dielectric constant of the blocking oxide, more voltage is applied across the tunnel oxide. Taking a write curve as an example, more and more charge enters into the nitride as the pulse widths get longer. As this occurs, there is less of a potential difference across the tunnel oxide, yet the same voltage is being applied to the device. Therefore, the voltage starts to increase across the blocking oxide as the pulse width gets longer. This increase in field is enough to break down the oxide at voltages above 12V. This could be mainly due to the increase dielectric constant, or a challenge with the oxide itself such as poor interfaces between the Al₂O₃ and HfO₂ layers.

Moving on to the write curves in Figure 4.15 and 4.17, we start to see some other interesting effects. Again, the aluminum oxide gridded capacitors has a few more high
voltage curves than the hafnium/aluminum gridded capacitors due to the breakdown recently discussed. The main effect to notice is the initial starting voltages of the curves. They are not related to the final voltages recorded in Figure 4-9 and 4-10. Also, one should notice that the curves start to decrease and then begin to increase. This is a major flaw in the flatband voltage tracking circuit. As discussed at the start of the chapter, the circuit locks onto a fixed capacitor value and reports the voltage at that capacitance as the device is written or erased. Looking at the C-V curves in Chapter 3, one can see that there are two voltages that have the same capacitance. The one on the left side of the curve is the flatband voltage that we wish to track, but there is also another capacitance with a different voltage which is on the right side of the curves. When the voltages become high, we start shifting the C-V curve far to the left for an erase state, and far to the right for a write state. One can see that the initial erase states and the final write states are the same for each of the samples. For erasing a device, we first initialize the device with a 100ms write pulse. This action shifts the curve to the right. Then, the flatband voltage tracking circuit always starts its search for the correct capacitance at a voltage of zero. Since the C-V curve was shifted to the right, the first voltage it sees with the correct capacitance is on the left side of the curve, and it is the correct flatband voltage. The circuit locks onto this capacitance and tracks the associated voltage extremely well. Now if we wish to obtain a write curve, we first initialize the device with a 100ms erase pulse. This shifts the curve to the left. Again, the flatband voltage tracking circuit starts searching for the correct capacitance at zero voltage and the first capacitance it sees is the capacitance on the right side of the C-V curve and NOT the flatband voltage which we wish to track. It locks onto the incorrect capacitance and starts to output

175
Figure 4.19: Flatband shift for aluminum oxide gridded capacitors on the write state.

an incorrect voltage. For this reason, the initial write and final erase points do not match up. As the pulse widths get longer, the circuit jumps back onto the left side of the C-V curve and starts tracking the correct flatband voltage. This can be seen in the write curves as the dip at the start of the curves. This is why the initial erase and final write do match up, since the circuit, at some point, starts to track the correct voltage.

As discussed previously, we are not interested with the exact flatband voltage of the device, we are only interested in the change of the flatband voltage. Therefore, we need to augment the previous figures and depict the change in the flatband voltage as it is related to the pulse width. This is shown below in Figures 4.19, 4.20, 4.21, and 4.22
Figure 4.20: Flatband shift for hafnium/aluminum oxide gridded capacitors on the write state.

Figure 4.21: Flatband shift for aluminum oxide gridded capacitors on the erase state.
4.4 Simulation of Non Volatile Charge Trap Memories

Using the equations and theory presented in the first section of this chapter we have created simulation programs for non-volatile charge trap memories. These programs have been used to simulate the results in the previous section. We first discuss the simulation of transistor devices, and then show that the simulation programs can also be used for the gridded capacitor structure. The programs are written in Matlab with the source code provided for in Appendix F.

We will walk through the erase.m program. The first step is to define all the constants. Comments in the source code start with a %.
clear all;

%set up constants
q=1.6E-19;
hbar=1.055E-34;
KbTq=0.0259;
m0=9.1094E-31;
mox_e=0.35*m0; %effective electron mass
mox_h=0.42*m0; %effective hole mass
phi1_e=3.1; %oxide barrier heights for electron
phi1_h=3.65; %oxide barrier heights for hole
phi2_e=1.05; %nitride barrier heights for electron
phi2_h=1.85; %nitride barrier heights for hole
phi3_e=2.98; %Al2O3 barrier height for electron
Xot=24E-10; %tunnel oxide thickness
Xn=44E-10; %nitride thickness
Xob=150E-10; %blocking oxide thickness
ep0=8.8542E-12; %vacuum permittivity
epSi=11.7*ep0; %permittivity of Si
epox=3.9*ep0; %permittivity of oxide
epn=6.5*ep0; %permittivity of nitride
epob=7.9*ep0; %permittivity of Al2O3
Na=5E22; %substrate doping
ni=1.202E16; %intrinsic doping
phif=KbTq*log(Na/ni); %bulk feimi level
Eg=1.12; %energy gap of Si
mn_e=(epox/epn)^2*phi1_h*mox_e/(phi1_e-phi2_e); %effective hole mass in the nitride
mn_h=(epox/epn)^2*phi1_h*mox_h/(phi1_h-phi2_h); %effective hole mass in the nitride
Xeff=Xot+(epox/epob)*Xob+(epox/epn)*Xn;
Ceff=epox/Xeff;
phigs=-0.304; %gate to semiconductor workfunction difference
Nt=4E25; %trap density
phit=1; %trap energy
gama=2; %nitride centroid coefficient (2: center of the nitride; 1: SiO2, SiN interface; infinite: SiN, blocking oxide interface)

Next, we set up some initial conditions. Two initial conditions are required to obtain accurate simulation results. First, we need to know what programming voltage will be used. Secondly, we need to know the initial threshold voltage of the device. The
initial threshold voltage is needed to calculate the amount of charge that is already present in the device. The programs assume that the device is already written or erased into a known state.

%initial conditions

\[ \text{Vp} = -7; \quad \text{\%erase voltage (set as needed)} \]
\[ \text{Vth0} = 1; \quad \text{\%initial threshold voltage} \]
\[ \text{phis} = -\left( \frac{\text{Eg}}{2} - \text{phif} \right); \quad \text{\%surface potential at accumulation} \]

Next we calculate the electric fields across the tunnel and oxide layers. We use the starting threshold and programming voltage to determine the amount of charge already present in the nitride region. With the calculation of the electric fields we calculate the associated tunnel currents. The tunneling current calculations take place in separate programs which act like a method or function call in other programming languages. The tunneling current program takes in variables for the electric field, layer thicknesses, and barrier heights and returns the tunnel current represented by those variables. The tunneling program calculates which type of tunneling process (FN, Direct, or MFN). The trap assisted tunneling current is represented by equation 4.11.

\[ \text{Qn0} = \frac{\phi_{\text{igs}} + 2\phi_{\text{if}} + \left( 4\epsilon_{\text{Si}}qN_{\text{a}}\phi_{\text{if}} \right)^{1/2}}{\epsilon_{\text{eff}}} - \frac{V_{\text{th0}}}{X_{\text{ob}}/\epsilon_{\text{ob}} + X_{\text{n}}/(\gamma_{\text{a}}\epsilon_{\text{nn}})}; \quad \text{\%initial nitride charge} \]
\[ \text{Eot0} = \frac{\left( \text{Vp} + \left( X_{\text{ob}}/\epsilon_{\text{ob}} + X_{\text{n}}/(\gamma_{\text{a}}\epsilon_{\text{nn}}) \right) \text{Qn0} - \phi_{\text{igs}} - \phi_{\text{is}} \right)}{X_{\text{eff}}}; \quad \text{\%initial tunnel oxide electric field} \]
\[ \text{Eob0} = \text{Eot0} \left( \frac{\epsilon_{\text{ox}}}{\epsilon_{\text{ob}}} \right) - \frac{\text{Qn0}}{\epsilon_{\text{ob}}}; \quad \text{\%initial blocking} \]
\[ \text{Eot0} = -\text{Eot0}; \]
\[ \text{Eob0} = -\text{Eob0}; \]
\[ \text{Jot0} = \text{Jbtb}(\text{Eot0}, \text{Xot}, \text{Xn}, \text{mox}, \phi_{\text{1}} \_h, \phi_{\text{2}} \_h); \quad \text{\%Initial band to band tunneling through the tunnel oxide} \]
Job0 = Jbtb(Eob0,Xob,Xn,mox_e,phi1_e,phi2_e); %Initial band to band tunneling through the blocking oxide
Jtat0 = Jtat(Eot0,Xot,mox_h,mn_h,phi1_h,phi2_h,phit,Nt); %Initial trap assisted tunneling through the tunnel oxide
a=(Xob/epob + Xn/(gama*epn))/Xeff;

Then we open a file text file to write our results to. We print out column headers as well as the initial conditions. We also define a variety of array variables. The main variable in this section is Time. This array depicts how many points will be calculated. For these simulations we take 10 points for each factor of time.

Fid=fopen(’al203_21_9_doping2_7v_8.txt’,’wt’); %Open writable text file
fprintf(Fid,’tp, dVth, Eot, Eob, Jot, Job, Jtat, loops, intgr
’); %Write headings into text file
fprintf(Fid,’%e, %e, %e, %e, %e, %e, %e, %d, %e
’, 0, 0, Eot0, Eob0, Jot0, Job0, Jtat0, 0, 0); %Write the initial conditions into the text file

% initialize array size variable
n = 0;
Eot = [Eot0];
Time = [1E-7:1E-7:1E-6,2E-6:1E-6:1E-5, 2E-5:1E-5:1E-4, 2E-4:1E-4:1E-3,
2E-3:1E-3:1E-2, 2E-2:1E-2:1E-1, 2E-1:1E-1:1];
dVth = [0];
signal = 0;

Next, we have the main iteration part of the simulation program. This loop shown in its completion below goes through each time point and calculates the electric fields present in at the device.

for t = Time
%initialize variables
n=n+1;

Eot1 = Eot0;
Eob1 = Eot1*(epox/epob) + (-Eot1*Xeff + phigs +phis - Vp) / (Xob + (epob/epn)*(Xn/gama)); %Calculate blocking oxide field from tunnel oxide field
Ess = CalE4(Eot0,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_h,phi2_h,phi3_e);
%Calculate minimum electrical field across the tunnel oxide
Eot3 = Ess;
Eob3 = Eot3*(epox/epob) + (-Eot3*Xeff + phigs +phis - Vp) / (Xob + (epob/epn)*(Xn/gama)); %Calculate blocking oxide field from this tunnel oxide
%Now have a minimum electric field and a higher electric field due to the applied programming voltage
%Set some initial variables
Diff=100;
Accuracy=10;
loops(n)=0;
if signal == 0
    while abs(Diff) > Accuracy
        Eot2 = Eot3 + (Eot1-Eot3)/2; %Calculate new tunnel oxide field that is midway between the low and high electrical fields
        Eob2 = Eot2*(epox/epob) + (-Eot2*Xeff + phigs + phis - Vp) / (Xob + (epob/epn)*(Xn/gama)); %Calculate the new blocking oxide field
        %Calculate the new tunneling currents
        Jot2 = Jbtb(Eot2,Xot,Xn,mox_h,phi1_h,phi2_h);
        Job2 = Jbtb(Eob2,Xob,Xn,mox_e,phi1_e,phi2_e);
        Jtat2 = Jtat(Eot2,Xot,mox_h,mn_h,phi1_h,phi2_h,phit,Nt);
        intgr =
            Simpson_erase6(Eot0,Eot2,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_h);
            %Looking for this equation to be less than the set accuracy, at which point we would have a solution and the correct electric fields. IF Diff is positive, replace the high E-field with new %one, else replace the low E-field with the new one. Repeat till %solution.
        Diff = intgr + a*t;
        if Diff > 0
            Eot1 = Eot2;
        else
            Eot3 = Eot2;
        end
    end
    loops(n)=loops(n)+1; %This is here if the above loop does not converge after 50 loops and just uses the minimum field.
%Never had the erase state get to this point, combined program will %look like write state.
if loops(n)>50
    Diff=0;
    signal = 1;
    Eot2 = Ess;
For each time, T, we wish to calculate the electric field across the gate stack. Once we calculate the new electric field after the carriers have tunneled into the nitride we can obtain the change in threshold voltage from the change in electric field. The first part calculates two tunneling electric fields. The first is the maximum electric field, which is our original electric field, since it is assumed that we start in a state with carriers already in the nitride. The second electric field is a minimum electric field. This calculation narrows the range in which we need to search for a solution.
The inner loop calculates the actual electric field by using Simpson’s rule. Simpson’s rule is a Newton-Cotes formula for approximating the integral of a function using quadratic polynomials. Simpson’s rule states:

\[
\int_{X_0}^{X_2} f(x)dx = \int_{X_0}^{X_0+2h} f(x)dx \approx \frac{1}{3} h(f_0 + 4f_1 + f_2) \tag{4.31}
\]

This approximation is used to solve the integral in 4.28. The calculation takes place in the Simpson write or Simpson erase Matlab files. The upper and lower electric fields are used as the bounds of the integral. We are looking for the difference between the integral calculation and the other side of the equation to be less than a set accuracy. If the difference is positive the upper electric field is set to the electric field between the lower and upper bounds. If it is negative, the lower electric field is switched out and the calculation is repeated. The loop runs until a solution is found. If after 50 times through the inner loop no convergence occurs the loop is terminated and the lower electric field is used. This is a safety value in case the integration gets stuck between two points.

