

NANOCRYSTAL MEMORY SCALING: FROM MATERIAL SELECTION TO
PERFORMANCE IMPROVEMENT

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Tuo-Hung Hou

May 2008

© 2008 Tuo-Hung Hou

NANOCRYSTAL MEMORY SCALING: FROM MATERIAL SELECTION TO PERFORMANCE IMPROVEMENT

Tuo-Hung Hou, Ph.D.

Cornell University 2008

Below the 65nm technology node, the present Flash memory technology is facing daunting scaling challenges. Smart and heterogeneous integration of materials throughout the entire device structure is required to facilitate the feature-size scaling without compromising the memory performance. The metal nanocrystal (NC) memory is promising for realizing high-density nonvolatile storage, while providing unique advantages on low-voltage operation and superior cycling lifetime. We present in this work a combined experimental and modeling study on the metal NC memory. A physical model based on the three-dimensional (3D) electrostatics and the one-dimensional (1D) Wentzel-Kramers-Brillouin (WKB) tunneling current calculation is established. The optimization strategies including NC array, gate dielectrics, and charge storage nodes are further detailed to achieve efficient program/erase (P/E) at $\pm 4V$. Beyond the metal NC memory, in the efforts of realizing hybrid molecular / Si electronics, we show programmable and quantized redox states of C_{60} molecules in a nonvolatile memory cell at room-temperature. C_{60} may also be employed as a double-junction resonant tunnel barrier in nonvolatile memories with improved tunneling asymmetry between P/E and retention.

BIOGRAPHICAL SKETCH

Born 1975 in Chia-Yi, Taiwan, Tuo-Hung Hou has got great influence from his family. He learned from his parents the importance of education, the necessity of diligence and perseverance, and the optimism amidst adversity. He also has enjoyed literature, philosophy, history, and geography since a very young age inspired by his three elder sisters. He was a brilliant and hard-working student in school, and completed his secondary education one year ahead of schedule.

Tuo-Hung made the first important yet risky decision in his young life by choosing to study electronic engineering in National Chiao Tung University (NCTU) over other medical schools, which were overwhelmingly more popular among his friends. The then 17-year-old by no means knew anything about electronic engineering except that he thought his father's job responsible for heavy-duty electric generators in a sugar plant was pretty cool. He soon realized the scale of transistors and integrated circuits are VERY different from electric generators. Fortunately, the major suited him well in the end. He developed strong interests particularly in the area of solid state devices. He graduated in 1996 with a B.S. and was the first-ranked applicant admitted to Institution of Electronics in NCTU for the graduate study. He then joined the research group of Professor Tan-Fu Lei working on high-sensitive magnetic-field sensors, thin-film transistors, and nickel silicide processes.

To experience firsthand the vibrant silicon industry in Taiwan, after receiving his M.S. degree in 1998, and another two years serving as a second lieutenant in an armed reconnaissance helicopter unit in the army, Tuo-Hung joined the advanced module technology division in Taiwan Semiconductor Manufacturing Company (TSMC) in 2000. In 2001, he was assigned to International SEMATECH, Austin, TX as a TSMC assignee working on the new-generation gate dielectric development until 2003. During his tenure with TSMC, he had contributed to several key front-end

processes such as spike annealing, low thermal budget spacer, and high- κ / metal gate stack through atomic layer chemical deposition. Moreover, he actively involved in the reliability improvement of ultra-thin gate oxide and the investigation of mobility degradation/charge trapping behaviors in high- κ dielectrics.

In 2004, after very intense four years in the goal-driven silicon industry, Tuo-Hung decided to resume his academic pursuits on the long-dreamed-of Ph.D. in the School of Electrical and Computer Engineering, Cornell University. His research at Cornell has centered on the combined theoretical and experimental study of nonvolatile-charge-based devices in heterogeneous integration with nanocrystals, fullerenes, and carbon nanotubes. His works on the design optimization of nanocrystal memory, the redox states of molecular floating gate and the molecule-embedded tunnel oxide are hoped to not only benefit the scaling of nonvolatile memory, but also shed light on the potential of hybrid molecular/Si electronics. After his graduation in May 2008, he is going to continue his academic career as a professor in Taiwan.

