ABSTRACT

SCHINKE, DANIEL JOHANNES. Computing with Novel Floating Gate Devices. (Under the direction of Dr. Paul Franzon.)

MOSFETs employing programmable floating gates have drawn much attention in the semiconductor memory market in the past decade and are found in many applications today such as computers, MP3 players, digital cameras, mobile phones or PDAs. Regardless of their enormous success, there has been little investigation of use of these nonvolatile devices outside of persistent storage memories. One reason for this is because they have been engineered to achieve long data retention time (>10 years) while operating voltage, switching speed and endurance are compromised in favor of retention time. Lower operating voltages and faster switching can be achieved by using band-gap engineered gate stacks, multiple metal floating gates, thinner oxides, and tunneling as the main programming mechanism. In principle, this also enables reduced stresses which results in increased endurance. Such devices can then be used as logic circuit elements in interesting applications with the goal to minimize power consumption, maximize performance, and provide high functionality.

This dissertation suggests that nanocrystal floating gate devices may be very useful for static applications such as a FPGA. Since FPGAs are typically reconfigured only a few times during their lifetime, fast switching programmable elements such as the SRAMs are not needed and only consume a large fraction of board area and power. They can be replaced by these nonvolatile devices to enable an area and power efficient FPGA board where a time-consuming configuration on every boot up is no longer required.

Furthermore, this dissertation presents the invention of a new unified memory device with merged dynamic and nonvolatile memory operations. It may also operate in both modes simultaneously as a 2-bit device. It is based on a dual floating gate structure that can potentially be built with high write cycle endurance in the dynamic memory domain by using direct tunneling for programming and/or metal nanocrystals in the floating gates. It is scalable to at least the 16 nm node and has potential for 3D stacking for ultra-high density. Applications of such a memory include instant-on computing, power proportional computing, and resilient computing. It could also be very useful in network on chips, synaptic processors or other applications that could benefit from a unified memory with multi-bit capability.

© Copyright 2011 by Daniel Johannes Schinke

All Rights Reserved

Computing with Novel Floating Gate Devices

by
Daniel Johannes Schinke

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina

2011

APPROVED BY:

Dr. Veena Misra

Dr. W. Rhett Davis

Dr. Paul Franzon
Chair of Advisory Committee

Dr. Lew Reynolds

DEDICATION

To my parents

BIOGRAPHY

Daniel Johannes Schinke was born in Cologne, Germany on July 19, 1984. He received his high school diploma in 2003 in Pulheim, Germany. He attended the Rheinisch-Westfälische Technische Hochschule (RWTH) in Aachen, Germany from 2003 to 2006 with the major in Electrical Engineering where he received his “Vordiplom” degree and completed one year of the “Hauptstudium”. In 2006 he transferred to North Carolina State University to attend the Masters program in Electrical Engineering. In 2008 he graduated with the Master of Science degree and during those two years Daniel was an active student athlete and competed with the collegiate tennis team in the division I. After graduation, he remained in North Carolina State University to pursue a Ph.D. degree in Electrical Engineering.

ACKNOWLEDGEMENTS

I would like to thank several people that have supported and assisted me throughout my graduate work and without them this achievement would not have been possible.

First of all, I would like to thank my advisor Dr. Paul Franzon for his excellent guidance, mentorship and constant support. His professionalism and enthusiasm were an inspiration and has developed me tremendously as an academic student.

I would like to express my gratitude to my committee members Dr. Veena Misra, Dr. W. Rhett Davis and Dr. Lew Reynolds. They were always accessible to me and have improved this dissertation with their feedback and support. Thank you Dr. Veena Misra for assisting in the fabrication of the unified memory device described in Chapter 5 in this dissertation.

I am very grateful to my research group, which includes Dr. Neil Di Spigna, W. Sheperd Pitts, Shivam Priyadarshi, Mihir Shiveshwarkar and Gary Charles. They have treated me with a lot of respect and it was a great pleasure working with them. Special thanks to Dr. Neil Di Spigna who always assisted me and gave me very helpful feedback throughout my entire work. He also fabricated the unified memory device which is greatly appreciated.

I wish to express my deepest appreciation to my parents Cecylia and Johannes and my three sisters Carolina, Anita and Susanne. They have always encouraged me and gave me their endless support. Lastly, I would like to thank my fiancée Cristina. Your love and support is completely overwhelming and you have my deepest gratitude.

TABLE OF CONTENTS

List of Tables	ix
List of Figures	x
Chapter 1 Introduction	1
Chapter 2 Literature Review	3
2.1 Introduction	3
2.2 SRAM	6
2.3 DRAM	7
2.4 Nonvolatile Memory	8
2.4.1 Continuous Floating Gate Device	9
2.4.2 Nanocrystal Floating Gate Device	14
2.4.2.1 Silicon vs. Metal Nanocrystals	17
2.4.2.2 Coulomb Blockade	20
2.4.2.3 Advanced Nanocrystal Floating Gate Devices	22
2.4.3 Emerging Nonvolatile Memory Technologies	26
2.4.3.1 FeRAM	26
2.4.3.2 MRAM, SPRAM	27
2.4.3.3 PCM	28
2.4.3.4 Comparison to Flash Memory	29
2.4.3.5 Future Prospects	30
2.5 Summary	32
References	34
Chapter 3 A SPICE Compatible Physical Model of Nanocrystal Floating Gate Devices for Circuit Simulation	40
3.1 Motivation	40
3.2 Introduction	41
3.3 Reference Model	43

3.4 Physical Model	46
3.4.1 Algorithm for Computing the Floating Gate Voltage	47
3.4.2 Surface Potential	54
3.4.2.1 Accumulation	55
3.4.2.2 Weak Inversion	56
3.4.2.3 Strong Inversion	59
3.4.2.4 Summary	60
3.4.3 Program and Erase Mode	61
3.5 Results and Discussion	63
3.5.1 Comparison to the Device Simulator	64
3.5.2 Comparison to Experimental Data from Fabricated Devices	65
3.5.2.1 Threshold Voltage Shift Comparison	66
3.5.2.2 Retention Time Comparison	67
3.5.2.3 Flat-Band Voltage Shift Comparison	69
3.5.3 HSPICE Circuit Simulation with Implementation of the Physical Model	71
3.5.4 Run Time Analysis	72
3.6 Summary	74
References	75
Chapter 4 Low Power Interconnect Design for FPGAs with Bidirectional Wiring Using Nanocrystal Floating Gate Devices	78
4.1 Introduction	78
4.2 Design and Driving Method of the Nanocrystal Floating Gate Device	81
4.3 Interconnect Design	89
4.3.1 Switch Box	89
4.3.2 Connection Block	93
4.4 Area, Speed and Power Analysis	96
4.5 Summary	99
References	100

Chapter 5 Unified Memory Device	103
5.1 Introduction	103
5.2 Device Structure and Operation	105
5.2.1 Dynamic Memory Operation	108
5.2.2 Nonvolatile Memory Operation	115
5.2.3 Combined Dynamic and Nonvolatile Memory Operation: 2-Bit Device	120
5.3 Fabricated Test Device	124
5.4 Advanced Technology Using a Back Gate and SOI for Selectivity and Improved Scalability	126
5.5 Physical Model	131
5.6 Temperature Dependency	136
5.7 Employing Discontinuous Floating Gates	138
5.8 Challenges and Potential Solutions	147
5.9 Applications	148
5.10 Summary	150
References	151
Chapter 6 Unified Memory Array	155
6.1 Motivation	155
6.2 Introduction	156
6.3 Circuit Design and Analysis	159
6.4 Comparison to DRAM	167
6.5 Implications on the Memory Hierarchy of the Computer Architecture	171
6.5.1 Unified Volatile and Nonvolatile Main Memory	173
6.5.2 Unified Solid State Disk Drive and Main Memory	174
6.6 Summary	175
References	176
Chapter 7 Conclusion and Future Work	178
7.1 Conclusion	178

7.2 Future Research Objectives	180
References	182

LIST OF TABLES

Table 2.1	Typical properties of SiO ₂ and HfO ₂	23
Table 2.2	Comparison of the conventional single-layer NCFG device and its modified versions	25
Table 2.3	Comparison between the Flash memory and emerging nonvolatile memory alternatives	30
Table 2.4	Description and key attributes of the most common memory Technologies	32
Table 3.1	Device geometry parameters used for the reference model in the device simulator	45
Table 3.2	Accuracy and simulation time measured for various time steps for both the device simulator and the physical model	73
Table 4.1	Comparison between the driving methods of the NCFG device, the continuous floating gate device and the 6T-SRAM cell	85
Table 4.2	Configuration of the NCFG devices through the controller	89
Table 4.3	Power consumption results for the 32-tap FIR Filter implemented as a FPGA	99
Table 5.1	Comparison results between the physical model and the reference model from the device simulator of the unified memory device for both the dynamic and nonvolatile (NV) memory mode	136
Table 5.2	Geometry parameters of the dual-layer NCFG device	141
Table 5.3	Dynamic and static memory operation of the designed discontinuous dual floating gate device	147
Table 6.1	Read time analysis	161
Table 6.2	States of a unified memory cell before and after the applied voltage envelope on the word line (WL) and select line (SL). The data is obtained from Sentaurus TCAD simulations together with circuit level HSPICE simulations	165

LIST OF FIGURES

Figure 2.1	Memory market growth over the past decade [2.2]	3
Figure 2.2	6T-SRAM cell	6
Figure 2.3	1T-1C DRAM cell	7
Figure 2.4	Schematic cross-section of a continuous floating gate device	10
Figure 2.5	Threshold voltage of an uncharged, programmed and erased device	11
Figure 2.6	Mechanisms of the charge transfer in a floating gate device during (a) the program mode and (b) the erase mode. Note that the band bending in the polysilicon floating gate and control gate is not shown	12
Figure 2.7	(a) Schematic cross-section of a NCFG device and (b) TEM image of Si nanocrystals on top of the SiO ₂ [2.17]	14
Figure 2.8	SILC susceptibility of (a) a continuous floating gate device and (b) a NCFG device	16
Figure 2.9	Band diagrams of a NCFG device with low and high work function metal in the (a) program, (b) retention, and (c) erase mode	18
Figure 2.10	Theoretical coulomb charging energy of Pd nanocrystals with varying diameter comparing to the thermal energy (3-4 k _B T) [2.26]	20
Figure 2.11	Threshold voltage of a NCFG device for 40 K, 77 K, and 300 K as a function of the control gate voltage [2.24]	21
Figure 2.12	Conventional NCFG device (I) and its modified versions (II-VI)	22
Figure 2.13	Schematic cross-section of a FeRAM cell with a stacked capacitor [2.41]	26
Figure 2.14	(a) Conventional MRAM cell [2.43] and (b) SPRAM cell [2.43]	27
Figure 2.15	Typical structure of the PCM [2.44]	28
Figure 2.16	Status of the widely known emerging nonvolatile memories [2.51]	31
Figure 3.1	A single-layer NCFG device with 45 nm gate length designed in the device simulator	44
Figure 3.2	Block diagram of the physical model	46
Figure 3.3	Design flow of computing the floating gate voltage V _{FG} in the Verilog-A module	48

Figure 3.4	Capacitor model between the control gate, floating gate and substrate surface nodes of the NCFG device	49
Figure 3.5	Band diagram across the Pt-SiO ₂ -Pd-SiO ₂ -Si stack when a high voltage is applied on the control gate	51
Figure 3.6	Transforming the Gaussian distribution of the ions in the channel to a constant distribution with equivalent doping density	57
Figure 3.7	Comparison of the Gaussian distribution of the ions implantation in the channel between the reference model and the physical model	58
Figure 3.8	Comparison of the surface potential V_{surface} between the reference model and the physical model for the control gate voltage V_{CG} ranging from -7 to 7 V	60
Figure 3.9	Band diagram of the NCFG device with $d_{\text{cox}} = 4$ nm and $d_{\text{tox}} = 2$ nm in (a) the program mode and (b) the erase mode. The thick solid line represents the potential	62
Figure 3.10	Comparison of ΔV_T between the physical model and the device simulator for (a) the program mode and a programming time of 10 μ s and (b) the erase mode and an erasing time of 1 ms	65
Figure 3.11	(a) Experimental results of V_T for programming and erasing a NCFG device with a nanocrystal surface coverage of 50% (inset) [3.26]. (b) The V_T curves are computed with the physical model	67
Figure 3.12	Comparison of the retention time results for (a) a fixed V_{CG} of 0 V and different d_{tox} [3.27] and (b) a fixed device ($d_{\text{tox}} = 4$ nm and $d_{\text{cox}} = 7$ nm) with different V_{CG} [3.27]	69
Figure 3.13	C-V comparison between the physical model and experimental results from [3.12]. The gate voltage is swept from -5 V to +5 V and back to -5 V	70
Figure 3.14	DC simulation in HSPICE of an inverter with the NCFG device using the physical model and replacing the NMOS transistor. This device is cycled consecutively through an uncharged, programmed, retention and erased state	72
Figure 4.1	Simplified FPGA architecture with a 4-channel routing network	79
Figure 4.2	Power consumption analysis of traditional SRAM-based FPGAs [4.6, 4.7] ...	80
Figure 4.3	Cross-section of a NCFG device in 45 nm technology	82

Figure 4.4	Driving method via word line (WL) and select line (SL) for (a) the new design with a NCFG device, (b) the design Actel uses with continuous floating gate devices, and (c) the 6T-SRAM cell	84
Figure 4.5	FPGA configuration	86
Figure 4.6	Voltage envelopes required for the designed device for (a) the program mode and (b) the erase mode	88
Figure 4.7	Switch box design for (a) a traditional bidirectional [4.15] and (b) a modern unidirectional SRAM-based FPGA [4.15]	91
Figure 4.8	NCFG-based switch box design with bidirectional routing switches attached at each output node	92
Figure 4.9	Connection block from the routing channel (RCh) to the input of the logic block for traditional and modern SRAM-based FPGAs in a 4-channel routing system	94
Figure 4.10	Connection block for transferring the signal in a 4-channel routing system by use of NCFG devices from (a) the source logic block output to the routing channel (RCh) and (b) the routing channel to the target logic block input	95
Figure 4.11	The logic block used in each FPGA architecture consists of an 8-input LUT, a multiplexer (Mux), and a Flip Flop	97
Figure 5.1	(a) Device structure. (b) Band diagram across the gate stack under flat-band condition. (c) Oxide material and thickness. (d) Metal electrode material and thickness	106
Figure 5.2	Band diagram of the device from Figure 5.1 for the dynamic memory mode for a control gate voltage of (a) 4 V and (b) -4 V	110
Figure 5.3	(a) Threshold voltage shift and (b) flat-band voltage shift when the device is in the nonvolatile uncharged state and consecutively cycled through the dynamic program, retention, and refresh mode	111
Figure 5.4	Charge redistribution on the floating gate pair with respect to time when the device is in the nonvolatile uncharged state and consecutively cycled through the dynamic (a) program, (b) retention, and (c) refresh mode	114
Figure 5.5	Band diagrams of the device from Figure 5.1 for the nonvolatile memory operation for a control gate voltage of (a) 8 V and (b) -8 V	115

Figure 5.6	(a) Threshold voltage shift and (b) flat-band voltage shift when the device is consecutively cycled through the nonvolatile program mode, charge settling after the program mode, erase mode, and charge settling after the erase mode	116
Figure 5.7	Charge redistribution on the floating gate pair with respect to time when the device is consecutively cycled through (a) the nonvolatile program mode, (b) charge settling after the program mode, (c) erase mode, and (d) charge settling after the erase mode	120
Figure 5.8	(a) Threshold voltage shift and (b) flat-band voltage shift when the device is in the nonvolatile charged state and consecutively cycled through the dynamic program, retention, and refresh mode	122
Figure 5.9	Charge redistribution on the floating gate pair with respect to time when the device is in the nonvolatile charged state and consecutively cycled through (a) the dynamic program, (b) retention, and (c) refresh mode	123
Figure 5.10	Fabrication of a dual floating gate test device (image courtesy of Neil Di Spigna, North Carolina State University). (a) Transmission electron micrograph (TEM) of the cross-section, and (b) flat-band voltage shift demonstrating the dynamic and nonvolatile memory operation of the device	125
Figure 5.11	Cross-section and doping profile of the new device in 16 nm technology ...	127
Figure 5.12	I-V simulations of the device from Figure 5.11	128
Figure 5.13	Comparison of the (a) charging and (b) discharging behavior between the device designed in 65 nm technology and the device designed in 16 nm	129
Figure 5.14	Capacitor model between the control gate, the floating gate pair and the substrate surface nodes of the unified memory device	132
Figure 5.15	Band diagram of the unified memory device in (a) the program mode and (b) the erase mode. The thick solid line represents the potential	134
Figure 5.16	Temperature dependency in the dynamic (a) program and (b) retention mode	138
Figure 5.17	(a) Cross-section of the dual-layer NCFG device and (b) its corresponding conduction band diagram under flat-band condition	140
Figure 5.18	Conduction band diagram of the dual layer NCFG device under low stress of (a) 4 V and (b) -4 V	142

Figure 5.19	Drain current before and after a voltage pulse of ± 4 V on (a) a linear and (b) a logarithmic scale	143
Figure 5.20	Retention time results for (a) the threshold voltage shift ΔV_T and (b) $\Delta(\Delta V_T)$	144
Figure 5.21	Conduction band diagram of the new device under high-stress for (a) the program mode with 13 V and (b) the erase mode with -13 V	145
Figure 5.22	Threshold voltage shift of (a) programming the device for 100 μ s and (b) erasing the device for 10 ms	146
Figure 5.23	Drain current after program and erase operation on a) a linear and b) a logarithmic scale	146
Figure 6.1	A conventional DRAM opposed to a hybrid DRAM/Flash main memory [6.3]	157
Figure 6.2	Memory array using the dual floating gate device	160
Figure 6.3	HSPICE input waveforms for the word lines (WL1-WL4) and the select lines (SL1-SL4)	162
Figure 6.4	HSPICE input waveforms for the clock (CLK) and word lines WL1-WL4 and output waveforms of the sense amplifiers SA1-SA4	163
Figure 6.5	(a) Leakage current vs. drain voltage for different temperatures [6.11]. (b) Cumulative bit failure vs. retention time for different temperatures [6.11]	169
Figure 6.6	(a) Traditional memory hierarchy. (b) Unified memory hierarchy	172

Chapter 1

Introduction

This dissertation deals with novel floating gate devices with the motivation to not only use them as pure storage elements but also as logic circuit elements within complementary metal-oxide-semiconductor (CMOS) logic. Device, circuit and architecture issues are simultaneously investigated because coupled multi-level research drives sophisticated innovations that single discipline research would not. As such, by appropriately engineering and optimizing these novel devices for certain implementations and finding new ideas, architectural and circuit issues can be solved.

The literature review in Chapter 2 gives an insight about existing volatile and nonvolatile memory technologies. Emerging memory devices that may play a significant role in the foreseeable future are also presented. The main focus is the floating gate device where device physics, charge transfer mechanisms and variations are studied.

In order to investigate the use of the floating gate devices in circuits a physical model of these devices is required that is compatible with the existing SPICE-based infrastructure. Chapter 3 presents a comprehensive, accurate, and configurable model of the nanocrystal floating gate device that requires short computation time and is therefore suitable for large and complex circuit simulations. It interacts dynamically with the rest of the circuit such that power consumption and speed analysis of the circuit is permitted.

Chapter 4 suggests that a field programmable gate array (FPGA) is a suitable static application of these nanocrystal floating gate devices. A new interconnect is presented that utilizes these devices with the goal to reduce area and power consumption while maintaining good performance compared to commercial FPGAs that use static random access memories (SRAMs) to configure the board. The new design also demonstrates high functionality.

The invention of a new unified memory device is described in Chapter 5. It is based on a dual floating gate structure that may be used for both dynamic and nonvolatile memory operation in a single device. The mode of operation is clearly distinguished by the applied voltages. Furthermore, it permits operation in both memory domains simultaneously and the device can operate as a 2-bit device. The device may have a huge impact for solving problems in instant-on computing, energy-proportional computing, and improving resiliency. The device may also have great potential for implementation in addressable memories, embedded look-up-tables, circuit routers in network on chips (NoCs), synaptic processors and other applications as they emerge.

The most obvious application of such a unified memory device is instant-on computing and is investigated in Chapter 6. A memory array is designed that can quickly switch between an active mode and hibernation. It can be brought in and out of partial sleep states to reduce the power consumption when operating in an idle mode. As the device has the ability for 2-bit operation, efficient check-pointing can be utilized (storing a state to enable later recovery on an error). This unified memory array has potential to outperform the conventional dynamic random access memory (DRAM) in several aspects.

Chapter 7 concludes the work and gives an insight into the future work.

Chapter 2

Literature Review

2.1 Introduction

The semiconductor memory market plays a significant role in the integrated circuit design and has in fact been one of the greatest driving forces in the advancement of solid-state technology. The memory chip capacity has boomed due to the development in technology and design of existing memories and every two years the chip capacity has doubled for diverse storage media [2.1]. In addition, new memory technologies are under extensive research to maintain this trend in the future. The rapid growth of the semiconductor memory market in the past decade is illustrated in Figure 2.1 [2.2]. It is forecasted that in 2012 the market will exceed \$100 billion.

Figure 2.1: Memory market growth over the past decade [2.2].

The semiconductor memory can be divided into two categories: the volatile and the nonvolatile memory. In the volatile memory, the information is saved only as long as the system power is on. The SRAM and the DRAM fall into this category. The SRAM is typically used in the cache memory of a central processing unit (CPU) due to its very high switching speed between two states. The DRAM instead is mainly used in the main memory due to its small cell size. It can be built with a high density, which enables overall high speed data access to recently used data.

In the nonvolatile memory, the information is stored even if the power is off. There are more than 30 technologies and technology variations that in part aim to replace the DRAM [2.3] in the main memory. The most common nonvolatile memory is the Flash memory, which is based on metal-oxide-semiconductor field-effective-transistors (MOSFETs) with programmable floating gates embedded in the dielectrics. These so-called floating gate devices have enjoyed enormous success as the basis for a scalable nonvolatile memory technology over the past years as illustrated in Figure 2.1. It is available in a NOR and NAND architecture and it is the dominant nonvolatile memory technology with a share of almost 100% in the nonvolatile memory market in 2004 [2.3]. Furthermore, it is estimated that the Flash memory will grow with a higher average annual rate than DRAM and SRAM, reaching \$50 billion in 2011 [2.4]. Next to its nature of being nonvolatile, it has other attractive features such as high density, very low idle power consumption, and low cost. It is in use for digital cameras, cellular phones and other portable electronic equipment. It has also been proposed to implement Flash memory in a hybrid main memory in combination with DRAM [2.5, 2.6] to prevent the slow data transfer between the DRAM and hard disk

and to consume less power. Regardless of this, there has only been limited investigation to use the floating gate devices outside of persistent storage memories to date [2.7-2.12]. One reason for this is because they have been optimized primarily towards maximizing retention time while trading off both endurance (number of program/erase cycles) and operating voltage. In addition, their switching speed is currently too low to be integrated e.g. in a Flash-only main memory.

There are other nonvolatile memories that have potential to mitigate these issues and aim for high density, high switching speed, high endurance, low power and long data retention simultaneously. They may have great potential to be the next generation memory technology and those that are already matured in mass production are the ferroelectric random access memory (FeRAM), the magnetic random access memory (MRAM) and the phase change memory (PCM). Moreover, there are additional devices that are still in the early stages of investigation, e.g. the resistive random access memory (RRAM), or other candidates such as the carbon nano tube (CNT) memory, molecular memory and polymer memory.

In this Chapter, all these aforementioned memory technologies are described in more detail. In particular, the physics, charge storage mechanism, etc. of the floating gate device is reviewed more comprehensively along with its variations and improvements in the past years.

2.2 SRAM

A typical SRAM cell is composed of six transistors as shown in Figure 2.2. It can be accessed via the word line (WL), which controls the NMOS transistors N1 and N4. These are used to transfer the data between the cell and the bit lines (BL and $\overline{BL}$) for read and write operations. If the word line is not enabled, N1 and N4 are disconnected from the bit lines, and the data is stored via the cross-coupled inverters that are formed by N2, N3, P1 and P2 as long as the power supply V_{DD} is high. Therefore, the SRAM is a volatile memory.

Figure 2.2: 6T-SRAM cell.

The main benefit from using a SRAM is that it offers the highest switching speed among all existing memories. It is capable of writing and reading a bit in just a few nanoseconds. However, its drawback is its big size, which results in a low density such that it is not feasible for a high density memory array. It is also inefficient in cost and power due to the transistor overhead. SRAMs are therefore typically used in the cache memory of a CPU where high speed is most crucial. Its capacity ranges from kilobytes (KB) for microcontrollers to several megabytes (MB) for powerful microprocessors.

2.3 DRAM

Unlike the 6T-SRAM cell, the DRAM cell typically consists of only one transistor and one capacitor (1T-1C) as illustrated in Figure 2.3. The transistor N1 basically functions as a pass transistor with the word line (WL) connected to the gate terminal and the bit line (BL) connected to the drain terminal. In the main memory the bit line is linked to numerous DRAM cells in a column and is then the input to a sense amplifier (not shown in Figure 2.3). The word line is linked to a row of these cells and during the operation mode only one word line at a time is asserted to read or write a row. The capacitor C1 functions as a storage node. During the write operation, the word line is driven high in order to switch on the transistor. The capacitor is then charged until a desired value, which depends on the voltage on the bit line. In the read mode, the word line is first turned off and the bit line is precharged to a voltage such that it is preferably exactly in the middle between a low and a high logic level. Next, the word line is enabled and the capacitor shares its charge with the bit line. This results in a slight change in the voltage on the bit line, which the sense amplifier is able to detect and to amplify to a fully low or fully high logic level.

Figure 2.3: 1T-1C DRAM cell.

The drawback of the DRAM is that the capacitor for each cell is leaky, and the cell needs to be refreshed periodically to avoid loss of the stored data. Therefore, it is a dynamic and volatile memory device. The JEDEC standard requires all rows in the DRAM to be refreshed at least every 64 ms. This periodic refresh results in significant power consumption. The write speed is also slower than in SRAMs and is typically on the order of tens of nanoseconds. In addition, scaling of the DRAM is difficult mainly because of the large capacitor with a capacitance of typically 20 – 25 fF. Reducing the size of the capacitor results in less stored charge and therefore a smaller change in voltage on the bit line that the sense amplifier is to detect. This could lead to bit errors. Furthermore, the transistor cannot scale with CMOS because its leakage current needs to be very low in order to achieve a minimal refresh period of 64 ms. Regardless of this, the DRAM is widely in use in the main memory of computers. The reasons for this are that they are cost efficient and they are manufactured with a very high density, which enables overall high speed data access. DRAM chips with a size in the gigabyte (GB) scale are already common for home computers and laptops.

2.4 Nonvolatile Memory

The nonvolatile memory has drawn much attention over the past years due to the booming market of portable electronic equipments where memory devices with a retention time of ~10 years are desired. The Flash memory is the dominant technology and offers a very high density on-chip and very low idle power consumption. However, the switching speed between two states is still low compared to the volatile memory and is therefore

typically used as the secondary memory storage or long term persistent storage where a slow speed may be tolerated. Nowadays, Flash memory is integrated e.g. in music players, smart cellular phones, digital cameras, digital assistants, hybrid and solid-state drives and other portable electronic consumer equipments. There are other nonvolatile memories technologies such as the FeRAM, MRAM, PCM or RRAM. However, the Flash memory is the only one that is manufactured in a high volume to date.

Flash memory is the pre-dominant nonvolatile memory technology with the fastest growing rate due in part to its compatibility with CMOS process [2.13] and its extensive development. It was invented in 1984 by Fujio Masuoka at Toshiba and Intel quickly recognized its great potential and was the first company that shipped a Flash memory chip in 1988 with an array size of 256 KB; today the capacity reaches 32 GB and higher. The Flash memory is based on the floating gate device that has a structure where a gate is embedded in the oxide of a transistor in order to store charge, and thus data. The device employs typically a continuous floating gate but in the past years a discontinuous floating gate device has been introduced where isolated nanocrystals are inserted in the oxide. This so-called nanocrystal floating gate (NCFG) device is in many aspects superior to its counterpart. Most novel NCFG devices are further modified using a high-k dielectric or a dual-layer structure to improve charge transfer and/or retention characteristics.

2.4.1 Continuous Floating Gate Device

The idea of a continuous floating gate embedded in the oxide of a transistor was first reported by Khang and Sze in 1967. Figure 2.4 shows a typical schematic cross-section of

this device. The upper electrode is the control gate, and the lower electrode is the conductive floating gate. The lower oxide that isolates the floating gate from the p-type silicon (Si) substrate is called the tunnel oxide. The upper oxide, which is inserted between the floating gate and the control gate, is called the control gate oxide, or sometimes the blocking oxide. The floating gate acts as the storage node in this nonvolatile memory device and is configured by applying the appropriate bias on the control gate, drain and source for an appropriate duration.

Figure 2.4: Schematic cross-section of a continuous floating gate device.

By applying a high voltage on the control gate, electrons tunnel from the channel through the tunnel oxide to the floating gate, thus programming the device. The resulting accumulated negative charge on the floating gate acts as a shield between the control gate and the channel, and results in a positive threshold voltage shift ΔV_T of the device as shown in Figure 2.5. An applied bias on the control gate in the operation mode, which is less than the new threshold voltage of the device, keeps the programmed device in the cut-off region, which leads to a minimal subthreshold leakage current. In order to remove the charge on the

floating gate, a high negative voltage needs to be applied on the control gate such that the electrons tunnel back to the substrate. Then, the threshold voltage of the device shifts back towards the initial value V_{T0} , thus erasing the device. Furthermore, ΔV_T may also become negative as shown in Figure 2.5 if the applied erasing voltage is too high and/or too long, enabling a positively charged floating gate. The amount of charge that is transferred between the floating gate and the substrate depends on several factors such as the control gate voltage, programming/erasing time, as well as tunnel and control gate oxide thicknesses. The materials that are chosen for the control gate and floating gate, as well as the oxides, also contribute significantly to the charge transfer.

Figure 2.5: Threshold voltage of an uncharged, programmed and erased device.

There are three different mechanisms for programming a continuous floating gate device which are illustrated in the band diagram in Figure 2.6(a), but the most common ones are Fowler-Nordheim tunneling (FNT) and channel hot electron injection (CHEI). The third mechanism is direct tunneling (DT) but it plays an insignificant role in a continuous floating

gate device due to its relatively thick tunnel oxide. However, as will be seen in the following Section 2.4.2, it contributes significantly to the charge transfer in a NCFG device, which uses an ultra-thin tunnel oxide. FNT is a uniform wave-mechanical tunneling process through a triangular oxide barrier and during the program mode it is applied for positive control gate voltages. It increases exponentially with an increase of the voltage or a decrease of the tunnel oxide thickness. In addition, it is strongly dependent on the barrier height in the band diagram. In the case of programming the device through CHEI, a relatively high positive voltage is also applied on the drain, which results in a large lateral electric field, especially at the drain side of the channel. This field is sufficiently high to heat up the electrons and gives rise to the impact ionization at the drain terminal. Due to the high transversal electric field between the control gate and the channel, these “hot” electrons are injected into the floating gate. Thus, this charge transfer mechanism is also known as drain side tunneling.