Once the desired accuracy is achieved, the loop ends and the final electric fields and tunnel currents, and change in threshold voltage at that particular time are calculated. The time is outputted to the console so the user knows the program is still running. The values are also exported to the file. Then start the process over for the next time T and continue until each point is calculated. Finally, the file that we were writing to is closed and the results are graphed.

```matlab
fclose(Fid); %Close file for writing

%Plot the results
Time = [0, Time];
figure;
semilogx(Time, dVth, 'r');hold on;
```

184
The simulation programs have accurately been used against production transistors [99]. The simulation programs have also been applied to non-volatile charge trap memories provided by Northrop Grumman Corporation. Figure 4.23 shows the results for a NVSM transistor with a 1.8nm silicon tunnel oxide, 18nm silicon nitride region, and a 4nm blocking oxide. It should be noted that the given thicknesses are not measured by TEM. As shown in Chapter 2 the actual results do not always line up with what is originally planned. It is important to have accurate thickness results. Figure 4.23 demonstrates that many combinations of gate stack thicknesses can lead to 'accurate' simulation results. In this instance, the likely gate stack is closer to 18/140/31Å than the originally reported 18/180/40Å.

Figure 4.23 shows what happens when the nitride layer is reduced, or only the voltage is decreased. As devices progress the applied voltage must be decreased. The voltage should be at or below 5V so that the memory can be easily integrated with low power systems. When the nitride layer is reduced by 10Å there is a significant shift to a quicker write speed. However, if the applied voltage is decreased by 1V there is a larger shift towards slower write speeds. 5V operation on these devices is not possible by reducing the nitride layer alone. The tunnel oxide is very thin and can not be reduced further without severely affecting retention characteristics, likewise with the blocking oxide. Due to these restrictions the only way to achieve 5V operation or lower is to investigate alternative high K oxides for use in the gate stack.
Figure 4.23: Write curve and simulations on a Northrop Grumman charge trap non-volatile memory. It is important to know the thicknesses of the gate stack for accurate simulations since many gate stack combinations can produce the same curve. To speed up the device, the nitride layer can be reduce. Reducing the voltage will slow the device down.
Figure 4.24: Erase curve and simulations on a Northrop Grumman charge trap non-volatile memory. The original program did not include electron back tunneling. Once it is added and exact match is shown.
Figure 4.24 shows the erase curve of the same Northrop device. The applied voltage is -10V. Again, the originally reported thickness is 1.8/18/4nm silicon oxide/silicon nitride/silicon oxide. However in the write curves we used 1.8/14/3.1nm. The simulation was done with these thicknesses. The original simulation in red shows a large difference at later times than the actual measurements.

This is due to electron back tunneling. In our original tunnel current calculations at the beginning of this chapter this tunneling was not taken into account with our equations. This current is easily added as another term in Equation 4.28.

\[
J_{BTn}(E_{OT}, E_n) = \frac{Q_n X_n}{\tau_0} \left[ \int_0^{X_t} e^{-\alpha_{ox} X_{ot}} e^{-\alpha_n X} dx + e^{-\alpha_{ox} X_{ot}} e^{-\alpha_n X_t} \right] (4.32)
\]

\[
\alpha_{ox} = \frac{2\sqrt{2q m^*_{oxe}(\phi_e + \phi_t)}}{\hbar} \left( 1 - \frac{1}{4} \frac{E_{ot} X_{ot}}{\phi_e + \phi_t} - \frac{1}{2} \frac{E_n X}{\phi_e + \phi_t} \right) (4.33)
\]

\[
\alpha^*_ox = \frac{2\sqrt{2q m^*_{oxe}(\phi_e + \phi_t)}}{\hbar} \left( 1 - \frac{1}{4} \frac{E_{ot} X_{ot}}{\phi_e + \phi_t} \right) (4.34)
\]

\[
\alpha_n = \frac{2\sqrt{2q m^*_{oxe}(\phi_t)}}{\hbar} \left( 1 - \frac{1}{4} \frac{E_n X}{\phi_e + \phi_t} \right) (4.35)
\]

With the addition current, the results shown in green in Figure 4.24 are obtained. This shows how easy it is to modify the program depending on the needs of the device. Each change in carriers can be modeled as a current which can easily be added as an additional term. Each additional term will cause the simulation programs to run longer, and so a balance between number of needed terms and accuracy needs to be determined by the user.
Figure 4.25: Experimental and simulation write and erase curves performed at 8V on a gridded capacitor with an aluminum oxide blocking oxide. The gate stack is 2.4/7.7/15nm silicon oxide/silicon nitride/aluminum oxide. The simulation curves are an excellent match to the experimental curves.

In continuation with the gridded capacitor structure, speed write and erase calculations were performed. Accurate gate stack thickness measurements from our TEM images must be obtained and those thicknesses are used in our calculations. Figure 4.25 shows the write and erase curves taken on a the aluminum oxide gridded capacitor. This capacitor has a gate stack of 2.4/7.7/15nm of silicon oxide tunnel oxide, silicon nitride charge trap material, and aluminum oxide blocking oxide.

Figure 4.25 shows excellent agreement between experimental and simulation curves on the gridded capacitor performed at 8V. Simulation of such devices can be challenging as the programs require a starting threshold voltage. The measurement of
Figure 4.26: Experimental and simulation erase curves performed at 7 and 8V on a gridded capacitor with an aluminum oxide blocking oxide. The gate stack is 2.4/7.7/15nm silicon oxide/silicon nitride/aluminum oxide. Curves are included for the 7V measurement with different threshold voltages. The original measurement of 0.2V was incorrect given poor agreement between the curves. With a 1V starting threshold voltage there is excellent agreement between the curves.

gridded capacitors does not measure an exact threshold voltage. Rather the measurement records the change in voltage at a certain capacitance. For simulation purposes it is important to know the exact threshold voltage for accurate simulations. This can be seen in the erase curves shown in Figure 4.26.

Figure 4.26 shows the erase curves at -8 and -7V. The 8V curve has excellent agreement but for the -7V curve the starting accuracy was poor. At first the threshold voltage inputed into the program was 0.2V which gives the purple X curve. The curve does not agree with the experimental results. However, when the threshold voltage
Figure 4.27: Experimental and simulation write curves performed at 7, 8, and 9V on a gridded capacitor with an aluminum oxide blocking oxide. The gate stack is 2.4/7.7/15nm silicon oxide/silicon nitride/aluminum oxide. The 9V curve is slightly off showing an issue with the starting threshold voltage value. When the threshold voltage was changed to 1V the simulation curve had perfect agreement with the experimental curve.

Figure 4.27 shows the write curves with simulation curves at 7, 8, and 9V. The simulation curves show good agreement with the experimental curves. The agreement could be improved with better calculation of the threshold voltages.

4.5 Summary

This chapter provides an in-depth analysis of speed write and erase measurements on the gridded capacitor structure. The first section examined the theory behind charge injection of carriers into the nitride. In the next section, speed write and
erase measurements were performed on gridded capacitors. We are able to perform measurements down to an applied voltage of 6V and up to an applied voltage of 12-14V. The theory of the first section has been placed into a simulation program that predicts speed write/erase curves. The program requires constants such as band heights and permittivities to be entered. An important parameter that needs to be supplied to the simulation program is the threshold voltage. Once the necessary parameters are supplied the program can calculate the threshold voltage at any time.
Chapter 5: Retention

Electron and hole traps hold the key to device performance degradation of different semiconductor devices. Leakage currents from bulk and interface traps limit charge retention of dynamic RAM cells, dynamic range, signal/noise ratio and charge transfer efficiency of charge coupled device arrays. For MOS transistors, traps in the gate dielectric lead to logic threshold instabilities. Mobility is reduced in polysilicon thin-film transistors when there are bulk traps in the polysilicon. But not all traps are detrimental to semiconductor devices. Gold traps have been used as a lifetime "killer" in silicon for fast-recovery power rectifiers and oxygen is used for intrinsic gettering of other undesirable impurities in silicon[82].

For memory devices, traps have led to the adoption of charge trap non-voltaile memory devices discussed in this dissertation. Silicon nitride has efficient charge trapping characteristics which are used in SONOS and MANOS devices. These devices have two memory states. One memory state is linked to excess electrons in the charge trap material (silicon nitride layer) while the other corresponds to excess holes in the charge trap material.

For SONOS and MANOS devices, traps in the silicon nitride are responsible for memory action and charge transport. The evidence of the nitride film being charge
negatively (positively) with the injection of electrons (holes) from the silicon substrate leads to two possibilities

1. The film contains non-interacting, close-compensating donor and acceptor traps in large densities of more than $10^{18} \text{cm}^{-3}$. The donor trap has two charge states ($D_D^+, D_D^0$) while the acceptor trap also has two charge states ($D_A^+, D_A^0$). In this scenario the trap interact with electrons and holes independent of each other as shown in Figure 5.1

2. The film contains amphoteric traps which have three charge states ($D^+, D^0, D^-$) and two transition energies ($E_D, E_A$). Each amphoteric trap interacts with both electrons and holes as shown in Figure 5.2. For amphoteric traps in silicon nitride, it is postulated that these traps arise from a trivalent Si center[82]

In this chapter, we examine the capture and emission processes via amphoteric traps provided in Roy and White[82] and previously in White and Chao[105]. We will
Figure 5.2: The amphoteric trap model has only one trap. This trap has three states, positive, neutral, or negative.

Figure 5.3: Band diagram depicting the different trap levels. Acceptor traps lie in the upper portion of the energy gap, while donor traps lie in the lower portion of the energy gap.
write out the basic formulism for the processes and discuss how they lead to retention simulations. Finally we will provide and examine retention measurement results on our gridded capacitors.

5.1 Retention Theory

Amphoteric traps have four capture processes, two for electron capture and two for hole capture, for the three different charge states. Likewise, the traps have four emission processes. These processes can be expressed with two coupled first-order rate equations for the occupancy functions $f^+(D^+ state)$, $f^-(D^- state)$, $f^0(D^0 state)$:

$$\frac{df^+}{dt} = \sigma_0^p \nu_{th} p_v f^0 - e_p^+ f^+ - \sigma_n^+ \nu_{th} n_c f^+ + e_n^0 f^0$$

(5.1)

$$\frac{df^-}{dt} = \sigma_0^n \nu_{th} n_c f^0 - e_n^- f^- - \sigma_p^- \nu_{th} p_v f^- + e_p^0 f^0$$

(5.2)

where $\nu_{th}$ is the thermal velocity of carriers in the nitride. For the capture cross-sections $\sigma_0^p$, $\sigma_p^-$, $\sigma_n^-$ and $\sigma_n^+$ and emission coefficients $e_p^+$, $e_p^0$, $e_n^0$, and $e_n^-$, the superscript denotes the charge state and the subscript denotes hole(p) or electron(n) process. $p_v(n_c)$ is the hole(electron) density in the valence(conduction) band of the nitride.