To my parents, Yi-Hsiung Hou and Hsiu-Ying Lee

ACKNOWLEDGMENTS

The past four years at Cornell have been splendid and memorable time in my life. Although there were times when endless homework assignments and abstruse quantum mechanics seemed too much for an old student, and times when unsuccessful experiments after sleepless efforts, harsh criticism on my papers from outside reviewers, and infamous bitter winters got me frustrated, in the end I find myself fortunate to be where I am and have enjoyed every bit of my stay here. None of these would have been possible without the help of so many wonderful people.

First, I would like to express my gratitude to my advisor, Prof. Edwin C. Kan. Over the years, I came to realize what a dedicated scientist, a passionate teacher, and a great person he is. I will always remember his patience of nurturing my research interests, his trust for developing my independence, his guidance to shape my career goal, his accommodation on my family life, and his encouragement when needed most. He is an outstanding role model of a great professor which I will endeavor to come close in coming years. It is also my great honor to have Prof. Sandip Tiwari and Prof. Sunil Bhave on my Ph.D. committee. Their inspiring questions and insightful suggestions taught me how to tackle difficult problems from fundamental concepts and fresh viewpoints.

I would like to thank Chungho Lee for his mentoring in my first year and teaching me all essential cleanroom skills. I learned firsthand how carefulness and perseverance could make big differences from a true experimentalist. His comprehensive Ph.D. research on the nanocrystal memory also laid the groundwork for my later theoretical study. I appreciate Udayan Ganguly very much for sharing all tricks in nanotube growth and fullerene deposition. All delightful chat and scientific brainstorming with him had been so much fun and kept me intellectually challenged. I also thank Hassan Raza for introducing me the beauty of quantum transport theories

and correcting my often naïve view of physics. I am grateful to my other group members: Venkat Narayanan, Jinsook Kim, Weiping Ni, and Blake Jacquot for their assistance and advice in my early days here, Jaegoo Lee for his four-year faithful friendship, Jonathan Shaw, Nini Muñoz, Keith Lyon, Daniel Ruebusch, Shantanu Rajwade, Xiaoyang Li, Fan Yu, Krishna Jayant, and Ebenezer Kwame Amponsah for their wonderful personalities and the pleasant time spending with me. I also would like to thank other fellow graduate students at Cornell, Jack Chung-Jr Huang, Chia-Chen Jason Fang, Ron-Bin Cheng, Po-Hsun Lin, Kai-Yi Chen, Jeng-Huei Chen, Yunju Sun, Jie Lu, Shriram Shivaraman, Hengky Chandrahilim, Wei Min Chan, and Chunhui Song for all their friendship and support.

I would like to acknowledge the financial support from Center for Nanoscale Systems at Cornell (CNS) for making my study possible and the staff members in Cornell Nanoscale science and technology Facilities (CNF) and CNS for facility maintenance and training. In particular, Michael Skvarla, Rob Ilic, Daniel Woodie, Garry Bordonaro, Jerry Drumheller, and Jonathan Shu are most appreciated.