Figure 2.6: Mechanisms of the charge transfer in a floating gate device during (a) the program mode and (b) the erase mode. Note that the band bending in the polysilicon floating gate and control gate is not shown.

In the erase mode FNT becomes the only significant mechanism for discharging the floating gate. Then, those electrons stored in the floating gate tunnel back to the substrate through the triangular barrier seen at the interface between the floating gate and the tunnel oxide. As indicated in Figure 2.6(b), CHEI does not apply for erasing the device due to the absence of a lateral electric field across the floating gate. Similar to the program mode, DT is only marginal compared to FNT due to the thick oxide.

It should also be noted, that programming or erasing the floating gate is not limited to charge transfer of electrons. FNT and DT of holes, as well as channel hot hole injection (CHHI) may also occur and are indicated in Figure 2.6(a) and (b) as well. However, their contribution to the overall charge transfer is small because holes are heavier than electrons and the barrier height in the valence band is higher than in the conduction band. Furthermore, charge transfer may also take place through the control gate oxide as indicated in Figure 2.6(a) and (b). This is not desired such that the control gate oxide should be designed sufficiently thick to mitigate this problem.

One of the main hazards of the continuous floating gate device is the formation of defects in the tunnel oxide. Defects occur after many program and erase cycles (typically 10^5 - 10^6 cycles) where electrons get trapped in the oxide, thus establishing a conductive path between the floating gate and the channel. This is also known as stress induced leakage current (SILC). It is fatal for the device because the entire charge on the floating gate leaks to the channel, resulting in a bit error and failure of the device. Therefore, a high quality and sufficiently thick tunnel oxide are essential for a reliable device with good endurance. The thickness of the tunnel oxide in continuous floating gate devices is typically 6-7 nm [2.14].

This in turn implies that the program/erase voltage is required to be high to enable a high electric field across the tunnel oxide. Alternatively, the program/erase time needs to be long to overcome the reduced tunneling rate. It is clear that scaling of the continuous floating gate device is limited due to SILC and this limit will be reached at the 45 nm technology node for the NOR Flash architecture and possibly 32 nm for the NAND architecture [2.15, 2.16].

2.4.2 Nanocrystal Floating Gate Device

Sandip Tiwari presented first the idea of a discontinuous floating gate with Si nanocrystals embedded in the oxide in 1995. NCFG devices have not only drawn much attention in the nonvolatile memory market since then but have recently been proposed for other applications such as a tunable gain operational amplifier, programmable reference circuit and DRAM [2.7, 2.8]. A schematic cross-section of this device is drawn in Figure 2.7(a). The nanocrystals have a diameter of just a few nanometers and an areal density of typically 10^{11} - 10^{13} cm⁻². The tunnel oxide is thinner than in the continuous floating gate device, resulting in a more efficient charge transfer and better scalability.

Figure 2.7: (a) Schematic cross-section of a NCFG device and (b) TEM image of Si nanocrystals on top of the SiO₂ [2.17].

Figure 2.7(b) shows a transmission electron microscopy (TEM) image of a fabricated NCFG where the Si nanocrystals have a diameter of ~ 4 nm in average and a density of $\sim 6 \times 10^{12} \text{ cm}^{-2}$ [2.17]. The program and erase characteristics of the NCFG are very similar to the continuous floating gate device. By applying a positive voltage on the control gate the electrons tunnel from the channel to the nanocrystals and are stored there. The stored charge leads to a positive ΔV_T of the device. In order to discharge the floating gate a negative voltage needs to be applied on the control gate such that the electrons tunnel back to the channel. The mechanisms of the charge transfer of a NCFG device are now typically uniform tunneling of electrons due to FNT and DT. DT is a quantum mechanical tunneling process through a rectangular oxide barrier and increases exponentially with a decrease of the oxide. It applies only for ultra-thin oxides with a thickness of less than ~ 4 nm [2.18]. It is also dependent on the barrier height and the electric field across the oxide. CHEI may also apply for a NCFG device but is not typical as it is highly localized at the drain terminal and therefore not a uniform tunneling. As such, only the outer nanocrystals that are close to the drain would be charged through CHEI but would not share the charge with the other nanocrystals if their inter-particle spacing is sufficiently thick. As in the case of a continuous floating gate device, hole tunneling may also occur in a NCFG device. But again, the tunneling rate of holes is very low compared to electrons due to the heavy mass of holes and the higher barrier in the valence band.

One of the great benefits of the NCFG device compared to the continuous floating gate device of equivalent dimensions is the improved susceptibility to SILC by storing the charge on isolated colloids of small dimensions. Figure 2.8(a) demonstrates how defects in

the tunnel oxide lead to charge leakage of the entire floating gate to the substrate for a continuous floating gate device. In this case, the device changes from a programmed state to an erased state, which results in a bit error. Since these defects are very difficult to remove the lifetime of the device is practically over. In the case of a NCFG device, defects in the tunnel oxide would result in charge leakage of only the nearby nanocrystals as illustrated in Figure 2.8(b). The other nanocrystals are not affected by this and maintain their charge. Therefore, even if part of the total charge on the discontinuous floating gate is lost due to SILC, the remaining charge may be sufficient to retain the device in a programmed state.

Figure 2.8: SILC susceptibility of (a) a continuous floating gate device and (b) a NCFG device.

This improves the reliability of the NCFG device compared to its counterpart and allows thinning of the tunnel oxide, which results in a more efficient charge transfer while maintaining a retention time of ~10 years. A higher tunneling rate is gained due to the increased electric field across the tunnel oxide, which enables a higher programming/erasing speed. Alternatively, the programming/erasing voltage can be reduced to achieve a similar charging/discharging behavior of the discontinuous floating gate as for the continuous

floating gate with an equivalent bias on the control gate. The operational voltage is also reduced, which results in a reduced power consumption. The endurance is improved and typically over 10^6 write and erase cycles can be achieved. In addition, the reduction of drain to floating gate coupling, and the increased gate capacity due to the thinner tunnel oxide, reduces the device susceptibility to drain induced barrier lowering (DIBL). This improves the channel length scalability of the NCFG device.

2.4.2.1 Silicon vs. Metal Nanocrystals

The NCFG has been improved by using nanocrystals in the floating gate composed of metal instead of Si. They are superior to their semiconductor counterparts due to higher density of states, stronger coupling with the channel, better scalability and the design freedom of the work function of the metal to optimize device characteristics [2.19-2.21]. High work function metals such as gold (Au), palladium (Pd) or platinum (Pt) with a work function of 5.10 eV, 5.12 eV and 5.65 eV, respectively [2.22], create a deep potential well in the floating gate for efficient data retention. It does not affect the program characteristics of the device, which is a key advantage because usually the program characteristics and the retention time trade off. In addition, it reduces the probability of electrons to tunnel further from the floating gate to the control gate significantly. The drawback is that erasing the device becomes more difficult due to the higher barrier between the floating gate and the tunnel oxide. These features can be observed in Figure 2.9(a), (b) and (c) which illustrates the band diagram of the device with low and high work function metal nanocrystals during the program, retention and erase mode, respectively.

Figure 2.9: Band diagrams of a NCFG device with low and high work function metal in the (a) program, (b) retention, and (c) erase mode.

In the program mode, the use of a low or high work function metal for the nanocrystals has theoretically no influence on the tunneling rate of electrons from the channel as long as the work function is higher than the electron affinity of the channel (typically in the range of 4.0 – 4.2 eV). The conduction band in both cases is lower than in the channel such that the electrons tunnel through the same barrier. In the retention state, the electrons stored in high work function metal nanocrystals face a higher barrier to tunnel back to the channel due to the deep potential well. Therefore, the charge leakage is less than in the case of electrons that are stored in low work function metal nanocrystals. Thus, nanocrystals with a high work function metal enable further thinning of the tunnel oxide while maintaining a retention time of ~10 years, which in turns improves the scalability and charge transfer of the device. The decrease in the tunneling rate from the floating gate to the channel in the erase mode due to the higher barrier is then by part compensated by the choice of a thinner tunnel oxide.

This method of work function engineering can also be used for the control gate. The use of a low or high work function metal in the control gate is also illustrated in Figure 2.9(a), (b) and (c). The choice of a high work function metal is preferred for better erase characteristics while retention and program characteristics are negligibly affected [2.23]. It significantly reduces the probability of electron tunneling from the control gate to the floating gate due to the higher barrier at the interface between the control gate and the oxide. This is important for the erase mode where high negative voltages are applied on the control gate. It also enables thinning of the control gate oxide as long as charge transfer between the floating gate and the control gate during both the program and erase mode is still negligible.

2.4.2.2 Coulomb Blockade

The coulomb blockade effect is an important phenomenon that has been demonstrated for metal and silicon nanocrystals. As the nanocrystals in the floating gate usually vary in size the tunneling electrons select the larger sized colloids due to their lower sub-band energy and lower coulombic barrier [2.24]. Each electron that tunnels from the substrate to the nanocrystal causes a raise of the electrostatic potential in this nanocrystal by $e^2/2C$ [2.25], where e is the electron charge and C the self-capacitance of the nanocrystal. If this so-called coulomb charging energy is greater than the thermal energy, thermal noise cannot charge up the colloid any further. Therefore, the coulomb blockade effect is strongly dependent on the temperature and also the size of the nanocrystal as shown in Figure 2.10 [2.26]. A smaller size results in a smaller self-capacitance and thus a larger coulomb charging energy.

Figure 2.10: Theoretical coulomb charging energy of Pd nanocrystals with varying diameter comparing to the thermal energy (3-4 $k_B T$) [2.26].

Once the larger sized nanocrystals are sufficiently charged such that their coulombic barrier is higher than in the smaller sized nanocrystals, the electrons begin to tunnel to these smaller colloids. Continuing this procedure, the entire floating gate will eventually exhibit a coulomb repulsion such that the tunneling of further electrons from the channel is blocked. These electrons can only be transferred by either raising the temperature to increase the thermal energy, or by raising the control gate voltage to increase the electric field across the tunnel oxide as illustrated in Figure 2.11 [2.24]. It can be observed, that the coulomb blockade effect is present for lower temperatures and leads to a “staircase” charging characteristics of the nanocrystals, in particular for 40 K. The tunnel and control gate oxide thicknesses are ~ 1.9 nm and ~ 9 nm, respectively. The nanocrystals have a size ranging from 2 nm to 10 nm in diameter.

Figure 2.11: Threshold voltage of a NCFG device for 40 K, 77 K, and 300 K as a function of the control gate voltage [2.24].

2.4.2.3 Advanced Nanocrystal Floating Gate Devices

Most recently, a wide range of variations of the NCFG device have been investigated and their schematic cross-sections are illustrated in Figure 2.12 [2.27-2.33]. The devices II-V imply the use of HfO_2 as an insulator (HfAlO , HfSiO , ZrO_2 or Al_2O_3 are also common) while the device VI consists of a dual-layer NCFG structure.

Figure 2.12: Conventional NCFG device (I) and its modified versions (II-VI).

The use of a high-k material in NCFG devices such as HfO_2 offers many advantages but also a few disadvantages towards SiO_2 . HfO_2 is already widely integrated in modern MOSFETs, especially in the CMOS process with 45 nm technology and below. Due to the

high dielectric constant the insulator can be physically thicker while keeping the effective oxide thickness (EOT) similar to its counterpart. Therefore, the gate leakage current is reduced by several orders of magnitude, which became a major problem in the last generations of the MOSFET with SiO₂ as the insulator. A comparison between these two insulators is shown in Table 2.1.

Table 2.1: Typical properties of SiO₂ and HfO₂.

Parameter	SiO ₂	HfO ₂
Dielectric Constant	3.9	25
Band Gap	9 eV	5.8 eV
Electron Affinity	0.9 eV	2.9 eV
Effective Electron Mass	0.4 m ₀	0.17 m ₀
Conduction Band Offset with Si	3.1 eV	1.1 eV

m₀ is the free electron mass and given by $\sim 9.109 \times 10^{-31}$ kg. In NCFG devices the higher dielectric constant of HfO₂ enables enhanced coupling of the surrounding electrodes [2.28, 2.34]. In the case of device II in Figure 2.12, a lower programming voltage can be achieved due to the smaller conduction band offset (i.e. barrier height) at the interface between the Si substrate and the HfO₂ tunnel oxide and the charge leakage can be reduced due to the physically larger thickness of the insulator [2.27-2.30, 2.35]. Furthermore, the effective electron mass is relatively low, which results in enhanced charge transfer. However, several problems occur when fabrication aspects are considered: HfO₂ on top of Si

adhesion causes various problems such as mobility degradation, poor thermal stability or poor interface quality whereas the interface between SiO_2 and Si is very good [2.36]. In addition, an interfacial SiO_2 growth between HfO_2 and Si is difficult to avoid [2.37], which results in less efficient charge transfer. These fabrication issues are prevented in the devices III, IV and V. Device III uses HfO_2 as the control gate oxide to enable a stronger coupling between the floating gate and the channel. Device IV has an additional HfO_2 layer in the tunnel oxide between the floating gate and the SiO_2 to improve both the retention time and charge transfer compared to a SiO_2 -only tunnel oxide. In device V the tunnel oxide is composed of a symmetric $\text{SiO}_2/\text{HfO}_2/\text{SiO}_2$ stack to obtain similar benefits as device IV with the additional feature of equal (or almost equal) program and erase characteristics. In these cases, the charge transfer is reduced due to the SiO_2 layer in the tunnel oxide compared to the HfO_2 -only device II but still superior to the conventional SiO_2 -only NCFG device I.

Device VI uses SiO_2 for both the tunnel and control gate oxide but it has a dual-layer NCFG structure with the ability to improve both retention time and memory window [2.32, 2.35]. The bottom layer is comprised of very small nanocrystals that are energetically unfavorable for charge storage. Most likely, the electrons will step through them and tunnel further to the top layer during the program mode. Then, the small nanocrystals in the bottom layer exploit coulomb blockade such that electron back-tunneling from the top layer is mostly suppressed, which results in an improved retention time. Furthermore, the top layer provides increased storage capacity in order to enlarge the memory window.

Table 2.2 summarizes the device characteristics of the conventional NCFG device and compares them to its modified versions.

Table 2.2: Comparison of the conventional single-layer NCFG device and its variations.

Figure 2.12 [Ref]		Tunnel Oxide	Control Gate Oxide	NC Diameter, Density	Prog. Time, Voltage	Voltage Shift	Charge Loss, Retention
[2.38]	3.5nm (SiO ₂)	8nm (SiO ₂)	3.2nm, 2e11cm ⁻² (Si)	1μs, 10V 10ms, 10V	ΔV _T =0.2V ΔV _T =1V	
	[2.39]	3nm (SiO ₂)	36nm (SiO ₂)	6nm, 4e11cm ⁻² (Au)	5s, 3V	ΔV _T =0.6V	
	[2.40]	2.8nm (SiO ₂)	6nm (SiO ₂)	6nm, 5e11cm ⁻² (Si)	1μs, 8V 10ms, 8V	ΔV _T =0.2V ΔV _T =0.6V
[2.27]	5.1nm (HfO ₂)	15nm (HfO ₂)	5e11cm ⁻² (Ni)	1s, 3V	ΔV _T =0.8V	20%, 5e4s
	[2.28]	5nm (HfAlO)	8nm (HfAlO)	7-10nm, 2e11cm ⁻² (Ge)	100ms, 7V	ΔV _T =1.5V	40%, 1e3s
	[2.29]	4.5nm (HfO ₂)	13nm (HfO ₂)	7.4nm, 1e11cm ⁻² (Si)	5s, 4V 100μs, 5V	ΔV _T =0.4V ΔV _T =0.2V	25%, 5e4s
[2.33]	4nm (SiO ₂)	24nm (Al ₂ O ₃)	2-3nm, 2.7e12cm ⁻² (Au)	16ms, 15V	ΔV _{FB} =2V	0.009V/s
[2.30]	1.5nm (SiO ₂) 2nm EOT (HfO ₂) 1.5nm (SiO ₂)	7nm (SiO ₂)		100μs, 6V	<i>Programmed</i>	10 years
	[2.31]	1.6nm (SiO ₂) 1.8nm EOT (HfO ₂) 1.6nm (SiO ₂)	7nm (SiO ₂)		100μs, 10V	<i>Programmed</i>	10 years
[2.31]	1.5nm (SiO ₂) 4nm EOT (ZrO ₂)	7nm (SiO ₂)		100μs, 7V	<i>Programmed</i>	10 years
[2.32]	2.4nm (SiO ₂)	2-3nm (SiO ₂) 20-30nm (SiO ₂)	5nm (Au) 1nm (C ₆₀)	-2...5V	ΔV _{FB} =3V	1.1e3s

2.4.3 Emerging Nonvolatile Memory Technologies

Next to the floating gate device, there are other nonvolatile memory devices that are under extensive research. Their goal is to obtain long retention time, high endurance, high switching speed, and low power consumption while keeping its cell size small.

2.4.3.1 FeRAM

The FeRAM is based on a metal-ferroelectric-insulator-semiconductor (MFIS) structure and stores the data in a ferroelectric film as illustrated in Figure 2.13 [2.3, 2.41]. The ferroelectric material is typically a perovskite material and can be polarized by an electric field where the dipoles align themselves in the same direction of the field. A change in direction results in a displacement of the dipoles in the crystal structure of the ferroelectric material. The distribution of the charge is then also shifted. Therefore, the material shows hysteresis characteristics, which enables two stable states of the device. However, the FeRAM has several challenges: the read operation is destructive and is expected to have an endurance similar to the write operation. In addition, the cell size is relatively large, the cost per bit is high, and the compatibility with CMOS is poor [2.42].

Figure 2.13: Schematic cross-section of a FeRAM cell with a stacked capacitor [2.41].

2.4.3.2 MRAM, SPRAM

The MRAM technology stores the data by magnetic storage elements and uses a magnetic tunnel junction (MTJ). There are two kinds of MRAM: i) the conventional MRAM where the magnetic field is generated due to the current flow in the word line and the bit line as shown in Figure 2.14(a) [2.43], and ii) the advanced spin transfer torque RAM (SPRAM) where the write operation is performed by the current flow through the MTJ itself and a write word line is not required as shown in Figure 2.14(b) [2.43]. The MTJ is composed of two ferromagnetic plates which are separated by a thin oxide layer, composed of a magneto-resistive material [2.3, 2.43]. One plate is a permanent magnet and is set to a fixed polarity; while the other plate is a free layer whose polarity can be changed. The electric resistance of the MTJ changes with the orientation of the magnetic field. Usually, if the polarities of the two plates are the same, the resistance is very low, resulting in a logic '1'. If the polarities are opposite, the resistance is increased by orders of magnitude, which represents a logic '0'. The data is therefore stored in a magnetic state of bit and is then read by sensing the resistance.

Figure 2.14: (a) Conventional MRAM cell [2.43] and (b) SPRAM cell [2.43].

The MRAM offers high speed, excellent endurance, and low voltages. One issue is the high cost per bit which is related to its incompatibility with CMOS process. In addition, it shows thermal instability in its back-end process and requires a high write current. The SPRAM can operate with a reduced write current which enables better scalability [2.43]. Furthermore, a SPRAM with vertical MOSFET is proposed in [2.43] where the bit line, word line, transistor, and MTJ are stacked vertically. With this structure, the cell size may be as small as $4F^2$. Other key benefits can be summarized as high drivability, low power consumption, and good integration of logic-in-memory architecture.

2.4.3.3 PCM

The PCM, also known as PRAM or ovonic unified memory (OUM), is another potential candidate for the driving force in the nonvolatile memory market in the foreseeable future. A three-dimensional (3D) device structure is illustrated in Figure 2.15 [2.44].

Figure 2.15: Typical structure of the PCM [2.44].

It offers high endurance, high performance and high density along with promising scaling characteristics [2.3, 2.44, 2.45]. The PCM uses a thin film typically composed of a

chalcogenide material that is capable of reversible phase transitions between amorphous and crystalline by applying heat. It is placed between the top electrode (bit line) and a heating element that extends from the bottom electrode (word line). The memory cell is programmed through a relatively high current that heats up the chalcogenide material and leads to a thermally induced phase change. Reading occurs by sensing the change in resistance after a phase change with a low bias. The material changes from a low resistance to a much higher resistance depending on the state of the phase. One drawback is that the PCM requires a high programming current and it is sensitive to temperature variation, which may result in unintended phase-change. Therefore, it operates in a smaller temperature window compared to for example the Flash memory.

2.4.3.4 Comparison to Flash Memory

A comparison between the Flash memory and the aforementioned alternative nonvolatile memories is provided in Table 2.3. The SPRAM is not included in this comparison because it is not matured in mass production yet. Due to the improvements in endurance as well as program (or write) and erase speed in the new technologies it can be expected that the Flash memory may lose its dominant market share in the foreseeable future if the Flash technology will not be further investigated and new improvements and solutions found. Today, Flash memory is still in wide use due to their good compatibility with CMOS process, low cost, low idle power consumption and high density.

Table 2.3: Comparison between the Flash memory and emerging nonvolatile memory alternatives.

Attribute	Flash NOR [2.3, 2.46]	Flash NAND [2.3]	FeRAM [2.3, 2.41, 2.43, 2.46, 2.47]	MRAM [2.3, 2.46, 2.48]	PCM [2.3, 2.45, 2.46, 2.49]
Cell Size	$10F^2$	$4-5F^2$	$15-100F^2$	$10-30F^2$	$8-20F^2$
Endurance	10^{5-6} cycles	10^{5-6} cycles	10^{8-12} cycles	10^{9-16} cycles	10^{9-12} cycles
Write Time	1 μ s	200 μ s / page	30-200 ns	10-30 ns	10-100 ns
Erase Time	1 s / sector	2 ms / block	30-200 ns	30 ns	100-120 ns
Read Time	20-60 ns	60 ns / serial	20-80 ns	10-30 ns	20-100 ns
Scalability	Fair	Fair	Poor	Poor	Good
Multi-Bit	Possible	Possible	No	No	Difficult
Cost/Bit	Medium	Low	High	High	Medium
Maturity	High	High	Medium	Medium	Medium
Process	Full custom or +10 masks	Full custom or +10 masks	+ 2-3 masks back-end process	+4-6 masks back-end process	+ 2-3 masks back-end process

2.4.3.5 Future Prospects

Another important nonvolatile memory is the RRAM. It has drawn increasing attention due to an easier fabrication process and the ability of 3D stacking for ultra-high density [2.50]. It relies on a filament in an ultra-thin dielectric that is formed or broken by appropriately applied biases [2.50, 2.51]. If it is formed, a conductive path exists between the two plates that surround the dielectric, leading to a low resistance. If the filament is broken

the cell acts as a high resistance. The switching mode between low and high resistance can be either unipolar or bipolar. There has been little demonstrated to date. However, a bipolar switching speed of 20 ns to set, and 60 ns to reset, with a data retention time of $>10^5$ s is reported in [2.51]. One of the greatest challenges of the RRAM is to control the formation and rupture of the filament. It is still in the beginning of development and it is a challenge to improve its reliability and retention. In addition, it suffers from a high on-resistance for CMOS integration.

Other nonvolatile memories that have not reached the maturity level yet to be integrated as the next generation nonvolatile memory, but should be mentioned, are the CNT memory, molecular memory, polymer memory and probe storage. The concepts for these memories are relatively new. A roadmap of the technology advancements from December 2007 of the emerging nonvolatile memories [2.52] is illustrated in Figure 2.16.

Figure 2.16: Status of the widely known emerging nonvolatile memories [2.52].

2.5 Summary

The memory is a growing field in the integrated circuit design and high efforts are being made to keep this trend by improving existing technologies and finding new solutions for the future. This Chapter provides an overview of the current volatile and nonvolatile memory technologies. Those that are well (or fairly well) established are summarized in Table 2.4.

Table 2.4: Description and key attributes of the most common memory technologies.

Technology	Description	Advantages	Disadvantages
SRAM	Charge is stored via two cross-coupled inverters.	Very high speed	Volatile; large area; low density; cost and power inefficient
DRAM	Charge is stored on a capacitor that shares the charge with the bit line.	High speed; high density; cost efficient	Volatile; constant refresh required; high power consumption
Continuous FG Device	Electrons from the channel charge the continuous floating gate mainly due to FNT and CHEL.	Nonvolatile; very high density; cost efficient; low idle power consumption; multi-bit capability	Low endurance; slow write/erase; high write/erase voltage

Table 2.4 (Continued).

NCFG Device	Electrons from the channel charge the discontinuous floating gate mainly due to FNT and DT.	Improved reliability, scalability and charge transfer compared to the continuous FG device	
FeRAM	Charge is stored in a ferroelectric film that is polarized by an electric field.	Nonvolatile; high endurance	Poor scalability; high cost/bit; read destructive
MRAM, SDRAM	The electric resistance in a magnetic tunnel junction changes with the orientation of a magnetic field.	Nonvolatile; high speed; very high endurance	Poor scalability; high cost/bit; high write current
PCM	The electric resistance of a chalcogenide film changes by applying heat due to reversible phase transitions between amorphous and crystalline.	Nonvolatile; high speed; very high density; high endurance; good scalability; multi-bit capability	High write current; slow erase speed; temperature sensitive

Table 2.4 (Continued).

RRAM	A filament in an ultra-thin dielectric is formed or broken by appropriate biases on the upper and lower electrode.	Nonvolatile; high speed; ultra-high density (3D stacking); good scalability	High on-resistance; good reliability and retention is a challenge; difficult control of the formation/rupture of the filament
------	--	---	---

References

- [2.1] C.-H. Wu, "A Time-Predictable System Initialization Design for Huge-Capacity Flash-Memory Storage Systems," *In: Proceedings of the 6th IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, pp. 13-18, 2008.
- [2.2] A. Padovani, "Modeling and Reliability of Innovative Flash Memories," *Università degli Studi di Ferrara, Ph.D. Dissertation*, 2008.
- [2.3] R. Bez and A. Pirovano, "Non-Volatile Memory Technologies: Emerging Concepts and New Materials," *Materials Science in Semiconductor Processing*, vol. 7, pp. 349-355, 2004.
- [2.4] A. Rani, "Simulation and Modeling of SONOS Non-Volatile Memory," *George Mason University, Master Thesis*, 2010.
- [2.5] T. Kgil and T. Mudge, "FlashCache: a NAND Flash Memory File Cache for Low Power Web Servers," *International Conference on Compilers, Architectures, and Synthesis for Embedded Systems*, pp. 103-112, 2006.

- [2.6] V.G. Moshnyaga, H. Vo, G. Reinman, and M. Potkonjak, "Reducing Energy of DRAM/Flash Memory System by OS-Controlled Data Refresh," *IEEE International Symposium on Circuits and Systems*, pp. 2108-2111, 2007.
- [2.7] S. Wang, J. Dai, E.-S. Hasaneen, L. Wang, and F. Jain, "Programmable Threshold Voltage Using Quantum Dot Transistors for Low-Power Mobile Computing," *IEEE International Symposium on Circuits and Systems*, pp. 3350-3353, 2008.
- [2.8] S.A. Rao, R.F. Steimle, M. Sadd, C.T. Swift, B. Hradsky, S. Straub, T. Merchant, M. Stoker, S.G.H. Anderson, M. Rossow, J. Yater, B. Acred, K. Harber, E.J. Prinz, B.E. White, Jr. and R. Muralidhar, "Silicon Nanocrystal Based Memory Devices for NVM and DRAM Applications," *Solid-State Electronics*, vol. 48, pp. 1463-1473, 2004.
- [2.9] A. Basu, S. Brink, C. Schlottmann, S. Ramakrishnan, C. Petre, S. Koziol, F. Baskaya, C.M. Twigg, and P. Hasler, "A Floating-Gate-based Field-Programmable Analog Array," *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 1781-1794, 2010.
- [2.10] V. Srinivasan, G.J. Serrano, J. Gray, and P. Hasler, "A Precision CMOS Amplifier Using Floating-Gate Transistors for Offset Cancellation," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 280-291, 2007.
- [2.11] E. Ozalevli, H.-J. Lo, and P.E. Hasler, "Binary-Weighted Digital-to-Analog Converter Design Using Floating-Gate Voltage References," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 55, pp. 990-998, 2008.
- [2.12] B.P. Degnan, C.J. Duffy, and P.E. Hasler, "Crossbar Switch Matrix for Floating-Gate Programming Over Large Current Ranges," *In: Proc. of 2010 IEEE Int. Symp. on Circuits and Systems*, pp. 861-864, 2010.
- [2.13] Y. Zhu, "Engineering the Nanocrystal Floating Gates for Nonvolatile Flash Memory Devices," *University of California, Riverside, Ph.D. Dissertation*, 2007.
- [2.14] *International Technology Roadmap for Semiconductors (ITRS)*.
- [2.15] Y.S. Shin, M. Div, S.E. Co, and S.K. Gyeonggi-Do, "Non-Volatile Memory Technologies for Beyond 2010," *IEEE Symposium on VLSI Circuits - Digest of Technical Papers*, pp. 156-159, 2005.
- [2.16] G. Atwood, "Future Directions and Challenges for ETox Flash Memory Scaling," *IEEE Transactions on Device and Materials Reliability*, vol. 4, pp. 301-305, 2004.
- [2.17] M.L. Ostraat, J.W. De Blauwe, M.L. Green, L.D. Bell, M.L. Brongersma, J. Casperson, R.C. Flagan, and H.A. Atwater, "Synthesis and Characterization of Aerosol Silicon Nanocrystal Nonvolatile Floating-Gate Memory Devices," *Applied Physics Letter*, vol. 79, pp. 433-435, 2001.