The occupancy functions satisfy the relation in Equation 5.3 as the traps have to be in one of the three states.

$$f^0 + f^- + f^+ = 1$$

(5.3)

We can now write the trapped charge in relation to our defined occupancy functions.

$$\rho_N(x, t) = q N_T(x) [f^+(x, t) - f^-(x, t)] = q N_T(x) [1 - f^0(x, t) - 2f^-(x, t)]$$

(5.4)
Previously, White and Chao [4] have solved the Equations 5.1 and 5.2 with the 1D continuity equation for the case of electron injection and capture only process, i.e. the emission processes are neglected and initially the nitride is free of any electrons \( f^+(x, 0) = 1 \).

Roy and White[82] originally only considered the emission processes as charge losses.

\[
\frac{df^+}{dt} = -e_p^+ f^+ + e_n^0 f_0^0 \approx -e_p^+ f^+ \tag{5.5}
\]

\[
\frac{df^-}{dt} = -e_n^- f^- + e_p^0 f_0^0 \approx -e_n^- f^- \tag{5.6}
\]

They further neglected tunnel emission from the neutral state, the SI conduction(valence) band is aligned with the trap level for hole(electron) emission and Poole-Frenckel emission to the nitride band since the traps are deep within the silicon nitride layer. These approximations allowed Roy and White to decouple the emission equations and arrive at the solutions:

\[
f^+(x, t) = f^+(x, 0) \exp[-e_p^+(x)t] \tag{5.7}
\]

\[
f^-(x, t) = f^-(x, 0) \exp[-e_n^-(x)t] \tag{5.8}
\]

These results are consistent with thin tunnel oxides and thick blocking oxides. However, the repeated thinning of the device layers has allowed injection of holes back into the silicon nitride layer. Roy and White [82] go on further to present a simple retention model for thick devices.
Figure 5.4: Energy band diagram of a MONOS charge trap memory at zero bias with electrons in the trapping material. Electrons can tunnel out from traps into the conduction band of the substrate or can rearrange themselves in the nitride through Frenkle-Poole emission. Holes can back tunnel into the nitride material from the valence band of the substrate.
Figure 5.4 shows the energy band diagram for a MONOS structure at zero bias with excess electrons trapped in the charge trapping silicon nitride region. The electron emission processes include thermal excitation from the negative charged state into the nitride conduction band, back-tunneling into the silicon conduction band from the negative charges state, and the injection of holes from the silicon substrate to the neutral charge state.

It is also possible for the same electron emission process to take place at the boundary with the blocking oxide. A good charge trapping NVSM has a large enough blocking oxide to minimize these processes. It should also be noted that with the replacement of silicon oxide to higher-K oxides the potential exists to have trap assisted tunneling through the oxide as high-K oxides have much larger trap densities than silicon oxide.

5.2 Retention Results

Retention measurements were completed on the gridded capacitor structure pictured in Figure 5.5 with the results shown in Figure 5.6
Figure 5.5: Gate stack diagram and TEM of the device under test. We used a 2.4/7.7/15nm MANOS device.

Figure 5.6 was completed on a capacitor with a 100 $\mu$m grid spacing and a 250$\mu$m capacitor radius. The measurement was completed with an 8V write pulse of 2.5E-3 seconds and a -8V erase pulse of 7.5E-3 seconds. After the pulse is applied read measurements are completed out to 1E4 seconds with no refresh of the device. The erase curve is solidly flat out to 1E4 seconds. The write curve, however, starts to decay after 10 seconds in a linear fashion. The retention of these particular devices can be improved since the nitride processing issues discussed previously can be eliminated. More importantly is that the measurements were taken on a gridded capacitor device. The gridded capacitor structure is capable of completing the full test suite of speed, retention and endurance measurement on non-volatile charge trap based devices.
Figure 5.6: Retention curves on a gridded capacitor with aluminum oxide blocking layer. The erase state shows excellent retention characteristics. The positive state starts to lose charge after a second. This is due to the poor silicon nitride process producing traps near the interface. The thickness of the device is 2.4/7.7/15nm.
Chapter 6: Other Material Considerations

6.1 Gridded Capacitors in III-V Materials

The gridded capacitors structure is not limited to silicon based semiconductor devices. To expand the use cases of gridded capacitors we have fabricated devices in III-V material systems. The gate stacks fabricated are shown below in Figure 6.1 and 6.2.

Our study focused on the lattice matched material systems GaAs on GaAs, and InGaAs on InP. This allows for the investigation of both major competing high electron mobility III-V materials. These materials have been chosen for their ease of growth over a large area, high availability, wide use, and high mobility (8500 cm/s for GaAs and 12,000 cm/s for InGaAs). Both p-type and n-type wafers will be investigated as well as applications in MOSFETs, Figure 6.2, and nonvolatile memory systems, Figure 6.1.

6.2 Fabrication of Devices

Gallium Arsenide (GaAs) and Indium Phosphide (InP) substrate wafers were purchased. The GaAs wafers are three inches in diameter while the InP wafers are two inches in diameter. Four wafers of each substrate type were purchased. The four
Figure 6.1: Cross section of the III-V charge trap non volatile memory stack. Aluminum oxide has been used for both the blocking and tunneling oxide.

Figure 6.2: Cross section of the III-V transistor stack. An aluminium/hafnium oxide layer has been used as the insulating later.
Wafers are broken down into two n-type wafers and two p-type wafers. Each of these sets will then have a MOSFET gate stack or a nonvolatile memory gate stack deposited.

To start the fabrication process an 1 $\mu$m active layer has been deposited on all wafers. For gallium arsenide (GaAs) the active layer is a high quality GaAs layer while for indium phosphide (InP) wafers the active layer is a high quality InGaAs layer. After the active layer is grown the wafers are prepared for implantation.

The implants have been calculated by Monte Carlo analysis using the SRIM (The Stopping and Range of Ions in Matter) program. This program allows us to calculate the energy and dose required to implant a total ion concentration of 1E18 cm$^{-3}$. The output of a SRIM analysis on a GaAs active layer implanted with carbon at 50keV is shown in Figure 6.3 below.

The longer axis out of the page depicts the depth of the ions, which is also shown in red. The shorter axis, shown in green, depicts the lateral range of the implant. A projected range of around 1100 was used to keep the implant energy low. The lower energy will lessen damage to the active layer material. Also, the lower implant energy will cause less lateral straggle. This will keep the implants more defined to the grid lines. The units of the graph are in \((\text{Atoms/cm}^3)/(\text{Atoms/cm}^2)\). Although these units seem strange, when you multiply by an implantation dose \((\text{Ions/cm}^2)\), you will end with the impurity concentration \((\text{atoms/cm}^3)\).

Therefore, it is easy to find the needed implantation dose by taking the desired impurity concentration and dividing by SRIM output data. SRIM calculations were completed on all wafer material types and implants with the results shown in Table 6.4.
Figure 6.3: SRIM ion distribution analysis on a GaAs active layer implanted with carbon at 50keV.
Figure 6.4: Ion doses used to implant InGaAs and GaAs wafers.

<table>
<thead>
<tr>
<th>Material</th>
<th>Implant</th>
<th>Energy</th>
<th>Dose</th>
</tr>
</thead>
<tbody>
<tr>
<td>GaAs</td>
<td>Si</td>
<td>105keV</td>
<td>1.33E+13</td>
</tr>
<tr>
<td>GaAs</td>
<td>C</td>
<td>50keV</td>
<td>1.45E+13</td>
</tr>
<tr>
<td>InGaAs</td>
<td>Si</td>
<td>120keV</td>
<td>1.54E+13</td>
</tr>
<tr>
<td>InGaAs</td>
<td>Mg</td>
<td>90keV</td>
<td>1.66E+13</td>
</tr>
</tbody>
</table>

N-type
P-type
Figure 6.5: Silicon oxide sacrificial protective layer and the silicon nitride masking layer used while implantation is completed.

The implants are completed by INNOViON which provide implantation services on a wide variety of substrates and doses.

Since the III-V compound semiconductor interface with dielectrics is one of the most problematic areas of compound devices, which are also under study, it is imperative that the surface be kept as clean and damage free prior to the deposition of the gate dielectrics. We developed an implant process that minimized surface damage of III-V compound materials. The process starts after the active layer has been grown. A 100nm silicon dioxide layer is deposited by PECVD directly onto the compound substrate. This layer will be a sacrificial protective layer. On top of the SiO$_2$ layer, a 1um thick silicon nitride layer is also deposited by PECVD. This is shown in Figure 6.5. The silicon nitride layer is the masking layer which will protect the substrate from the ion implantation step.

Next a photolithography step is completed to mask out of the grid lines. Photoresist is applied to the wafers and patterned with the grid. The grid lines are then
Figure 6.6: An SF6 plasma is used to create the grid lines in the silicon nitride masking layer.

etched through the silicon nitride layer with a low damage SF6 plasma, shown in Figure 6.6. The etch lasts two minutes and has a over-etch built in. The over etch is to slightly thin down the silicon dioxide for the ion implantation step.

The wafers are then sent out to receive the implantation. The implantation is done through the SiO$_2$ layer. This is to minimize surface roughness from the damaging ion beam. We only want to dope the grid lines, therefore the areas with the silicon nitride still present will protect the substrate from the ion implantation, shown in Figure 6.7.

Once the wafers are received from the ion implantation, the silicon nitride as well as the silicon dioxide is striped off. Usually, a hot phosphoric acid is used to chemically etch silicon nitride. However, this etch will completely etch away GaAs, as well as
Figure 6.7: Ion implantation is completed through the silicon oxide protective layer.

InP. Therefore, we use the low damage SF$_6$ to etch away the rest of the silicon nitride. The plasma will slightly attack the silicon dioxide but not the underlying compound materials.

On the day selected to deposit the gate stack, a buffered HF solution was used to remove the protective silicon dioxide layer. The silicon dioxide layer is made thin to reduce the time to clean the wafers. Also, the shorter the chemical etch, the shorter time the wafer is exposed to contaminants before the gate deposition.

Once the wafers are implanted and the mask is removed an activation anneal is completed. This anneal is critical for the implanted ions to become part of the lattice and electrically active rather than interstitial defects. To accomplish this, wafers are subjected to high temperatures (1000°C) where the wafer lattice can rearrange itself. This anneal poses a challenge for the compound wafers. At temperatures above
500°C the arsenic atoms will completely break free of the lattice and sublimate, leaving behind a pure gallium layer. This is unacceptable for the fabrication of devices. To prevent this evaporation the wafers were encased with both a SiO$_2$ and SiN layer. These layers will prevent the arsenic atoms from leaving the surface. After the anneal the protection layers can be etched away leaving a pristine doped surface.

To minimize further damage from excess high temperature the anneal was performed through a Rapid Thermal Annealer, or RTA. The AG Associated 410 RTA uses high powered lamps to quickly raise the temperature of wafers placed in the tool. To complete the activation anneal requires a temperature of 1000°C for 15 seconds, with a one minute ramp up and one minute ramp down cycle.

Upon completion of the activation anneal the gate stacks can be deposited by Nanotech Wests PicoSun SUNALE R-150B Atomic Layer Deposition (ALD) Tool. In preparation of this step, multiple test wafers were run through the tool to fine tune the correct gate stack thicknesses. It is extremely important to know the growth per pulse of the ALD tool. Each ALD behaves slightly differently which can produce
different results. For our devices Al$_2$O$_3$ and HfO$_2$ dielectrics were used. Al$_2$O$_3$ is the most studied dielectric that is deposited by ALD and this dielectric has been extensively studied at the lab. These studies have revealed a growth per pulse of 0.86Å. The halfnium dioxide has been used less extensively and therefore require a few runs to quantify the growth per pulse. These results are outlined in Table 6.9.