Special thanks to my family for their everlasting love. Without their positive support and cheerful encouragement, I might not have faith to resume my study four years ago. My parents, Yi-Hsiung Hou and Hsiu-Ying Lee, have struggled for their whole life to provide me the most and the best they can. I am also grateful to my elder sisters, Syu-Jyuan Hou, Syu-Ling Hou, and Nai-Chi Hou, for taking good care of my parents and their sincere concern while we are oceans apart. Last but not least, my warmest thanks to my wife, Yu-Ching Chen, for sharing the bitterness and happiness in life. It has been her unconditional love, sacrifice, and company making Ithaca our sweet home even in the most miserable winter. Thanks to my son, Brian Bo-Syuan Hou. His charm and laugh have kept me entertained every day and made this study seem short.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgments	vi
Table of Contents	viii
List of Figures	xii
List of Tables	xvii
List of Abbreviations	xviii
List of Symbols	xxi
CHAPTER 1: INTRODUCTION	1
1.1 <i>Flash Memory Scaling</i>	1
1.2 <i>Material Selection in Flash Memory</i>	3
1.2.1 Charge Storage Medium	3
1.2.2 Tunnel Oxide	11
1.2.3 Control Oxide and Control Gate	13
1.2.4 Sensing Channel	16
1.2.5 Summary of Material Selection in Flash Memory	19
1.3 <i>Chapter Organization</i>	19
<i>REFERENCE</i>	21
CHAPTER 2: DESIGN OPTIMIZATION OF METAL NANOCRYSTAL MEMORY	
— PART I: NANOCRYSTAL ARRAY ENGINEERING	29
2.1 <i>Abstract</i>	29
2.2 <i>Introduction</i>	29
2.3 <i>Device Modeling</i>	31

2.3.1	Simulation Structures	31
2.3.2	3D Electrostatics and 1D Tunneling Calculation	33
2.3.3	P/E and Retention Dynamics	35
2.4	<i>Results and Discussion</i>	38
2.4.1	NC Spacing Effect	38
2.4.2	NC Diameter Effect	44
2.4.3	Figure of Merit for NC array design	46
2.5	<i>Conclusion</i>	48
	<i>REFERENCE</i>	49
CHAPTER 3: DESIGN OPTIMIZATION OF METAL NANOCRYSTAL MEMORY		
– PART II:	GATE STACK ENGINEERING	53
3.1	<i>Abstract</i>	53
3.2	<i>Introduction</i>	53
3.3	<i>Device Modeling</i>	54
3.4	<i>Results and Discussion</i>	55
3.4.1	High- κ Dielectric as Control Oxide	55
3.4.2	High- κ Dielectric as Tunnel Oxide	62
3.4.3	Figure of Merit	66
3.4.4	NC Memory Design for Low P/E Voltage Operation	67
3.5	<i>Conclusion</i>	69
	<i>REFERENCE</i>	70
CHAPTER 4: MODELING OF MULTI-LAYER NANOCRYSTAL MEMORY		
4.1	<i>Abstract</i>	73
4.2	<i>Introduction</i>	73

4.3	<i>Device Modeling</i>	75
4.4	<i>Results and Discussion</i>	77
4.4.1	Program Characteristics	77
4.4.2	Retention Characteristics	81
4.4.3	Inter-NC Oxide	84
4.4.4	Work-Function Offset of Nanocrystals	85
4.4.5	Nanocrystal Size	86
4.4.6	Multiple layer of NCs	90
4.5	<i>Conclusion</i>	91
	<i>REFERENCE</i>	93
CHAPTER 5: FERMI-LEVEL PINNING IN NANOCRYSTAL MEMORIES		...97
5.1	<i>Abstract</i>	97
5.2	<i>Introduction</i>	97
5.3	<i>Device Fabrication</i>	99
5.4	<i>Fermi-Level Pinning Theory in NC Memory</i>	99
5.5	<i>Conclusion</i>	105
	<i>REFERENCE</i>	106
CHAPTER 6: PROGRAMMABLE MOLECULAR ORBITAL STATES OF C ₆₀ FROM INTEGRATED CIRCUITS	109
6.1	<i>Abstract</i>	109
6.2	<i>Introduction</i>	109
6.3	<i>Device Fabrication</i>	110
6.4	<i>Redox States of C₆₀ Molecules</i>	111
6.5	<i>Conclusion</i>	116