- [2.18] K.F. Schuegraf, C.C. King, and C. Hu, "Ultra-Thin Silicon Dioxide Leakage Current and Scaling Limit," *IEEE Symposium on VLSI Technology - Digest of Technical Papers*, pp. 18-19, 1992.
- [2.19] Z. Liu, C. Lee, V. Narayanan, G. Pei, and E.C. Kan, "Metal Nanocrystal Memories - Part I: Device Design and Fabrication," *IEEE Transactions on Electron Devices*, vol. 49, pp. 1606-1613, 2002.
- [2.20] Z. Liu, C. Lee, V. Narayanan, G. Pei, and E.C. Kan, "Metal Nanocrystal Memories – Part II: Electrical Characteristics," *IEEE Transaction on Electron Devices*, vol. 49, pp. 1614-1622, 2002.
- [2.21] K.S. Seol, S.J. Choi, J.-Y. Choi, E.-J. Jang, B.-K. Kim, S.-J. Park, D.-G. Cha, I.-Y. Song, J.-B. Park, Y. Park, and S.-H. Choi, "Pd-Nanocrystal-Based Nonvolatile Memory Structures with Asymmetric SiO₂/HfO₂ Tunnel Barrier," *Applied Physics Letters*, vol. 89, 2006.
- [2.22] N. Miura, T. Nanjo, M. Suita, T. Oishi, Y. Abe, T. Ozeki, H. Ishikawa, T. Egawa, and T. Jimbo, "Thermal Annealing Effects on Ni/Au Based Schottky Contacts on n-GaN and AlGaIn/GaN with Insertion of High Work Function Metal," *Solid-State Electronics*, vol. 48, pp. 689-695, 2004.
- [2.23] S. Jeon, J.H. Han, J.H. Lee, S. Choi, H. Hwang, and C. Kim, "High Work-Function Metal Gate and High-k Dielectrics for Charge Trap Flash Memory Device Applications," *IEEE Transactions on Electron Devices*, vol. 52, pp. 2654-2659, 2005.
- [2.24] S. Tiwari, F. Rana, K. Chan, L. Shi, and H. Hanafi, "Single Charge and Confinement Effects in Nano-Crystal Memories," *Applied Physics Letters*, vol. 69, pp. 1232-1234, 1996.
- [2.25] K. Yano, T. Ishii, T. Hashimoto, T. Kobayashi, F. Murai, and K. Seki, "Room-Temperature Single-Electron Memory," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1628-1638, 1994.
- [2.26] W.S. Pitts, N. Di Spigna, and P. Franzon, "Nanocrystal Film Devices and Circuits," *SRC Review Meeting*, 2008.
- [2.27] J.J. Lee and D.-L. Kwong, "Metal Nanocrystal Memory with High-k Tunneling Barrier for Improved Data Retention," *IEEE Transactions on Electron Devices*, vol. 52, pp. 507-511, 2005.

- [2.28] J.H. Chen, Y.Q. Wang, W.J. Yoo, Y.-C. Yeo, G. Samudra, D.S. Chan, A.Y. Du, and D.-L. Kwong, "Nonvolatile Flash Memory Device Using Ge Nanocrystals Embedded in HfAlO High-k Tunneling and Control Oxides: Device Fabrication and Electrical Performance," *IEEE Transactions on Electron Devices*, vol. 51, pp. 1840-1848, 2004.
- [2.29] J.J. Lee, X. Wang, W. Bai, N. Lu, and D.-L. Kwong, "Theoretical and Experimental Investigation of Si Nanocrystal Memory Device with HfO₂ High-k Tunneling Dielectric," *IEEE Transactions on Electron Devices*, vol. 50, pp. 2067-2072, 2003.
- [2.30] J. Buckley, B. De Salvo, G. Ghibaudo, M. Gely, J.F. Damlencourt, F. Martin, G. Nicotra, and S. Deleonibus, "Investigation of SiO₂/HfO₂ Gate Stacks for Application to Non-Volatile Memory Devices," *Solid-State Electronics*, vol. 49, pp. 1833-1840, 2005.
- [2.31] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, and K. De Meyer, "VARIOT: a Novel Multilayer Tunnel Barrier Concept for Low-Voltage Nonvolatile Memory Devices," *IEEE Electron Device Letters*, vol. 24, pp. 99-101, 2003.
- [2.32] K. Afshari, "Nonvolatile Memory with Multi-Stack Nanocrystals as Floating Gates," *National Nanotechnology Infrastructure Network Electronics*, pp. 38-39, 2007.
- [2.33] C.-C. Wang, Y.-K. Chiou, C.-H. Chang, J.-Y. Tseng, L.-J. Wu, C.-Y. Chen, and T.-B. Wu, "Memory Characteristics of Au Nanocrystals Embedded in Metal-Oxide-Semiconductor Structure by Using Atomic-Layer-Deposited Al₂O₃ as Control Oxide," *Journal of Physics D: Applied Physics*, pp. 1673-1677, 2007.
- [2.34] S. Das, K. Das, R.K. Singha, A. Dhar, and S.K. Ray, "Improved Charge Injection Characteristics of Ge Nanocrystals Embedded in Hafnium Oxide for Floating Gate Devices," *Applied Physics Letters*, vol. 91, 2007.
- [2.35] P. Dimitrakis and P. Normand, "Semiconductor Nanocrystal Floating-Gate Memory Devices," *Materials Research Society Symposium Proceedings*, vol. 830, 2004.
- [2.36] D. Misra, H. Iwai, and H. Wong, "High-k Gate Dielectrics," *Electrochemical Society Interface*, 2005.
- [2.37] T.-H. Hou, C. Lee, V. Narayanan, U. Ganguly, and E.C. Kan, "Design Optimization of Metal Nanocrystal Memory – Part II: Gate-Stack Engineering," *IEEE Transactions on Electron Devices*, vol. 53, pp. 3103-3109, 2006.
- [2.38] A. Campera and G. Iannaccone, "Modelling and Simulation of Charging and Discharging Processes in Nanocrystal Flash Memories During Program and Erase Operations," *Solid-State Electronics*, vol. 49, pp. 1745-1753, 2005.

- [2.39] C. Lee, U. Ganguly, V. Narayanan, T.-H. Hou, and J. Kim, "Asymmetric Electric Field Enhancement in Nanocrystal Memories," *IEEE Electron Device Letters*, vol. 26, pp. 879-881, 2005.
- [2.40] C.M. Compagnoni, D. Ielmini, A.S. Spinelli, and A.L. Lacaita, "Modeling of Tunneling P/E for Nanocrystal Memories," *IEEE Transactions on Electron Devices*, vol. 52, pp. 569-576, 2005.
- [2.41] K. Takeuchi, "Ferroelectric-Gate FET for Flash Memory & SRAM (CMOS) from Design/System Perspective," *ITRS ERD/ERM Memory Workshop Presentation*, April 2010.
- [2.42] L. Baldi, "Technology Limitations FeRAM," *ITRS ERD/ERM Memory Workshop Presentation*, April 2010.
- [2.43] T. Endoh, "Emerging Research Memory Technologies Spin Transfer Torque MRAM (SPRAM) and its Application," *ITRS ERD/ERM Memory Workshop White Paper*, April 2010.
- [2.44] M.K. Qureshi, V. Srinivasan, and J.A. Rivers, "Scalable High Performance Main Memory System Using Phase-Change Memory Technology," *Proc. of the 36th Int. Symp. on Computer Architecture*, pp. 24-33, 2009.
- [2.45] H.-S.P. Wong, S. Raoux, S.B. Kim, J. Liang, J.P. Reifenberg, B. Rajendran, M. Ashegi, and K.E. Goodson, "Phase Change Memory," *ITRS ERD/ERM Memory Workshop White Paper*, April 2010.
- [2.46] H. Hidaka, "Embedded Flash Memory," in *Embedded Memories for Nano-Scale VLSIs*, K. Zhang, Ed., ed New York, NY: Springer US, 2009, pp. 177-240.
- [2.47] G.R. Fox, R. Bailey, W.B. Kraus, F. Chu, S. Sun, and T. Davenport, "The Current Status of FeRAM," in *Ferroelectric Random Access Memory: Fundamentals and Applications*, H. Ishiwara, M. Okuyama, and Y. Arimoto, Eds., Berlin, Germany: Springer-Verlag, pp. 139-148, 2004.
- [2.48] L. Tran, "STT-MRAM Challenges," *ITRS ERD/ERM Memory Workshop Presentation*, April 2010.
- [2.49] L. Tran, "Phase Change Memory Challenges," *ITRS ERD/ERM Memory Workshop Presentation*, April 2010.
- [2.50] Y. Wang, Q. Liu, S. Long, W. Wang, Q. Wang, M. Zhang, S. Zhang, Y. Li, Q. Zuo, J. Yang, and M. Liu, "Investigation of Resistive Switching in Cu-Doped HfO₂ Thin Film for Multilevel Non-Volatile Memory Applications," *Nanotechnology*, vol. 21, 2010.

- [2.51] N. Xu, L.F. Liu, X. Sun, C. Chen, Y. Wang, D.D. Han, X.Y. Liu, R.Q. Han, J.F. Kang, and B. Yu, "Bipolar Switching Behavior in TiN/ZnO/Pt Resistive Nonvolatile Memory with Fast Switching and Long Retention," *Semiconductor Science Technology*, vol. 23, 2008.
- [2.52] V. Ermolov, M. Heino, A. Kärkkäinen, R. Lehtiniemi, N. Nefedov, P. Pasanen, Z. Radivojevic, M. Rouvala, T. Ryhänen, E. Seppälä, and M.A. Uusitalo, "Significance of Nanotechnology for Future Wireless Devices and Communications," *IEEE 18th Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, 2007.

Chapter 3

A SPICE Compatible Physical Model of Nanocrystal Floating Gate Devices for Circuit Simulation

3.1 Motivation

The majority of the single-layer NCFG research has been done at the device level. Circuit level research is still in its early stages due to the lack of a fast computing physical model for circuit simulations. To facilitate the circuit designer using NCFG devices and enable them to explore the advantages of the features of these devices as described in Chapter 2, a compact circuit simulation model is required. Several NCFG device models have been proposed in the past. Many use a numerical approach either by incorporating Monte Carlo simulations [3.1-3.3] or by solving Schrödinger-Poisson equations [3.4-3.8], which need to be solved numerically and require long computation time. Therefore, these device models are not suitable for dynamic simulations of large and complex circuits because the simulation time would be unacceptable. Alternatively, a circuit simulation model is reported in [3.9] which is easy to integrate in the existing SPICE-based infrastructure. However, it is not a physical model and therefore it cannot be used for a circuit's power consumption computation and speed analysis.

The motivation in this Chapter is to develop a fast, accurate, dynamic and

configurable physical model of the NCFG device for circuit simulation. The main contribution is a Verilog-A module that captures the physical behaviors of programming and erasing the device. It uses only explicit formulas without any correction factors or curve fitting and can be used in any SPICE simulator supporting Verilog-A. It interacts dynamically with the rest of the circuit and includes charge leakage, which enables power consumption analysis. This Chapter is important to bridge the gap between device and circuit level research.

3.2 Introduction

This Chapter describes a comprehensive and accurate physical model of NCFG devices, which is suitable for both DC and transient circuit simulations. It is based on uniform direct tunneling and Fowler-Nordheim tunneling of electrons. The physical behavior of charging and discharging the device is implemented in a Verilog-A module to provide portability across various SPICE-based circuit simulators. It has the control gate voltage as the input state variable and computes the floating gate voltage of the device. This floating gate voltage is then connected to the gate terminal of a predictive NMOS transistor which is used for modeling conduction channel behavior in order to obtain the current-voltage (I-V) characteristics.

This model utilizes solely analytical and physical equations, resulting in fast computation appropriate for circuit simulation. It is parameterized in terms of process and geometry to provide configurability.

These parameters are listed below:

- Gate length (nm)
- Gate width (nm)
- Control gate oxide thickness (nm)
- Dielectric constant in the control gate oxide
- Tunnel oxide thickness (nm)
- Dielectric constant in the tunnel oxide
- Control gate work function (eV)
- Floating gate work function (eV)
- Nanocrystal diameter (nm)
- Nanocrystal density (cm^{-2})
- Bulk doping (cm^{-3})
- Ion implanted channel doping (cm^{-3})
- Ion implanted channel depth (nm)
- Free electron mass in silicon (kg)
- Effective electron mass in the tunnel oxide (kg)
- Effective electron mass in the control gate oxide (kg)
- Surface interface charge (C)
- Temperature (K)

This allows circuit designers to explore the trade-offs in design choices of the NCFG device such as speed, power consumption and retention time. In addition, the physical model

interacts dynamically with the rest of the circuit throughout the entire transient circuit simulation. Therefore, charge leakage is included and power consumption of the circuit can be analyzed. Also, DC circuit simulations are possible once the charge on the floating gate is known and inserted in the model's netlist. The charge on the floating gate can be extracted by running a previous transient simulation. As DC circuit simulations are time independent, sweeping the control gate voltage will not affect the charge on the floating gate.

The accuracy of the physical model is benchmarked to a reference model with varying tunnel oxide thicknesses from Silvaco's numerical device simulation framework ATLAS. Experimental results from a variety of fabricated nanocrystal devices are also compared in terms of threshold voltage shift, flat-band voltage shift and retention time. In addition, application of this model in shifting the switching threshold of an inverter is shown in a DC simulation in HSPICE. A run time analysis of the physical model is also provided to demonstrate its fast computation.

3.3 Reference Model

The NCFG device is created in the device simulator to obtain a benchmark for the physical model not only for the computation of the threshold voltage shift but also of the current densities, charge on the floating gate, surface potential, etc. Figure 3.1 shows a structure of a device with a gate length of 45 nm. For the control gate a high work function metal is preferred for better erase characteristics while program characteristics are negligibly affected [3.10]. As described in Chapter 2 it significantly reduces the probability of electron

tunneling from the control gate to the floating gate, which enables vertical scaling of the control gate oxide. Therefore, platinum (Pt) is chosen in the reference model which has a very high work function of 5.65 eV [3.11]. Metal is also used for the nanocrystals in the floating gate because they are superior to their semiconductor counterparts. Furthermore, controlled palladium (Pd) nanocrystal deposition has been demonstrated with very good charge storage [3.12] and is proposed for use in the floating gate. Pd's relatively high work function of 5.12 eV [3.11] creates a deep potential well for efficient data retention [3.13]. The control gate and tunnel oxide of the NCFG device are both composed of SiO_2 , which has a dielectric constant of $\epsilon_r = 3.9$. The p-type substrate is uniformly doped with boron (B) at a concentration of $5 \times 10^{16} \text{ cm}^{-3}$, and B ions are diffused into the channel with a Gaussian distribution at a concentration of $4.5 \times 10^{17} \text{ cm}^{-3}$. The n-type source and drain are highly doped with arsenic (As).

Figure 3.1: A single-layer NCFG device with 45 nm gate length designed in the device simulator.

The device is projected 1 μm deep into the third dimension, which is the default in the device simulator and corresponds to the control gate width. The geometry parameters for the device are listed in Table 3.1. Note also that a very fine mesh is chosen throughout the entire device for accurate simulation results. Especially at the material interfaces, where electrons begin to tunnel, the mesh is as thin as a 100th of a nanometer.

Table 3.1: Device geometry parameters used for the reference model in the device simulator.

Parameter	Reference Model
Tunnel Oxide Thickness	2 nm – 3.2 nm
Control Gate Oxide Thickness	4 nm
Nanocrystal Length	3 nm
Nanocrystal Height	2 nm
Nanocrystal Spacing	3 nm
Control Gate Length	45 nm
Control Gate Width	1 μm
Control Gate Thickness	50 nm

The tunnel oxide thickness d_{tox} chosen in this reference model is in the range of 2 nm to 3.2 nm. For thin oxides, the programming/erasing voltage can be reduced to achieve a similar threshold voltage shift ΔV_T as in the case for thicker oxides. Alternatively, a shorter programming/erasing time can be chosen. The drawback is increased charge leakage, which results in a reduced retention time. The control gate oxide thickness d_{cox} is 4 nm. Reducing this thickness would result in significant charge leakage from the floating gate to the control

gate through the control gate oxide, which is not desired [3.14]. The Pd nanocrystals have a dimension of 3 nm in length, 2 nm in height, and the inter-particle spacing is 3 nm. For simplicity, it is assumed that the nanocrystals are equally separated throughout the entire floating gate.

3.4 Physical Model

The concept that is used for the physical model is shown in Figure 3.2. The block diagram consists of a Verilog-A module attached to a predictive NMOS transistor [3.15]. The physical model is approximated by assuming that the floating gate of the device can be treated as the gate of a traditional MOS device with very similar effects from the point of view of typical MOS characteristics, as in continuous floating gate devices [3.16]. Therefore, the floating gate and threshold voltage of the device replace the gate and threshold voltage of a MOS transistor [3.17]. This approximation is justified by the accurate comparison results of the physical model to the reference model and experimental data from fabricated devices in Section 3.5.

Figure 3.2: Block diagram of the physical model.

The input for the Verilog-A module is the control gate voltage V_{CG} . The module contains the physical behavior of programming and erasing the device, and it refreshes the voltage on the floating gate V_{FG} dynamically with respect to time such that charge leakage and unintended programming during operation mode is included. This voltage is then applied to the gate terminal of any SPICE compatible NMOS transistor model, which is designed with an oxide thickness equal to d_{tox} of the NCFG device in order to obtain the correct I-V characteristics. In this Chapter, a predictive NMOS transistor is used for ease of simulations.

3.4.1 Algorithm for Computing the Floating Gate Voltage

The key element of the physical model is the computation of the floating gate voltage V_{FG} . The algorithm for computing V_{FG} at each time step during the transient simulation is shown in Figure 3.3. It accounts for both programming and erasing the device, and includes charge leakage. This design flow is implemented in the Verilog-A module of Figure 3.2.

Note that for DC simulations this algorithm is only computed with one iteration and the user needs to know the charge on the floating gate beforehand (e.g. by running a previous transient simulation) and insert it manually in the model's netlist. During the control gate voltage sweep any charge transfer through the oxide is not accounted for, as the DC simulations are time independent. This may be different in a real life situation where for example a sweep to a higher control gate voltage for a longer time may result in significant charging of the floating gate.

Figure 3.3: Design flow of computing the floating gate voltage V_{FG} in the Verilog-A module.

For the transient simulation it is assumed that at the starting point $t = 0$ s there is no charge Q_{FG} on the floating gate. V_{CG} is the only input state variable and a constant time step Δt is chosen by the user. A smaller Δt results in a more accurate final result, but leads to a longer simulation time. Given this information, the voltages across the tunnel oxide V_{tox} and control gate oxide V_{cox} are calculated, which is further explained in Section 3.4.3 along with the derivation of V_{FG} .

Next, the tunnel current densities in the tunnel oxide and the control gate oxide are computed. Depending on the control gate voltage V_{CG} and the charge on the floating gate Q_{FG} , tunnel current flows either to the floating gate, or out of the floating gate. The following equations account for both the inward and outward tunnel current densities of the

floating gate as indicated in Figure 3.4:

$$J_{sub-FG} = J_{in,sub-FG} - J_{out,sub-FG} \quad (3.1)$$

which describes the tunnel current density in the tunnel oxide from the substrate to the floating gate and

$$J_{FG-CG} = J_{out,FG-CG} - J_{in,FG-CG} \quad (3.2)$$

which describes the tunnel current density in the control gate oxide from the floating gate to the control gate.

Figure 3.4: Capacitor model between the control gate, floating gate and substrate surface nodes of the NCFG device.

The physical model includes the carrier transport through the tunnel and control gate oxide due to uniform direct tunneling (DT) and Fowler-Nordheim tunneling (FNT) of electrons from conduction band to conduction band as shown in Figure 3.5. These are the dominant charge transfer mechanisms when the source and drain terminals of the device are grounded. If a voltage is applied on the drain terminal along with a high control gate voltage, the floating gate may also be charged due to channel hot electron injection (CHEI). CHEI is

highly localized at the drain terminal and is not applicable for uniform tunneling [3.5]. It is not included in the physical model and the model only applies for uniform tunneling in the program mode with no lateral field across the channel, i.e. no or minimal impact ionization is present. This implies that the model does not take charge transfer between the nanocrystals into account, which may occur when the device is programmed due to CHEI where the nanocrystals that are in close proximity to the drain terminal are highly charged compared to the adjacent nanocrystals. By utilizing uniform tunneling, all the nanocrystals in the floating gate are similarly charged and simulations of the reference model in the device simulator have demonstrated a slight deviation between the outer and inner nanocrystals of less than 3%. The physical model assumes that all nanocrystals are equally charged or discharged.

In the operation mode and the erase mode the model can be used without the constraint of keeping the drain and source terminal grounded. CHEI is negligible for NCFG devices with today's deep submicron geometries when biased with low operational supply, drain and source voltages [3.16]. Numerical device simulations have shown that the charge transfer due to CHEI for a device with $d_{\text{tox}} = 2 \text{ nm}$, $V_{\text{CG}} = 2 \text{ V}$, and a drain and source voltage of 1.1 V and 0 V , respectively, is six orders of magnitude less compared to the charge transfer due to uniform tunneling for the same device with $V_{\text{CG}} = 6 \text{ V}$ and the drain and source terminal grounded. Therefore, the impact of CHEI in the operation mode can be considered minimal. In the erase mode CHEI is not applicable due to the absence of a lateral field in the floating gate.

Figure 3.5: Band diagram across the Pt-SiO₂-Pd-SiO₂-Si stack when a high voltage is applied on the control gate.

DT is dominant for ultra-thin oxides (less than ~ 3 nm) and low electric fields across the oxides and is used solely for $qV_{tox} < \phi_{b,Si}$ and $qV_{cox} < \phi_{b,Pd}$, in the case of programming the device. Furthermore, it is the main contribution to the charge leakage of a programmed device in retention state, where zero or low voltages are applied on the control gate. It is rather insignificant for the thicker control gate oxide but may still apply, especially if a device with a retention time of longer than 10 years is considered. The electrons tunnel through the SiO₂ barrier of rectangular shape and the current density of DT is given by [3.16]

$$J_{DT} = \frac{q^3}{8\pi h \phi_b} \left(\frac{m_{Si}}{m_{ox,n}} \right) E^2 \theta \quad (3.3)$$

with

$$\theta = \exp \left\{ - \frac{8\pi \sqrt{2m_{ox,n}} [\phi_b^{3/2} - (\phi_b - qV)^{3/2}]}{3hqE} \right\} \quad (3.4)$$

where E is the electric field across the oxide. $m_{\text{si}} \approx m_0$ represents the mass of the free electron in the silicon and $m_{\text{ox},n} = 0.4m_0$ [3.18] is the effective mass of the tunneling electron in the oxide, which is assumed to be constant for both the control gate and tunnel oxides. ϕ_b is the conduction band barrier height between the interfaces of the materials, which are illustrated in Figure 3.5 along with the tunneling mechanisms DT and FNT. For the Si/SiO₂ interface a barrier height of $\phi_{b,\text{Si}} = 3.1$ eV [3.18] is used. For the interface between the metals and the SiO₂ it is determined by the difference between the work function of the metal and the energy level corresponding to the SiO₂ conduction band with respect to the vacuum level, which is given by 0.9 eV [3.19]. Hence, the barrier height at the Pd/SiO₂ interface is given by $\phi_{b,\text{Pd}} = 4.22$ eV and at the Pt/SiO₂ interface by $\phi_{b,\text{Pt}} = 4.75$ eV.

FNT is applied for $qV_{\text{tox}} < \phi_{b,\text{Si}}$ and $qV_{\text{cox}} < \phi_{b,\text{Pd}}$ as well, but plays the dominant role for high electric fields $qV_{\text{tox}} > \phi_{b,\text{Si}}$ and $qV_{\text{cox}} > \phi_{b,\text{Pd}}$, or thicker oxides. Therefore, it plays a major role when programming the device with a relatively high positive control gate voltage V_{CG} , and erasing the device with a relatively high negative V_{CG} . In the case of FNT, the electrons tunnel through the SiO₂ barrier with the shape of a triangle and the current density is given by [3.16]

$$J_{\text{FN}} = A \cdot E^2 e^{\frac{-B}{E}} \quad (3.5)$$

with

$$A = \frac{q^3}{8\pi h \phi_b} \left(\frac{m_{\text{Si}}}{m_{\text{ox},n}} \right) \quad B = \frac{8\pi \sqrt{2m_{\text{ox},n} \phi_b^3}}{3hq} \quad (3.6)$$

Equations 3.3-3.6 account for erasing the device as well, where the barrier heights are now given by $\phi_{b,\text{Pd}}$ and $\phi_{b,\text{Pt}}$. The conditions for DT are $|qV_{\text{tox}}| < \phi_{b,\text{Pd}}$ for the tunnel oxide and

$|qV_{\text{cox}}| < \phi_{\text{b,Pt}}$ for the control gate oxide. Again, FNT applies for these conditions as well but it is dominant for $|qV_{\text{tox}}| > \phi_{\text{b,Pd}}$ and $|qV_{\text{cox}}| > \phi_{\text{b,Pt}}$. Thus, DT and FNT are applied for all four tunnel current densities: the current densities into the floating gate from the control gate and substrate ($J_{\text{in,FG-CG}}$ and $J_{\text{in,sub-FG}}$), and the current densities out of the floating gate to the control gate and substrate ($J_{\text{out,FG-CG}}$ and $J_{\text{out,sub-FG}}$).

The next step is to compute the new charge on the floating gate Q_{FG} after one time step Δt with the known total current density into the floating gate $J_{\text{sub-FG}} - J_{\text{FG-CG}}$. It is the sum of the charge after the previous Δt plus the additional charge of the current Δt [3.20]:

$$Q_{\text{FG}} = Q_{\text{FG-}prev} - (J_{\text{sub-FG}} - J_{\text{FG-CG}})Area_{\text{dot}}R_{\text{dot}}\Delta t \quad (3.7)$$

where $Area_{\text{dot}}$ is the surface area of the nanocrystal. When comparing to the reference model, the nanocrystals have a form of cuboids, while comparing to experimental results they have a form of spheres. R_{dot} is the floating gate surface portion covered by the nanocrystals. A variation in the number of nanocrystals in the floating gate can be modeled with R_{dot} , however the model does not take into account a potential displacement of the nanocrystals and assumes a uniform distribution. In theory, the physical model may also be used for a continuous floating gate device with a filling rate of $R_{\text{dot}} = 100\%$. However, continuous floating gate devices are typically programmed due to FNT and CHEI; the latter is not included in the model. Thus, it is left to the designer to use the model for continuous floating gate devices while programming it solely due to uniform tunneling mechanisms.

Finally, the last step of the design flow is to update the output V_{FG} which is derived in Section 3.4.3 by means of the band diagrams. In addition, if it is desired, the threshold

voltage shift ΔV_T can be determined as [3.21]

$$\Delta V_T = -\frac{Q_{FG}R_{dot}}{C_{cox}} \quad (3.8)$$

The algorithm starts again at the top of the design flow for the next time step Δt and the iterations continue until the transient circuit simulation ends. The key advantage of this approach is that V_{FG} and ΔV_T of the device are always dynamically refreshed throughout the entire transient circuit simulation. Thus, this physical model includes charge leakage, and possible unintended charging of the device during operation mode, as well as other potential disturb conditions. Therefore, an accurate analysis of the power consumption is enabled. Due to the use of only explicit formulas where numerical computations are not included, the computation is fast and the physical model is suitable for transient circuit simulations. In the next Section 3.4.2 it is described how the surface potential can be analytically computed, without solving Schrödinger-Poisson equations numerically.

3.4.2 Surface Potential

The surface potential V_{surface} at the Si/SiO₂ interface needs to be known to determine the voltage across the tunnel oxide V_{tox} , which is required for both DT and FNT in $J_{\text{sub-FG}}$. Very accurate computations of V_{surface} can be obtained by solving Schrödinger-Poisson equations through an iterative numerical approach, such as in the device simulator [3.22] or in other model papers [3.1-3.8]. However, this requires heavy computations, which makes it unsuitable for circuit simulation.

In this Chapter an explicit model for surface potential calculation is used. The

analytical formulas for the surface band bending in the accumulation, weak inversion, and strong inversion region are well known for MOSFETs [3.23, 3.24] and it is assumed that they are also valid for NCFG devices where the voltage on the floating gate V_{FG} acts as the gate voltage of a MOSFET. Therefore, the voltage difference between the gate voltage and the flat-band voltage of a MOSFET is replaced by the voltage difference between V_{FG} and a modified flat-band voltage V_{FB}^* . Since the nanocrystals are surrounded by SiO_2 , the contact potentials at the Pd/SiO_2 interfaces cancel out everywhere. Therefore, the flat-band voltage is independent of the Pd/SiO_2 contact potential and modified to

$$V_{FB}^* = -\varphi_F - \frac{Q_o' d_{tox}}{\epsilon_{ox}} \quad (3.9)$$

which includes the Fermi potential in the channel φ_F and the charge per unit area Q_o' at the Si/SiO_2 interface that the user may specify in the parameter list. In this Chapter, it is assumed that there is no charge trapped at the interface. The surface band bending in the accumulation region, the weak inversion region and strong inversion region is then analytically computed with the formulas from [3.23, 3.24].

3.4.2.1 Accumulation

First, the band bending in the accumulation region is being derived. The NCFG device is in the accumulation region when the condition $V_{FG} < -\varphi_F$ is met, which is the case for erasing the device and may also apply when the device is in retention state (considering a negative charge and thus negative voltage on the floating gate). The band bending is then computed as [3.23]

$$\psi_{accu} = -\varphi_t \ln \left[1 - \frac{\psi_{accu}^*}{\varphi_t} + \left(\frac{V_{FG} - V_{FB}^* - V_{FG}^* - \psi_{accu}^*}{\gamma \sqrt{\varphi_t}} \right)^2 \right] \quad (3.10)$$

Here, $\varphi_t = \kappa T/q$ is the thermal voltage and the functions ψ_{accu}^* and V_{FG}^* are given by

$$\psi_{accu}^* = \frac{\Gamma(V_{FG} - V_{FB}^* - V_{FG}^*)}{\sqrt{1 + \left[\frac{\Gamma(V_{FG} - V_{FB}^* - V_{FG}^*)}{4\varphi_t} \right]^2}} \quad (3.11)$$

and

$$V_{FG}^* = \frac{1}{2} (V_{FG} - V_{FB}^* + \sqrt{(V_{FG} - V_{FB}^*)^2 + 4K^2}) \quad (3.12)$$

where $\Gamma = (1 + \gamma/\sqrt{2\varphi_t})^{-1}$ ensures correct physical behavior around flat-band condition and $K = 0.1$ V is a smoothing constant [3.23]. The equation for the body parameter γ is given in the following Section 3.4.2.2. The accuracy of modeling the surface potential $V_{surface}$ in the accumulation region is demonstrated in Figure 3.8 in the Section 3.4.2.4 along with the surface potential in the weak and strong inversion region.

3.4.2.2 Weak Inversion

The NCFG device is in the weak inversion region when the conditions $V_{FG} > -\varphi_F$ and $\psi_{weak} < 2\varphi_F$ are met and the surface band bending is given by [3.24]

$$\psi_{weak} = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{FG} - V_{FB}^*} \right)^2 \quad (3.13)$$

The body parameter γ is defined as

$$\gamma = \frac{\sqrt{2q\epsilon_{Si}(N_{eff} + N_A)d_{tox}}}{\epsilon_{ox}} \quad (3.14)$$

where N_A is the p-type doping in the substrate and N_{eff} the effective doping of the channel,

such that ion diffusion into the channel is taken into account. N_{eff} can be found by transforming the Gaussian distribution of the ions in the channel with the doping concentration N_{Channel} to a constant distribution N_{eff} with equivalent area under the curve as shown in Figure 3.6.