It is important to note that these tests were completed on silicon wafers. The initial growth on different substrates such as GaAs and InGaAs may act different for the first 3-5 pulse cycles. After the initial dielectric layer is deposited the growth per cycle will match the above readings.

The gate stack was deposited using OSUs Nanotech West Picosun SUNALE R-150B Atomic Layer Deposition Tool shown in Figure 6.10. The deposition of the MOSFET devices required a 5nm Al$_2$O$_3$ layer, followed by a 4nm HfO$_2$ layer, which required 59 Al$_2$O$_3$ cycles and 91 HfO$_2$ cycles. The aluminum oxide layer was increased from 2-3nm to 5nm to give a larger thickness with a large barrier height to suppress
gate leakage. The deposition of the memory devices required a 3nm Al₂O₃ blocking oxide, a 9nm HfO₂ charge trapping region, and a 10nm Al₂O₃ blocking oxide. These devices required 35 cycles of Al₂O₃, 205 cycles for the HfO₂ layer, and 117 cycles for the Al₂O₃ layer.

The gate deposition was immediately followed by the metallization step to reduce contamination of the wafers. Metal deposition was completed using the CHA Solution System E-Gun Evaporator at Nanotech West, shown in Figure 6.11. The front side deposition consisted of a 300nm layer of aluminum. This layer was reduced to 300nm from 500nm to improve the aluminum etch step. The back side deposition consisted of a 30nm layer of titanium, followed by a 200nm layer of platinum. The Ti/Pt system is well studied in making high quality ohmic contacts to GaAs and InP.

Following the metal depositions, the wafers underwent a metal anneal, in Nanotech Wests Rapid Thermal Anneal (RTA). This tool allows us to heat the wafer to 420°C in 30 seconds. The anneal takes place at 420°C for 60 seconds in an atmosphere consisting of 5% hydrogen and 95% nitrogen, which is also referred to as forming gas. This step anneals out any damage in the wafers that might have taken place during the metal deposition.

We then defined the capacitors on the front side of the wafers with a photolithography step. Shipley S1813 photoresist is applied to both sides of the wafer. The back side metal needs to be protected during the final etch step hence the extra resist coating. The wafers are then exposed with the capacitor mask. This mask defines the various capacitors on the wafer. Afterwards the wafers are developed in MF 319 developer.
Figure 6.10: An image of the Picosun ALD machine while open at The Ohio State University Nanotech Lab.
Figure 6.11: An image of the metal deposition machine at The Ohio State University Nanotech Lab.
The final step in fabrication of gridded capacitors is the etching of the aluminum layer. There are many different ways to etch aluminum. One of the most common is with phosphoric, nitric, and acetic acid, also referred to as PAN etch. This solution etches aluminum at roughly 200nm per minute. It is an extremely fast etch. Previous experiments using this etch resulted in poor capacitors. The etch rate was too quick to control and caused over etching of the aluminum layer. This problem was solved by using CD-26 developer. This developer can also be used to develop the S1813 resist, however it also etches aluminum at the rate of 50nm per 3.5 minutes. The develop and etch step could be combined by just using CD-26 developer. However, to make sure the photolithography step was successful in defining the capacitors, the two steps were separated. MF 319 developer does not etch aluminum. Since CD-26 etches aluminum at a much slower rate, a thinner aluminum layer was deposited. Afterwards, the photoresist is stripped using acetone. The final gridded capacitor cross section is shown in Figure 6.12 and 6.13.
Figure 6.13: Final gate stacks cross section on the III-V charge trap non volatile memory stack.

A top-view of the gridded capacitor masks are shown in Figure 2.2, and Figure 2.3. The wafer consists of many grids, as shown in Figure 2.2. Each grid consists of different size capacitors, shown in Figure 2.3. There are seven different grid spacings on the wafer. We are currently testing to see which grid spacing provides the best results. If there are two few grid lines, the surface may not invert. However, we want as few grid lines as possible to reduce the area of the implanted regions.

6.3 III-V Gridded Capacitor Results

Upon completion of the gridded capacitor wafer testing started. The wafer sits on the probe station. The device is connected to the Boonton 728D Differential Capacitance Meter as well as our custom voltage tracking circuit. These are then connected to the computer through custom programs in the LabVIEW™ environment. All tests are controlled through the LabVIEW™ interface. Tests that were to be completed include Capacitance-Voltage tests to investigate the interface of the devices, as well
as mobility measurements. Additional test are completed on non-volatile devices. These tests include write/erase speed, retention, and endurance. At the start of testing the wafers immediately showed shorting issues. None of the capacitors showed any promise and it was determined that the aluminum gate had shorted through the device. To combat the shorted capacitors two wafers were reworked. An etch was completed on both wafers to remove the aluminum metal as well as the gate stack.

Since all wafers were used in the first fabrication and testing sequence, the wafers needed to be stripped and clean. This required an aluminum etch in CD-26 developer, which etches aluminum at roughly 25 Å per minute. The wafers were left in CD-26 developer for 45 minutes to over etch and remove all traces of aluminum. The dielectric layers were removed with a buffered oxide etch. This etch is able to remove both aluminum oxide and hafnium oxide layers. The wafers were left to etch for 1 hour to remove the gate stack layers. The etch does not significantly effect the compound substrate. Once the wafers were cleaned we could redeposit the gate stack layers.

The gate stack was deposited using OSUs Nanotech West Picosun SUNALE R-150B Atomic Layer Deposition Tool. The deposition of the MOSFET devices required a 5nm Al₂O₃ layer, followed by a 4nm HfO₂ layer, which required 59 Al₂O₃ cycles and 91 HfO₂ cycles. The aluminum oxide layer was increased from 2-3nm to 5nm to give a larger thickness with a large barrier height to suppress gate leakage. The deposition of the memory devices required a 3nm Al₂O₃ blocking oxide, a 9nm HfO₂ charge trapping region, and a 10nm Al₂O₃ blocking oxide. These devices required 35 cycles of Al₂O₃, 205 cycles for the HfO₂ layer, and 117 cycles for the Al₂O₃ layer. These were the same deposition cycles as used in our previous runs.
The gate deposition was immediately followed by the metallization step to reduce contamination of the wafers. Metal deposition was completed using the CHA Solution System E-Gun Evaporator at Nanotech West. Only the front of the wafers needed to be redeposited with nickel. The deposition consisted of a 300nm layer of nickel. Upon removing the two wafers from the evaporator the MOSFET device wafers nickel gate metal began to peel. Approximately 50-60% of the deposited metal simply lifted off of the wafer. The NVSM wafers had no such peeling. The MOSFET device wafers have a top dielectric layer of hafnium oxide, while the NVSM wafers have a top dielectric layer of aluminum oxide. It appears the nickel metal has difficulty adhering to hafnium oxide. This is currently not discussed in literature and needs to be further examined. The MOSFET wafer continued through the final fabrication steps for further testing.

Next, the capacitors were defined through a final photolithography step. Shipley S1813 photoresist was applied to both sides of the wafer. The back side metal needed to be protected during the final etch step hence the extra resist coating. The wafers were then exposed with the capacitor mask. This mask defined the various capacitors on the wafer. Afterwards the wafers were developed in MF 319 developer. Finally, the nickel was etched with Transene Nickel Etchant TFB, which etches nickel at roughly 30Å per second.

Upon completion of the compound gridded capacitors, the devices were evaluated through the use of Capacitance-Voltage(C-V) measurements. A C-V curve completed on a 50um capacitor radius and 30um grid spacing is shown below in Figure 6.14.

Figure 6.14 displays two C-V measurements on the same device, one at 100kHz and another at 1MHz on the NVSM dielectric stack. The stack consisted of an

218
Figure 6.14: C-V measurement on a 50um capacitor radius and 30um grid spacing at 1MHz and 100kHz. A slight C-V curve is seen.
aluminum oxide tunneling layer, a hafnium oxide charge trap region, and a aluminum oxide blocking layer the NVSM device structure. The measurement was completed with a bias sweep from -6 volts to 6 volts. This range was selected due to the devices breaking down on voltages higher than 8 volts.

One characteristic of these curves is how spread out the dip appears in voltage. Typically a C-V curve has a narrow dip as the device goes from accumulation through depletion and then back to inversion. The dip in capacitance is caused by the depletion region forming at the surface. The depletion capacitance then reduces the overall capacitance at that voltage. In accumulation a large number of holes are at the surface and the measured capacitance is the oxide capacitance. With the gridded capacitor, the grid supplies electrons to form an inversion layer and the capacitance then increases to the oxide capacitance as the voltage is increased. With the devices shown in Figure 6.14 a slight dip is observed that is very spread out. This spreading out is caused by a poor quality interface and interface traps. Since these devices were reworked, the interface quality is not as good as a fresh untouched wafer.

Another characteristic of these devices is that at higher frequency conditions the C-V curve flattens out. At 1MHz the dip is much less pronounced than at 100kHz. This is most likely due to series resistance in the wafers. The substrate material consists of semi-insulating GaAs. This selection was made due to the fact that all compound transistors are currently fabricated in semi-insulating wafers. These wafers have a high series resistance >20k ohms from the gate dielectrics to the substrate. A high series resistance is a limiting factor in the characterization of gridded capacitors. Semiconducting compound wafers are available, and are used mostly for solar cell applications. These wafers can be purchased to reduce the series resistance in
Figure 6.15: C-V measurement were scanned from high to low (blue) and low to high (red). There is a distinct shift in the curve which leads to voltage shift of 3.2V

the substrate. The semiconducting wafers can then be grown with a smaller semi-insulating region to mimic currently compound wafer production.

Figure 6.15 shows C-V sweeps of a NVSM gate stack gridded capacitor. The sweeps were completed from low to high, and then from high to low bias. This test mimics a write and erase operation to quickly see how much of a threshold voltage shift is possible. When starting a sweep at high bias voltages, electrons tunnel into the hafnium charge trap region. This shifts the C-V curve to the write. When the sweep starts at low bias voltages holes are injected into the charge trap material and this shifts the C-V curve to the left. For these devices a shift of 3.2 volts was observed.
for a -6V to 6V bias sweeps. This is a much larger voltage shift than what is seen in typical silicon NVSM devices, and at much lower voltages.

6.4 Summary

In conclusion, compound gridded capacitors have been fabricated in GaAs wafers. The wafers were reworked due to the gate metal (aluminum) spiking through and shorting out the gate dielectrics. The reworked wafers were completed with a nickel gate. The devices show a large spread out C-V curve due to poor interface quality due to reworking of the wafers. They also have a large series resistance component due to the use of semi-insulating wafers. The series resistance can be reduced in future tests by using semi-conducting wafers and growing a smaller semi-insulating compound material on top. The initial non-volatile memory devices show a large shift in the C-V from opposite bias sweeps. This indicates a large amount of charge is tunneling into the device at low bias voltages. The 3 volt shift is much larger than typical silicon NVSM devices and at much lower voltages. This shows the potential for fast NVSM devices at very low voltages.
Chapter 7: Conclusions

7.1 Conclusions from This Dissertation

The gridded capacitor structure uses grid lines doped with the opposite carrier type than the substrate. With both carrier types available in the substrate numerous tests can be carried out on non-volatile semiconductor memory (NVSM) gate stacks as well as gate stacks for use in insulated gate FET devices without the need for source/drain contacts. The structure requires only two masks, a grid line definition mask and a capacitor area definition mask on a suitably prepared gridded substrate. This allows for a rapid fabrication process to study many different gate stacks in a short period of time. NVSM tests that can be carried out on gridded capacitors include, write and erase speed curves, retention, and endurance tests. Write and erase speed tests have been demonstrated on gridded capacitors with a gate stack of 2.4nm SiO$_2$ tunnel oxide, 7.7nm Si$_3$N$_4$ charge trapping layer, 15nm Al$_2$O$_3$ blocking oxide or a 15nm HfO$_2$ / Al$_2$O$_3$ blocking oxide and a metal gate.