<i>REFERENCE</i>	117
CHAPTER 7: NONVOLATILE MEMORY WITH MOLECULE-ENGINEERED TUNNEL BARRIERS	119
7.1 <i>Abstract</i>	119
7.2 <i>Introduction</i>	119
7.3 <i>Device Fabrication</i>	121
7.4 <i>Resonant Tunnel Barrier</i>	121
7.5 <i>Improved Tunneling Asymmetry in NC Memory</i>	127
7.6 <i>Conclusion</i>	129
<i>REFERENCE</i>	130
CHAPTER 8: CONCLUSION	133
8.1 <i>Summary of Major Contributions</i>	133
8.2 <i>Suggestions for Future Work</i>	133
<i>REFERENCE</i>	137

LIST OF FIGURES

Figure 1.1	Energy band diagram representation at P/E and retention in the nonvolatile memory cells.....	5
Figure 1.2	SEM plane-view image of Au NCs, and STEM cross-sectional image of Au nanocrystals embedded in SiO ₂	7
Figure 1.3	Cross-sectional views of the 3D electrostatic potential contours in NC memory unit cells with a metal NC and a Si NC.....	8
Figure 1.4	Calculated tunneling current from the WKB approximation for three tunnel oxide with the same 2-nm EOT.....	11
Figure 1.5	Energy band diagram representation of the crested barrier, the asymmetric layered barrier, the bandgap-engineered ONO, and the double tunnel junction.....	12
Figure 1.6	Cross-sectional views of the 3D electrostatic potential contours in the NC memory unit cell with SiO ₂ and HfO ₂ as the control oxide.....	15
Figure 1.7	Potential contour and cutline for a bottom-gate CNT FET.....	18
Figure 2.1	Schematics of the metal NC unit cell and the corresponding capacitor model.....	32
Figure 2.2	Calculated band diagrams as a function of stored electrons under program, retention, and erase in a metal NC memory.....	34
Figure 2.3	Flatband voltage shift versus program pulse time in a metal NC memory.....	38
Figure 2.4	Cross-sectional views of the 3D electrostatic potential contours in NC memory cells with the stand-alone NC and the hexagonal close-packed NC array.....	40
Figure 2.5	Electrostatic potential profiles inside the tunnel oxide along the cut line	

	of the least-action tunneling path in Fig. 2.4.....	40
Figure 2.6	Oxide electric field at the SiO ₂ /Si interface and tunneling current density into NCs as a function of the NC spacing.....	41
Figure 2.7	Single-electron charging energy and 3D channel-control factor R_{3D} as a function of the NC spacing.....	41
Figure 2.8	Flatband voltage shift versus program pulse time as a function of the NC spacing.....	43
Figure 2.9	Room-temperature retention characteristics as a function of the NC spacing.....	43
Figure 2.10	Oxide electric field at the SiO ₂ /Si interface and tunneling current density into NCs as a function of the NC diameter.....	45
Figure 2.11	Single-electron charging energy and 3D channel-control factor R_{3D} as a function of the NC diameter.....	45
Figure 2.12	Flatband voltage shift versus program pulse time as a function of the NC diameter	47
Figure 2.13	Room-temperature retention characteristics as a function of the NC diameter.....	47
Figure 2.14	Program time versus retention time for various designs of the NC array.....	48
Figure 3.1	Flatband voltage shift versus program pulse time as a function of the thickness of SiO ₂ control oxide.....	56
Figure 3.2	Single-electron charging energy and 3D channel-control factor R_{3D} as a function of the control oxide EOT and dielectric constant.....	58
Figure 3.3	Cross-sectional views of the 3D electrostatic potential contours in the NC memory unit cell with SiO ₂ and HfO ₂ as the control oxide.....	59
Figure 3.4	Flatband voltage shift versus program pulse time with SiO ₂ and HfO ₂	