Figure 3.6: Transforming the Gaussian distribution of the ions in the channel to a constant distribution with equivalent doping density.

The general equation of the Gaussian function is given by

$$f(y) = ae^{-\frac{y^2}{2c^2}} \quad (3.15)$$

and its integral by

$$\int_0^\infty ae^{-\frac{y^2}{2c^2}} = \frac{ac\sqrt{2\pi}}{2} \quad (3.16)$$

By using equation 3.15 in order to model the Gaussian distribution of the implanted ions in the channel, the following assumptions are made: first, at the Si/SiO₂ interface there is a

maximum concentration of implanted ions $N_{Channel}$:

$$f(0) = N_{Channel} \rightarrow a = N_{Channel} \quad (3.17)$$

and second, at the channel junction y_{max} the number of implanted ions equals the number of uniformly doped acceptors in the substrate N_A :

$$f(y_{max}) = N_A \rightarrow c = \frac{y_{max}}{\sqrt{2 \ln\left(\frac{N_{Channel}}{N_A}\right)}} \quad (3.18)$$

Therefore, using equation 3.17 and 3.18 the Gaussian distribution of the implanted ions in the channel follows the equation

$$N(y) = N_{Channel} e^{-\frac{y^2}{2c^2}} \quad (3.19)$$

and is plotted in Figure 3.7. It can be observed that it is modeled very accurately compared to the reference model.

Figure 3.7: Comparison of the Gaussian distribution of the ions implantation in the channel between the reference model and the physical model.

Now that the function of the Gaussian distribution for the implanted ions is known, N_{eff} can be calculated with equation 3.16, using the constants from equations 3.17 and 3.18. It is divided by the channel junction depth y_{max} , such that

$$N_{\text{eff}} = \frac{N_{\text{channel}} c \sqrt{2\pi}}{2y_{\text{max}}} \quad (3.20)$$

This method is important for an accurate computation of the body parameter γ which is used for the band bending formulas in each region of the NCFG device. In operation mode the device may be in the weak inversion region if no charge or a small negative or positive charge is stored on the floating gate. In the next Section 3.4.2.3 the strong inversion region is briefly described which is important for programming the device.

3.4.2.3 Strong Inversion

By raising the control gate voltage further, the device will eventually be in the program mode where electrons in the channel tunnel to the floating gate. As the floating gate voltage also increases along with the control gate voltage, the device will reach the strong inversion region once the condition $\psi_{\text{weak}} > 2\phi_F$ is fulfilled. At this point the surface band bending is saturating and is given by [3.24]

$$\psi_{\text{strong}} = 2\phi_F + \Delta\phi \quad (3.21)$$

where $\Delta\phi$ is an approximate explicit expression which accounts for any surface potential difference between weak inversion and strong inversion. It enables a smooth transition between these two regions and is defined as

$$\Delta\phi = \frac{2\phi_t}{n} \ln \left(1 + \frac{\psi_{\text{weak}} - 2\phi_F}{2\phi_t} \right) \quad (3.22)$$

where n represents the inverse of the slope in the weak inversion region in Figure 3.8, defined as

$$n = 1 + \frac{y}{2\sqrt{\psi_{weak}}} \quad (3.23)$$

3.4.2.4 Summary

Using these explicit formulas for band bending in accumulation, weak and strong inversion regions of the device, the surface potential is given by [3.24]

$$V_{surface} = (\psi_{accu}, \psi_{weak}, \psi_{strong}) + \varphi_F \quad (3.24)$$

They are implemented in the design flow of Figure 3.3 for evaluating V_{FG} . Therefore, $V_{surface}$ is also dynamically updated for every cycle. The results for an uncharged device with a tunnel oxide thickness of 2.6 nm are plotted in Figure 3.8.

Figure 3.8: Comparison of the surface potential $V_{surface}$ between the reference model and the physical model for the control gate voltage V_{CG} ranging from -7 to 7 V.

This illustrates how accurate the surface potential is modeled in comparison to the numerical device simulator that solves Schrödinger-Poisson equations. The shapes of the curves are very similar and the maximum deviation compared to the reference model in the accumulation and strong inversion regions, where erasing and programming of the device take place, is less than 3.45% and 3.71%, respectively. These close matching results for the surface potential are the prerequisites for modeling accurately the charge transfer of the NCFG device during program, erase, and operation mode.

3.4.3 Program and Erase Mode

A positive voltage on the control gate programs the device. Figure 3.9(a) shows a typical band diagram in the program mode across the Pt-SiO₂-Pd-SiO₂-Si stack. Electrons uniformly tunnel from the channel to the floating gate under the condition $V_{FG} > V_{surface}$. As the floating gate is being programmed, the tunneling electrons contribute to an increasing negative Q_{FG} . Therefore, the floating gate voltage V_{FG} is lowered, which results in a decreased voltage across the tunnel oxide V_{tox} . This causes a reduction of the tunneling current from the channel to the floating gate. Furthermore, V_{cox} increases during programming, which means that the probability of electrons to tunnel further from the floating gate to the control gate increases. When this tunneling current is in balance with the current from the channel to the floating gate, the threshold voltage shift ΔV_T is saturating. Additionally, as V_{FG} decreases during the program mode, it is likely that in the operation mode $V_{FG} < V_{surface}$ such that the electrons tend to tunnel back to the channel, which results in charge leakage. As the physical model updates dynamically during the transient circuit

simulation, those described scenarios are always accounted for.

Erasing of the device occurs when a negative voltage is applied on the control gate. A typical band diagram across the Pt-SiO₂-Pd-SiO₂-Si stack in the erase mode is illustrated in Figure 3.9(b). Electrons tunnel uniformly from the floating gate to the substrate as long as the condition $V_{FG} < V_{surface}$ is met. In the case where the device has been previously programmed and the floating gate negatively charged, V_{FG} will become even lower in the erase mode. Thus, the back-tunneling of electrons is accelerated due to the increased V_{tox} compared to an uncharged floating gate under the same applied V_{CG} . Analogous to the program mode, electrons may tunnel through the control gate oxide if V_{cox} is sufficiently high. Based on DT and FNT equations the high field may overcome the relatively thick control gate oxide and enable electron tunneling through the SiO₂ barrier. This is not desired such that scaling of d_{cox} and V_{CG} is limited.

Figure 3.9: Band diagram of the NCFG device with $d_{cox} = 4$ nm and $d_{tox} = 2$ nm in (a) the program mode and (b) the erase mode. The thick solid line represents the potential.

The tunneling current densities for DT and FNT are unique functions of the voltages across the tunnel and control gate oxides [3.25], which are derived from the band diagrams as

$$V_{tox} = |V_{FG} - V_{surface}| \quad (3.25)$$

$$V_{cox} = |(V_{CG} + \varphi_m) - V_{FG}| \quad (3.26)$$

Considering charge neutrality on the floating gate ($V_{tox}C_{tox} = V_{cox}C_{cox}$), V_{FG} is derived as

$$V_{FG} = \frac{(V_{CG} + \varphi_m)C_{cox} + V_{surface}C_{tox}}{C_{cox} + C_{tox}} + \frac{Q_{FG} + Q_{FG}^*}{C_{cox} + C_{tox}} \quad (3.27)$$

The second term takes into account the transferred charges from the channel and Q_{FG}^* represents the initial charge trapped on the nanocrystals which is required for DC simulations. The oxide capacitances C_{cox} and C_{tox} are calculated based on the geometry of the nanocrystals and φ_m is the contact potential of the control gate. Note that equation 3.27 differs from those reported in [3.17, 3.21] which do not consider Q_{FG}^* , φ_m , and $V_{surface}$. Note also that the equations derived for $V_{surface}$ and V_{FG} are interdependent. Thus, the value for $V_{surface}$ calculated from the previous time step is used in computing V_{FG} at the current time step. The physical model now includes all scenarios of electron tunneling due to DT and FNT. The results of its accuracy and speed are investigated in the following Section 3.5.

3.5 Results and Discussion

The physical model is compared with the reference model from the device simulator and experimental results [3.12, 3.26, 3.27]. A DC simulation of an inverter with the NCFG device is demonstrated. In addition, a timing comparison between the physical model and the reference model is provided.

3.5.1 Comparison to the Device Simulator

Figure 3.10(a) and (b) demonstrate a close matching of the physical model to the device simulator in the charging and discharging behavior of the NCFG device with different d_{tox} of 2 nm, 2.6 nm and 3.2 nm. The process and geometry parameters (shown in Table 3.1) are identical for both models and no curve fitting is applied. A ΔV_T vs. V_{CG} plot of programming the device for various d_{tox} and a programming time of 10 μs is illustrated in Figure 3.10(a). Simulations have shown that the plots for different times of 0.1 μs , 1 μs and 100 μs look very similar. For $d_{\text{tox}} = 3.2$ nm FNT is dominating, whereas for $d_{\text{tox}} = 2$ nm DT is dominating. In the case of $d_{\text{tox}} = 2.6$ nm a minimal discrepancy can be observed for lower biases on the control gate. This offset may be due to the combination of DT and FNT, where the device simulator demonstrates a slightly lower programming V_{CG} than the physical model. For high fields FNT is dominating and the curves merge together again.

Figure 3.10(b) also demonstrates good agreement between the physical model and the device simulator when applying erase voltages for 1 ms. The floating gate has been precharged in order to obtain an initial ΔV_T of 2 V. Due to the deep asymmetric potential well in the floating gate the electrons require more time and/or energy to tunnel through the higher SiO_2 barrier compared to the program mode. Therefore, a longer erasing time of 1 ms is chosen. Similar agreement for other erasing times of 0.1 ms, 10 ms and 100 ms has been observed. The curves show a slight offset for a lower negative V_{CG} where the physical model predicts less erasing current. Again, it is anticipated that this may be due to the combination of DT and FNT. For similar voltages across the oxides for the program and erase conditions,

the DT component is more significant in the erase condition since the barrier is higher. The maximum deviation can be observed for $d_{\text{tox}} = 2.6 \text{ nm}$ with V_{CG} ranging from -7 to -9 V. It is assumed that in this case the tunneling probabilities of DT and FNT are evenly distributed. For higher negative voltages only FNT is present and the curves align very well.

Figure 3.10: Comparison of ΔV_T between the physical model and the device simulator for (a) the program mode and a programming time of $10 \mu\text{s}$ and (b) the erase mode and an erasing time of 1 ms.

3.5.2 Comparison to Experimental Data from Fabricated Devices

Comparisons between the physical model and various experimental results are illustrated in Figures 3.11-3.13 in terms of threshold voltage shift ΔV_T , retention time and flat-band voltage shift ΔV_{FB} . The geometry and process parameters in the model are matched in each case to the corresponding fabricated device in order to provide a fair comparison.

3.5.2.1 Threshold Voltage Shift Comparison

Figure 3.11(a) shows V_T curves of a fabricated NCFG device with respect to time for different programming voltages ranging from 8 V to 14 V and an erasing voltage of -14 V [3.26]. The control gate is composed of n^+ -polysilicon and the nanocrystals of Si, which have a surface coverage of $\sim 50\%$. The device uses a d_{tox} and d_{cox} of ~ 4 nm and ~ 6 nm, respectively. The V_T curves are saturating for longer times because the tunnel currents from the channel to the floating gate and from the floating gate to the control gate are in balance [3.26]. Furthermore, Figure 3.11(a) indicates that a greater ΔV_T can be achieved by applying a lower V_{CG} for a longer time. The reason for this is that with lower voltage biases, the voltage across the control gate oxide V_{cox} is also reduced which makes it more difficult for electrons to tunnel further from the floating gate to the control gate. Only for longer times, when sufficient electrons are stored on the floating gate, the voltage on the floating gate drops to a point where V_{cox} is large enough to enable electron tunneling to the control gate.

The device is then simulated using the physical model with the equivalent parameters. The barrier height at the interface between the Si nanocrystals and the oxide is now similar to the barrier height at the interface between the channel and the oxide, which is chosen to be $\phi_{\text{b,Si}} = 3.1$ eV. For the control gate a work function of 4.17 eV is assumed, and therefore a barrier height of $\phi_{\text{b,n-Poly}} = 3.27$ eV is used at the interface between the control gate and the oxide. The V_T curves obtained from this model are represented in Figure 3.11(b). It can be observed that they match nicely with the experimental results. Note that the curve for an applied control gate voltage of 8 V starts at a slightly higher initial V_{T0} . It is assumed that in

the experimental investigations, the device has not been completely erased before being programmed again. This slight change in V_{T0} has also been adjusted in the predictive NMOS transistor of the model to provide a one-to-one comparison. The maximum offset has been measured for the 8 V programming curve at ~100 ms where the physical model predicts a threshold voltage of 1.8 V, which is ~20% less than the experimental data which is measured at ~2.25 V.

Figure 3.11: (a) Experimental results of V_T for programming and erasing a NCFG device with a nanocrystal surface coverage of 50% (inset) [3.26]. (b) The V_T curves are computed with the physical model.

3.5.2.2 Retention Time Comparison

To test the retention time, experimental data and its corresponding extrapolation line are extracted from [3.27] and compared to simulation results from the physical model. The

retention time is measured at a charge loss of 25% of a programmed NCFG device. The fabricated devices use n^+ -polysilicon for the control gate and contain Si nanocrystals with a diameter of ~ 6 nm and a density of $\sim 3 \times 10^{11} \text{ cm}^{-2}$. This results in a surface coverage of $\sim 8.5\%$. The device is then modeled with the equivalent geometry parameters and a barrier height of $\phi_{b,\text{Si}} = 3.1$ eV is used again for the interface between the channel and the oxide and between the nanocrystals and the oxide. A barrier height of $\phi_{b,n\text{-Poly}} = 3.27$ eV is chosen between the control gate and the oxide. Figure 3.12(a) and (b) illustrate the experimental data along with their extrapolation lines from [3.27]. The results obtained from the physical model are directly inserted in each graph. In Figure 3.12(a) the retention time is represented for different d_{tox} ranging from 2.5 nm to 6.5 nm with a fixed V_{CG} of 0 V. The data points are obtained from various fabricated devices with $d_{\text{cox}} = 7$ nm for tunnel oxides thicknesses of $d_{\text{tox}} = 3$ nm and $d_{\text{tox}} = 4$ nm; whereas $d_{\text{cox}} = 6$ nm is chosen for a relatively thin tunnel oxide of $d_{\text{tox}} = 2.8$ nm. Figure 3.12(b) represents the retention time of a NCFG device with $d_{\text{tox}} = 4$ nm and $d_{\text{cox}} = 7$ nm for V_{CG} ranging from -7 V to 0 V. Experimental data points are given for $V_{\text{CG}} \in \{-6 \text{ V}, -5 \text{ V}, -4 \text{ V}, -3 \text{ V}\}$. Computing the retention time in the physical model results in a similar curve as the extrapolation line in each graph. The maximum deviation has been observed in Figure 8(b) where the retention time for the experimental data point $V_{\text{CG}} = -4$ V reaches 4000 s; whereas the physical model predicts a V_{CG} of -3.7 V to achieve this retention time, resulting in a deviation of 7.5%.

Figure 3.12: Comparison of the retention time results for (a) a fixed V_{CG} of 0 V and different d_{tox} [3.27] and (b) a fixed device ($d_{tox} = 4$ nm and $d_{cox} = 7$ nm) with different V_{CG} [3.27].

3.5.2.3 Flat-Band Voltage Shift Comparison

A comparison between the physical model and a fabricated nanocrystal MOSCAP [3.12] is shown in Figure 3.13 in terms of the flat-band voltage shift ΔV_{FB} . Gold (Au) nanocrystals are embedded in the oxide, which have a work function of 5.10 eV [3.11]. They have a diameter of ~ 10 nm and a density of $\sim 4 \times 10^{11} \text{ cm}^{-2}$, which results in a surface coverage of $\sim 31\%$. The substrate of the device is n-type and the control gate is composed of molybdenum (Mo) with a work function of 4.60 eV [3.11].

A voltage sweep on the control gate from -5 V to 5 V and back in steps of 0.05 V is performed. In each step the applied bias lasts for ~ 0.5 s. The capacitance-voltage (C-V) curves in Figure 3.13 show hysteretic characteristics during the sweep. This outcome is

expected since a negative voltage first attracts holes to the floating gate and repels electrons, thus positively charging it and shifting the flat-band voltage to the left. This charge increases during the sweep to the point where V_{CG} reaches a value such that $V_{FG} > V_{surface}$. Then, the positive charge on the floating gate will start to leak away and eventually it will be negatively charged once V_{CG} is relatively high, thus shifting V_{FB} to the right. This time, the charge remains negative during the reverse sweep until negative voltages are applied again. The gap in the hysteresis curve is interpreted as the flat-band voltage shift ΔV_{FB} and has a value of roughly 0.8 V in the fabricated device, whereas in the physical model ΔV_{FB} is at 0.85 V.

Note that the slopes are slightly different. This may be due to the assumptions made in this Chapter that the silicon is uniformly doped and that there are no trapped charges in the Si/SiO₂ interface. In reality, trapped charges, material impurities and defects in the silicon are inevitable, and cause a reduction in the slope of the C-V curve [3.28].

Figure 3.13: C-V comparison between the physical model and experimental results from [3.12]. The gate voltage is swept from -5 V to +5 V and back to -5 V.

3.5.3 HSPICE Circuit Simulation with Implementation of the Physical Model

To demonstrate SPICE compatibility the physical model is implemented in HSPICE and tested in a DC simulation of an inverter, where the NMOS transistor is replaced by a NCFG device. The device is designed with thin oxides of $d_{\text{tox}} = 2$ nm and $d_{\text{cox}} = 4$ nm in order to demonstrate quick charging, discharging and charge leakage. It contains nanocrystals composed of Pd with a diameter of 3 nm and a spacing of 3 nm between them. Therefore, a nanocrystal density of $\sim 2.78 \times 10^{12} \text{ cm}^{-2}$ and a surface coverage of $\sim 20\%$ is obtained. The predictive NMOS transistor in the physical model is a BSIM4.0 model with a gate length of 45 nm and a gate width of 120 nm. The PMOS transistor has a gate length of 45 nm as well, and a gate width of 180 nm. The DC simulation results of the inverter with different consecutive input voltage pulses are illustrated in Figure 3.14. The NCFG device is initially uncharged and then programmed for 500 μs by a voltage pulse of 3 V. After the applied voltage is taken, the device is kept in a retention state for 1 s with $V_{\text{CG}} = 0$ V where a significant amount of charge is leaking through the ultra-thin tunnel oxide. In the last step an erasing voltage pulse of -1 V is applied for 100 ms. The voltages used in this particular simulation are relatively low such that oxide breakdown of the PMOS transistor (and thus simulation failure) is avoided. The supply voltage is set to 1 V.

Figure 3.14: DC simulation in HSPICE of an inverter with the NCFG device using the physical model and replacing the NMOS transistor. This device is cycled consecutively through an uncharged, programmed, retention and erased state.

3.5.4 Run Time Analysis

The speed of the physical model is simulated on one NCFG device with oxide thicknesses of $d_{\text{cox}} = 4 \text{ nm}$ and $d_{\text{tox}} = 2 \text{ nm}$ in HSPICE. The applied control gate is constantly biased with 4.5 V for 100 μs . As a reference, the numerical solver ATLAS is used. A fine mesh as thin as 100th of a nanometer is defined at all material interfaces in order to obtain theoretically accurate results. The mesh of the entire device consists of a total of 8790 grid points and 17220 triangles. The robustness of this mesh has been verified by running multiple simulations with a different mesh for a device with varying tunnel oxide thicknesses and varying input voltages. In comparison, if a mesh as thin as a 50th of a nanometer is used at all material interfaces, the grid points and triangles are reduced to 6794 and 13286,

respectively, which results in an average decrease in the computation time of 20%, and in an average deviation in ΔV_T of 1.3%. The simulations are executed multiple times for different time steps on a 2.66 GHz Intel Core i7 server with 12 GB DDR3 and the average results are presented in Table 3.2. The device simulator fully utilizes the 8 parallel process threads of this CPU; whereas the physical model is only utilizing a single process thread.

Table 3.2: Accuracy and simulation time measured for various time steps for both the device simulator and the physical model.

Time Step	Reference Model		Physical Model	
	ΔV_T	Sim. Time	ΔV_T	Sim. Time
1×10^{-5} s	2.1602 V	47 s	2.2548 V	< 1 s
1×10^{-6} s	2.0892 V	6:12 min	2.0449 V	1 s
1×10^{-7} s	2.0823 V	57:19 min	2.0317 V	2 s
1×10^{-8} s	2.0816 V	9:50 h	2.0304 V	4 s

Table 3.2 clearly shows that the computation time of the physical model is significantly reduced compared to its counterpart while only slightly sacrificing accuracy (2.43% deviation for a time step of 1×10^{-7} s, or 100 ns). The results also suggest use of a time step of 100 ns in order to achieve relatively fast and accurate simulations of complex circuits that include a large number of NCFG devices. Reducing this time step will result in longer simulation times while only gaining slightly better accuracy.

3.6 Summary

A fully parameterized and configurable physical model of NCFG devices is described. Due to explicit analytical formulas, no heavy computations are required, which makes the model suitable for circuit simulations in the existing SPICE-based infrastructure. The speed of this model is demonstrated as it is shown that the charging simulation of one device during a voltage pulse of 4.5 V and 100 μ s takes only approximately 2 s while arriving at a value within 2.43% compared to the reference model from Silvaco's device simulator ATLAS, which solves Schrödinger-Poisson equations numerically. The physical model interacts dynamically with the rest of the circuit so that charge transfer through the oxide with respect to time is accounted for. Comparisons to the reference model and various experimental results demonstrate good matching in threshold voltage shift, flat-band voltage shift and retention time. A wide range of devices with different oxide thicknesses, voltage pulses and materials for the nanocrystals and the control gate are investigated. In addition, a DC simulation of an inverter with an integrated NCFG device is presented in HSPICE.

The results in accuracy, speed and flexibility suggest that the developed physical model can be used in large and complex circuit simulations. Therefore, this Chapter is important in bridging the gap between device and circuit level research.

References

- [3.1] D. Corso, C. Pace, F. Crupi, and S.A. Lombardo, "Single-Electron Program/Erase Tunnel Events in Nanocrystal Memories," *IEEE Transactions on Nanotechnology*, vol. 6, pp. 35-42, 2007.
- [3.2] R. Gusmeroli, A.S. Spinelli, C.M. Compagnoni, and D. Ielmini, "Threshold-Voltage Statistics and Conduction Regimes in Nanocrystal Memories," *IEEE Electron Device Letters*, vol. 27, pp. 409-411, 2006.
- [3.3] J.S. Sim, J. Kong, J.D. Lee, and B.-G. Park, "Monte Carlo Simulation of Single-Electron Nanocrystal Memories," *Japanese Journal of Applied Physics*, vol. 43, 2004.
- [3.4] A. Campera and G. Iannaccone, "Modelling and Simulation of Charging and Discharging Processes in Nanocrystal Flash Memories During Program and Erase Operations," *Solid-State Electronics*, vol. 49, pp. 1745-1753, 2005.
- [3.5] C.M. Compagnoni, D. Ielmini, A.S. Spinelli, and A.L. Lacaita, "Modeling of Tunneling P/E for Nanocrystal Memories," *IEEE Transaction on Electron Devices*, vol. 52, pp. 569-576, 2005.
- [3.6] P. Singaraju and R. Venkat, "Modeling of the Charging Dynamics in Silicon Nanocrystal Nonvolatile Flash Memory Cells," *Physica E: Low-Dimensional Systems and Nanostructures*, vol. 41, pp. 9-14, 2008.
- [3.7] E.-S. Hasaneen, E. Heller, R. Bansal, W. Huang, and F. Jain, "Modeling of Nonvolatile Floating Gate Quantum Dot Memory," *Solid-State Electronics*, vol. 48, pp. 2055-2059, 2004.
- [3.8] P. Coli and G. Iannaccone, "Evaluation of Performance and Perspectives of Nanocrystal Flash Memories Based on 3D Quantum Modeling," *In: Proceedings of the 2001 1st IEEE Conference on Nanotechnology*, pp. 140-145, 2001.
- [3.9] G.W. Ling, C. Tupei, A. Sutiono, S. Tao, S. Yujing, and V. Richie, "Simple Compact Transient Models of Nanocrystal Flash Memory Cell," *IEEE International Conference on Communications, Circuits and Systems*, pp. 980-984, 2007.
- [3.10] S. Jeon, J.H. Han, J.H. Lee, S. Choi, H. Hwang, and C. Kim, "High Work-Function Metal Gate and High-k Dielectrics for Charge Trap Flash Memory Device Applications," *IEEE Transactions on Electron Devices*, vol. 52, pp. 2654-2659, 2005.

- [3.11] N. Miura, T. Nanjo, M. Suita, T. Oishi, Y. Abe, T. Ozeki, H. Ishikawa, T. Egawa, and T. Jimbo, "Thermal Annealing Effects on Ni/Au Based Schottky Contacts on n-GaN and AlGaIn/GaN with Insertion of High Work Function Metal," *Solid-State Electronics*, vol. 48, pp. 689-695, 2004.
- [3.12] N. Di Spigna, "Electronic Devices and Interface Strategies for Nanotechnology," *North Carolina State University, Ph.D. Dissertation*, 2006.
- [3.13] K.S. Seol, S.J. Choi, J.-Y. Choi, E.-J. Jang, B.-K. Kim, S.-J. Park, D.-G. Cha, I.-Y. Song, J.-B. Park, Y. Park, and S.-H. Choi, "Pd-Nanocrystal-Based Nonvolatile Memory Structures with Asymmetric SiO₂/HfO₂ Tunnel Barrier," *Applied Physics Letters*, vol. 89, 2006.
- [3.14] K.F. Schuegraf, C.C. King, and C. Hu, "Ultra-Thin Silicon Dioxide Leakage Current and Scaling Limit," *IEEE Symposium on VLSI Technology - Digest of Technical Papers*, pp. 18-19, 1992.
- [3.15] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45nm Early Design Exploration," *IEEE Transactions on Electron Devices*, vol. 53, pp. 2816-2823, 2006.
- [3.16] G. Groeseneken, H.E. Maes, J. Van Houdt, and J.S. Witters, "Basics of Nonvolatile Semiconductor Memory Devices," in *Nonvolatile semiconductor memory technology*, W. D. Brown and J. Brewer, Eds., NJ: Wiley-IEEE Press, pp. 1-81, 1997.
- [3.17] E.-S. Hasaneen, A. Rodriguez, B. Yarlagadda, F. Jain, E. Heller, W. Huang, J. Lee, and F. Papadimitrakopoulos, "Nonvolatile Quantum Dot Memory (NVQDM) in Floating Gate Configuration: Device and Circuit Modeling," *In: Proc. of the 3rd IEEE Conf. on Nanotechnology*, vol. 2, pp. 741-744, 2003.
- [3.18] W.-C. Lee and C. Hu, "Modeling CMOS Tunneling Currents Through Ultrathin Gate Oxide Due to Conduction- and Valence-Band Electron and Hole Tunneling," *IEEE Transactions on Electron Devices*, vol. 48, pp. 1366-1373, 2001.
- [3.19] S. Zafar, C. Cabral, R. Amos, and A. Callegari, "A Method for Measuring Barrier Heights, Metal Work Functions and Fixed Charge Densities in Metal/SiO₂/Si Capacitors," *Applied Physics Letters*, vol. 80, pp. 4858-4860, 2002.
- [3.20] V. Beyer, J. von Borany, and M. Klimenkow, "A Transient Electrical Model of Charging for Ge Nanocrystal Containing Gate Oxides," *Journal of Applied Physics*, vol. 101, 2007.

- [3.21] B. De Salvo, A. Fernandes, G. Ghibaudo, B. Guillaumot, T. Baron, and G. Reimbold, "Investigation of Dynamic Memory Effects in Si-dot Devices," *In: Proceeding of the 30th European Solid-State Device Research Conference*, pp. 280-283, 2000.
- [3.22] "ATLAS User's Manual," *Silvaco International*.
- [3.23] F. Prégaldiny, C. Lallement, and D. Mathiot, "Accounting for Quantum Mechanical Effects from Accumulation to Inversion, in a Fully Analytical Surface-Potential-Based MOSFET Model," *Solid-State Electronics*, vol. 48, pp. 781-787, 2004.
- [3.24] Y. Tsididis, *Operation and Modeling of the MOS Transistor*, 2nd ed. New York: Oxford, 2003.
- [3.25] Y. Shi, T.P. Ma, S. Prasad, and S. Dhanda, "Polarity Dependent Gate Tunneling Currents in Dual-Gate CMOSFETs," *IEEE Transaction on Electron Devices*, vol. 45, pp. 2355-2360, 1998.
- [3.26] R.A. Rao, R.F. Steimle, M. Sadd, C.T. Swift, B. Hradsky, S. Straub, T. Merchant, M. Stoker, S.G.H. Anderson, M. Rossow, J. Yater, B. Acred, K. Harber, E.J. Prinz, B.E. White, Jr. and R. Muralidhar, "Silicon Nanocrystal Based Memory Devices for NVM and DRAM Applications," *Solid-State Electronics*, vol. 48, pp. 1463-1473, 2004.
- [3.27] C.M. Compagnoni, D. Ielmini, A.S. Spinelli, A.L. Lacaita, C. Previtali, and C. Gerardi, "Study of Data Retention for Nanocrystal Flash Memories," *IEEE 41st Annual International Reliability Physics Symposium Proceedings*, pp. 506-512, 2003.
- [3.28] B.G. Streetman and S. Banerjee, *Solid State Electronic Devices*, 5th ed. Englewood Cliffs: Prentice Hall, 1999.

Chapter 4

Low Power Interconnect Design for FPGAs with Bidirectional Wiring Using Nanocrystal Floating Gate Devices

4.1 Introduction

A FPGA is a reconfigurable circuit architecture and consists of logic blocks, connections blocks, switch boxes and routing channels as illustrated in Figure 4.1. Each component (except the routing channel) is configurable by programming elements. They interface via horizontal and vertical routing channels throughout the entire board to form the desired circuit functionality. The path that a signal travels from one logic block to another is determined by the programmable memory elements in the switch boxes and connection blocks.

The major issues of a FPGA are the power consumption and the increasing cost in high volume applications. Industries pursuit of Moore's law has led to a significant increase in transistor density but also in power consumption primarily due to the subthreshold and gate leakage current of transistors [4.1]. The FPGA is especially affected by this trend because it uses significantly more transistors per function than fixed-logic solutions, such as

an application-specific integrated circuit (ASIC) [4.2]. Regardless of this, FPGAs are found in a wide range of complex applications that require flexibility and performance due to their reconfigurable circuit architectures. Low non-recurring engineering cost (NRE), fast time to market and simplicity are the benefits from choosing a FPGA. ASICs, on the other hand, consume significantly less power and area [4.2] and are more cost efficient for high volume applications [4.3]. Therefore, they are the preferred choice for mobile devices or other battery-powered applications [4.4, 4.5]. As transistors continue to scale, it is important to find new methods for FPGAs to be more energy efficient. This Chapter presents how NCFG devices can be used to reduce the gap to ASICs in terms of power consumption and area.