For the aluminium oxide sample at 6V the memory has a cross-over time of 1s and a memory window of 1.2V, whereas, for the composite aluminum oxide, hafnium oxide sample a 0.12s cross-over time and a memory window of 1.6V was observed. At 7V the aluminum oxide sample has a 0.17s cross-over time and a memory window of
2.4V, with the composite aluminium oxide hafnium oxide sample having a 0.05s cross-over time and memory window of 3.0V. Finally, with an applied voltage of 8V, the aluminum oxide sample has 80ms cross-over time and a memory window of 4.1V while the composite sample has a cross-over time of 10ms and the same memory window of 4.1V. The composite aluminum oxide hafnium oxide blocking oxide provides an increase in speed performance as well as a large memory window at smaller voltages.

Capacitance-voltage tests showed that the optimal grid spacing was between 50um and 100um. Too many grids will offset the substrate doping, while too few will not allow an inversion layer to form. The optimal capacitor radii is 250um. An accurate dielectric constant of 7.8 for aluminum oxide was observed with a capacitor radii of 250um. The dielectric constant of the aluminum/hafnium blocking oxide blend was seen to be 14. The addition of hafnium greatly increases the dielectric constant of the hybrid layer.

A simulation program has also been developed to match and predict speed write/erase curves for charge trap non-volatile memory devices. Providing the simulation program with material properties such as K value, permittivities, band offsets, etc, allows the program to calculate various tunneling currents to supply a final speed write/erase curve. This simulation program is able to match experimental curves from full transistor devices as well as gridded capacitors.

In addition, a method was presented to determine the channel mobility with a two terminal gridded capacitor. A substrate is implanted with minority carriers in a grid structure. The grid acts as a minority carrier source. The substrate can be doped in any manner to simulate a transistor channel. This allows for the study of mobility
with respect to channel doping as well as electric field. The effects of new high-K gate stacks on the channel mobility can also be investigated. In the conductance measurements, we are able to determine the channel mobility with excellent agreement between the transmission line AC mobility \((322 \, \text{cm}^2/\text{V} \cdot \text{s})\) and the conductance mobility \((318 \, \text{cm}^2/\text{V} \cdot \text{s})\) for transistors. This method was applied to MANOS gridded capacitors and obtained a mobility of \(334 \, \text{cm}^2/\text{V} \cdot \text{s}\).

Gridded capacitors are two terminal devices requiring only two mask steps. This allows for rapid fabrication and evaluation gate stacks as well as channel doping schemes. The structure also requires no source/drain contacts.

### 7.2 Recommendations for Future Research

- Fabricate gridded capacitor devices with other high-K materials such as zirconium oxide, lanthium oxide, and zinc oxides. The goal for using these materials would be to obtain programming voltages below 5V while maintaining retention of 10 years or more.

- Fabricate gridded capacitor devices on other material substrates. The structure could be used to fabricate memory devices on III-V substrates such as GaAs and InP, or to better characterize oxides on these substrates. Another direction is to use the gridded capacitor structure with carbon nanotubes. The gridded capacitor does not require source/drain contacts and could therefore help characterize the mobility of carbon nanotubes.

- Further improve the speed write/erase programs for speed and accuracy. Make it easier to change gate stack layers and thicknesses for the addition of high-K materials.
• Investigate nanolaminates and how they could be used to improve device characteristics. Also formulate a model of the nanolaminate layers and add it to the current simulation programs.

• Develop temperature dependent speed write/erase model by adding temperature related parameters into the simulation.

• Develop a retention simulation program from the amphoteric trap model. Compare the results against non-volatile charge trap memories and gridded capacitors. Combine the retention and speed write/erase program into one to create a BSIM model for NVSMs.
Appendix A: MANOS Transistor Fabrication

1. Starting Material

   (a) 3-inch (100) p-type Si wafers, boron-doped, 15-30 -cm resistivity

2. N-Well Formation

   (a) RCA clean with HF dip

   (b) Oxidation, 1000 (Dry, 1000C, 160 min)

   (c) Photolithography (mask: NW = N-well)

   (d) BHF etch to obtain 200 pad oxide (14 min)

   (e) N-well implantation (phosphorus, 100KeV, 4.8x1012 cm-2)

   (f) Plasma PS strip

   (g) Chemical PR strip

   (h) RCA clean

   (i) Anneal (1100C, 60 min, N2, 2 slpm)

   (j) Oxidation (dry, 1100C, 15 min)

   (k) N-well drive in anneal (1100C, 18 hours 45 min, N2, 2 slpm)
3. LOCOS (Field Oxide Step)

(a) RCA clean

(b) LPCVD Nitride 1000 (725°C, 0.3 Torr, SiCl2H2:NH3 = 20:100 sccm, 65 min)

(c) Photolithography (mask: AD = active device)

(d) Plasma etch nitride (0.3 Torr, 300 Watts, CF4, 12 sccm)

(e) Chemical PR strip

(f) Photolithography (mask: FI = Field Implant)

(g) Field Implantation (Boron, 40 KeV, 5x1014 cm-2)

(h) Plasma PR strip

(i) Chemical PR strip

(j) RCA clean

(k) Field oxidation / implant activation (wet, 900°C, 180 min, 30 min anneal in N2)

(l) Field oxidation (wet, 1000°C, 180 min)

(m) Etch oxynitride over the nitride (10:1 BHF, 1 min)

(n) Etch nitride 1000 (phosphoric acid, 180°C, 60 min)

(o) Etch buffer oxide 1000 (10:1 BHF, 2 min)

(p) RCA clean

(q) Oxidation to remove residual nitride, 1000 (dry, 1000°C, 11 min)

(r) Etch oxide (10:1 BHF, 3 min)
4. Channel Doping (Threshold Voltage Adjust)

(a) RCA clean

(b) Pad oxide 200 (wet, 900C, 15 min)

(c) Blanket threshold voltage adjust implantation (Boron, 50 KeV, 1.7x10^{12} \text{ cm}^{-2})

which provides adjustment for both the n-channel and p-channel devices

(d) RCA clean

(e) Implant activation anneal (900C, 30 min, N2)

5. Source/Drain Formation

(a) RCA clean

(b) Gate oxide 350 (dry, 1000C, 52 min, O2, 15 min anneal in N2)

(c) Polysilicon gate 3500 (LPCVD, 625C, 0.8 Torr, 25 min, 282 sccm Silane)

(d) POCL3 diffusion (900C, N2 and O2 flow with POCL3 bubbler at 19C for 20 min, 20 min drive-in with N2 only)

(e) Etch P-glass (100:1 BHF, 4 min)

(f) Photolithography (mask: PY = Polysilicon)

(g) Plasma etch polysilicon (150 mTorr, 300W, 45 sccm SF6, 1 min)

(h) Chemical PR strip

(i) Photolithography (mask: NN = N+)

(j) N-Source/Drain implantation (phosphorous, 40 KeV, 5x10^{15} \text{ cm}^{-2})

(k) Plasma PR strip
(l) Chemical PR strip

(m) Photolithography (mask: PP = P+)

(n) P-Source/Drain implantation (boron, 32 KeV, 5x10^15 cm^-2)

(o) Plasma PR strip

(p) Chemical PR strip

(q) Etch polysilicon gate (HF:HNO3:CH3COOH = 1:50:20, about 4 min)

(r) RCA clean

(s) Implantation anneal and drive-in (1000C, 180 min, N2, 120 min dry oxidation in O2 to remove residues from wet polysilicon etch

(t) Etch oxide (10:1 BHF, till hydrophobic)

6. Gate Formation

(a) RCA clean with HF dip

(b) Tunnel oxide 20 (TWO, 850C, 30 min, 30 sccm Ar, 30 min anneal in Ar) 
   Triple Wall Oxidation (TWO) Furnace

(c) Silicon nitride 80 (LPCVD, 680C, 20 min, SiCl2H2:NH3 = 10:100 sccm)
   Low Pressure Chemical Vapor Deposition (LPCVD) Furnace

(d) Aluminum oxide 80 (ALD at IBM) Atomic Layer Deposition (ALD)

(e) Rapid Thermal Anneal (RTA) (900C, 60 sec in N2, the 30 sec in O2)

(f) Aluminum E-beam evaporation 2900 Formation of metal for MANOS

(g) Photolithography (mask: PY)

(h) Etch Aluminum (PAN etch, 43C, 49 sec)
(i) Chemical PR strip

(j) Etch the ANO stack in the source/drain regions (100:1 BHF for 3 min for ALD Al2O3; Plasma etch CF4:O2 = 25:2 sccm, 100W, 340 mTorr for 30 sec for silicon-rich nitride; 100:1 BHF for 3 min for tunnel SiO2

(k) PECVD Intermediate oxide 4000 (300C, 30 min)

7. Metal Contact

(a) Photolithography (mask: CW = Contact Window)

(b) Etch oxide 4000 (10:1 BHF, 4 min 30 sec)

(c) Chemical PR strip

(d) Aluminum E-beam evaporation 7000 (40 wt% Si in Al)

(e) Photolithography (mask: MET = Metal)

(f) Etch Aluminum (PAN etch, 43C, 65 sec)

(g) Si dust (precipitates) removal (HNO3:H2O:NH4F = 189:96:7.5ml, dip wafers into the solution for 7 seconds)

(h) Chemical PR strip

(i) Organic Clean (acetone for 5 min, methanol for 5 min)

(j) Post Metal Anneal (375C, 30 min, D2:N2 = 200:1800 sccm)
Appendix B: Gridded Capacitor Fabrication

1. Starting Material

   (a) 3-inch (100) p-type Si wafers, boron-doped, 15-30 -cm resistivity

2. Grid Formation

   (a) RCA clean with HF dip

   (b) Wet Oxidation (Expected thickness = 5000, T=1000C)

   (c) Photolithography (Mask: Implantation Grid)

   (d) Etch Silicon Dioxide (5000, 10:1 BHF, 12 minutes)

   (e) Chemical Photoresist Strip

   (f) RCA clean

   (g) POCl3 Diffusion (900C, N2 and O2 flow with POCl3 bubbler at 19 C for
       20 min, then N2 only for 20 min)

   (h) P-glass etch (100:1 BHF, 150 seconds)

   (i) Etch Silicon Dioxide (5000, 10:1 BHF, 12 minutes)

3. Gate Dielectric Formation
(a) RCA clean with HF dip

(b) Tunnel Oxide 25 -45 (TWO, 900C, 38seconds 5:21min, 30 sccm O2 in 3000 sccm Ar, 30 min anneal in Ar)

(c) Silicon Nitride 70 (LPCVD, 680C, 15 min, SiCl2H2:NH3 = 10:100 sccm)

(d) Blocking Oxide ALD Al2O3 or Al2O3/HfO2 7:3 Northrop Grumman

(e) Rapid thermal anneal (900C, 60 sec in N2)

4. Metallization

(a) Organic clean (acetone for 5 min, methanol for 5 min)

(b) E-Beam Aluminum Deposition (Expected thickness = 5000)

(c) Photolithography (Mask: Capacitors)

(d) Etch Aluminum (PAN etch, 43C, 60 seconds)

(e) Plasma Etch Backside: SF6 (5min), CF4+O2 (1 min)

(f) BHF back surface (1min)

(g) E-Beam Aluminum Deposition on Backside (Expected Thickness =5000)

(h) Chemical Photoresist Strip

(i) Organic clean (acetone for 5 min, methanol for 5 min)

(j) Post-metal Anneal (400C, 30 min, D2:N2 = 200:1800 sccm)
Appendix C: C-V Measurement

1. Setup

(a) Connect Agilent 4192A LF Impedance Analyzer to computer via GPIB

(b) Connect Agilent 4192A LF Impedance Analyzer to probe stand (high to gate low to substrate)

(c) Place ground/substrate probe to back of wafer and leave gate probe in air

(d) Turn LF 4192A on switch to capacitance and conductance readings

(e) Put meter in parallel measurement mode

(f) Zero meter by hitting [2nd] [8]