	as the control oxide.....	61
Figure 3.5	Room-temperature retention characteristics with SiO ₂ and HfO ₂ as the control oxide.....	61
Figure 3.6	Flatband voltage shift versus program pulse time with SiO ₂ and SiO ₂ / HfO ₂ bi-layer as the tunnel oxide.....	63
Figure 3.7	Calculated band diagram of SiO ₂ and SiO ₂ / HfO ₂ bi-layer tunnel oxide during program.....	64
Figure 3.8	Room-temperature retention characteristics with SiO ₂ and SiO ₂ / HfO ₂ bi-layer as the tunnel oxide.....	65
Figure 3.9	Calculated band diagram of SiO ₂ and SiO ₂ / HfO ₂ bi-layer tunnel oxide at retention.....	65
Figure 3.10	Program time versus retention time for various gate stack designs...66	
Figure 3.11	P/E and retention characteristics of a scaled NC memory design operated at ± 4 V.....	68
Figure 3.12	P/E characteristics of the scaled NC memory design as a function of P/E voltage.....	68
Figure 4.1	Schematic of the unit cell of the DL NC memory, assuming azimuthal symmetry around the central axis in the z direction.....	75
Figure 4.2	Measured and calculated flatband voltage shift versus program pulse time for both SL and DL devices.....	78
Figure 4.3	Calculated energy band diagrams of the DL NC device at program along the least-action path and the corresponding evolution of charge states.....	80
Figure 4.4	Calculated room-temperature retention time for both SL and DL devices.....	81
Figure 4.5	Calculated energy band diagrams of the DL NC device at retention	

	along the least-action path and the corresponding evolution of charge states.....	82
Figure 4.6	Calculated room-temperature retention time for DL devices with variations of inter-NC oxide.....	84
Figure 4.7	Calculated program transients and room-temperature retention time for DL devices as a function of NC2 diameter.....	87
Figure 4.8	Calculated Coulomb charging energy E_{CH} and total blockade energy including quantum confinement energy E_{QM} as a function of the NC1 diameter.....	88
Figure 4.9	Calculated room-temperature retention time for DL devices as a function of NC1 diameter.....	89
Figure 4.10	Calculated room-temperature retention time for DL and quad-layer devices.....	91
Figure 5.1	High-frequency CV measurements of Au NC capacitors with Al_2O_3 control oxide and evaporated/PECVD SiO_2 control oxide.....	100
Figure 5.2	Flatband voltage shift as a function of program voltage in Au NC and C_{60} capacitors.....	103
Figure 5.3	Dipole formation at the NC–dielectric interface and the corresponding 1D energy-band diagrams of Au NC capacitors with high density of interface states.....	104
Figure 6.1	High frequency CV measurements of C_{60} capacitors with Al_2O_3 control oxide and SiO_2 control oxide.....	111
Figure 6.2	Flatband voltage shift as a function of program voltage measured at C_{60} capacitors with Al_2O_3 control oxide and SiO_2 control oxide.....	113
Figure 6.3	Energy band/level diagram representation of C_{60} capacitors with Al_2O_3 control oxide and SiO_2 control oxide, respectively.....	114

Figure 7.1	Schematics of heterogeneous gate stacks (S1-S5) examined in Chapter 7, and energy band/level diagram representation of the C ₆₀ -embedded tunnel barriers under high and low electric fields.....	122
Figure 7.2	Tunneling current and CV measurement of the C ₆₀ -embedded tunnel barrier.....	123
Figure 7.3	Temperature-dependent tunneling current measurement of the C ₆₀ -embedded tunnel barrier.....	124
Figure 7.4	High frequency CV sweeps for memory cells S3 and S4 in Chapter 7.	127
Figure 7.5	Retention and P/E characteristics of metal NC memories with the C ₆₀ -embedded tunnel barrier.	128

LIST OF TABLES

Table 1.1	Material selections in flash memory.....	20
Table 5.1	Preferable polarity of charge storage from the multiple CV measurements of metal NC memories available in the literature...	102