Figure 4.1: Simplified FPGA architecture with a 4-channel routing network.

The biggest portion of a FPGA is the interconnect which is comprised of the switch boxes, connection blocks and routing channels. They are configured by programmable memory elements, which are typically SRAMs (e.g. Xilinx, Altera, Lattice), but also continuous floating gate devices have been introduced (e.g. Actel). Figure 4.2 illustrates the typical power distribution among traditional SRAM-based FPGAs [4.6, 4.7]. In general 60-70% of dynamic and static power is dissipated in the interconnect; while it occupies a similar fraction of die area [4.8, 4.9]. Therefore, the first target to reduce power consumption in a FPGA must be the switch box and the connection block. A great deal of research has been done to find solutions to saving power. The most common ones at the circuit level can be summarized as: dual- V_T (high threshold voltage for reduced subthreshold leakage in non-critical paths or SRAMs) [4.10, 4.11], dual- t_{ox} (thicker oxide for reduced gate current leakage in non-critical paths or SRAMs) [4.10], dual- V_{DD} (low- V_{DD} for low power and high- V_{DD} for high speed) [4.11], sleep-transistors (switch off power supply in circuit blocks that are unused) [4.12], routing channel segmentation (tracks with different lengths to bypass switch boxes) [4.13] or low swing signaling [4.14]. However, these solutions typically trade off complexity, area or speed for reduced power consumption.

Figure 4.2: Power consumption analysis of traditional SRAM-based FPGAs [4.6, 4.7].

Figure 4.2 (Continued).

In the recent years modern SRAM-based FPGAs have been introduced that use a single driver at each wire to eliminate power-hungry tri-states and reduce the quantity of SRAMs, multiplexers and buffers [4.15]. The major drawback of this unidirectional wiring architecture is the increased complexity due to the reduced availability of paths compared to the traditional SRAM-based interconnect with bidirectional wiring. In this Chapter both the traditional and the modern SRAM-based FPGAs are compared to the new NCFG-based design.

4.2 Design and Driving Method of the Nanocrystal

Floating Gate Device

The NCFG device is a nonvolatile device whose charge is stored on isolated storage nodes. Therefore, one major advantage over SRAMs is that a configuration on every boot up is not required. A physical model of this device for circuit simulation has been developed in Chapter 3 and is used for simulating the FPGA with the new interconnect design in HSPICE.

In this Chapter, the device is designed with a gate length of 45 nm, a gate width of 120 nm, and a total oxide thickness of 10.6 nm as illustrated in Figure 4.3. The nanocrystals are composed of palladium (Pd) and have a hemispherical shape with a height of 2 nm, a diameter of 3 nm, and a density of $\sim 1 \times 10^{12} \text{ cm}^{-2}$. It is assumed that the spacing in between the nanocrystals is constant throughout the entire floating gate. The tunnel oxide is composed of SiO_2 and has a thickness of 3.6 nm for efficient data retention of ~ 10 years. A high work function metal is preferred for the control gate to prevent electron tunneling from the control gate to the floating gate in the erase mode, while not affecting the program mode [4.16]. This allows the control gate oxide, which is also composed of SiO_2 , to be scaled down to 5 nm. Simulations have shown that negligible charge transfer takes place in the control gate oxide for the voltages used in this Chapter.

Figure 4.3: Cross-section of a NCFG device in 45 nm technology.

There are other emerging memory technologies such as RRAM, MRAM, FeRAM or PCM as described in Chapter 2. They could also be used as the programmable memory elements in FPGAs. However, their fabrication concepts are still relatively new and cost

inefficient. Due to their poor compatibility with CMOS process it would be difficult to integrate them within CMOS logic. Floating gate devices instead are fabricated in a front-end bulk process with a well developed technology set. They show good compatibility with CMOS process and are relatively cost efficient compared to the other nonvolatile technologies. This is beneficial as the objective in this Chapter is to embed NCFG devices within the CMOS logic on the FPGA board and use them not only as memory elements but also as logic circuit elements.

The NCFG device is integrated into the FPGA as shown in Figure 4.4(a). The source and drain are connected to the routing channel of the FPGA. The select line is connected to the bulk, and the word line to the control gate. They both are used for programming and erasing the device via uniform tunneling. This is different than how Actel drives the continuous floating gate devices in their FPGAs as shown in Figure 4.4(b) [4.17]. The floating gate is shared between a sensing transistor and a switching transistor. The sensing transistor is used for programming the floating gate by means of two select lines mainly via drain side tunneling. The switching transistor connects or separates the routing channel and is also used to erase the floating gate. Figure 4.4(c) illustrates how the SRAMs are integrated into the FPGA. It is the typical 6T-SRAM structure where the sixth transistor is used as the switching transistor. Compared to the floating gate devices it is the fastest switching memory element with the lowest programming voltage. However, it does not retain its state when the power supply is off and needs to be configured on every boot up. Furthermore, it consumes a large portion of area (and thus power) compared to its counterparts.

Figure 4.4: Driving method via word line (WL) and select line (SL) for (a) the new design with a NCFG device, (b) the design Actel uses with continuous floating gate devices, and (c) the 6T-SRAM cell.

The new design differs from that of Actel because the latter requires two transistors based on 130 nm Flash-based CMOS technology [4.17] instead of one with a gate length of 45 nm. In addition, three global metal wires are used in order to drive the continuous floating gate devices, whereas the new design requires only two. Therefore, the new design is more area efficient which enables a dense integration of the NCFG devices on the FPGA board. Furthermore, on the device level the NCFG device offers many advantages over its counterpart while retaining long retention time such as [4.18]:

- Higher reliability
- Improved scalability
- Less program/erase time
- Lower program/erase voltage and power

A comparison between the three designs of Figure 4.4 in terms of area, program speed, program voltage, and non-volatility is provided in Table 4.1.

Table 4.1: Comparison between the driving methods of the NCFG device, the continuous floating gate device and the 6T-SRAM cell.

Design	Area	Program Speed	Program Voltage	Non-Volatility
NCFG Device	1 Transistors	1-100 μ s	9-11 V	Yes
Cont. FG Device	2 Transistors	1-10 ms	12-15 V	Yes
6T-SRAM Cell	6 Transistors	few ns	V_{DD}	No

Similar to conventional FPGAs [4.19], the new NCFG-based FPGA requires a controller in order to configure the NCFG devices as shown in Figure 4.5. A bitstream is generated by a FPGA CAD flow (e.g. Xilinx's ISE for the Virtex5 board) and supplied to the controller. It contains the information for the controller to serially load the select line shift register for the desired circuit functionality. Once the contents in the select shift register are valid, the controller asserts one word line; thus configuring the NCFG devices that are driven by this word line. After this is accomplished, the select line shift register is loaded with the corresponding contents for the next word line assertion. This algorithm repeats until each word line has been asserted one time such that all NCFG devices on the FPGA board are configured. The entire configuration process for the NCFG devices consumes more time and power compared to the SRAMs which require a lower word line voltage and a configuration time in the nanoseconds range for each cell. However, the new architecture may still be

beneficial considering that NCFG devices are nonvolatile memory elements. Thus, as a FPGA needs to be reconfigured only a few times during its lifetime, the configuration process in the nonvolatile architecture only needs to be executed a few times as well. Therefore, overall the new architecture may be less time and power consuming compared to the SRAM-based FPGA that needs to be configured at every boot up.

Figure 4.5: FPGA configuration.

During the operation mode a word line voltage of 1.4 V is applied to all (erased and previously programmed) NCFG devices. This is necessary to attain the full benefit from the NCFG devices for the new circuit topologies and to guarantee a full swing from G_{ND} to V_{DD} on the output node of an erased device. As shown in the new circuit topologies in Section 4.3 this full swing voltage is the input to a buffer, such that its NMOS and PMOS transistors

cannot be turned on simultaneously. It has been simulated that for a gate width of 120 nm the on-current (I_{on}) of an erased device is $\sim 100 \mu A$. As this operational voltage is applied to the programmed devices as well, they require a threshold voltage shift ΔV_T of at least 1.4 V in order to stay in the cut-off region. The drain current of a programmed device that has a ΔV_T of 1.4 V and a word line voltage of 1.4 V is therefore considered the off-current (I_{off}) and has a value of $\sim 3 \text{ nA}$, resulting in an I_{on}/I_{off} ratio of 3.33×10^5 , which meets the ITRS criterion [4.20]. Note that I_{off} can be reduced by programming the device with a higher and/or longer voltage, resulting in a higher ΔV_T and thus saving static power due to the reduced subthreshold leakage current of the programmed device.

To program and erase the device different voltage envelopes (voltage amplitude and duration) are possible that the designer can choose from. Figure 4.6(a) shows the curves for programming the device for word line voltages ranging from 5 V to 8 V while the select line is kept at -3 V. The threshold voltage shift ΔV_T of 1.4 V is indicated as well. In Figure 4.6(b) the erasing characteristics of the device are demonstrated for word line voltages ranging from -4 V to -7 V while the select line is kept at 5 V. Note that a word line voltage of -7 V is required to completely erase the device in less than 10 ms.

Of course, a controller that handles high voltages and that also supplies a lower operational voltage requires a novel design with an area overhead compared to a traditional controller for SRAM-based FPGAs. However, this area (and thus power) penalty may be more than compensated by the new interconnect design that achieves small area, low power, high speed, and high functionality simultaneously.

Figure 4.6: Voltage envelopes required for the designed device for (a) the program mode and (b) the erase mode.

For example, to program the device such that its ΔV_T is on the order of 1.4 V a voltage pulse of 7 V on the word line and -3 V on the select line for $\sim 10 \mu\text{s}$ is required. Alternatively, 8 V on the word line and -3 V on the select line for $\sim 1 \mu\text{s}$ can be chosen. If the FPGA needs to be reconfigured, the programmed devices can be completely erased by applying -7 V on the word line and 5 V on the select line for a time period of ~ 10 ms. To avoid any change in the threshold voltage shift of a device during the program and erase mode the voltage on the select line must be 0 V such that the voltage across the gate stack is too low to exploit any noteworthy tunneling. Table 4.2 summarizes the appropriate voltages supplied by the controller to enable the different modes of the NCFG devices in this Chapter.

Table 4.2: Configuration of the NCFG devices through the controller.

Mode	Word Line	Select Line
Operation Mode	1.4 V	0 V
Program Mode (10 μ s)		
- charge floating gate	7 V	-3 V
- no change in the floating gate	7 V	0 V
Erase Mode (10 ms)		
- discharge floating gate	-7 V	5 V
- no change in the floating gate	-7 V	0 V

The NCFG device has now been designed appropriately for a static application such as the FGPA with long charge retention of ~ 10 years. Its configuration and operation method has been established in a higher level circuit architecture using a controller and shift registers. In the following Section 4.3 the new circuit topologies for the interconnect are demonstrated by using these novel devices.

4.3 Interconnect Design

4.3.1 Switch Box

In the interconnect of a FPGA the signal propagates from the output of a logic block to the input of a target logic block via routing channels that span horizontally and vertically between the logic blocks. Switch boxes are used to interface these routing channels at their intersections which enable signal propagation in every direction. They are a critical

component in the interconnect where high speed and low power consumption are strongly desired. SRAMs (SR) are used to configure the tri-states and the multiplexers in the switch box to pass through or block the incoming signal from the other routing channels or the connection block (ConB). Figure 4.7(a) represents the circuit design for the switch box that is used in traditional FPGAs with bidirectional wiring [4.15]. It consists of twelve SRAMs, four multiplexers, four tri-state buffers, and four buffers; thus it occupies a large fraction of die area and power. Each multiplexer uses pass-transistors for fast operation which results in a slight signal attenuation. Therefore, the tri-state buffers are required to restore full swing of the signal as it propagates to the next circuit component. The speed of the propagating signal from one switch box to the adjacent one is determined by the delay of one buffer, one multiplexer, one tri-state buffer and the capacitance of the routing channel. In contrast, a modern FPGA exploits unidirectional switch boxes that comprise two routing channels as shown in Figure 4.7(b) [4.15]. Therefore, the adjacent routing channels alternate the propagation direction such that the signal flows down/left on even-numbered routing channels and up/right on odd-numbered ones. This design consists of eight SRAMs, four multiplexers and eight buffers. Thus, the overall logic area is reduced, which results in a more energy efficient switch box compared to its bidirectional counterpart. The signal propagates through one multiplexer and two buffers to the next switch box.

Figure 4.7: Switch box design for (a) a traditional bidirectional [4.15] and (b) a modern unidirectional SRAM-based FPGA [4.15].

A schematic of the new design is shown in Figure 4.8. The switch box itself is comprised solely of six NCFG devices that function as logic circuit elements. In order to

provide a path the NCFG device needs to be erased. If it is programmed, the path is blocked and the signal cannot propagate through. Due to its small dimension it is not capable of driving the signal through longer paths with a large capacity. Therefore, routing switches are inserted in between the switch boxes. In order to create a bidirectional routing switch a NCFG device is connected in series with a buffer. This structure is then duplicated, mirrored, and placed in parallel with the original structure. Routing switches are necessary to maintain full swing and full signal strength throughout the entire interconnect.

Figure 4.8: NCFG-based switch box design with bidirectional routing switches attached at each output node.

It may occur that an entire switch box is to be disconnected from the routing network. In this case it is beneficial to have all its NCFG devices erased - including those in the

adjacent routing switches. This allows that all the corresponding nodes are pulled down to G_{ND} . Otherwise floating nodes will exist in the circuit, which may lead to a high crowbar current in the buffers of the routing switches and waste power.

In the new design, the propagation delay of the signal from one switch box to the adjacent one is determined by the delay of two NCFG devices, one buffer and the capacitance on the routing channel. Due to the fewer transistors and the ability of bidirectional routing, the new switch box design offers high efficiency in speed, power and functionality, which cannot be achieved simultaneously with the SRAM-based designs from Figure 4.7(a) and (b). Details of the speed and power consumption are given in Section 4.4

4.3.2 Connection Block

The connection block in a FPGA has two functions: first to pass the signal from the output of the source logic block to the desired routing channel, and second to establish a connection from the routing channel to the input of the target logic block. In the case of the bidirectional and unidirectional SRAM-based interconnects, the logic block output pins are directly connected to the multiplexers in the switch boxes, which can be observed in Figure 4.7(a) and (b). The signal first travels to the input of these multiplexers and is then passed onto the desired routing channel.

In order to transfer the signal from the routing channel to the input of the target logic block the design shown in Figure 4.9 is used for both SRAM-based interconnect designs. Each channel is connected to the input pins of a multiplexer which is configured by two SRAMs in the case of a 4-channel routing system. Buffers are inserted in the paths between

the routing channels and the multiplexer to avoid signal attenuation [4.2]. The output of the multiplexer is then connected to the input of the target logic block.

Figure 4.9: Connection block from the routing channel (RCh) to the input of the logic block for traditional and modern SRAM-based FPGAs in a 4-channel routing system.

In the new design a technique of passing the signal directly from the logic block output to the switch box as in its SRAM-based counterparts is not feasible due to the absence of multiplexers. Therefore, the logic block output is first connected to the routing channels with the use of NCFG devices as shown in Figure 4.10(a). One NCFG device for each routing channel is then either erased to pass the signal through, or programmed to block the signal. Level restorers are not required as the output nodes of the NCFG devices provide a full swing that is maintained on the routing channel due to the routing switches.

The connection block design for the input of the target logic block for the new interconnect is shown in Figure 4.10(b). Buffers for each channel are not necessary because the routing channels are already buffered in the routing switches. Instead, NCFG devices are again used as switches for each routing channel whose output nodes are tied together and connected to a buffer. This buffer then drives the incoming signal to the input of the logic

block. By rule, at any time only one NCFG device should be erased to pass the signal through while the others are all programmed. If none of the channels are to be connected to the logic block, such that their corresponding NCFG devices are all programmed, an additional NCFG device is used to pull the input node of the buffer down to G_{ND} in order to avoid a floating node. As illustrated in Figure 4.10(b) it is not connected to any routing channel but the source terminal is connected to the G_{ND} rail.

Figure 4.10: Connection block for transferring the signal in a 4-channel routing system by use of NCFG devices from (a) the source logic block output to the routing channel (RCh) and (b) the routing channel to the target logic block input.

4.4 Area, Speed and Power Analysis

The main advantages of the new design are the savings in area, and thus power, and the improvement in speed while maintaining high functionality. When equal buffers are used for each design, the total gate area in the interconnect is reduced by 87% compared to the bidirectional and 63% compared to the unidirectional SRAM-based interconnect. This is mainly due to the absence of the large multiplexers and 6T-SRAM cells. Compared to a SRAM, the NCFG device has the advantage of non-volatility, smaller area and two orders of magnitude lower power consumption in the operation mode, as it has been simulated. The latter arises mainly due to the permanent crowbar current from V_{DD} to G_{ND} that exists in the SRAM cell and the reduction in transistor count and thus leakage current. This is crucial considering that the SRAM block is one of the largest contributor to the static power consumption in Xilinx' FPGAs with 40/45 nm technology [4.21]. A speed analysis has shown that the modern unidirectional SRAM-based design demonstrates the highest speed if the signal passes through only a very few switch boxes. This is a result of the logic block output being directly connected to the switch box; while a connection block from the logic block output to the routing channels as in the new design does not exist. However, for cascaded switch boxes the new design demonstrates the highest speed. It has been simulated that a signal requires 590 ps to pass through 10 switch boxes in an unclocked routing network, whereas in the bidirectional and unidirectional SRAM-based designs the signal requires 825 ps and 638 ps, respectively.

To measure the power consumption a 32-tap FIR Filter is designed in Cadence Virtuoso and simulated in HSPICE for the bidirectional and the unidirectional SRAM-based FPGAs, as well as the NCFG-based one. They use the switch box and connection block designs described in Section 4.3. The logic block topology is illustrated in Figure 4.11. It utilizes SRAM-based LUTs and is, along with the clock distribution, the same for each FPGA architecture.

Figure 4.11: The logic block used in each FPGA architecture consists of an 8-input LUT, a multiplexer (Mux), and a Flip Flop.

Any difference in the result is therefore explicitly due to the modifications in the switch box and connection block. The entire FPGA contains 96 logic blocks, 72 switch boxes and 96 connection blocks that are connected through a routing system with 8 routing channels. The 32-tap FIR Filter design is based on predictive BSIM4.0 CMOS with 45 nm gate length technology; the supply voltage is set to 1 V and the clock frequency to 500 MHz. A summary of the power consumption results is presented in Table 4.3. The power

consumption analysis is based on the operation mode, such that the corresponding programming elements in the FPGA architectures are already configured. The wire capacitance is not taken into account for the power consumption analysis in HSPICE simulations. In fact, it can be assumed that the wire capacitance is less in the NCFG-based FPGA compared to its counterparts due to the significantly reduced gate area in the interconnect. This would result in further power savings of the new design which is not included in the following power consumption analysis. The SRAM-based designs are simulated using high- V_T CMOS in the SRAMs while the rest of the circuit uses low- V_T CMOS for higher speed. High- V_T instead of low- V_T CMOS in the SRAMs greatly reduces the static and dynamic power consumption by approximately 445 μ W. Both SRAM-based FPGAs are still inferior compared to the NCFG-based design, which uses the developed physical model for the NCFG devices from Chapter 3 with the parameters given in Section 4.2 and also low- V_T CMOS for the rest of the circuit. Compared to the traditional bidirectional FPGA, the static and dynamic power consumption is reduced by 58% and 34%, respectively. In contrast to the modern unidirectional FPGA, the static power consumption is slightly reduced by 8% while the dynamic power consumption is very similar. These improvements in power consumption, especially compared to the traditional FPGA, are mainly due to the smaller area and thus fewer transistors in the design, and due to the absence of the crowbar current from V_{DD} to G_{ND} in the SRAMs. The drawback by choosing the new design is that the configuration of the NCFG devices is relatively long and requires a relatively high amount of power. However, these nonvolatile devices retain their state even if the power is off unlike the SRAMs that need to be configured on every boot up.

Furthermore, FPGAs are in general only reconfigured a few times during their lifetime such that this drawback may only be marginal.

Table 4.3: Power consumption results for the 32-tap FIR Filter implemented as a FPGA.

Power Consumption	Traditional SRAM-based FPGA (bidirectional)	Modern SRAM-based FPGA (unidirectional)	New NCFG-based FPGA (bidirectional)
Static	374 μ W	171 μ W	157 μ W
Dynamic	1246 μ W	832 μ W	826 μ W

4.5 Summary

In this Chapter, new circuit topologies for the switch box and connection block are described for use in an energy efficient and high speed FPGA with bidirectional wiring. Power-hungry and large SRAMs are replaced by NCFG devices. These nonvolatile devices retain their state while the system power is off and do not need to be configured at every boot up. The NCFG-based FPGA is benchmarked against both a traditional bidirectional and a modern unidirectional SRAM-based FPGA using a 32-tap FIR Filter designed in Cadence Virtuoso and simulated in HSPICE using predictive BSIM4.0 CMOS with 45 nm gate length technology. Compared to a traditional bidirectional and the modern unidirectional SRAM-based interconnects the total gate area is reduced by 87% and 63%, respectively. The new design also achieves better performance and reduced power consumption compared to the traditional bidirectional SRAM-based FPGA. The signal propagation delay through an

unlocked routing system with 10 switch boxes is 235 ps shorter, resulting in a decrease by 28%. Furthermore, simulations demonstrate a reduction of 58% in static and 34% in dynamic power. Compared to a modern unidirectional SRAM-based FPGA, the power consumption and propagation delay are slightly reduced. Here, the main benefit in the new design is that functionality and complexity are greatly improved as a result of doubling the available routing channels by exploiting bidirectional wiring. Therefore, the potential benefits from choosing the NCFG-based design can be summarized as small area, low power consumption, high speed and high functionality, which typically trade off and cannot be achieved simultaneously by the SRAM-based counterparts. Compared to previous designs that use continuous floating gate devices in FPGAs, the approach described in this Chapter requires less overhead, lower voltages, and offers improved reliability.

References

- [4.1] R. Shen, S.X.-D. Tan, and J. Xiong, "A Linear Statistical Analysis for Full-Chip Leakage Power with Spatial Correlation," *Proc. of the 20th Symp. on VLSI*, pp. 227-232, 2010.
- [4.2] Y. Lin, F. Li, and L. He, "Power Modeling and Architecture Evaluation for FPGA with Novel Circuits for Vdd Programmability," *Proc. of the ACM/SIGDA 13th Int. Symposium on Field-Programmable Gate Arrays*, pp. 199-207, 2005.
- [4.3] S.M. Qasim, A.A. Telba, and A.Y. AlMazroo, "FPGA Design and Implementation of Matrix Multiplier Architectures for Image and Signal Processing Applications," *Int. J. of Comp. Sci. and Netw. Secur.*, vol. 10, pp. 168-176, 2010.
- [4.4] T. Tuan and B. Lai, "Leakage Power Analysis of a 90nm FPGA," *In: Proc. of the IEEE Custom Integrated Circuits Conference*, pp. 57-60, 2003.

- [4.5] J.H. Anderson and F.N. Najm, "Active leakage power optimization for FPGAs," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 423-437, 2006.
- [4.6] A. Lodi, L. Ciccarelli, and R. Giansante, "Combining low-leakage techniques for FPGA routing design," *In: Proc. of the ACM/SIGDA 13th Intl. Symp. on FPGAs*, pp. 208-214, 2005.
- [4.7] V. George and J.M. Rabaey, *Low-energy FPGAs: architecture and design*. Boston, MA: Kluwer, 2001.
- [4.8] A. Rahman and V. Polavarapuv, "Evaluation of Low-Leakage Design Techniques for Field Programmable Gate Arrays," *In: Proc. of the ACM/SIGDA 12th Int. Symp. on FPGAs*, pp. 23-30, 2004.
- [4.9] J. Lamoureux and W. Luk, "An Overview of Low-Power Techniques for Field-Programmable Gate Arrays," *NASA/ESA Conf. on Adaptive Hardware and Systems*, pp. 338-345, 2008.
- [4.10] B. Amelifard, F. Fallah, and M. Pedram, "Reducing the Sub-threshold and Gate-tunneling Leakage of SRAM Cells using Dual-Vt and Dual-Tox Assignment," *Proc. of the Conf. on Design, Automation and Test in Europe*, pp. 995-1000, 2006.
- [4.11] F. Li, Y. Lin, L. He, and J. Cong, "Low-Power FPGA Using Pre-Defined Dual-Vdd/Dual-Vt Fabrics," *Proc. of the ACM/SIGDA 12th Int. Symp. on FPGAs*, pp. 42-50, 2004.
- [4.12] P.-H. Yuh, C.-L. Yang, C.-F. Li, and C.-H. Lin, "Leakage-Aware Task Scheduling for Partially Dynamically Reconfigurable FPGAs," *ACM Trans. on Des. Autom. of Electron. Syst.*, vol. 14, art. no. 52, 2009.
- [4.13] K. Siozios, D. Soudris, and A. Thanailakis, "Designing a General-Purpose Interconnection Architecture for Field Programmable Gate Arrays," *J. of Low Power Electron.*, vol. 4, pp. 34-47, 2008.
- [4.14] M. Meijer, R. Krishnan, and M. Bennebroek, "Energy-Efficient FPGA Interconnect Design," *Proc. of the Conf. on Design, Automation and Test in Europe*, pp. 42-47, 2006.
- [4.15] G. Lemieux, E. Lee, M. Tom, and A. Yu, "Directional and Single-Driver Wires in FPGA Interconnect," *IEEE Int. Conf. on Field-Programmable Technology*, pp. 41-48, 2004.

- [4.16] S. Jeon, J.H. Han, J.H. Lee, S. Choi, H. Hwang, and C. Kim, "High Work-Function Metal Gate and High-k Dielectrics for Charge Trap Flash Memory Device Applications," *IEEE Transaction on Electron Devices*, vol. 52, pp. 2654-2659, 2005.
- [4.17] *Actel IGLOO Handbook*, November 2009.
- [4.18] J. De Blauwe, "Nanocrystal Nonvolatile Memory Devices," *IEEE Transactions on Nanotechnology*, vol. 1, pp. 72-77, 2002.
- [4.19] I. Kuon, A. Egier, and J. Rose, "Design, Layout and Verification of an FPGA Using Automated Tools," *In: Proc. of the ACM/SIGDA 13th Intl. Symp. on FPGAs*, pp. 215-226, 2005.
- [4.20] *International Techonology Roadmap for Semiconductors (ITRS)*.
- [4.21] M. Klein, "Power Consumption at 40 and 45nm," *Xilinx White Paper: Spartan-6 and Virtex-6 Devices*, 2009.

Chapter 5

Unified Memory Device

5.1 Introduction

Memory technologies that aim for high switching speed, long data retention, high density, high endurance and low power consumption are currently being investigated and some of them are matured in mass production such as the FeRAM, MRAM, or PCM as described in Chapter 2. However, they are built in a back-end process and suffer from poor compatibility with CMOS and high cost per bit. In addition, they have not reached the maturity level yet to be manufactured in high volume. Resistive memories (RRAM) are two terminal devices that have potential for high density because they can be built vertically within a crossbar. The crossbar elements have potential for 3D stacking. To build a memory today, each RRAM device has to be integrated with a transistor. This limits the RRAM to a single layer structure. Also while ‘unified’ operation is discussed, there has been little demonstrated to date. Furthermore, a silicon on insulator (SOI) nanocrystal based memory is reported in [5.1] that has the features of a unified memory, i.e. a single device with merged dynamic and nonvolatile memory operation. A single-layer NCFG employs a storage node for nonvolatility and a floating body capacitor region in the SOI employs a storage node for the dynamic memory. However, it requires high voltages and needs to be refreshed approximately every 5 ms in the dynamic memory operation. A bulk FinFET-based unified

memory is reported in [5.2] that uses oxide/nitride/oxide layers for the nonvolatile memory operation and a floating body for the dynamic memory operation. The floating body is formed by a buried n-well layer in the bulk substrate that allows hole accumulation. This device shows an improved data retention in the dynamic memory mode of 8-30 ms.

In this Chapter the invention of a scalable dual metal floating gate device technology is described that can be used as the device basis for a unified memory. The required voltages to operate this device in either memory mode are potentially lower and the data retention, especially in the dynamic memory mode, is potentially longer than its floating gate alternatives described above. Furthermore, it permits both dynamic and nonvolatile memory operation simultaneously, thus it can operate as a 2-bit device. It can be built in a front-end bulk process with a technology set that is mature and well understood. It is band engineered to operate at modest voltages compared with most floating gate devices. With band gap engineered gate stacks, multiple metal floating gates, selective use of thinner oxides, and using tunneling as the main charge transfer mechanism, lower operating voltages are possible, and in principle, stresses reduced and endurance increased.

The state of the device is determined by the amount and mix of charges stored on the two floating gates. Data stored dynamically can be made nonvolatile in a single cycle operation by adding charge from the channel to the floating gates. It can then continue operating as a dynamic memory while storing a different bit pattern in the nonvolatile mode. A back gate may be used to create a select line but this function could be provided through the substrate node (at some voltage penalty). Since it mainly uses direct tunneling in the dynamic memory mode, and not Fowler-Nordheim tunneling or hot carrier injection, it can

be expected that it can achieve high endurance. Nanocrystal metals can be used in the floating gates to further increase endurance at the expense of reduced charge storage.

The device is scalable to at least the 16 nm node and could potentially be packed as an $8F^2$ cell. For ultra-high density, the unified memory device can be monolithically 3D stacked using deposited layers of Indium Gallium Zinc Oxide (IGZO) amorphous semiconductors. Recently, IGZO semiconductors have been used to make high performance thin film transistors [5.3] and have demonstrated high uniformity and quality over large areas [5.4]. IGZO based transistors can achieve better performance than amorphous silicon transistors, and have sufficient performance to enable a nonvolatile memory device [5.5, 5.6]. Fundamentally, there is no limit on 3D stacking beside wafer cost. A stack of four devices has potential to achieve densities equivalent to the 8 nm node. Then, since each cell actually stores two bits, the array density can be as high as 0.4 Tbit/cm^2 . Of course, the area required for peripheral circuits would most likely lead to actual densities half of this.

5.2 Device Structure and Operation

By using a dual floating gate structure, the device may operate as a dynamic memory, a nonvolatile memory, and as both a dynamic and nonvolatile memory concurrently. The device may be switched rapidly between these modes of operation. The state of the device modifies the threshold voltage of the device, which may be detected through an appropriate external circuit. Details of a candidate device designed for the 65 nm node in Sentaurus TCAD is shown in Figure 5.1.

Figure 5.1: (a) Device structure. (b) Band diagram across the gate stack under flat-band condition. (c) Oxide material and thickness. (d) Metal electrode material and thickness.