(g) Place gate probe on capacitor

(h) Turn on computer and open cgvsbias.vi

2. Measurement

(a) Set voltage range (usually -5V to 5V)

(b) Set voltage step rate (usually .05V)

(c) Set frequency up to 13 MHz

(d) Set oscillation amplitude (usually 300mV)
(e) Set sweep direction (high to low or low to high)

(f) Enter device name and ID number

(g) Click on arrow to run

(h) Save file
Appendix D: LVR Measurement

1. Setup

(a) Connect HP 33120A 15Mhz Function/Arbitrary Waveform Generator, Keithley 6514 System, Electrometer, and AD Convertor to computer via GPIB

(b) Connect Keithley 6514 System Electrometer to gate probe

(c) Connect HP 33120A 15 MHz Function/Arbitrary Waveform Generator to substrate probe

(d) Connect HP 33120A 15Mhz Function/Arbitrary Waveform Generator to AD Convertor

(e) Turn on computer and open lvr 6514.vi

2. Measurement

(a) Set voltage sweep range (-10V to 10V)

(b) Select voltage step (usually 100mV)

(c) Enter device name and ID number

(d) Click save (nothing visual will happen)

(e) Set run delay to 30 seconds
(f) Click on arrow to run program

*NOTE* Save button activates the save routine once the program is started. It will activate a save window before the measurement takes place. It will save the PREVIOUS run and not the run you are currently performing.
Appendix E: Speed Write/Erase Measurement

1. Setup

(a) Perform a C-V measurement on the device first

(b) Connect Nation Instruments SCB-68 to computer via GPIB

(c) Connect Nation Instruments SCB-68 wires to circuit board (all are labeled)

(d) Connect various power supply cables to board (labeled)

(e) Test high cable from board goes to test high port on Boonton 72BD Capacitance Meter

(f) Wafer high cable from board goes to gate probe

(g) Substrate probe goes to test low on Boonton 72BD Capacitance Meter

(h) Turn on computer and open speederase2.vi and speedwrite2.vi

2. Measurement

(a) For erase measurement in speederase2.vi set Erase Pulse Width Min (sec) to 1E-7

(b) For erase measurement in speederase2.vi set Erase Pulse Width Max (sec) to 1E0
(c) For erase measurement in speederase2.vi set Write Pulse Width to 1E-1

(d) For erase measurement in speederase2.vi set Read Delay to 1.0E-3

(e) Set Points per decade to 4

(f) Select file for saving the data

(g) For erase measurement in speedwrite2.vi set Write Pulse Width Min (sec) to 1E-7

(h) For erase measurement in speedwrite2.vi set Write Pulse Width Max (sec) to 1E0

(i) For erase measurement in speedwrite2.vi set Erase Pulse Width to 1E-1

(j) For erase measurement in speedwrite2.vi set Read Delay to 1.0E-3

(k) Set Points per decade to 4

(l) Select file for saving the data

(m) Run speedwrite2.vi first then run speederase2.vi complete 3 cycles to obtain good data
Appendix F: Matlab Write/Erase Simulation Programs

F.1 Erase Program

```matlab
clear all;

q=1.6E-19;
hbar=1.055E-34;
KbTq=0.0259;
m0=9.1094E-31;
m0_e=0.35*m0; %effective electron mass
m0_h=0.42*m0; %effective hole mass
phi1_e=3.1; %oxide barrier heights for electron
phi1_h=3.65; %oxide barrier heights for hole
phi2_e=1.05; %nitride barrier heights for electron
phi2_h=1.85; %nitride barrier heights for hole
phi3_e=2.98; %Al2O3 barrier height for electron
Xot=24E-10; %tunnel oxide thickness
Xn=44E-10; %nitride thickness
Xob=150E-10; %blocking oxide thickness

ep0=8.8542E-12; %vacuum permittivity
epSi=11.7*ep0; %permittivity of Si
epox=3.9*ep0; %permittivity of oxide
epn=6.5*ep0; %permittivity of nitride
epob=7.9*ep0; %permittivity of Al203
Na=5E22; %substrate doping
ni=1.202E16; %intrinsic doping

phif=KbTq*log(Na/ni); %bulk fermi level
Eg=1.12; %energy gap of Si

mn_e=(epox/epn)^2*phi1_h*m0_e/(phi1_e-phi2_e); %effective hole mass in the nitride

mn_h=(epox/epn)^2*phi1_h*m0_h/(phi1_h-phi2_h); %effective hole mass in the nitride
```
\[ X_{\text{eff}} = X_{\text{ot}} + (\frac{ep_{\text{ox}}}{ep_{\text{ob}}})X_{\text{ob}} + (\frac{ep_{\text{ox}}}{ep_{\text{n}}})X_{\text{n}}; \]
\[ C_{\text{eff}} = \frac{ep_{\text{ox}}}{X_{\text{eff}}}; \]
\[ \phi_{\text{hgs}} = -0.304; \quad \% \text{gate to semiconductor workfunction difference} \]
\[ N_t = 4 \times 10^{25}; \quad \% \text{trap density} \]
\[ \phi_{\text{ht}} = 1; \quad \% \text{trap energy} \]
\[ g_{\alpha} = 2; \quad \% \text{nitride centroid coefficient (2: center of the nitride; 1: SiO}_2, \text{SiN interface; infinite: SiN, blocking oxide interface)} \]

%initial conditions

\[ V_p = -7; \quad \% \text{erase voltage (set as needed)} \]
\[ V_{\text{th}0} = 1; \quad \% \text{initial threshold voltage (set for now combined program will automatically figure this out)} \]
\[ \phi_{\text{is}} = -(E_{\text{g}}/2 - \phi_{\text{if}}); \quad \% \text{surface potential at accumulation} \]
\[ Q_{\text{n}0} = (\phi_{\text{hgs}} + 2\phi_{\text{if}} + (4*ep_{\text{Si}}*q_{\text{Na}}*\phi_{\text{if}})^{1/2}/C_{\text{eff}} - V_{\text{th}0}) / (X_{\text{ob}}/ep_{\text{ob}} + X_{\text{n}}/(g_{\alpha}*ep_{\text{n}})); \quad \% \text{initial nitride charge} \]
\[ E_{\text{ot}0} = (V_p + (X_{\text{ob}}/ep_{\text{ob}} + X_{\text{n}}/(g_{\alpha}*ep_{\text{n}}))*Q_{\text{n}0} - \phi_{\text{hgs}} - \phi_{\text{is}}) / X_{\text{eff}}; \quad \% \text{initial tunnel oxide electric field} \]
\[ E_{\text{ob}0} = E_{\text{ot}0}(ep_{\text{ox}}/ep_{\text{ob}}) - Q_{\text{n}0}/ep_{\text{ob}}; \quad \% \text{initial blocking} \]
\[ E_{\text{ot}0} = -E_{\text{ot}0}; \]
\[ E_{\text{ob}0} = -E_{\text{ob}0}; \]
\[ \% \text{Erase so the fields are in the opposite direction} \]
\[ E_{\text{b}} = -E_{\text{b}}; \]
\[ \% \text{Same as above} \]
\[ J_{\text{ot}0} = J_{\text{btb}}(E_{\text{ot}0}, X_{\text{ot}}, X_{\text{n}}, \text{mox}_h, \phi_{1_h}, \phi_{2_h}); \quad \% \text{Initial band to band tunneling through the tunnel oxide} \]
\[ J_{\text{ob}0} = J_{\text{btb}}(E_{\text{ob}0}, X_{\text{ob}}, X_{\text{n}}, \text{mox}_e, \phi_{1_e}, \phi_{2_e}); \quad \% \text{Initial band to band tunneling through the blocking oxide} \]
\[ J_{\text{tat}0} = J_{\text{tat}}(E_{\text{ot}0}, X_{\text{ot}}, \text{mox}_h, \text{mn}_h, \phi_{1_h}, \phi_{2_h}, \phi_{\text{ht}}, N_t); \quad \% \text{Initial trap assisted tunneling through the tunnel oxide} \]
\[ a = (X_{\text{ob}}/ep_{\text{ob}} + X_{\text{n}}/(g_{\alpha}*ep_{\text{n}}))/X_{\text{eff}}; \]

\text{Fid=fopen('al203_21_9_doping2_7v_8.txt','wt');} \quad \% \text{Open writable text file} \]
\text{fprintf(Fid,'tp, dV_{\text{th}}, E_{\text{ot}}, E_{\text{ob}}, J_{\text{ot}}, J_{\text{ob}}, J_{\text{tat}}, loops, intgr\n');} \quad \% \text{Write headings into text file} \]
\text{fprintf(Fid,'%e, %e, %e, %e, %e, %e, %e, %e\n', 0, 0, E_{\text{ot}0}, E_{\text{ob}0}, J_{\text{ot}0}, J_{\text{ob}0}, J_{\text{tat}0}, 0, 0);} \quad \% \text{Write the initial conditions into the text file} \]
\text{% initialize array size variable} \]
\text{n = 0;} \]
\text{E_{\text{ot}} = [E_{\text{ot}0}];}
dVth = [0];
signal = 0;

%range of variable is a = x:y:z ... for a = x to a = z in steps of y
for t = Time
    %initialize variables
    n=n+1;
    Eot1 = Eot0;
    Eob1 = Eot1*(epox/epob) + (-Eot1*Xeff + phigs + phis - Vp) / (Xob + (epob/epn)*(Xn/gama)); %Calculate blocking oxide field from tunnel oxide field
    Ess = CalE4(Eot0,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_h,phi2_h,phi3_e,phi2_e,phit,Nt,phigs,phis,Vp); %Calculate minimum electrical field across the tunnel oxide
    Eot3 = Ess;
    Eob3 = Eot3*(epox/epob) + (-Eot3*Xeff + phigs + phis - Vp) / (Xob + (epob/epn)*(Xn/gama)); %Calculate blocking oxide field from this tunnel oxide
    %Now have a minimum electric field and a higher electric field due to the applied programming voltage
    %Set some initial variables
    Diff=100;
    Accuracy=10;
    loops(n)=0;
    if signal == 0
        while abs(Diff) > Accuracy
            Eot2 = Eot3 + (Eot1-Eot3)/2; %Calculate new tunnel oxide field that is midway between the low and high electrical fields
            Eob2 = Eot2*(epox/epob) + (-Eot2*Xeff + phigs + phis - Vp) / (Xob + (epob/epn)*(Xn/gama)); %Calculate the new blocking oxide field
            %Calculate the new tunneling currents
            Jot2 = Jbtb(Eot2,Xot,Xn,mox_h,phi1_h,phi2_h);
            Job2 = Jbtb(Eob2,Xob,Xn,mox_e,phi1_e,phi2_e);
            Jtat2 = Jtat(Eot2,Xot,mox_h,mn_h,phi1_h,phi2_h,phit,Nt);
            intgr = Simpson_erase6(Eot0,Eot2,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_h,phi2_h,
            phi1_e,phi2_e,phit,Nt,phigs,phis,Vp,t);
            %Looking for this equation to be less than the set accuracy, at which point we would have a solution and the correct electric fields. IF Diff is positive, replace the high E-field with new one, else replace the low E-field with the new one. Repeat till
%solution.
Diff = intgr + a*t;
if Diff > 0
    Eot1 = Eot2;
else
    Eot3 = Eot2;
end
loops(n)=loops(n)+1; %This is here if the above loop doesn't
    converge after 50 loops and just uses the minimum field.
%Never had the erase state get to this point, combined program will
%look like write state.
if loops(n)>50
    Diff=0;
    signal = 1;
    Eot2 = Ess;
    Eob2 = Eot2*(epox/epob) + (-Eot2*Xeff + phigs + phis - Vp) /
        (Xob + (epob/epn)*Xn/gama));
    Jot2 = Jbtb(Eot2,Xot,Xn,mox_h,phi1_h,phi2_h);
    Job2 = Jbtb(Eob2,Xob,Xn,mox_e,phi1_e,phi2_e);
    Jtat2 = Jtat(Eot2,Xot,mox_h,mn_h,phi1_h,phi2_h,phit,Nt);
end
end
else
    %Repeated section is a hold over from previous people. Will sort
    %this out in combined program.
    Eot2 = Ess;
    Eob2 = Eot2*(epox/epob) + (-Eot2*Xeff + phigs + phis - Vp) /
        (Xob + (epob/epn)*Xn/gama));
    Jot2 = Jbtb(Eot2,Xot,Xn,mox_h,phi1_h,phi2_h);
    Job2 = Jbtb(Eob2,Xob,Xn,mox_e,phi1_e,phi2_e);
    Jtat2 = Jtat(Eot2,Xot,mox_h,mn_h,phi1_h,phi2_h,phit,Nt);
end
%Print result to screen
fprintf('%e
', x);
Eot = [Eot, Eot2];
%dVth = [dVth, Xeff*(Eot2 - Eot0)];
fprintf(Fid,'%e, %e, %e, %e, %e, %e, %e, %e
', t, xeff*(Eot2 - Eot0), Eot2, Eob2, Jot2, Job2, Jtat2, loops(n), intgr);
end
fclose(Fid); %Close file for writing
%Plot the results
Time = [0, Time];
F.2 Write Program

clear all;