LIST OF ABBREVIATIONS

0D	Zero-Dimensional
1D	One-Dimensional
1T1MTJ	One-Transistor-One-Magnetic-Tunnel-Junction
1T1R	One-Transistor-One-Resistor
2D	Two-Dimensional
3D	Three-Dimensional
ALD	Atomic Layer Deposition
BER	Bit Error Rate
C_{60}^0	neutral C_{60}
C_{60}^{1-}	C_{60} Monoanion
C_{60}^{2-}	C_{60} Dianion
C_{60}^{3-}	C_{60} Trianion
CMOS	Complementary Metal-Oxide-Semiconductor
CNF	Cornell Nanoscale science and technology Facility
CNT	Carbon NanoTubes
CNS	Center for Nanoscale Systems
CR	Coupling Ratio
CNL	Charge Neutrality Level
CV	Capacitance versus Voltage
DC	Direct Current
DL	Double-Layer
DRAM	Dynamic Random Access Memory
DT	Direct Tunneling
EOT	Equivalent Oxide Thickness

FG	Floating Gate
F-P	Frenkel-Poole
FEM	Finite Element Method
FET	Field Effect Transistor
FinFET	FET with a Fin-like channel
F-N	Fowler-Nordheim
FRAM	Ferroelectric Random Access Memory
HOMO	Highest Occupied Molecular Orbital
IV	Current versus Voltage
LBL	Large Bandgap dielectric Layer
LOCOS	Local Oxidation of Silicon
LUMO	Lowest Unoccupied Molecular Orbital
MIGS	Metal-Induced Gap States
MRAM	Magnetoresistive Random Access Memory
MO	Molecular Orbital
MOS	Metal-Oxide-Semiconductor
NC	Nanocrystal
NC1	lower-layer Nanocrystal in the DL NC memory
NC2	upper-layer Nanocrystal in the DL NC memory
ONO	Oxide-Nitride-Oxide
P/E	Program/Erase
PECVD	Plasma Enhanced Chemical Vapor Deposition
PRAM	Phase-change Random Access Memory
RAM	Random Access Memory
SBL	Small Bandgap dielectric Layer
SCE	Short Channel Effect

SEM	Scanning Electron Microscopy
SILC	Stress-Induced Leakage Current
SL	Single-Layer
SoC	System-on-a-Chip
STEM	Scanning Transmission Electron Microscopy
SOI	Silicon-On-Insulator
SONOS	Silicon-Oxide-Nitride-Oxide-Silicon
SRAM	Static Random Access Memory
SRH	Schockley-Read-Hall
TANOS	TaN-Al ₂ O ₃ -Nitride-Oxide-Silicon
TEM	Transmission Electron Microscopy
TFT	Thin-Film Transistor
TMA	TriMethylAluminum
WF	Work Function
WKB	Wentzel-Kramers-Brillouin

LIST OF SYMBOLS

ε	Permittivity
ε_{conl}	Control oxide permittivity
ε_{tunl}	Tunnel oxide permittivity
ε_{NC}	NC permittivity
σ	Tunneling capture cross section
σ_{NC}	NC tunneling capture cross section
\hbar	Planck's constant
e	Elemental charge
g_i	Degeneracy of the i^{th} valley of the Si inversion layer
k	Correction constant translating the capacitance ratio to ΔV_{FB}
m^*	Electron effective mass
m_0	Free electron mass
m_{di}	Electron density-of-states effective mass in the i^{th} valley of the Si inversion layer
t_{ox}	Oxide thickness
t_{PE}	P/E time
t_R	Retention time
t_Q	Time required for the Q^{th} electron being stored in NC
Φ	Electrostatic potential
Φ_{NC}	NC work function
ΔE_{C60}	C ₆₀ potential energy change respect to the Si channel
ΔE_c	Conduction band offset between Si and dielectric barrier
ΔE_v	Valence band offset between Si and dielectric barrier
ΔV_{FB}	Flatband voltage shift