A cross-section of the device and its band diagram are shown in Figure 5.1(a) and (b), respectively. Two floating gates (FG_{TOP} and FG_{BOT}) are separated from the control gate by a thick top oxide (or control gate oxide) tox₃, separated from each other by a thin middle oxide (or inter floating gate oxide) tox₂ and separated from the transistor channel by a bottom oxide (or tunnel oxide) tox₁ of intermediate thickness. The selected materials for the oxides and metal electrodes are listed in Figure 5.1(c) and (d), respectively, along with their key parameters and thicknesses.

The basics of the device operation are as follows. Two floating gates may each be programmed via a tunneling current. Free charges may be transferred between the two floating gates to program and/or refresh the device by applying a relatively low control gate voltage. This charge redistribution occurs rapidly via the thin inter floating gate oxide tox_2 , resulting in one positively and one negatively charged floating gate and is considered the dynamic storage element. This also implies that the device can only store the dynamic charge very briefly when the system power is switched off, making it a volatile memory mode. Next, the two floating gates may also be programmed more slowly through a relatively thick tunnel oxide tox_1 via tunneling of electrons from the channel by applying a relatively high programming voltage on the control gate, similar to traditional floating gate devices, and is considered the nonvolatile memory operation. Thus, the two memory operations are clearly distinguished by the applied voltage envelope and the device may be switched rapidly between the two modes of operation. The device also shows multi-bit capability when exploiting both memory modes simultaneously. The state of the device is determined by its threshold voltage, which may be detected through an appropriate external circuit. Note that in any operation a charge transfer between the floating gate pair and the control gate is not desired such that a physically thick control gate oxide tox_3 is chosen for this design. The device has numerous possible configurations:

- A planar device as shown in Figure 5.1 with continuous floating gates.
- A SOI device that permits better scaling, and the introduction of a back gate that enables selectivity of the device and may further improve scalability.

- A planar device with NCFGs. Nanocrystal gates may increase reliability at the expense of less charge transfer especially between the two floating gates, thus changing programming voltage and time in the dynamic memory mode.
- A vertical device with the gate stack on each of one or both side walls.
- A circular device with a surround gate stack.
- Two devices connected serially – one device optimized for write, and then biased into a low leakage state, and another device optimized for read.

5.2.1 Dynamic Memory Operation

For dynamic programming, relative low voltages applied on the control gate enable a transfer of free charges between the floating gate pair to program and/or refresh the dual floating gate device. This charge redistribution occurs rapidly via a thin high-k oxide, such as HfSiO, resulting in one positively and one negatively charged floating gate, thus changing the threshold voltage of the device. It is considered the dynamic memory domain as the subsequent charge leakage is relatively fast and a refresh is required. This implies that it is also a volatile memory mode. The device utilizes direct tunneling under low electric fields ($< 0.2 \text{ V/nm}$) in this mode where endurance and reliability factor into design criteria. The electrons are not exposed to scattering events with the lattice atoms in the oxide [5.7, 5.8]. This produces less damage to the oxide than the alternatives of Fowler-Nordheim tunneling and hot carrier injection [5.7-5.9], especially in low-field stress.

The amount of current is determined by the collective heights and widths of the barriers in the band diagram under bias conditions. Band engineering permits a decrease in

the actual height of the barrier for fast programming and an increase in the actual thickness of the barrier for maximizing retention time. In order to optimize these trade-offs, the choice of the materials for the floating gates as well as for the dielectrics is important. High-k insulators, such as HfO_2 , Al_2O_3 , ZrO_2 , HfAlO , or HfSiO have a lower band gap and a higher electron affinity than low-k insulators, such as SiO_2 , resulting in a lower barrier especially in the conduction band. In addition, the effective electron mass in these high-k insulators is lower than in its low-k counterpart. Therefore, the tunneling rate of the charges, especially the electrons, is enhanced for fast programming. The trade-off is that the charge retention time is shortened because the charge leakage through the same insulator occurs rapidly. Regardless of this charge leaking, high-k insulators permit physically thicker oxides for the same capacitance and improve the retention-to-program time ratio [5.10-5.14].

Choosing appropriate materials for the floating gates also assists with optimization of the device, in particular for the dynamic memory operation. A high work function metal creates a deep potential well for the electrons and the charge leakage may be reduced. On the other hand it is also more difficult for the electrons to tunnel through the high barrier in the program mode. As such, the programming voltage may be larger to maintain a short programming time. An example of such a programming voltage includes 4-5 V for the low work function metal opposed to 7-8 V for the high work function metal. For the dynamic memory operation, a low work function metal, such as TaN, TiN, Al, Ta or Mg, may be chosen for the bottom floating gate FG_{BOT} to enable rapid charge transfer to the top floating gate FG_{TOP} . The idea is to enable this very quick charge transfer from FG_{BOT} to FG_{TOP} through the low barrier for fast dynamic program operation (50 ns for the designed device

from Figure 5.1). FG_{TOP} may then be comprised of a high work function metal, such as Pt, Au, or Pd, to reduce the electron back-tunneling rate and improve the charge retention time; thus enabling a longer refresh period. For the designed device from Figure 5.1 it takes ~ 1 s until all the charges transfer back to FG_{BOT} such that the device is back in its uncharged state.

Figure 5.2(a) and (b) show the band diagram across the Mo-HfO₂-Pt-HfSiO-Mg-SiO₂-Si gate stack when the device is dynamically programmed with 4 V applied on the control gate and erased with -4 V applied on the control gate, respectively. These diagrams are directly taken from the device simulator.

Figure 5.2: Band diagram of the device from Figure 5.1 for the dynamic memory mode for a control gate voltage of (a) 4 V and (b) -4 V.

It can be observed that the electrons face a higher barrier at the interface between FG_{TOP} and tox_2 in the erase mode compared to the barrier at the interface between FG_{BOT} and tox_2 in the program mode. Thus, it is more difficult for those electrons that are stored in FG_{TOP} to tunnel to FG_{BOT} . This is beneficial in one sense because the charge retention after a

dynamic program is increased. However, this also indicates that a dynamic erase is slower than a dynamic program. Note that the voltage drop across tox_2 in either case is relatively low ($< 0.2 \text{ V/nm}$) so that the electrons tunnel mainly due to uniform direct tunneling and oxide damage is greatly suppressed due to the low-field stress.

Figure 5.3(a) and (b) show simulation results to indicate the threshold voltage shift ΔV_T and flat-band voltage shift ΔV_{FB} , respectively. The device is first in an uncharged state and consecutively cycled through the dynamic program, retention, and refresh mode. In Figure 5.4(a)-(c) the charge redistribution of the two floating gates is demonstrated for each cycle to help explain the reasons for the voltage shifts.

Figure 5.3: (a) Threshold voltage shift and (b) flat-band voltage shift when the device is in the nonvolatile uncharged state and consecutively cycled through the dynamic program, retention, and refresh mode.

Figure 5.3 (Continued).

The states of the device in the dynamic mode for each cycle are as follows:

- **Step 1: Initial uncharged device.** The device has zero net charge and there is no charge redistribution between the floating gate pair.
- **Step 2: Dynamic program.** A positive applied control gate voltage enables charge transfer from FG_{BOT} to FG_{TOP} , leaving behind a positively charged FG_{BOT} and a negatively charged FG_{TOP} , which can be observed in Figure 5.4(a). For a programming time of 50 ns under a control gate bias of 5 V the charges on the floating gate pair have a value of $\pm 6.73 \times 10^{-16}$ C/ μm , which results in a negative threshold/flat-band voltage shift of approximately -330 mV. Thus, the shift is *negative*, which first seems misleading considering that in traditional floating gate devices a positive control gate voltage results in a *positive* shift. The reason for this unusual behavior is that in the dynamic program mode there is only a redistribution of

the free charges between the two floating gate layers. There is no charge transfer from the channel so that the net charge of the device is the same. However, as the electrons from FG_{BOT} tunnel to FG_{TOP} , a positively charged FG_{BOT} is obtained due to the electron loss, and a negatively charged FG_{TOP} is obtained due to the electron overshoot. As FG_{BOT} is in closer proximity to the channel than FG_{TOP} , the I-V characteristics of the channel are more controlled by FG_{BOT} . Therefore, the positively charged FG_{BOT} with a higher internal voltage relative to the FG_{BOT} of an uncharged device results in a negative shift. It can also be observed in Figure 5.4(a) that the two floating gates can be further charged by biasing the control gate for a longer time. After $\sim 10 \mu s$ the charge redistribution is saturating as then the voltages on the floating gate pair are in balance.

- **Step 3: Dynamic retention.** When the device is left in retention state, the charges on the two floating gates move back to zero with respect to time as shown in Figure 5.4(b). The shift becomes gradually smaller and after 300 ms the charges on the floating gate pair are reduced to $\pm 1.98 \times 10^{-16} \text{ C}/\mu\text{m}$. At this point, the negative shift has decreased from -330 mV to approximately -110 mV. It can also be observed in Figure 5.4(b) that after $\sim 1 \text{ s}$ the device is back in its initial uncharged state.
- **Step 4: Dynamic refresh.** In order to keep the device in the dynamically programmed mode the device needs to be refreshed by applying another voltage pulse on the control gate. Figure 5.4(c) illustrates the charge redistribution in the refresh mode after the device has been left in the retention mode for 300 ms. As the floating gate pair is still charged with $\pm 1.98 \times 10^{-16} \text{ C}/\mu\text{m}$, another voltage pulse of 50 ns is not

required to bring the charges back to (or close to) $\pm 6.73 \times 10^{-16} \text{ C}/\mu\text{m}$. A control gate voltage of 5 V for 40 ns is sufficient to move the shift back to -330 mV, and the curves for step 2 and step 4 are (almost) overlapping.

Figure 5.4: Charge redistribution on the floating gate pair with respect to time when the device is in the nonvolatile uncharged state and consecutively cycled through the dynamic (a) program, (b) retention, and (c) refresh mode.

5.2.2 Nonvolatile Memory Operation

The device may operate as a nonvolatile memory by applying a relatively high and long voltage pulse on the control gate compared to the dynamic memory, which enables charge transfer from the channel as in traditional floating gate devices. The band diagrams across the Mo-HfO₂-Pt-HfSiO-Mg-SiO₂-Si gate stack for programming and erasing the device under a control gate bias of ± 8 V are shown in Figure 5.5(a) and (b), respectively.

Figure 5.5: Band diagrams of the device from Figure 5.1 for the nonvolatile memory operation for a control gate voltage of (a) 8 V and (b) -8 V.

The device utilizes Fowler-Nordheim tunneling to move charge from the channel to the floating gates for the nonvolatile memory operation. However, it may also utilize channel hot carrier injection by applying a relatively high voltage on the drain terminal. In the simulations for programming and erasing the device, the drain and source terminal are always grounded in this Chapter and mainly uniform tunneling is applied. Figure 5.6(a) and (b) demonstrate the threshold voltage shift ΔV_T and flat-band voltage shift ΔV_{FB} of the

device when it is cycled consecutively through the nonvolatile program mode, charge settling after the program mode, erase mode, and charge settling after the erase mode. Figure 5.7(a)-(d) demonstrate the corresponding charging characteristics on each floating gate.

Figure 5.6: (a) Threshold voltage shift and (b) flat-band voltage shift when the device is consecutively cycled through the nonvolatile program mode, charge settling after the program mode, erase mode, and charge settling after the erase mode.

The states of the device in the nonvolatile mode for each cycle are as follows:

- **Step 1: Initial uncharged device.** The device has zero net charge and there is no charge redistribution between the floating gate pair.
- **Step 2: Nonvolatile program.** While programming the device into a nonvolatile charged state with a high positive voltage of 9 V, the electrons from the channel begin to tunnel through the tunnel oxide tox_1 and add negative charges on FG_{BOT} and FG_{TOP} after a time period of ~ 100 ns as it can be observed in Figure 5.7(a). The electrons tunnel first to FG_{BOT} , which then by part tunnel further to FG_{TOP} . The voltage levels on both floating gates then begin to drop with respect to time. In this nonvolatile program mode the threshold and flat-band voltage of the device first just slightly shift to the right relative to the initially uncharged device. This is because immediately after the applied voltage the majority of the electrons from the channel are stored in FG_{TOP} . In addition, free charges from FG_{BOT} have also transferred to FG_{TOP} very similar to the dynamic memory operation, which results in a strongly negatively charged FG_{TOP} . Even though the net charge of the device is now negative, FG_{BOT} is positively charged as shown in Figure 5.7(a), which in combination with a strongly negatively charged FG_{TOP} results in a relatively small positive shift.
- **Step 3: Charge settling after the nonvolatile program.** Immediately after the applied voltage pulse is removed, the electrons stored in FG_{TOP} will partially tunnel back to FG_{BOT} as shown in Figure 5.7(b) until a voltage balance between both layers is reached, which can be described as charge settling. The slightly positive shift becomes larger as FG_{BOT} is being increasingly negatively charged with respect to

time. After ~ 1 s, the electron back-tunneling saturates and the shift remains constant which indicates that at this point a voltage balance between the floating gate pair is reached. However, it can be observed in Figure 5.7(b) that the charge curves for both floating gates do not merge together (even for a longer time) as it is the case in the dynamic memory operation. The reason for this is that in the dynamic memory operation with a nonvolatile uncharged state the net charge in the device is always zero. However, in the nonvolatile charged state of the device the net charge is less than zero due to the added electrons in the floating gate pair from the channel. If no voltages are applied, the device is in the deep accumulation region, which results in a negative surface potential in the range of -0.6 V to -0.7 V. Therefore, in order to obtain a voltage balance on the floating gate pair, FG_{TOP} requires more negative charges than FG_{BOT} , which explains the gap between the charges in Figure 5.7(b) after ~ 1 s. Note that charge leakage to the channel is mainly prevented due to the sufficiently thick tunnel oxide tox_1 .

Step 4: Nonvolatile erase. In order to transfer the electrons that are stored in the floating gate pair back to the channel, a relatively high and long negative voltage needs to be applied on the control gate. Then, the device is in the erase mode and ideally the amount of charges that transferred in the program mode from the channel to the floating gate pair is now being removed in order to obtain zero net charge of the device. In the erase mode with for example $V_{CG} = -8.5$ V for $30 \mu s$, the electrons in FG_{BOT} tunnel back to the channel before the electrons from FG_{TOP} begin to tunnel to FG_{BOT} as illustrated in Figure 5.7(c). This is due to the higher barrier at the

interface between FG_{TOP} and tox_2 compared to the barrier at the interface between FG_{BOT} and tox_1 . Another reason is that the electric field across tox_1 is stronger due to the low-k insulator compared to the electric field across the high-k insulator tox_2 , which accelerates the electron back-tunneling.

- **Step 5: Charge settling after the nonvolatile erase.** Immediately after the erase mode, the device is in fact dynamically programmed because the majority of the stored electrons are still stored in FG_{TOP} and FG_{BOT} is positively charged. This is because in the nonvolatile erase mode first the charges from FG_{BOT} transfer back to the channel before the charges in FG_{TOP} tunnel back to FG_{BOT} . In order to obtain the initially uncharged device again, the device needs to be left in the charge settling mode such that the electrons tunnel back from FG_{TOP} to FG_{BOT} until a voltage balance between the floating gate pair is reached, which takes ~ 1 s as illustrated in Figure 5.7(d). Of course the charge settling time may be shortened by applying a negative voltage on the control gate, but care must be taken not to apply a negative voltage too high that may enable charge transfer from the floating gate pair to the channel as then the net charge of the device becomes positive. The charge settling after the erase mode is then very similar to the dynamic charge leakage (step 3 in Section 5.2.1). Note that in the simulations a high negative voltage of -8.5 V for the erase mode (step 4) was purposely chosen to demonstrate that the device may also obtain a positive net charge. In order to obtain a device with zero net charge as the initially uncharged device (i.e. the curves for step 1 and step 5 are overlapping), a control gate voltage of approximately -8.3 V for 30 μ s is required.

Step 2: $V_{CG}=9V$, $30\mu s$ (after Step 1: Initial State)

(a)

Step 3: $V_{CG}=0V$, $3s$ (after Step 2)

(b)

Step 4: $V_{CG}= -8.5V$, $30\mu s$ (after Step 3)

(c)

Step 5: $V_{CG}= 0V$, $1s$ (after Step 4)

(d)

Figure 5.7: Charge redistribution on the floating gate pair with respect to time when the device is consecutively cycled through (a) the nonvolatile program mode, (b) charge settling after the program mode, (c) erase mode, and (d) charge settling after the erase mode.

5.2.3 Combined Dynamic and Nonvolatile Memory Operation: 2-Bit Device

The device may operate in simultaneous/concurrent states by first programming it with a nonvolatile state and then super-imposing a dynamic state onto the device. In this

mode of operation, there may be considered four combined states for the device. The (dynamic, nonvolatile) pairing of the states may be represented by the array ('00', '01', '10', '11'), where each state produces a different threshold voltage for the device. Thus, the device can operate as a 2-bit device which may be beneficial for certain applications (e.g. check-pointing as described in Chapter 6).

Figure 5.8 (a) and (b) show simulation results for the threshold voltage shift ΔV_T and flat-band voltage shift ΔV_{FB} of the device, respectively. The device is in a nonvolatile charged state and consecutively cycled through the dynamic program, retention, and refresh. The plots are similar to those obtained for the dynamic memory operation when the device is in an uncharged nonvolatile state in Section 5.2.1. However, as expected, all the curves are shifted to the right due to the added negative charges in the floating gate pair. Thus, step 1 in Figure 5.8(a) and (b) can be interpreted as a dynamically uncharged and nonvolatile charged device, putting it in the state '01'. A short and low voltage envelope of 5 V and 50 ns enables a decrease of this positive shift (step 2) by ~ 310 mV, putting the device into the state '11'. Similar to the dynamic memory operation for a nonvolatile uncharged state, the back-tunneling of electrons from FG_{TOP} to FG_{BOT} occurs quickly and after 300 ms the majority of the electrons leaked away (step 3). This leads to an increase of the positive shift, but is still ~ 110 mV less than the shift of the dynamically uncharged and nonvolatile charged device of step 1. Again, the device may be refreshed by applying 5 V for 40 ns on the control gate (step 4), resulting in a shift equal (or almost equal) to the one obtained in step 2.

Figure 5.8: (a) Threshold voltage shift and (b) flat-band voltage shift when the device is in the nonvolatile charged state and consecutively cycled through the dynamic program, retention, and refresh mode.

The charge redistribution on the floating gate pair for each cycle is illustrated in Figure 5.9(a)-(c). Unlike the dynamic program mode for a nonvolatile uncharged device, for a device in a nonvolatile charged state the initial charge levels on the floating gate pair are negative and not zero. This is because the net charge is always less than zero when the device is in a nonvolatile charged state. Other than that, the charge redistribution behavior between the floating gate pair is analogous to the behavior described in Section 5.2.1.

Figure 5.9: Charge redistribution on the floating gate pair with respect to time when the device is in the nonvolatile charged state and consecutively cycled through (a) the dynamic program, (b) retention, and (c) refresh mode.

5.3 Fabricated Test Device

A key enabling technology for fabricating thin-film devices such as the unified memory is the use of atomic layer deposition (ALD). In the most recent years, scaling of floating gate devices has been accompanied by introducing new materials for the insulator and metal for the floating gates and control gates, which have to be deposited with tight control of thickness and composition. In [5.15] a transparent, flexible and high performance nonvolatile memory is proposed that uses ALD techniques for the Pt nanocrystals and the control gate oxide. Furthermore, [5.16] reports the fabrication of an ultra-thin metal floating gate and control gate due to ALD. The device also utilizes ALD for the $\text{HfO}_2/\text{Al}_2\text{O}_3/\text{HfO}_2$ control gate oxide and has demonstrated excellent programming/erasing and retention characteristics. This indicates the feasibility to fabricate the new unified memory device consisting of either ultra thin metals or nanocrystals. Ultra thin metals or nanocrystals are attractive since they minimize the height of the memory stack which in turn can reduce cell-to-cell interference and improve scalability. Physical vapor deposition (PVD) methods may be used to form nanocrystal/thin films of other relevant metals such as palladium (Pd).

Figure 5.10(a) shows a TEM image of a first generation test device. The control gate is composed of Pd with a thickness of 122 nm. The top and bottom floating gate are also composed of Pd with a thickness of 4.6 nm and 5.4 nm, respectively. The control gate oxide consists of HfAlO with a thickness of 27.7 nm whereas the inter floating gate is composed of HfO_2 and relatively thin with 4.5 nm. SiO_2 is used for the tunnel oxide with a thickness of 7.2 nm. C-Vs were measured as the control gate voltage was swept from a negative voltage, to a positive one and then back to the negative voltage. The resulting shift in the flat-band

voltage is shown in Figure 5.10(b). As anticipated, at low applied voltages, the control gate is not able to draw up charge from the channel, but can only redistribute charge between the two floating gates. A positive voltage at the control gate draws electrons to the top floating gate, leaving behind a slight positive charge on the bottom floating gate. This bottom gate has greater control on the substrate due to its closer proximity, resulting in a small negative flat-band voltage shift. This represents the dynamic operation mode of the device. Eventually as the control gate voltage increases, the field is strong enough to draw up charge from the substrate, resulting in a net increase in the electron charge on the floating gates. This negative charge on the floating gates causes a positive voltage shift and is indicative of the more traditional nonvolatile operation mode.

Figure 5.10: Fabrication of a dual floating gate test device (image courtesy of Neil Di Spigna, North Carolina State University). (a) Transmission electron micrograph (TEM) of the cross-section, and (b) flat-band voltage shift demonstrating the dynamic and nonvolatile memory operation of the device.

5.4 Advanced Technology Using a Back Gate and SOI for Selectivity and Improved Scalability

The new unified memory device is scalable to at least the 16 nm node and has been designed in Sentaurus TCAD on SOI with a back gate as illustrated in Figure 5.11. It could potentially be packed as an $8F^2$ cell. The back gate is used to create an input for a select line but it is optional. It could be built using 3D techniques or in a SOI process using a backside aligned step. Alternatively it can be implemented using a well contact (like it is in Flash memory devices). The back gate may be used to provide flexibility in choices of voltages to change the operational state of the device for specific applications. It may further improve scalability of the device by biasing it with a low negative voltage in order to suppress DIBL.

The control gate in this design is now composed of aluminum (Al) along with the back gate and the drain and source contacts in order to obtain a relatively low threshold voltage of an uncharged device. Using high work function metals would result in a high threshold voltage device that may be unsuitable for circuit design with high performance. The material and thickness selection for the insulators and the floating gates are equivalent to the design in Figure 5.1. The SOI is uniformly doped with boron (B) at a concentration of $1.0 \times 10^{15} \text{ cm}^{-3}$ which is typical for SOI transistors. The drain and source are highly doped with arsenic (As) with a concentration of $7.0 \times 10^{19} \text{ cm}^{-3}$ to reduce the contact resistances.

Figure 5.11: Cross-section and doping profile of the new device in 16 nm technology.

The device I-V characteristics, taken from the device simulator, are shown in Figure 5.12. It shows the drain current as a function of the control gate voltage in a linear and logarithmic scale while the drain voltage is constantly 0.7 V (typical supply voltage for the 16 nm technology node) and the source is grounded. In addition, the drain current is shown as a function of the drain voltage for different control gate voltages of 1.5 V, 2 V, and 2.5 V while the source terminal is attached to ground. Note that the back gate was not included in these simulations.

Figure 5.12: I-V simulations of the device from Figure 5.11.

While the device is in an uncharged state, a drain current of $\sim 100 \mu A/\mu m$ can be achieved for a control gate voltage of $\sim 1.9 V$. For an applied bias of $1.2 V$ the drain current is $\sim 30 \mu A/\mu m$. The off-current (I_{off}) of the device ($V_{CG} = 0 V$, $V_{DS} = 0.7 V$) is approximately $140 nA/\mu m$. One major issue of scaling the device is DIBL such that the use of SOI in this technology node is inevitable. The back gate can be used in the read mode to suppress DIBL by applying a negative voltage on it, which may result in further improved scalability.

Regardless of the significant improvement of SOI on the technology scaling, the charging and discharging characteristics of the floating gate pair are weakly dependent on the horizontal scaling of the device assuming that the device employs uniform tunneling mechanisms. It can be observed in Figure 5.13(a) and (b) that the threshold voltage shift ΔV_T in the dynamic program and retention mode of the device is very similar for a 16 nm gate length device compared to the device from Figure 5.1 which is based on 65 nm technology. In Figure 5.13(a) the device is in the dynamic program mode and the voltage across the gate stack for both devices is in the range of 3 V to 5 V (note that the 65 nm device does not include a back gate). In Figure 5.13(b) the device is left in the retention mode where no control gate voltage is applied. Here, the device has been previously dynamically programmed to obtain a threshold voltage shift ΔV_T of -250 mV, -350mV, and -450 mV.

Figure 5.13: Comparison of the (a) charging and (b) discharging behavior between the device designed in 65 nm technology and the device designed in 16 nm.

Figure 5.13 (Continued).

Furthermore, the charge transfer of the scaled device from Figure 5.11 is similar if the voltage across the gate stack is equal or approximately equal. For example, for 5 V on the control gate and 0 V on the back gate the charge transfer is very similar to the case where the control gate is biased with 3 V and the back gate with -2 V. The purpose of the back gate is to selectively charge/discharge the device in both dynamic and nonvolatile memory operation, which is beneficial for certain applications such as the unified main memory described in Chapter 6. For example, if an entire row of these devices are driven with the same control gate voltage of 3 V, then those devices that are to be dynamically programmed require a back gate voltage of -2 V, while the other devices that are to maintain the uncharged state require a back gate voltage of +2 V. This is analogous for the nonvolatile memory operation. To switch an entire row of these devices from a nonvolatile uncharged to

a nonvolatile charged state, examples of choices for voltage envelopes may include 7 V on the control gate and -2 V on the back gate for 30 μ s; while a back gate bias of +2 V leaves selected devices in the row in a nonvolatile uncharged state as the voltage across the gate stack is too low to enable charge tunneling from the channel. To switch a device from a nonvolatile charged state back to the nonvolatile uncharged state, the control gate needs to be biased with -7 V and the back gate with +2 V for 14 μ s. Those devices in the row that are to remain in the nonvolatile charged state require a back gate voltage of -2 V.

5.5 Physical Model

A physical model of the unified memory device is developed and can be used for circuit simulation across various SPICE-based circuit simulators. The concept for this model is equivalent to the concept used for modeling the single-layer NCFG device described in Chapter 3. A Verilog-A module with the control gate voltage V_{CG} as the input captures the programming and erasing characteristics for the dual floating gate device for any memory operation. The output is then the voltage on the bottom floating gate V_{FG_BOT} which is connected to the gate terminal of a predictive NMOS transistor in order to obtain the correct I-V characteristics. The oxide thickness of this transistor is equal to the tunnel oxide thickness tox_1 of the unified memory device. Only analytical equations are used with no correction factors or curve fitting, and the voltages on the two floating gates are updated after every time step Δt with respect to time, providing a fast computing and dynamic model. The difference to the model of the single-layer NCFG device is that now there is an additional

floating gate layer and oxide layer. The capacitor model of the unified memory device is illustrated in Figure 5.14. Note that the capacitors C_1 , C_2 , and C_3 not only differ due to the different oxide thicknesses but also due to the different choices for the oxide materials, e.g. the capacitors C_{cox} and C_{tox} for the NCFG reference model from Chapter 3 differ due to their unequal thicknesses, but they are both composed of SiO_2 . The device structure from Figure 5.1 serves as the reference model for this physical model.

Figure 5.14: Capacitor model between the control gate, the floating gate pair and the substrate surface nodes of the unified memory device.

The tunnel current densities for each oxide in both vertical directions are also included in Figure 5.14. The direction of the tunnel currents depend on the control gate voltage V_{CG} and the charges on the floating gate pair. The tunnel current density into each floating gate is the difference between the inward and outward tunnel current density, and given by

$$J_{FG_BOT} = J_{in,FG_BOT} - J_{out,FG_BOT} \quad (5.1)$$

$$J_{FG_TOP} = J_{in,FG_TOP} - J_{out,FG_TOP} \quad (5.2)$$

where each inward and outward tunnel current density is separated into:

$$J_{in,FG_BOT} = J_{in,sub-FG_BOT} + J_{out,FG_BOT-FG_TOP} \quad (5.3)$$

$$J_{out,FG_BOT} = J_{out,sub-FG_BOT} + J_{in,FG_BOT-FG_TOP} \quad (5.4)$$

$$J_{in,FG_TOP} = J_{in,FG_BOT-FG_TOP} + J_{out,FG_TOP-CG} \quad (5.5)$$

$$J_{out,FG_TOP} = J_{out,FG_BOT-FG_TOP} + J_{in,FG_TOP-CG} \quad (5.6)$$

These current densities are determined by the direct tunneling and Fowler-Nordheim tunneling equations which are given in Chapter 3. The barrier heights ϕ_b at the Si/SiO₂, Mg/HfSiO, and Pt/HfO₂ interfaces that are essential for the program mode are chosen by $\phi_{b,Si/SiO_2} = 3.15$ eV, $\phi_{b,Mg/HfSiO} = 1.25$ eV, and $\phi_{b,Pt/HfO_2} = 2.75$ eV, respectively. In the erase mode, the barrier heights at the Mo/HfO₂, Pt/HfSiO, and Mg/SiO₂ interfaces are selected by $\phi_{b,Mo/HfO_2} = 1.7$ eV, $\phi_{b,Pt/HfSiO} = 3.1$ eV, and $\phi_{b,Mg/SiO_2} = 2.9$ eV, respectively. The effective electron mass in the HfSiO insulator is chosen to be $0.2m_0$ [5.17, 5.18] and in the HfO₂ insulator $0.17m_0$ [5.19], where m_0 is the free electron mass. Note that this model is also completely parameterized in terms of process and geometry; thus it provides configurability (i.e. the model is not fixed to the device with the parameters given in Figure 5.1).

The electric fields, and therefore voltages, across each insulator need to be computed as the direct tunneling and Fowler-Nordheim tunneling equations are dependent on them. They can be derived from the band diagram of the device shown in Figure 5.15(a) and (b) where the device is in the program mode and erase mode, respectively.

Figure 5.15: Band diagram of the unified memory device in (a) the program mode and (b) the erase mode. The thick solid line represents the potential.

The voltage drops across the tunnel oxide V_1 , inter floating gate oxide V_2 , and control gate oxide V_3 are then given by

$$V_1 = |V_{FG_BOT} - V_{surface}| \quad (5.7)$$

$$V_2 = |V_{FG_TOP} - V_{FG_BOT}| \quad (5.8)$$

$$V_3 = |(V_{CG} + \phi_m) - V_{FG_TOP}| \quad (5.9)$$

where ϕ_m represents the contact potential between the control gate and the control gate oxide.