%set up constants
q=1.6E-19;
hbar=1.055E-34;
KbTq=0.0259;
m0=9.1094E-31;
mox_e=0.35*m0; %effective electron mass
mox_h=0.42*m0; %effective hole mass
phi1_e=3.1; %oxide barrier heights for electron
phi1_h=3.8; %oxide barrier heights for hole
phi2_e=1.05; %nitride barrier heights for electron
phi2_h=1.85; %nitride barrier heights for hole
phi3_e=0.7; %nitride to Al2O3 barrier heights for electron
Xot=40E-10; %tunnel oxide thickness
Xn=44E-10; %nitride thickness
Xob=150E-10; %blocking Al2O3 thickness
ep0=8.8542E-12; %vacuum permittivity
epSi=11.7*ep0; %permittivity of Si
epox=3.9*ep0; %permittivity of oxide
epn=6.5*ep0; %permittivity of nitride
epob=7.9*ep0; %permittivity of Al2O3
Na=5E23; %substrate doping
ni=1.202E16; %intrinsic doping
phif=KbTq*log(Na/ni); %bulk feimi level
mn_e=(epox/epn)^2*phi1_h*mox_e/(phi1_e-phi2_e); %effective hole mass in
the nitride
mn_h=(epox/epn)^2*phi1_h*mox_h/(phi1_h-phi2_h); %effective hole mass in
the nitride
Xeff=Xot+(epox/epob)*Xob+(epox/epn)*Xn;
Ceff=epox/Xeff;
phiggs=-0.304; %gate to semiconductor workfunction difference
Nt=4E25; %trap density
phit=0.8; %trap energy
gama=2; %nitride centroid coefficient (2: center of the nitride;
1: SiO2, SiN interface; infinite: SiN, blocking oxide interface)
%initial conditions
Vp = 7;  \ \ %programming voltage
Vth0 = -2.5;  \ \ %initial threshold voltage
phis = 2*phif;  \ \ %surface potential at strong inversion
Qn0 = \ (phig + 2*phif + (4*epSi*q*Na*phif)^(1/2)/Ceff - Vth0) / (Xob/epob + Xn/(gama*epn));  \ %initial nitride charge
Eot0 = \ (Vp + (Xob/epob + Xn/(gama*epn))*Qn0 - phigs - phis) / Xeff;  \ %initial tunnel oxide electric field
Eob0 = Eot0*(epox/epob) - Qn0/epob;
Jt10 = Jtat(Eot0,Xot,mox_e,phi1_e,phi2_e);
Jtat0 = Jtat(Eot0,Xot,mox_e,mn_e,phi1_e,phi2_e,phi1_e,phil_e,Nt);
a=(Xob/epob + Xn/(gama*epn))/Xeff;

Fid=fopen('al203_40_8_150.txt','w');
fprintf(Fid,'tp, dVth, Eot, Eob, Jot, Jtat, Job, loops, integr
');
fprintf(Fid,'%e, %e, %e, %e, %e, %e,%e, %d,%e
', 0, 0, Eot0, Eob0, Jot0, Jtat0, 0,0,0);

%initialize array size variable
n = 0;
Eot = [Eot0];
Time = [1E-7:1E-7:1E-6,2E-6:1E-6:1E-5, 2E-5:1E-5:1E-4, 2E-4:1E-4:1E-3,
2E-3:1E-3:1E-2, 2E-2:1E-2:1E-1];
dVth = [0];
Eot2 = Eot0;
%range of variable is a = x:y:z ... for a = x to a = z in steps of y
for t = Time
   %initialize variables
   n=n+1;
   gama=gama.*(-log(t)/log(1E-7));
   if t==1e-3
      gama=20000;
   end

   Eot1=Eot2;
   Eob1=Eot1 - (Eot1*Xeff + phigs + phis - Vp) / (Xob +
   (epox/epn)*(Xn/gama));
   Eot4=Eot2;
   %Eot3=5E8;
   Eot3 = CalE_write(Eot0,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phil_e,phi2_e,phi3_e,phi1_e,phi1_e,phil_e,Nt,phigs,phis,Vp,gama,t);
Eob3 = Eot3 - (Eot3 * Xeff + phigs + phis - Vp) / (Xob + 
(epox / epn) * (Xn / gama));

Diff = 100;
Accuracy = 1;
loops(n) = 0;
while abs(Diff) > Accuracy
        Eot2 = Eot3 + (Eot1 - Eot3) / 2;
        Eob2 = Eot2 - (Eot2 * Xeff + phigs + phis - Vp) / (Xob + 
(epox / epn) * (Xn / gama));
        intgr = 
                Simpson_write(Eot0, Eot2, Eob0, Eob2, Xot, Xn, Xob, mox_e, mn_e, phi1_e, phi2_e,
                                phi3_e, phit, Nt, Xeff, phigs, phis, Vp, epox, epob, epn, gama, t);
        Diff = intgr + a*t;
        if Diff > 0
                Eot1 = Eot2;
        else
                Eot3 = Eot2;
        end
        loops(n) = loops(n) + 1;
        if loops(n) > 30
                %Diff = 0;
                fprintf('approaching\n');
                Eot1 = Eot4;
                Eot2 = Eot4;
                Diff = 100;
                while Diff > 0
                        Eot2 = Eot2 - 1E2/t;
                        Eob2 = Eot2 - (Eot2 * Xeff + phigs + phis - Vp) / (Xob + 
(epox / epn) * (Xn / gama));
                        intgr = 
                                Simpson_write(Eot0, Eot2, Eob0, Eob2, Xot, Xn, Xob, mox_e, mn_e, phi1_e,
                                                phi2_e, phi3_e, phit, Nt, Xeff, phigs, phis, Vp, epox, epob, epn, gama, t);
                        Diff = intgr + a*t;
                        end
                Eot3 = Eot2;
                loops(n) = 0;
                Diff = 100;
                while abs(Diff) > Accuracy
                        Eot2 = Eot3 + (Eot1 - Eot3) / 2;
                        Eob2 = Eot2 - (Eot2 * Xeff + phigs + phis - Vp) / (Xob + 
(epox / epn) * (Xn / gama));
                        intgr = 
                                Simpson_write(Eot0, Eot2, Eob0, Eob2, Xot, Xn, Xob, mox_e, mn_e, phi1_e, phi2_e,
                                                phi3_e, phit, Nt, Xeff, phigs, phis, Vp, epox, epob, epn, gama, t);
Diff = intgr + a*t;
if Diff > 0
Eot1 = Eot2;
else
Eot3 = Eot2;
end
loops(n)=loops(n)+1;
if loops(n)>100
Diff=0;
end
end
end
fprintf('%e
',t);
Eot = [Eot, Eot2];
dVth = [dVth, Xeff*(Eot0 - Eot2)];
Jot2 = Jbtb(Eot2,Xot,Xn,mox_e,phi1_e,phi2_e);
Job2 = Jbtb2(Eob2,Xob,mox_e,phi2_e,t);
Jtat2 = Jtat(Eot2,Xot,mox_e,mn_e,phi1_e,phi2_e,phit,Nt);
fprintf(Fid,'%e, %e, %e, %e, %e, %e, %d, %e, %e
', t, Xeff*(Eot0 - Eot2)+Vth0, Eot2, Eob2, Jot2, Jtat2, Job2, loops(n), intgr);
end
fclose(Fid);

Time = [0, Time];
figure;
semilogx(Time, dVth, 'r');hold on;
grid on;

F.3 Jbtb Program

function f = Jbtb(Eot,Xot,Xn,mox,phi1,phi2,phit) %calculate the band-to-band tunneling current in dual dielectric case
q = 1.6E-19;
hbar = 1.055E-34;
if Eot >= phi1 ./ Xot
   Ct = (q.^2).*(Eot.^2) ./ (16 .* (pi.^2) .* hbar .* phi1);
   Pox = exp (-4 .* ((2.*q.*mox).^(0.5)) .* (phi1.^(1.5)) ./ (3 .* hbar .* Eot));
   Pn = 1;
elseif \( E_{ot} \geq \frac{(\phi_1 - \phi_2)}{X_{ot}} \)

\[
C_t = \frac{(q^2)(E_{ot}^2)}{16 \pi^2 \hbar} \left( \frac{\phi_1^{0.5} - (\phi_1 - E_{ot}X_{ot})^{0.5}}{(-2)} \right)^{-2};
\]

\[
P_{ox} = \exp \left( \frac{-4 \sqrt{2m_{ox}q}}{3 \hbar E_{ot}} \left( \phi_1^{1.5} - (\phi_1 - E_{ot}X_{ot})^{1.5} \right) \right);
\]

\( P_n = 1; \)

elseif \( E_{ot} \geq \frac{(\phi_1 - \phi_2)}{(X_{ot} + 0.6 \cdot X_n)} \)

\[
C_t = \frac{(q^2)(E_{ot}^2)}{16 \pi^2 \hbar} \left( \frac{\phi_1^{0.5} - (\phi_1 - E_{ot}X_{ot})^{0.5} + \left(\frac{\phi_1}{\phi_1 - \phi_2}\right)^{0.5} \cdot (\phi_1 - \phi_2 - E_{ot}X_{ot})^{0.5}}{(-2)} \right)^{-2};
\]

\[
P_{ox} = \exp \left( \frac{-4 \sqrt{2m_{ox}q}}{3 \hbar E_{ot}} \left( \phi_1^{1.5} - (\phi_1 - E_{ot}X_{ot})^{1.5} \right) \right);
\]

\[
P_n = \exp \left( \frac{-4 \sqrt{2m_{ox}q}}{3 \hbar E_{ot}} \left( \frac{\phi_1}{\phi_1 - \phi_2} \right)^{0.5} \cdot (\phi_1 - \phi_2 - E_{ot}X_{ot})^{1.5} \right);
\]

else

\( C_t = 0; \)

\( P_{ox} = 0; \)

\( P_n = 0; \)

end

\[
f = C_t \cdot P_{ox} \cdot P_n;
\]