Once the tunnel current densities are known, the new charge on the bottom floating gate Q_{FG_BOT} and top floating gate Q_{FG_TOP} after one time step Δt can be computed. In each case, it is the sum of the charge of the previous time step plus the additional charge during the current time step:

$$Q_{FG_BOT} = Q_{FG_BOT_prev} + J_{FG_BOT} Area_{FG} \Delta t \quad (5.10)$$

$$Q_{FG_TOP} = Q_{FG_TOP_prev} + J_{FG_TOP} Area_{FG} \Delta t \quad (5.11)$$

The final step is to update/refresh the voltages on the floating gate pair. Note that they are interdependent such that the value for V_{FG_BOT} calculated from the previous time step is used in computing V_{FG_TOP} at the current time step:

$$V_{FG_TOP} = \frac{(V_{CG} + \phi_m)C_3 + V_{FG_BOT}C_2}{C_2 + C_3} + \frac{Q_{FG_TOP}}{C_2 + C_3} \quad (5.12)$$

$$V_{FG_BOT} = \frac{V_{FG_TOP}C_2 + V_{Surface}C_1}{C_1 + C_2} + \frac{Q_{FG_BOT}}{C_1 + C_2} \quad (5.13)$$

V_{FG_BOT} is then the output of the Verilog-A module and drives the gate of the predictive transistor model. Table 5.1 demonstrates good matching of the physical model to the simulation results of the unified memory device from Figure 5.1 which is consecutively cycled through the dynamic program, retention and refresh mode. In these simulations the back gate is included. Results for switching the device from the dynamic memory mode to the nonvolatile memory mode, and back to the dynamic memory mode are also included. For the transfer from the dynamic to the nonvolatile mode two steps are considered: first, the threshold voltage shift is measured immediately after the applied voltage envelope of 7 V on the control gate and -2 V on the back gate for 30 μ s. In the second step, the applied voltages are removed ($V_{CG} = V_{BG} = 0$ V) and part of the electrons that are stored in the top floating gate tunnel back to the bottom floating gate, as described earlier. The final threshold voltage shift is then measured after a time period of 1 s.

Table 5.1: Comparison results between the physical model and the reference model from the device simulator of the unified memory device for both the dynamic and nonvolatile (NV) memory mode.

	Dyn. State	NV State	V_{CG}	V_{BG}	t	Device Simulator			Physical Model		
						Q_{FG_BOT}	Q_{FG_TOP}	ΔV_T	Q_{FG_BOT}	Q_{FG_TOP}	ΔV_T
Dynamic Program	1	0	3V	-2V	50ns	6.7e-16C	-6.7e-16C	-0.336V	6.1e-16C	-6.1e-16C	-0.331V
Dynamic Retention	1	0	0V	0V	300ms	2.0e-16C	-2.0e-16C	-0.111V	1.8e-16C	-1.8e-16C	-0.100V
Dynamic Refresh	1	0	3V	-2V	40ns	6.6e-16C	-6.6e-16C	-0.333V	5.8e-16C	-5.8e-16C	-0.324V
Dyn. → NV Transfer	0	1	7V	-2V	30 μ s	2.1e-15C	-3.2e-15C	+0.178V	2.2e-15C	-3.1e-15C	+0.066V
			0V	0V	1s	-3e-16C	-7.7e-16C	+1.526V	-2.2e-16C	-8.5e-16C	+1.468V
NV → Dyn. Transfer	1	0	-7V	2V	13.6 μ s	7.1e-16C	-7.1e-16C	-0.361V	6.4e-16C	-6.5e-16C	-0.352V

5.6 Temperature Dependency

Simulations in the device simulator are performed to measure the temperature dependency of the dual layer floating gate device. In particular, the dependency in the dynamic memory domain is investigated, where the new device may show improved programming and retention characteristics compared to the DRAM. The tunneling mechanism in the dynamic memory domain between the floating gate pair is mainly direct tunneling. Direct tunneling is essentially weakly dependent on variation in temperature [5.20]. Unlike the number of available electrons for tunneling from the channel that are temperature dependent [5.21], the number of electrons in the floating gates metal remains constant, such that an overall weakly temperature dependent device should be expected.

The simulation results for the dynamic program and retention mode under different temperature environments of 25° C, 85° C, and 105° C are shown in Figure 5.16(a) and (b), respectively. It can be observed that with increasing temperature, the charge transfer through the inter floating gate oxide is enhanced in both vertical directions. Thus, there is a trade-off between the dynamic program mode and the dynamic retention mode, and it can be stated that with increasing temperature the dynamic programming characteristics are improved whereas the dynamic retention characteristics are worsened, which results in a shorter refresh period. For example, a threshold voltage shift ΔV_T of approximately -330 mV is obtained when biasing the control gate with $V_{CG} = 5$ V for 50 ns at 25° C (room temperature). If the temperature is raised to 105° C, the programming time can be shortened to roughly 35 ns to achieve a very similar ΔV_T as can be observed in Figure 5.16(a).

However, the charge leakage increases by raising the temperature as demonstrated in Figure 5.16(b). If the device is previously programmed such that it has in the retention state initially a ΔV_T of -0.45 V, ΔV_T remains less than -0.1 V up to ~500 ms at 25° C. By raising the temperature to 85° C or even 105° C, the retention time is reduced to ~48 ms and ~32 ms, respectively. This is still better than for example Micron's DRAM where the refresh rate is 64 ms at room temperature, and doubled when the temperature is raised to 85° C [5.22].

Figure 5.16: Temperature dependency in the dynamic (a) program and (b) retention mode.

5.7 Employing Discontinuous Floating Gates

Variations of this dual floating gate structure that replace the continuous floating gates with NCFGs are here-in investigated. Isolated storage nodes may increase device endurance,

and thus reliability, due to decreased susceptibility to SILC compared to continuous floating gate devices of equivalent dimensions. This allows thinning of the tunnel oxide which results in an enhanced charge transfer. For the nonvolatile memory operation the benefits are described in Chapter 2 and are summarized here as lower programming/erasing voltage, reduced programming/erasing time, and improved scalability.

Furthermore, the device could benefit from using a discontinuous floating gate pair for the dynamic memory operation due to less susceptibility to SILC in the inter floating gate oxide tox_2 . The device already exploits mainly direct tunneling under low-stress in this memory domain, and with the additional feature of discontinuous floating gates the number of program and erase cycles may be further increased. However, the disadvantage of discontinuous floating gates is that they store less charge than the continuous variant, changing programming times and voltages.

A schematic cross-section of a device employing a discontinuous floating gate pair in the 65 nm node is created in the device simulator and illustrated in Figure 5.17(a). It involves a dual NCFG structure with an HfO_2 insulator sandwiched between them. The control gate, source and drain are composed of molybdenum (Mo), which has a work function of 4.60 eV [5.23]. Palladium (Pd), which has a work function of 5.12 eV [5.23], is used for the nanocrystals in both the top and bottom floating gate for similar programming and erasing characteristics in the dynamic memory operation. The bottom and top insulators tox_1 and tox_3 are composed of SiO_2 . The conduction band diagram under flat-band condition is shown in Figure 5.17(b). Within Figure 5.17(b), the first region on the left represents the

Mo layer, followed by the other represented layers of the Mo-SiO₂-Pd-HfO₂-Pd-SiO₂-Si stack from left to right.

Figure 5.17: (a) Cross-section of the dual-layer NCFG device and (b) its corresponding conduction band diagram under flat-band condition.

The geometry parameters and doping profile chosen for this device are listed in Table 5.2. The device is projected 1 μm deep into the third dimension, which is the default in the device simulator. In addition to the deep potential well in the nanocrystals, the SiO₂ tunnel oxide tox₁ is 4 nm thick in order to achieve a retention time of at least 10 years. The HfO₂ insulator tox₂ between the top and bottom NCFG layers is ultra thin with a physical thickness of 1.5 nm, which in addition to the high dielectric constant results in a very strong coupling between those two layers. For the SiO₂ control gate oxide tox₃ a thickness of 5 nm is chosen such that charge transfer between the nanocrystals and the control gate is reduced even for high programming and erasing voltages. For example, an electron tunneling current through the control gate oxide may be on the order of roughly 1×10^{-20} A/ μm . During the dynamic or nonvolatile memory operation, the tunneling current through tox₂ or tox₁ is in the range of

$1 \times 10^{-8} \text{ A}/\mu\text{m} - 1 \times 10^{-12} \text{ A}/\mu\text{m}$. As such, the tunneling current, and thus the charge transfer through tox_3 may be considered very minimal. Furthermore, it can be observed that the top layer nanocrystals are smaller than the bottom ones. If all the nanocrystals were the same size, a worst case scenario may occur where none of the bottom nanocrystals align with the top nanocrystals due to variation in the fabrication process. In this case, the charge transfer may be suppressed compared to an ideal scenario where the top and bottom nanocrystals are perfectly aligned. Therefore, the bottom nanocrystals are chosen to be bigger to average out the alignment with the nanocrystals in the top layer.

Table 5.2: Geometry parameters of the dual-layer NCFG device.

Parameter	Value
Channel Length	65 nm
tox_1 Thickness (SiO_2)	4 nm
tox_2 Thickness (HfO_2)	1.5 nm
tox_3 Thickness (SiO_2)	6 nm
Top Nanocrystals Height	2 nm
Top Nanocrystals Length	3 nm
Top Nanocrystals Spacing	3 nm
Bottom Nanocrystals Height	3.5 nm
Bottom Nanocrystals Length	5 nm
Bottom Nanocrystals Spacing	5 nm

Figure 5.18(a) and (b) illustrate the conduction band diagrams of the device during an applied V_{CG} of ± 4 V. The first region on the left represents the Mo layer, followed by the other represented layers of the Mo-SiO₂-Pd-HfO₂-Pd-SiO₂-Si stack from left to right.

Figure 5.18: Conduction band diagram of the dual layer NCFG device under low stress of (a) 4 V and (b) -4 V.

As expected, the simulation results represent one layer being negatively charged due to electron overshoot, while the other layer is being positively charged due to electron loss. Figure 5.19 illustrates that ΔV_T shifts slightly to the left or to the right depending on whether the applied voltage is positive or negative, respectively. For example, if the applied control gate voltage is positive, the voltage in the bottom layer increases due to electron loss and a negative ΔV_T is obtained (blue line). Analogously, if the applied control gate voltage is negative, the voltage in the bottom layer decreases due to electron overshoot, which results in a positive ΔV_T (red line). The applied voltage in both cases has a duration of 10 ns. Sufficient charge transfer in this short period while the control gate is biased relatively low with ± 4 V is only possible due to the ultra-thin inter floating gate oxide tox_2 . The bottom

floating gate has greater control on the I-V characteristics of the device due to the closer proximity to the channel. The difference of the two shifts $\Delta(\Delta V_T)$ is found within the simulation to be 223 mV. The simulations have shown that electron tunneling through the tunnel oxide tox_1 or control gate oxide tox_3 does not occur in this dynamic memory mode as the applied voltage pulses are too low and too short.

Figure 5.19: Drain current before and after a voltage pulse of $\pm 4 \text{ V}$ on (a) a linear and (b) a logarithmic scale.

Similar to the device employing a continuous floating gate pair, once the device switches to the retention mode, the electron back-tunneling from negatively to positively charged layer occurs rapidly. The results in the retention time of the dynamic charge storage for this particular device are shown in Figure 5.20(a). In Figure 5.20(b) it can be observed that a refresh of the device needs to be performed at least every $80 \mu\text{s}$ to keep $\Delta(\Delta V_T)$ above 100 mV. This is significantly shorter compared to the designed devices from Figure 5.1 and

Figure 5.11 that have achieved a refresh rate of ~ 300 ms. The main reason for this is the ultra-thin inter floating gate oxide tox_2 with a thickness of just 1.5 nm. To refresh the device, another voltage pulse of ± 4 V for 10 ns is not required. As simulated, a pulse width of 7.8 ns is sufficient to obtain the original $\Delta(\Delta V_T)$ of 223 mV.

Figure 5.20: Retention time results for (a) the threshold voltage shift ΔV_T and (b) $\Delta(\Delta V_T)$.

The nonvolatile memory operation of the device is exploited by applying a high control gate voltage V_{CG} . The conduction band diagrams of programming and erasing the device with ± 13 V are shown in Figure 5.21(a) and (b). It can be observed that the voltage drop across the HfO_2 insulator is very small compared to the voltage drop across the SiO_2 insulators. This indicates that programming and erasing the device for nonvolatile memory application will be very similar to a conventional single-layer NCFG device with equivalent tunnel and control gate oxide thicknesses.

Figure 5.21: Conduction band diagram of the new device under high-stress for (a) the program mode with 13 V and (b) the erase mode with -13 V.

An example of a high V_{CG} to put the device in a nonvolatile charged state may include a voltage in the range of 11.5 V – 13 V for the designed dual-layer NCFG device. Charge transfer takes place between the channel and the nanocrystals, which results in a large ΔV_T in the range of 0.5 V to 2 V, depending on the applied voltage envelope. The SiO_2 tunnel oxide is relatively thick and has a higher barrier compared to the HfO_2 layer such that charge leakage is significantly reduced and a retention time of more than 10 years can be expected. In Figure 5.22(a) the ΔV_T is plotted for different V_{CG} pulses for programming the device for 100 μs and in Figure 5.22(b) for erasing the device for 10 ms. In the latter case the device has been previously programmed in order to obtain a ΔV_T of approximately 2 V. Erasing the device is more difficult than programming because the electrons face a higher barrier at the interface between FG_{BOT} and tox_1 than at the interface between the substrate and tox_1 . This explains the longer erasing time.

Figure 5.22: Threshold voltage shift of (a) programming the device for 100 μs and (b) erasing the device for 10 ms.

Figure 5.23(a) and (b) show plots of the drain current in a linear and logarithmic scale, respectively. First, the drain current is illustrated for an initially uncharged device (dashed black line). Then the device is programmed (red line) such that a ΔV_T of $\sim 2 \text{ V}$ is obtained, and lastly the programmed device is being erased (blue line) resulting in a slightly negative ΔV_T . In addition, it can be observed in Figure 5.23(b) that the off-current (I_{off}) of the programmed device is less than 0.1 pA at $V_{\text{CG}} = 0 \text{ V}$ under these simulation conditions.

Figure 5.23: Drain current after program and erase operation on a) a linear and b) a logarithmic scale.

It has been shown by simulation results in the device simulator that the dual-layer NCFG device also includes both volatile and nonvolatile memory operation in one single device, which are differentiated by and controlled via the applied voltage. Table 5.3 summarizes the simulation results described above.

Table 5.3: Dynamic and static memory operation of the designed dual-layer NCFG device.

Operation	Voltage Envelope	Voltage Shift	Retention
Dynamic	10 ns, ± 4 V	$\Delta V_T = \pm 0.11$ V	$\Delta(\Delta V_T) \rightarrow 0.1$ V in 80 μ s
Nonvolatile	100 μ s, 13 V	$\Delta V_T = 2$ V	10 years

5.8 Challenges and Potential Solutions

Several challenges, beyond normal execution, exist in this invention of the unified memory device that needs to be addressed. These include the following:

- **Endurance.** When operating in the dynamic memory mode, high endurance is highly desirable, e.g. DRAMs have an endurance of up to 10^{16} cycles. Even in the nonvolatile mode, good endurance is desirable as it would be difficult to implement the load remapping schemes used in normal logic. Potential methods to improve endurance over that normally achieved in Flash devices include the following:
 - **Use of direct tunneling in the dynamic memory mode.** Direct tunneling is used rather than using Fowler-Nordheim tunneling or the hot electron effect. This is less stressful on the oxide.

- **Lower voltages and electric fields.** Lower voltages are used when compared with typical Flash memory. Again this reduces oxide stress.
- **Discontinuous floating gates.** The use of a discontinuous floating gate pair may reduce the device susceptibility to SILC, improving endurance.
- **Robust dielectrics.** Trap-free dielectrics will increase lifetime.
- **Dynamic erase trades off with dynamic retention.** If the floating gate pair is composed of one low and one high work function metal, a long dynamic retention time can be achieved at the cost of a relatively slow dynamic erase compared to a dynamic program. The dynamic erase can be made faster by using two low work function floating gates but then the dynamic retention time is reduced. Using a physically thicker inter floating gate oxide with a high electron affinity and a narrow band gap may improve both the dynamic erase (and also dynamic program) due to the lower barrier height and the dynamic retention due to the thicker barrier.
- **3D scalability.** The device may be 3D stacked using deposited layers of IGZO amorphous semiconductors. The amount of possible layers is mainly an engineering issue, not a fundamental science one. The stresses are low, so in principal it is the engineering trade-offs that decide each issue.

5.9 Applications

The architectural implications of this new device could be huge. The most obvious application of this is instant-on computing, wherein the entire memory system hibernates

whenever the computer is idle. This application is investigated in Chapter 6, along with the features of check-pointing (where the unified memory is utilized as a 2-bit device) and energy proportional computing. Check-points could be taken very quickly, either on a regular schedule or when power down. The unified memory could also enable the memory array to be placed into partial hibernate states, so that when the computer is in a low activity state (which it is most of the time) its power consumption would be more proportional to its activity.

Another good fit for this device includes a NoC. The device can be used to build ‘just in time’ (JIT) networks wherein a circuit path is setup in advance of a packet transition and torn down afterwards. This concept has been recently investigated using photonic circuits [5.24]. Since the dynamic programming mode of the device is fast (10 - 50 ns, depending on the programming voltage, oxide thickness and selected materials) the device can be used as a packet circuit switch for NoC by operating it as a JIT switch, i.e. the routing information is sent ahead of the data packet(s). Unused switches may be virtually disconnected from the network by converting the device in a nonvolatile charged state by applying a high voltage pulse, thus saving power due to the very low leakage currents.

The gate’s analog storage capacity also gives it potential to operate as a neural circuit element. Synaptic processors have the potential to outperform conventional computing especially in pattern recognition applications. However, achieving a dense useful Synapse array with conventional electronics is difficult, as a significant amount of analog circuitry is needed. The dual floating gate structure has potential to avoid the need for analog circuits and thus can achieve denser implementations.

5.10 Summary

A new device is demonstrated based on a dual floating gate structure that can be employed as a unified memory, permitting combined dynamic and nonvolatile storage in the same device. Both memory operations are clearly differentiated by the applied control gate bias and the device can be converted quickly between the two states by applying appropriate voltages. In addition, it allows dynamic operation while being in a nonvolatile charged state. Thus, it can operate as a 2-bit device. For the dynamic memory operation, a low control gate voltage in the range of 4 V to 5 V enables very fast charge transfer in the range of 10-50 ns between the floating gate pair. This charge redistribution results in a slight negative threshold voltage shift ΔV_T in the range of -0.1 V to -0.4 V. During the dynamic retention mode of the device, the electrons rapidly tunnel back to the source layer until the initial uncharged state of the device is reached. In order to maintain a slight ΔV_T of less than -0.1 V, a quick refresh pulse is required to bring ΔV_T back to higher negative voltages. Depending on how this device is engineered and how strong it has been previously programmed, the refresh rate may be on the order of 300 ms at room temperature.

A first generation test device has been fabricated that demonstrates both dynamic and nonvolatile memory operation in a single device. The device can potentially be built with high write cycle endurance in the dynamic memory domain by using direct tunneling and not Fowler-Nordheim tunneling or hot electron injection. The floating gates may be comprised of metal nanocrystals to further improve endurance. Simulations in Synopsys' numerical device simulator Sentaurus TCAD have shown that it can be scaled down to the 16 nm node.

Moreover, 3D stacking is possible with the use of IGZO amorphous semiconductors, resulting in ultra-high density. Temperature analysis in the device simulator has demonstrated that for increasing temperature the retention time is shortened whereas the programming characteristics are improved in the dynamic memory domain. At 105° C a control gate voltage of 5 V for 60 ns may result in a ΔV_T of approximately -0.45 V. This shift will decay with respect to time and after ~32 ms the shift is reduced to -0.1 V at this high temperature.

A true unified memory can have a huge impact on the power consumption and resiliency of computers and it may have multiple uses in computing. Applications of such a memory device include instant-on computing, power proportional computing, and resilient computing. They could also be very useful in synaptic computing applications or NoCs.

References

- [5.1] S.-W. Ryu, J.-W. Han, C.-J. Kim, S. Kim, and Y.-K. Choi, "Unified Random Access Memory (URAM) by Integration of Nanocrystal Floating Gate for Nonvolatile Memory and a Partially Depleted Floating Body for Capacitorless 1T-DRAM," *Solid-State Electronics*, vol. 53, pp. 389-391, 2009.
- [5.2] J.-W. Han, S.-W. Ryu, S. Kim, C.-J. Kim, J.-H. Ahn, S.-J. Choi, J.S. Kim, K.H. Kim, G.S. Lee, J.S. Oh, M.H. Song, Y.C. Park, J.W. Kim, and Y.-K. Choi, "A Bulk FinFET Unified-RAM (URAM) Cell for Multifunctioning NVM and Capacitorless 1T-DRAM," *IEEE Electron Device Letters*, vol. 29, pp. 632-634, 2008.
- [5.3] K. Nomura, H. Ohta, A. Takagi, T. Kamiya, M. Hirano, and H. Hosono, "Room-Temperature Fabrication of Transparent Flexible Thin-Film Transistors Using Amorphous Oxide Semiconductors," *Letters of Nature*, vol. 432, pp. 488-492, 2004.

- [5.4] *Society for Information Display 2010 conference*. Toshiaki Ari, et al. Sony Corporation Atugi-shi Japan, 69.3 “Highly Reliable Oxide Semiconductor for AMOLED Display” and Yeon Ma Go et al. Samsung Corporation Samsung Mobile Display, Yongin-City, Gyeonggi-Do, South Korea, 69.3 Amorphous Oxide TFT Backplanes for Large Sized AMOLED TVs.
- [5.5] S. Sarkar, A. Suresh, F.B. Myers, J.F. Muth, and V. Misra, "Modulating Indium Gallium Zinc Oxide Transistor Characteristics with Discrete Redox States of Molecules Embedded in the Gate Dielectric," *Applied Physics Letter*, vol. 92, art. no. 223304, 2008.
- [5.6] A. Suresh, S. Novak, P. Wellenius, V. Misra, and J.F. Muth, "Transparent Indium Gallium Zinc Oxide Transistor Based Floating Gate Memory with Platinum Nanocrystals in the Gate Dielectric," *Applied Physics Letter*, vol. 94, art. no. 123501, 2009.
- [5.7] K.F. Schuegraf, C.C. King, and C. Hu, "Ultra-Thin Silicon Dioxide Leakage Current and Scaling Limit," *IEEE Symposium on VLSI Technology*, pp. 18-19, 1992.
- [5.8] P.P. Apte and K.C. Saraswat, "Correlation of Trap Generation to Charge-to-Breakdown (Qbd): A Physical-Damage Model of Dielectric Breakdown," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1595-1602, 1994.
- [5.9] R.A. Rao, R.F. Steimle, M. Sadd, C.T. Swift, B. Hradsky, S. Straub, T. Merchant, M. Stoker, S.G.H. Anderson, M. Rossow, J. Yater, B. Acred, K. Harber, E.J. Prinz, B.E. White, Jr. and R. Muralidhar, "Silicon Nanocrystal Based Memory Devices for NVM and DRAM Applications," *Solid-State Electronics*, vol. 48, pp. 1463-1473, 2004.
- [5.10] P. Dimitrakis and P. Normand, "Semiconductor Nanocrystal Floating-Gate Memory Devices," *Materials Research Society Symposium Proceedings*, vol. 830, 2004.
- [5.11] J.J. Lee and D.-L. Kwong, "Metal Nanocrystal Memory with High-k Tunneling Barrier for Improved Data Retention," *IEEE Transaction on Electron Devices*, vol. 52, pp. 507-511, 2005.
- [5.12] J.H. Chen, Y.Q. Wang, W.J. Yoo, Y.-C. Yeo, G. Samudra, D.S. Chan, A.Y. Du, and D.-L. Kwong, "Nonvolatile Flash Memory Device Using Ge Nanocrystals Embedded in HfAlO High-k Tunneling and Control Oxides: Device Fabrication and Electrical Performance," *IEEE Transaction on Electron Devices*, vol. 51, pp. 1840-1848, 2004.

- [5.13] J. Buckley, B. De Salvo, G. Ghibaudo, M. Gely, J.F. Damlencourt, F. Martin, G. Nicotra, and S. Deleonibus, "Investigation of SiO₂/HfO₂ Gate Stacks for Application to Non-Volatile Memory Devices," *Solid-State Electronics*, vol. 49, pp. 1833-1840, 2005.
- [5.14] J.J. Lee, X. Wang, W. Bai, N. Lu, and D.-L. Kwong, "Theoretical and Experimental Investigation of Si Nanocrystal Memory Device with HfO₂ High-k Tunneling Dielectric," *IEEE Transactions on Electron Devices*, vol. 50, pp. 2067-2072, 2003.
- [5.15] S. Novak, B. Lee, X. Yang, and V. Misra, "Platinum Nanoparticles Grown by Atomic Layer Deposition for Charge Storage Memory Applications," *Journal of the Electromechanical Society*, vol. 157, pp. H589-H592, 2010.
- [5.16] S. Jayanti, X. Yang, D.J. Lichtenwalner, and V. Misra, "Technique to Improve Performance of Al₂O₃ Interpoly Dielectric Using a La₂O₃ Interface Scavenging Layer for Floating Gate Memory Structure," *Applied Physics Letter*, vol. 96, pp. 092905 - 092905-3, 2010.
- [5.17] T. Erlbacher, M.P.M. Jank, M. Lemberger, A.J. Bauer, and H. Ryssel, "HfSiO/SiO₂- and SiO₂/HfSiO/SiO₂-Gate Stacks for Non-Volatile Memories," *Thin Solid Films*, vol. 516, pp. 7727-7731, 2008.
- [5.18] M. Ľapajna, K. Hušková, J.P. Espinos, L. Harmatha, and K. Fröhlich, "Precise Determination of Metal Effective Work Function and Fixed Oxide Charge in MOS Capacitors with High-K Dielectric," *Materials Science in Semiconductor Processing*, vol. 9, pp. 969-974, 2006.
- [5.19] Y.-C. Yeo, T.-J. King, and C. Hu, "Direct Tunneling Leakage Current and Scalability of Alternative Gate Dielectrics," *Applied Physics Letter*, vol. 81, pp. 2091-2093, 2002.
- [5.20] W. Wang, T. Lee, and M.A. Reed, "Mechanism of Electron Conduction in Self-Assembled Alkanethiol Monolayer Devices," *Phys. Rev. B*, vol. 68, art. no. 035416, 2003.
- [5.21] G. Groeseneken, H.E. Maes, J. Van Houdt, and J.S. Witters, "Basics of Nonvolatile Semiconductor Memory Devices," in *Nonvolatile semiconductor memory technology*, W. D. Brown and J. Brewer, Eds., ed NJ: Wiley-IEEE Press, 1997, pp. 1-81.
- [5.22] M. Ghosh and H.-H.S. Lee, "Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs," *In: 40th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 134-145, 2007.

- [5.23] N. Miura, T. Nanjo, M. Suita, T. Oishi, Y. Abe, T. Ozeki, H. Ishikawa, T. Egawa, and T. Jimbo, "Thermal Annealing Effects on Ni/Au Based Schottky Contacts on n-GaN and AlGaN/GaN with Insertion of High Work Function Metal," *Solid-State Electronics*, vol. 48, pp. 689-695, 2004.
- [5.24] A. Shacham, K. Bergman, and L.P. Carloni, "Photonic Networks-on-Chip for Future Generations of Chip Multiprocessors," *IEEE Transactions on Computers*, vol. 57, pp. 1246-1260, 2008.

Chapter 6

Unified Memory Array

6.1 Motivation

The invented dual floating gate device can perform both dynamic and nonvolatile memory operation while the read mode is non-destructive. Furthermore it can continue operating as a dynamic memory device while storing a different bit pattern in the nonvolatile mode and it can be read as a 2-bit device. A well suited application of this unified memory is therefore instant-on computing, wherein the entire memory system hibernates whenever the computer is idle. It would create a memory that can be converted between an active mode and a hibernate mode through a row-by-row read-write cycle.

The implementation of these dual floating gate devices in a memory array is investigated in this Chapter and compared to the conventional DRAM. Additional beneficial features such as check-pointing and energy proportional computing are discussed. Since state saves are quick, the overhead of check-pointing is greatly reduced, and can easily be done on power interrupts, etc. The architectural implications of such a unified memory array could be huge and are here-in briefly described for improving computing efficiency in terms of lower power, higher performance, and new functionality.

6.2 Introduction

Certain storage systems in a CPU environment mix the DRAM main memory and the hard drive, which suffer from slow data transfer. In addition, the DRAM (typically composed of 1T-1C cells) consumes a large portion of overall system power due to the short periodic refresh requirement of at least every 64 ms. This is due to the leaky capacitor and the leakage current of the transistor, which cause the charge stored in the capacitor to gradually decrease with respect to time. Therefore, the maximal leakage current of the transistor should be on the order of 10^{-14} A [6.1]. As next generation transistors deal increasingly with short channel effects, such as DIBL which increases the subthreshold leakage current, scaling of the transistor is limited and DRAM cannot scale with CMOS technology. It is also difficult to scale the capacitor as it is required to maintain a capacity of 20-25 fF to enable sufficient charge sharing with the bit line, producing a voltage change large enough that the sense amplifier is able to detect and amplify. The ITRS forecasts that the DRAM 22 nm half pitch won't be produced earlier than 2016 [6.2]. Solutions to continue scaling beyond 22 nm are still unknown yet.

Recently, a hybrid main memory has been introduced, which consists of a smaller first-level DRAM augmented by a second-level Flash memory. The Flash memory consists of continuous floating gate devices, which consume orders of magnitude less idle power than DRAM. In addition, integrating Flash memory into the conventional system-level memory hierarchy enables quick warm up of file buffer caches. This hybrid DRAM/Flash memory system is proposed for low power web servers [6.3], and it is already in commercial use, e.g.

for modern cell phones [6.4]. They use the same concept, where the data is stored in the volatile DRAM, and then gradually transmitted to the nonvolatile Flash memory. Once the information is stored in the Flash memory it can be loaded back to the DRAM during boot time. Figure 6.1 shows a memory hierarchy with the hybrid DRAM/Flash architecture as the main memory replacing the conventional DRAM [6.3]. The benefits of the hybrid main memory compared to a conventional DRAM-only main memory are less loading time, smaller DRAM, and especially orders of magnitude less idle power consumption. However, Flash memory technologies will have difficulty scaling beyond the 16 nm node. To overcome this issue a bit-cost scalable (BICS) structure in which a NAND Flash structure is built vertically is investigated [6.5]. This has potential to outscale any proposed nonvolatile alternative, including emerging devices. However, it is only suited for slow nonvolatile memory operation. In contrast, the dual floating gate device is fast and intended as a merged dynamic and nonvolatile memory device.

Figure 6.1: A conventional DRAM opposed to a hybrid DRAM/Flash main memory [6.3].

The dual floating gate device may improve the main memory of a CPU beyond existing conventional and hybrid memory architectures. During active mode, the memory utilizes dynamic memory behavior. When the power is off, a high voltage pulse can be used on a row-by-row basis to save the information due to the nonvolatile charge storage mode. A controller can be used to provide the device with both low and high voltages. While the system is booting up, the device can be converted back from nonvolatile to dynamic memory operation and the main memory is rapidly loaded with the stored nonvolatile value. No time consuming data transfer between the main memory and the hard disk or Flash memory is required. In addition, as the main memory size increases, unused memory cells do not need to be refreshed and may be fully programmed, resulting in a very low leakage current, and may therefore reduce power consumption. These potential benefits make the device suitable for an overall high speed and very dense main memory. Furthermore, the device may exploit its multi-bit capability, enabling fast in-situ check-pointing and rollback/recovery. The check-point can be saved in-place, rather than being written to a disk through a narrow channel. The memory then continues operating as normal. If an error is detected, the memory state can be rolled back to the saved state in a single cycle and the computer resiliency is greatly improved. One competing technology is 3D stacking of DRAM and Flash memory using multiple wafers/die and through silicon vias (TSV) [6.6-6.8]. It would increase the transfer rate between volatile and nonvolatile storage, which permits check-pointing. However, the effective transfer rate in the new technology is much faster. In addition, this technology could also be potentially 3D stacked to create a memory array with ultra-high density, which may also further improve performance.

6.3 Circuit Design and Analysis

The designed memory array that utilizes the dual floating gate device is shown in Figure 6.2. Each device is shown with 4 terminals, including a back gate. The control gate is driven by the word line, and the drain terminal by the bit line. The back gate is connected to the select line to provide selectivity of each device and the source terminal is attached to a reference voltage V_{ref} . The overall structure is similar to a NOR Flash. The resistor R_{BL} and capacitor C_{BL} for each bit line represent the wire resistance and capacitance, respectively. The sense amplifier is designed such that it can detect the different voltages on the bit lines due to the modified threshold voltages of the new devices, depending on their states. If the array utilizes solely dynamic memory operation during active mode, as well as the conversion to a hibernation mode during power off and back to the active mode during power on, a buffer can be used for the sense amplifier. In addition, the device can store two bits, by having one bit stored in a dynamic state, while keeping another bit stored in the nonvolatile state. Each combination of bits could be detected through a quaternary logic sense amplifier, or by performing two successive read cycles at different voltages. The clock signal is attached to the gate terminal of a NMOS transistor whose drain terminal is connected to its corresponding bit line and is located in close proximity to the sense amplifier. If the clock signal goes high, the bit lines are pulled down to G_{ND} . This occurs very fast compared to the precharge mode in a conventional DRAM. If the clock signal goes low, the corresponding NMOS transistor is off and one of the word lines turns on to read an entire row. Due to the close proximity to the sense amplifier and the fast precharge mode an asynchronous clock

may be used with a shorter low signal. The clock frequency could then be increased to achieve better performance.

Figure 6.2: Memory array using the dual floating gate device.

The values for R_{BL} and C_{BL} for each column are calculated from Intel's 45 nm technology data sheet that are given by $R = 3.3 \, \Omega/\mu\text{m}$ and $C = 0.2 \, \text{fF}/\mu\text{m}$ for the lower metal [6.9]. As this array could potentially be packed with a cell size of $8F^2$ it is assumed that the bit line length per cell is equal to $2F$, which includes the gate width of the device. The values R_{BL} and C_{BL} are therefore dependent on the number of rows and are listed in Table 6.1. Of course, parasitic capacitances and resistances are also present in the word lines. In this case, a π -RC network structure is used between each cell for each row in the design (not shown in

Figure 6.2); thus the signal propagation delay for those columns that are further away from the controller is accounted for. It is assumed that the word line length per cell is equal to $4F$ which includes the length of the gate, drain and source. The read time is then measured in HSPICE simulations, which is interpreted as the minimal required word line pulse width to generate logic ‘0’s and ‘1’s at the output of the sense amplifiers with a minimum pulse width of 200 ps in this Chapter. Note that for the read time analysis in Table 6.1 the number of columns is always equal to the number of rows.

Table 6.1: Read time analysis.

No. of Rows	Bit Line Length	R_{BL}	C_{BL}	Read Time
4	0.36 μm	1.2 Ω	0.07 fF	0.2 ns
8	0.72 μm	2.4 Ω	0.14 fF	0.2 ns
16	1.44 μm	4.8 Ω	0.29 fF	0.22 ns
32	2.88 μm	9.6 Ω	0.57 fF	0.25 ns
64	5.76 μm	19.2 Ω	1.15 fF	0.31 ns
128	11.52 μm	38.4 Ω	2.3 fF	0.43 ns
256	23.04 μm	76.8 Ω	4.6 fF	0.68 ns
512	46.08 μm	153.6 Ω	9.2 fF	1.18 ns
1024	92.16 μm	307.2 Ω	18.4 fF	2.18 ns

The proposed unified memory array is designed in Cadence Virtuoso based on predictive BSIM4.0 CMOS with 45 nm gate length technology. The dual floating gate device uses the parameters given in Figure 5.11 of Chapter 5 with an increased gate length of

45 nm to allow circuit simulation in HSPICE and a gate width of 90 nm. A physical model of the device has been developed in Chapter 5 and is used for circuit simulation in HSPICE.

First, a small 4x4 array is constructed to test functionality of the design when it operates as a DRAM. The waveforms of the simulation results are shown in Figure 6.3 and Figure 6.4. The array is first in the dynamic write mode where each word line (WL1-WL4) is driven high one at a time with a voltage of 3 V for 50 ns as illustrated in Figure 6.3. The voltages for each select line SL1-SL4 are arbitrarily chosen with either +2 V or -2 V. The red numbers indicate whether the correspondent cell is written with a ‘1’ (the device is being dynamically programmed) or left at ‘0’ (the device remains in its uncharged state).

Figure 6.3: HSPICE input waveforms for the word lines (WL1-WL4) and the select lines (SL1-SL4).

The memory array is now configured and next the read mode is tested for which the output waveforms are shown in Figure 6.4. Due to the small size of the array the clock frequency can be relatively high with 2 GHz and the supply voltage, as well as the reference voltage V_{ref} , is set to 1 V. The word lines WL1-WL4 are consecutively addressed cycle-by-cycle with a read voltage of 1.2 V when the clock signal goes low. The correct functionality of the design is then demonstrated by means of the output waveforms of the sense amplifiers SA1-SA4. It can be observed that the dynamically written cells from Figure 6.3 correspond to a logic '1' at the output of the sense amplifiers whereas those cells with the uncharged devices correspond to a logic '0'.

Figure 6.4: HSPICE input waveforms for the clock (CLK) and word lines WL1-WL4 and output waveforms of the sense amplifiers SA1-SA4.

Note that the same simulation for the dynamic read mode has been performed after a retention time of 300 ms as well, which showed equivalent output waveforms as Figure 6.4. Another simulation run after a retention time of 400 ms however has demonstrated failure of the array: all output waveform of the sense amplifiers represented a logic '0'. A refresh beforehand would have been necessary to obtain the correct waveforms.

Next, the size of the array is increased to 32 KB (128 rows x 256 columns). The array is cycled row-by-row with a clock frequency of 1 GHz and the supply voltage and reference voltage V_{ref} are set to 1 V. The design is tested on correct functionality of both the dynamic and nonvolatile memory mode and a speed and energy consumption analysis is performed. The energy consumption in the read mode is measured to be 10.1 pJ per bit read. This includes the memory array, the sense amplifiers, the clock network and the capacitances and resistances of the bit lines as well as for the word lines. In order to write the entire memory array into the dynamic state '1' the word lines are biased with 3 V on a row-by-row basis while -2 V is applied on all select lines, which takes 6.4 μ s and consumes 9.8 nJ per bit written.

In Table 6.2, the states of the new memory array cell are listed, along with the voltages and time constants required to change the threshold voltage from the current state of the device to the desired one, as obtained from device and circuit simulation. Each transition is effected by applying the voltage envelopes to the word and select lines. The native threshold voltage of the device is $V_{T0} \sim 0.4$ V.

Table 6.2: States of a unified memory cell before and after the applied voltage envelope on the word line (WL) and select line (SL). The data is obtained from Sentaurus TCAD simulations together with circuit level HSPICE simulations.

	Before Applied Voltage Envelope			Voltage Envelope			After Applied Voltage Envelope		
	Dyn. State	NV State	ΔV_T	WL	SL	Time	Dyn. State	NV State	ΔV_T
Dynamic Write	0	0	0 V	3 V	-2 V	50 ns	1	0	-0.33 V
	0	1	+1.52 V	3 V	-2 V	50 ns	1	1	+1.21 V
	1	0	-0.33 V	-3 V	2 V	10 μ s	0	0	0 V
	1	1	+1.21 V	-3 V	2 V	10 μ s	0	1	1.52 V
Dynamic Retain	0	0	0 V	3 V	2 V	50 ns	0	0	0 V
	0	1	+1.52 V	3 V	2 V	50 ns	0	1	+1.52 V
	1	0	-0.33 V	-3 V	-2 V	10 μ s	1	0	-0.33 V
	1	1	1.21 V	-3 V	-2 V	10 μ s	1	1	1.21 V
Dynamic Refresh	1	0	-0.11 V	3 V	-2 V	40 ns	1	0	-0.33V
	1	1	+1.41 V	3 V	-2 V	40 ns	1	1	+1.21V
Nonvolatile Write	0	0	0 V	7 V	-2 V	30 μ s	0	1	+1.52 V*
	0	1	+1.52 V	-7 V	2 V	14 μ s	0	0	0 V*
	1	0	-0.33 V	7 V	-2 V	30 μ s	1	1	+1.21 V
	1	1	+1.21 V	-7 V	2 V	14 μ s	1	0	-0.33 V

Table 6.2 (Continued).

Dynamic/	0	0	0 V	1.2 V	0 V		0	0	0V
Nonvolatile	1	0	-0.33 V	1.2 V	0 V	0.31 ns	1	0	-0.33V
	0	1	+1.52 V	2.7 V	0 V	to	0	1	+1.52V
Read	1	1	+1.21 V	2.7 V	0 V	2.18 ns	1	1	+1.21V

* Followed by writing a dynamic '0'.

Obviously multiple voltage supplies are needed for correct operation. The bits are read by detecting the threshold voltage of each transistor. In the simulations above, this was done by biasing a buffer to detect the threshold voltage shift of each cell. The states of the memory are as follows:

- **Dynamic Write:** It takes 50 ns for the device to write a dynamic '1', similar to a DRAM. Changing a dynamic state from '1' to '0' takes 10 μ s. The latter could be made faster by applying higher voltages as long as the nonvolatile mode is inactive.
- **Dynamic Retain:** This is essential for those cells in a row that are to retain their states during a dynamic write or refresh.
- **Dynamic Refresh:** Dynamic refresh is only required if a '1' is written in the dynamic state. After 300 ms at room temperature, ΔV_T has been reduced to approximately -0.11 V due to charge leakage. The sense amplifier is designed such that it can detect any ΔV_T less than -0.1 V. Therefore, a read-write cycle for a quick refresh is required every 300 ms.

- **Nonvolatile Write:** Each device changes its nonvolatile state through a read-write cycle one row at a time. This takes 30 μs to change from '0' to '1', and 14 μs to change from '1' to '0', which is still faster than traditional Flash memory write and erase. Of course, the write data could be externally sourced too.
- **Dynamic/Nonvolatile Read:** Since this operates more like a SRAM than a DRAM, reads should be very fast: 1-4 ns including array overhead. The read voltage for a device in the nonvolatile uncharged state is with 1.2 V less than that for current DRAMs. Note that the read is non-destructive and no write-back is needed. If the device operates as a 2-bit device each combination of bits could be detected through a quaternary logic sense amplifier, or by performing two successive read cycles at different voltages. The read voltage for a device that is in the nonvolatile charged state is then 2.7 V (assuming a ΔV_T of the device of ~ 1.5 V).

6.4 Comparison to DRAM

Using the dual layer floating gate device to create a unified main memory offers several advantages towards a conventional DRAM. Similar to the single-layer floating gate device, fabrication costs of the new device may be reduced compared to conventional DRAM due to the absence of the capacitor. In addition to cost, the density of the array is expected to be higher due to the better scalability of the dual floating gate devices, especially when fabricated on SOI and potentially using nanocrystal metal gates. Another advantage of the unified memory array is that the read mode is very fast and a relatively low word line voltage

of 1.2 V is required if the device is in a nonvolatile uncharged state compared to 1.3 V in Intel's latest 32 nm DRAMs. The read voltage may be reduced for scaled devices. Depending on the size of the array (higher capacitance for larger arrays due to transistor overhead and added wire capacitance) the read time is in the range of 0.31 ns for 64 rows to 2.18 ns for 1024 rows, resulting in a clock frequency in the range of 230 MHz to 1.61 GHz. Furthermore, an asynchronous clock signal may be used with a shorter low signal to enable faster operations. Another savings comes about as the read is non-destructive, i.e. no following write cycle is needed as in the conventional DRAM. The energy consumption in the read mode is 10.1 pJ per bit read. 3D circuit sub-arraying may lead to further power reduction. The soft error bit rate may be statistically low because the sense amplifier needs to detect a relatively large voltage change of greater than 100 mV, similar to a conventional DRAM. Writes require higher voltages, depending on whether the storage is dynamic or nonvolatile. However, the refresh period (i.e. the retention time of the dynamic memory mode of the device) is 300 ms at room temperature which is longer than in a conventional DRAM with a minimal refresh period requirement of 64 ms. This may result in overall improved energy efficiency in the write mode of the unified memory array compared to its counterpart. One drawback of the new design is the transition of a dynamic '1' to a dynamic '0' as the dynamic erase is relatively slow. It could be made faster by applying higher voltages but care must be taken not to enable the nonvolatile erase mode. Alternatively, the inter floating gate oxide could be composed of a high-k oxide with a high electron affinity and a narrow band gap, which is built physically thicker to improve the ratio between dynamic program/erase and retention characteristics. Another option is to use a lower work

function metal for the top floating gate to enhance the transfer rate of electrons from the top floating gate to the bottom floating gate due to the lower barrier. However, this would result in a reduced dynamic retention time.

Similar to the dual layer floating gate device the DRAM cell is also temperature dependent. The leakage increases with temperature and the refresh rate must be doubled at 85° C in Micron's DRAM [6.10], resulting in a refresh period of 32 ms. In contrast, as described in Chapter 5 the memory array that employs the dual floating gate devices may obtain a refresh period of 32 ms when operating at 105° C. The leakage current and the cumulative bit failure for different temperatures for the DRAM cell are illustrated in Figure 6.5(a) and (b), respectively [6.11].

Figure 6.5: (a) Leakage current vs. drain voltage for different temperatures [6.11].

(b) Cumulative bit failure vs. retention time for different temperatures [6.11].

The key benefit of the unified memory design however is the capability of instant-on computing where the array can switch rapidly between the active mode and the hibernate mode. In principle, a memory switch to hibernate could be done in approximately 30 ms, assuming 1024 rows in each memory bank, and all the banks in all the memories can change state in parallel. A switch back to an active mode would take about half this time.

A more subtle use of this capability would be to actively control partial hibernation as one method to achieve energy proportional computing. When the computer is only running background tasks, portions of the memory can be “frozen” so that they are not consuming power, and when they are needed, they can be “unfrozen” in only a few cycles. This could be a valuable tool for helping solve the problem of energy proportional computing [6.12]. In [6.12], data taken from Google servers showed that the processors were consuming 60% of their peak power when only operating at 10% of their peak throughput. Furthermore for most of the time, the processors are operating at 60% or less of peak capacity. Since the CPU already has many fast-switching partial sleep modes, much of this “wasted power” is due to power consumption in the memory, since the DRAM cannot be put into a partial sleep mode quickly and easily. In contrast, a unified memory can be placed in a partial sleep in a few milliseconds, depending on how that mode is organized. Recovery is similarly fast.

In addition, the memory greatly enhances the resiliency of the computer – its ability to withstand and recover from faults. A fast save can be used to capture the entire state of the memory when a transient fault is detected, or upon a power-down. If check-pointing is used, this memory can decrease its performance impact significantly. For 1024 rows in each bank a check-point can be taken in ~30 ms by conducting a read-nonvolatile write cycle on each

row. The memory can then continue operating as a 2-bit device, with a backup store in the nonvolatile portion. Check-pointing will not only be faster but also more energy efficient.

6.5 Implications on the Memory Hierarchy of the Computer Architecture

A single memory cell that can behave as either a DRAM cell or a Flash memory cell has powerful implications for computing. Two major computing applications of this new technology with significant potential for improving computing efficiency in terms of lower power, higher performance, and new functionality are i) a unified volatile (dynamic) and nonvolatile main memory for executing programs and ii) a unified solid state disk drive and main memory.

Figure 6.6(a) shows a traditional memory hierarchy, which uses volatile storage in the processor caches and the main memory whereas nonvolatile storage is used in the solid state disk drive and the hard drive. The caches are typically comprised of SRAM and the main memory of DRAM. Compared to nonvolatile memory they offer fast read/write times and low read/write power consumption and they are therefore used to hold the data of programs that are currently executing. Regardless of their relatively low power consumption during read and write operations, they are power-inefficient when they are not being accessed. The SRAM cell is comprised of six transistors, which results in a large amount of leakage current. In addition, a permanent crowbar current from the supply voltage to ground is existent in the cross-coupled inverters, which leads to further power waste. The DRAM requires a refresh

at least every 64 ms to hold the data. A read or refresh requires high power in the peripheral circuits. In addition, transferring the data between the main memory and the hard disk or more recently Flash memory (e.g. when the computer is transitioning between active mode and hibernation) is not only time-consuming as the entire DRAM image needs to be copied to slow nonvolatile storage but it is also power-inefficient. This is analogous for check-pointing. The computer's file system is stored in a nonvolatile memory including magnetic hard disk drives or Flash memory based solid state disk drives. They consume very low idle power but they are slow. Therefore, data and programs are transferred to the main memory for fast execution and the nonvolatile memories are used as the secondary memory storage or long term persistent storage where a slow speed can be tolerated.

Figure 6.6: (a) Traditional memory hierarchy. (b) Unified memory hierarchy.

Figure 6.6(b) shows a proposed memory hierarchy based on the new memory cell. Each cell can operate as either a volatile DRAM cell or a nonvolatile Flash memory cell and the selection can be varied arbitrarily at run-time. The computer architecture can now employ a unified memory with DRAM and Flash mixed together, where the amount of each is variable and determined by hardware/software policies. Two novel applications of this flexible memory fabric are here-in introduced. A deep investigation will be a major research task in the future work.

6.5.1 Unified Volatile and Nonvolatile Main Memory

A unified volatile/nonvolatile main memory for executing programs is created to reduce the total memory power consumed by active programs by intelligently dividing their data among DRAM-configured and Flash-configured rows. This is one significant improvement compared to the hybrid DRAM/Flash memory system that is briefly described in Section 6.2 where the size of the DRAM and Flash is fixed and not variable. The memory controller will keep track of how each row is used. Only rows employing DRAM-mode are refreshed. The memory controller will also implement policies to configure/reconfigure rows as DRAM or Flash, with the goals to minimize the total memory power consumption and maximize performance, durability, recoverability, and efficiency of system state transitions. As such, the goal of the unified volatile/nonvolatile memory system is to achieve the optimal balance between i) DRAM's high speed data access and ii) Flash memory's low idle power consumption, intrinsic recoverability, and zero-overhead during system transitions between

active mode and hibernation. Thus, while the traditional memory hierarchy relegated Flash memory to just file system storage, the new memory hierarchy exploits favorable Flash memory characteristics for executing programs. The key idea is to configure a row as either DRAM or Flash memory according to the frequency of writes and to reconfigure the row according to varying write behavior. Moreover, the overhead of row reconfiguration is kept low by “piggybacking” transitions between the DRAM and the Flash memory on top of normal memory accesses such as refreshes and writes.

6.5.2 Unified Solid State Disk Drive and Main Memory

In the unified solid state disk drive and main memory the flexible memory fabric enables another, unprecedented memory hierarchy innovation: storing part or all of the computer’s file system in the same storage fabric as active program data. A portion of the main memory can be configured as a solid state disk drive for holding files. In terms of memory row configuration, there is no distinction between Flash-configured rows for running programs and Flash-configured rows for file storage. The distinction is only apparent at the system level. The key idea is to redesign the memory controller and/or software device drivers to support two protocols for the same storage fabric: main memory protocol and solid state disk protocol. In this system, the formerly separate solid state disk drive is absorbed by the main memory. Eliminating a dedicated component reduces the system footprint contributing to further miniaturization of systems. There may be performance and energy efficiency improvements as well.

6.6 Summary

The use of the dual floating gate devices in the main memory is investigated. It would enable an instant-on computer where the computer can switch quickly between active mode and hibernation on a row-by-row basis. The memory may switch to a hibernate mode in ~30 ms, assuming 1024 rows in each memory bank that can change state in parallel with the other banks. A switch back to an active mode can be performed in ~14 ms. A circuit design is presented that resembles a NOR Flash and could be built very dense. The read-mode is fast, non-destructive and operates at 1.2 V, comparable to tomorrow's DRAM. The energy consumption in the read mode is 10.1 pJ per bit read. Depending on the number of rows the clock frequency may be in the order of 230 MHz for 1024 rows per memory bank or 1.61 GHz for 64 rows per bank. An asynchronous clock signal may permit faster operations. The dynamic write voltage is relatively high compared to DRAM but is compensated by 4-5x longer refresh period requirement. Writing an entire row takes 50 ns and consumes 9.8 nJ per bit written. The unified memory array may also operate in a larger temperature window, allowing operation at 105° C while the refresh is on the order of 32 ms whereas the DRAM may operate at maximal 85° C to achieve this refresh rate.

In addition, check-points could be taken very quickly, either on a regular schedule or when power down to decrease the performance impact of the memory significantly. For 1024 rows in each bank a check-point can be taken in ~30 ms. The memory can then continue operating as a 2-bit device, with a backup store in the nonvolatile state of bit. The unified main memory can also have an impact on energy proportional computing. Cells that

are unused (e.g. when the computer is only running background tasks) can be placed in a partial sleep in a few milliseconds to reduce power consumption. If they are needed again they can be activated in just a few cycles.

The use of a unified volatile/nonvolatile memory system offers new possibilities in the memory hierarchy of the computer architecture and opens a new research field. A unified main memory may reduce power consumption by dividing it into DRAM-configured rows for high speed data access and Flash-configured rows for low idle power consumption and intrinsic recoverability. In addition, a unified solid state disk drive and main memory can be created to store data from the computer's file system in the same storage fabric as active program data.

References

- [6.1] J. Barth, "Embedded DRAM in Nano-Scale Technologies," in *Embedded Memories for Nano-Scale VLSIs*, K. Zhang, Ed., ed New York, NY: Springer US, 2009, pp. 127-175.
- [6.2] *International Techonology Roadmap for Semiconductors (ITRS)*.
- [6.3] T. Kgil and T. Mudge, "FlashCache: a NAND Flash Memory File Cache for Low Power Web Servers," *International Conference on Compilers, Architectures, and Synthesis for Embedded Systems*, pp. 103-112, 2006.
- [6.4] V.G. Moshnyaga, H. Vo, G. Reinman, and M. Potkonjak, "Reducing Energy of DRAM/Flash Memory System by OS-Controlled Data Refresh," *International Symposium on Circuits and Systems*, pp. 2108-2111, 2007.

- [6.5] Y. Fukuzumi, Y. Matsuoka, M. Kito, M. Kido, M. Sato, H. Tanaka, Y. Nagata, Y. Iwata, H. Aochi, and A. Nitayama, "Optimal Integration and Characteristics of Vertical Array Devices for Ultra-High Density, Bit-Cost Scalable Flash Memory," *IEEE International Electron Devices Meeting*, pp. 449-452, 2007.
- [6.6] P. Kettner, B. Kim, S. Pargfrieder, and S. Zhu, "New Technologies for Advanced High Density 3D Packaging by Using TSV Process," *Int. Conf. on Electronic Packaing Technology and High Density Packaging*, 2008.
- [6.7] G. H. Loh, "3D Stacked Memory Architectures for Multi-Core Processors," *Int. Symp. on Computer Architecture*, pp. 453-464, 2008.
- [6.8] P. Ramm, A. Klumpp, and J. Weber, "3D Integration Technologies for MEMS/IC Systems," *IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 138-141, 2009.
- [6.9] P. Moon, V. Chikarmane, K. Fischer, R. Grover, T.A. Ibrahim, D. Ingerly, K.J. Lee, C. Litteken, T. Mule, and S. Williams, "Process and Electrical Results for the On-Die Interconnect Stack for Intel's 45nm Process Generation" in *Intel Technology Journal - Intel's 45nm CMOS Technology*, L. Chao, Ed., vol. 12, issue 02, 2008.
- [6.10] M. Ghosh and H.-H.S. Lee, "Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs," *In: 40th Annual IEEE/ACM Int. Symp. on Microarchitecture*, pp. 134-145, 2007.
- [6.11] K. Saino, S. Horiba, S. Uchiyama, Y. Takaishi, M. Takenaka, T. Uchida, Y. Takada, K. Koyama, H. Miyake, and C. Hu, "Impact of Gate-Induced Drain Leakage Current on the Tail Distribution of DRAM Data Retention Time," *IEDM Technical Digest*, pp. 837-840, 2000.
- [6.12] L.A. Barroso and U. Holzle, "The Case of Energy-Proportional Computing," *IEEE Computer*, vol. 40, pp. 33-37, 2007.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This dissertation argues that floating gate devices have great potential to be integrated within CMOS logic and therefore function as logic circuit elements rather than pure storage memories. Single-layer NCFG devices can be used in static applications such as the interconnect of a FPGA board where high efficiency in terms of power, area, speed and functionality, which typically trade off, has been achieved simultaneously. Power-hungry and large SRAMs are replaced by nonvolatile NCFG devices that retain their state while the system power is off and do not need to be configured at every boot up. In order to simulate this circuit topology utilizing NCFG devices, an accurate physical model of these devices has been developed beforehand. It uses only analytical formulas appropriate for fast circuit simulation and is implemented in a Verilog-A module to provide portability across existing SPICE-based infrastructure. The fully parameterized and configurable physical model interacts dynamically with the rest of the circuit during a transient simulation. As such, charge leakage or unintentional programming of the devices are always accounted for and power consumption analysis can be performed.

In addition, a new unified memory device based on a dual metal floating gate structure is invented. It is scalable to at least the 16 nm node and has potential for 3D

stacking. This device can be used as a dynamic or nonvolatile memory. It can be changed quickly between the two states by applying the appropriate voltages. The device operates very fast in read mode and in the dynamic memory domain it achieves a refresh period 4-5x longer than conventional DRAM. Normally floating gate devices are not used for dynamic applications due to their relatively low endurance and slow switching speed. By using direct tunneling, rather than Fowler Nordheim tunneling, and/or using nanocrystals instead of solid floating gates, there is potential for the program/erase cycle count to be greatly improved. Due to the use of an ultra-thin high-k oxide a transition from a dynamic '0' to a dynamic '1' can be performed in just 50 ns. Furthermore, the device permits both dynamic and nonvolatile memory operation simultaneously; thus it can operate as a 2-bit device.

The unified memory may have a dramatic impact on the power consumption and resiliency of computers and it has multiple uses in computing. It could enable instant-on computers where the computer can switch quickly between active mode and hibernation on a row-by-row basis. A memory array utilizing these dual floating gate devices has been designed. The physical model of these devices uses the same concepts as the model of the NCFG devices, but has been extended to account for the dual floating gate structure. The array may switch to a hibernate mode in ~30 ms, assuming 1024 rows in each memory bank. A switch from a hibernate mode to the active mode takes ~14 ms. Compared to a conventional DRAM the read-mode is fast and non-destructive. Depending on the number of rows in the array the clock frequency may be in the order of 230 MHz for 1024 rows per memory bank or 1.61 GHz for 64 rows per bank on the 45 nm technology node. In the read mode the word line is biased with 1.2 V, comparable to tomorrow's DRAM, and scales with

next generation technologies. Even though the dynamic write voltage is relatively high with 4-5 V, the overall power consumption is potentially lower due to the 4-5x longer refresh period requirement. Since the unified memory array can be quickly brought in and out of partial sleep states, power consumption can easily be further reduced. The memory array can be placed into partial hibernate states, so that when the computer is in a low activity state (which it is most of the time), its power consumption would be more proportional to its activity. Check-points could be taken very quickly, either on a regular schedule or when power down. In addition, a refresh rate of 32 ms can be achieved when operating at 105° C whereas the conventional DRAM may operate at maximal 85° C to achieve this refresh period [7.1].

The unified memory device may have a significant impact on the architectural implications in the memory hierarchy of a computer. For example, a unified main memory may reduce power consumption by dividing it into DRAM-configured rows for high speed data access and Flash-configured rows for low idle power consumption and intrinsic recoverability. The new device has also great potential for implementation in a circuit switched NoC or synaptic computer applications.

7.2 Future Research Objectives

The results and findings of this dissertation address several interesting research areas for the future work. These include (i) additional testing and improvements of the NCFG-based

FPGA and its fabrication and (ii) further investigation of the unified memory device where device, circuit and architecture issues will be solved simultaneously.

1. More FPGA designs with different functions and different complexities will be investigated, and both SRAM-based and NCFG-based FPGA designs will be compared. A controller that is capable of programming, erasing and operating NCFG devices will be designed. In addition, further investigation of novel architectures for the logic block as well as embedded memory blocks with the implementation of NCFG devices instead of SRAMs will be explored. The final step is to fabricate the three FPGA topologies for a direct experimental comparison and verification.
2. The invention of the unified memory offers several new research fields. The device will be fabricated on a scaled technology and with ultra-thin metals and inter floating gate oxide. Measurements will be performed to verify the properties of the device simulations in Sentaurus TCAD. The reliability and endurance will be tested and a deeper understanding of the trade-offs of the device will be developed. Moreover, a monolithically 3D stack using two devices will be carried out. The device will be simulated beyond the 16 nm node, using calibrated models, to investigate further scalability. On the circuit level, appropriate circuit functions will be investigated for using the unified memory device in addressable memories, embedded look-up-tables, circuit routers in NoC, synaptic processors and other applications as they emerge. Another interesting research field is the circuit design of a 3D memory array that enables reduced power consumption by minimizing wire length (and thus capacitance and power) through the 3D structure. Lastly, architectural advantages of the circuit

functions using these devices will be quantified where new ideas will improve the power efficiency and resiliency of computers.

References

- [7.1] M. Ghosh and H.-H.S. Lee, "Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs," *In: 40th Annual IEEE/ACM Int. Symp. on Microarchitecture*, pp. 134-145, 2007.