---

**F.4 Jtat Program**

```matlab
function f = Jtat(Eot,Xot,Xn,mox,mn,phi1,phi2,phit,Nt) %calculate the
\  trap-assisted-tunneling current

q = 1.6E-19;
\hbar = 1.055E-34;
KbTq = 0.0259; % kT/q
tau0 = 1E-13;
gamma = 0.7; % epox/epn

if (Eot >= (phi1-phi2-phi2) ./ (Xot + 0.6 .* Xn)) && (Eot <=
(phi1-phi2-phi2) ./ Xot)
    kox = sqrt(2.*mox.*q.*phi1) ./ hbar;
    kn = sqrt(2.*mn.*q.*phit) ./ hbar;
    phi3 = phi1-phi2-phi2;
    f = (q .*Nt ./ (tau0 .* (-2 .*kn + gamma .*Eot./KbTq))) .* exp (
        -2.*(kox - kn./gamma).*Xot - 2.*kn.*phi3/(gamma.*Eot));
else
    f = 0;
end
```
F.5 Simpson Erase Program

```matlab
function f = Simpson_erase6(Eot0,Eot2,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_ot,phi2_ot,phi1_ob,phi2_ob,phit,Nt,phigs,phis,Vp,t)

% set up constants
ep0=8.8542E-12;  % vacuum permittivity
epox=3.9*ep0;     % permittivity of oxide
epn=6.5*ep0;      % permittivity of nitride
epob=3.9*ep0;     % permittivity of oxide

Xeff=Xot+Xob+(epox ./ epn).*Xn;
Ceff=epox ./ Xeff;
Eob0 = Eot0 + (-Eot0 .* Xeff + phigs + phis - Vp) ./ (Xob +
    (epox./epn).*(Xn./2));
Eob2 = Eot2 + (-Eot2 .* Xeff + phigs + phis - Vp) ./ (Xob +
    (epox./epn).*(Xn./2));
fa = Jbtb(Eot0,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
    Jtat(Eot0,Xot,Xn,mox_h,mn_h,phi1_ot,phi2_ot,phit,Nt);
fa = 1 ./ fa;
fb = Jbtb(Eot2,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
    Jtat(Eot2,Xot,Xn,mox_h,mn_h,phi1_ot,phi2_ot,phit,Nt);
fb = 1 ./ fb;

n = 2^10;
dx = (Eot2 - Eot0) ./ n;
i = 1;
sum1 = 0;
while i < n
    Eot = Eot0 + i .* dx;
    Eob = Eot + (-Eot .* Xeff + phigs + phis - Vp) ./ (Xob +
        (epox./epn).*(Xn./2));
    temp1 = Jbtb(Eot,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
        Jtat(Eot,Xot,Xn,mox_h,mn_h,phi1_ot,phi2_ot,phit,Nt);
    temp1 = 1 ./ temp1;
    sum1 = sum1 + temp1;
i = i + 2;
end
i = 2;
sum2 = 0;
```

249
while i < n
    Eot = Eot0 + i .* dx;
    Eob = Eot + (-Eot .* Xeff + phigs + phis - Vp) ./ (Xob +
              (epox./epn).*(Xn./2));
    temp2 = Jbtb(Eot,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
            Jtat(Eot,Xot,Xn,mox_h,mn_h,phi1_ot,phi2_ot,phit,Nt);
    temp2 = 1 ./ temp2;
    sum2 = sum2 + temp2;
    i = i + 2;
end

f = (fa + 4.*sum1 + 2.*sum2 + fb) .* dx ./ 3;

F.6 Simpson Write Program

function f =
    Simpson_write(Eot0,Eot2,Eob0,Eob2,Xot,Xn,Xob,mox_e,mn_e,phi1_ot,phi2_ot,
                   phi3_e,phit,Nt,Xeff,phigs,phis,Vp,epox,epob,epn,gama,t)

m0=9.1094E-31;
mox_e=0.35*m0;    %effective electron mass
mox_h=0.42*m0;    %effective hole mass

fa = Jbtb(Eot0,Xot,Xn,mox_e,phi1_ot,phi2_ot) +
    Jtat(Eot0,Xot,mox_e,mn_e,phi1_ot,phi2_ot,phit,Nt)-
    Jbtb2(Eob0,Xob,mox_e,phi3_e,t)+ Jbtb3(Eot0,Xot,mox_h,3,t);
fa = 1 ./ fa;
fb = Jbtb(Eot2,Xot,Xn,mox_e,phi1_ot,phi2_ot) +
    Jtat(Eot2,Xot,mox_e,mn_e,phi1_ot,phi2_ot,phit,Nt)-
    Jbtb2(Eob2,Xob,mox_e,phi3_e,t)+ Jbtb3(Eot2,Xot,mox_h,3,t);
fb = 1 ./ fb;

n = 2^10;
dx = (Eot2 - Eot0) ./ n;
i = 1;
sum1 = 0;
while i < n
    Eot = Eot0 + i .* dx;
    Eob = Eot*(epox/epob) - (Eot*Xeff + phigs + phis - Vp) / (Xob +
              (epox/epn)*(Xn/gama));
    temp1 = Jbtb(Eot,Xot,Xn,mox_e,phi1_ot,phi2_ot) +
            Jtat(Eot,Xot,mox_e,mn_e,phi1_ot,phi2_ot,phit,Nt)-
            Jbtb2(Eob,Xob,mox_e,phi3_e,t)+ Jbtb3(Eot,Xot,mox_h,3,t);
temp1 = 1 ./ temp1;
sum1 = sum1 + temp1;
i = i + 2;
end
i = 2;
sum2 = 0;
while i < n
    Eot = Eot0 + i .* dx;
    Eob = Eot*(epox/epob) - (Eot*Xeff + phigs +phis - Vp) / (Xob + 
    (epox/epn)*(Xn/gama));
    temp2 = Jbtb(Eot,Xot,Xn,mox_e,phi1_ot,phi2_ot) + 
    Jtat(Eot,Xot,mox_e,mn_e,phi1_ot,phi2.ot,phit,Nt)-
    Jbtb2(Eob,Xob,mox_e,phi3_e,t)+ Jbtb3(Eot,Xot,mox_h,3,t);
    temp2 = 1 ./ temp2;
    sum2 = sum2 + temp2;
    i = i + 2;
end

f = (fa + 4.*sum1 + 2.*sum2 + fb) .* dx ./ 3;

F.7  Modified Simpson Write Program

function f =
    Simpson_writeh(Eot0,Eot2,Eob0,Eob2,Xot,Xn,Xob,mox_h,mox_e,mn_e,phi1_ot,
    phi2.ot,phi3_e,phit,Nt,Xeff,phigs,phis,Vp,epox,epob,epn,gama,t)

m0=9.1094E-31;
mox_e=0.35*m0;  %effective electron mass
mox_h=0.42*m0;  %effective hole mass

fa = Jbtb(Eot0,Xot,Xn,mox_h,phi1_ot,phi2.ot) + 
    Jtat(Eot0,Xot,mox_h,mn_e,phi1_ot,phi2.ot,phit,Nt)-
    Jbtb2(Eob0,Xob,mox_h,phi3_e,t)+ Jbtb3(Eot0,Xot,mox_e,3,t);
fa = 1 ./ fa;
fb = Jbtb(Eot2,Xot,Xn,mox_h,phi1_ot,phi2.ot) + 
    Jtat(Eot2,Xot,mox_h,mn_e,phi1.ot,phi2.ot,phit,Nt)-
    Jbtb2(Eob2,Xob,mox_h,phi3_e,t)+ Jbtb3(Eot2,Xot,mox_e,3,t);
fb = 1 ./ fb;

n = 2^-10;
dx = (Eot2 - Eot0) ./ n;
i = 1;
sum1 = 0;
while i < n
    Eot = Eot0 + i .* dx;
    Eob = Eot*(epox/epob) - (Eot*Xeff + phigs + phis - Vp) / (Xob +
    (epox/epn)*(Xn/gama));
    temp1 = Jbtb(Eot,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
            Jtat(Eot,Xot,mox_h,mn_e,phi1_ot,phi2_ot,phit,Nt)-
            Jbtb2(Eob,Xob,mox_h,phi3_e,t)+ Jbtb3(Eot,Xot,mox_e,3,t);
    temp1 = 1 ./ temp1;
    sum1 = sum1 + temp1;
    i = i + 2;
end
i = 2;
sum2 = 0;
while i < n
    Eot = Eot0 + i .* dx;
    Eob = Eot*(epox/epob) - (Eot*Xeff + phigs + phis - Vp) / (Xob +
    (epox/epn)*(Xn/gama));
    temp2 = Jbtb(Eot,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
            Jtat(Eot,Xot,mox_h,mn_e,phi1_ot,phi2_ot,phit,Nt)-
            Jbtb2(Eob,Xob,mox_h,phi3_e,t)+ Jbtb3(Eot,Xot,mox_e,3,t);
    temp2 = 1 ./ temp2;
    sum2 = sum2 + temp2;
    i = i + 2;
end

f = (fa + 4.*sum1 + 2.*sum2 + fb) .* dx ./ 3;

%----------------------------------------------------------

F.8 Calculate Write Electric Field Program

function f =
    CalE_write(Eot0,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_ot,phi2_ot,phi3_e,phit,
         Nt,phigs,phis,Vp,gama,t);

    %Calculate the tunnel oxide Electrical field when write/erase reaches
    %saturation. That is the tunneling currents reach equilibrium,
    %Jot+Jbtb_ot=Jbtb_ob.

    %set up constants
    ep0 = 8.8542E-12;  %vacuum permittivity
    epox = 3.9.*ep0;    %permittivity of oxide
    epn = 5.5.*ep0;     %permittivity of nitride

    Xeff=Xot+Xob+(epox ./ epn).* Xn;
    Ceff=epox ./ Xeff;
DiffI = 100;
AccuI = 1E-10;
Eot_s = Eot0;

while DiffI > AccuI
    Eot_s = Eot_s - 1E5;
    Eob_s = Eot_s - (Eot_s .* Xeff + phigs + phis - Vp) ./ (Xob + (epox ./ epn) .* (Xn./gama));
    DiffI = Jbtb(Eot_s,Xot,Xn,mox_e,phi1_ot,phi2_ot) + 
    Jtat(Eot_s,Xot,mox_e,mn_e,phi1_ot,phi2_ot,phit,Nt) - 
    Jbtb2(Eob_s,Xob,mox_e,phi3_e,t);
end

f = Eot_s;

F.9 Calculate Erase Electric Field Program

%For Al2O3
function f = CalE4(Eot0,Xot,Xn,Xob,mox_e,mox_h,mn_e,mn_h,phi1_ot,phi2_ot,
    phi1_ob,phi2_ob,phit,Nt,phigs,phis,Vp,t);

%Calculate the tunnel oxide Electrical field when write/erase reaches saturation. That is the tunneling currents reach equilibrium, Jot+Jbtb_ot=Jbtb_ob.

%set up constants
ep0 = 8.8542E-12;  \%vacuum permittivity
epox = 3.9.*ep0;  \%permittivity of oxide
epn = 6.5.*ep0;  \%permittivity of nitride
epob = 9.0.*ep0;  \%permittivity of Al2O3
Xeff=Xot+(epox ./epob).*Xob+(epox ./ epn) .* Xn;
Ceff=epox ./ Xeff;
phi1_e=3.1;
DiffI = 100;
AccuI = 1E-10;
Eot_s = Eot0;

%Find the tunnel E-field at which the tunnel currents are equal
while DiffI > AccuI
    Eot_s = Eot_s - 1E5;
    \%Eob_s = Eot_s - (Eot_s .* Xeff + phigs + phis - Vp) ./ (Xob + (epox ./ epn) .* (Xn./2));
end
Eob_s = Eot_s.*epox ./epob + (-Eot_s .* Xeff + phigs + phis - Vp) ./
(Xob + (epob ./ epn) .* (Xn./2));

%DiffI = Jbtb(Eot_s,Xot,Xn,mox_h,phi1_ot,phi2_ot) +
        Jtat(Eot_s,Xot,mox_e,mn_e,phi1_ot,phi2_ot,phit,Nt) -
        Jbtb2(Eob_s,Xob,mox_e,phi1_e,t);

DiffI =
        Jbtb(Eot_s,Xot,Xn,mox_h,phi1_ot,phi2_ot)-Jbtb(Eob_s,Xob,Xn,mox_e,
        phi1_ob,phi2_ob);

end

%.*epox ./epob
% +Jtat(Eot_s,Xot,mox_h,mn_h,phi1_ot,phi2_ot,phit,Nt)
%Output tunnel E-field to program
f = Eot_s;
Bibliography


[96] The Ohio State University. Fei helios nanolab 600 dualbeam (fib/sem), 2014. [Online; accessed 4-June-2014].