A_{NC}	NC unit cell area
C_{1_1D}	1D substrate-NC coupling capacitance using the parallel-plate approximation
C_{2_3D}	3D NC-gate coupling capacitance in the unit cell
C_{1_3D}	3D substrate-NC coupling capacitance in the unit cell
C_{FG}	Self-capacitance of the charge storage node
C_{NC}	NC unit cell diameter
D_{it}	Interface state density
D_{NC}	NC diameter
D_{NC1}	NC1 diameter
D_{NC2}	NC2 diameter
E	Electric field
E_1	Ground-state eigenenergy of the infinite square well
$E_{f,NC}$	Fermi level of the NC
$E_{f,Si}$	Fermi level of the Si substrate
E_{ij}	Energy level of the j^{th} subband in the i^{th} valley of the Si inversion layer
E_{CH}	Single-electron charging energy
E_{C60_HOMO}	C ₆₀ HOMO energy
E_{C60_LUMO}	C ₆₀ LUMO energy
E_{QM}	Quantum confinement energy
E_{Si_C}	Energy of Si conduction band edge
E_{Si_V}	Energy of Si valence band edge
J	Tunneling current density
$J_{E,NG}$	Electron tunneling current from NCs to control gate
$J_{E,NS}$	Electron tunneling current from NCs to Si substrate
$J_{E,SN}$	Electron tunneling current from the discrete 2D subbands of the Si

	inversion layer into NCs
$J_{H,SN}$	Hole tunneling current from the Si accumulation layer into NCs
$J_{net}(Q)$	Net electron current density flowing into NCs with Q electrons stored
N_{NC}	NC number density
Q	Number of charges stored in NC
Q_{NC1}	Number of charges stored in NC1
Q_{NC2}	Number of charges stored in NC2
R	Channel-control factor
R_{3D}	3D channel-control factor
S_{NC}	NC spacing
T_{conl}	Thickness of control oxide
T_{tunl}	Thickness of tunnel oxide
T_{IL}	Thickness of inter-NC oxide
T_{WKB}	WKB transmission probability
V_{th}	Threshold voltage
V_G	Gate voltage

CHAPTER 1

INTRODUCTION

1.1 Flash Memory Scaling

Battery-powered portable electronics, such as mobile phones, MP3 players, digital cameras *etc.*, have fuelled skyrocketing demand for nonvolatile Flash memory since late 1990's. The advance in technology is even more impressive. The Flash technology has demonstrated its outstanding scaling capability in the last decade. A two-fold increase in bit-density of NAND Flash has been realized every year for the past seven years [1]. Today 16-gigabit density with 50-nm design rule is in mass production. This trend far exceeds the projection of the Moore's law in logic integrated circuits. Therefore, Flash is arguably the present technology driver of the semiconductor industry. However, this great momentum, mainly relying on the straightforward geometrical shrinkage, has been expected to slow down for technology nodes of 40 nm and beyond due to several challenging roadblocks in device scaling [1]-[4]. First, the thickness of tunnel oxide is not easily scaleable in order for satisfactory charge retention, especially after many program/erase (P/E) cycles. The stress induced leakage current (SILC) gives rise to unacceptable statistical distribution in retention for a high-density memory array, which limits the thickness of tunnel oxide to be 7-8 nm [2], [3]. The non-scalable tunnel oxide deteriorates the short channel effects (SCE) and impedes further gate-length scaling. This is particularly severe in NOR-type Flash where the large drain voltage ($> 3.2\text{V}$) is necessary for hot-carrier programming. Second, the distance between adjacent floating-gates (FGs) has become extremely narrow due to aggressive scaling. As a result, the cell-to-cell interference is no longer negligible. This in part can be mitigated by reducing the FG height and by utilizing a low- κ spacer between FGs. However, these inevitably hurt

the coupling ratio (CR) necessary for decent P/E efficiency. In conventional designs, while the thickness of inter-poly oxide or so called control oxide is also reaching its scaling limit, the CR can still be engineered by the additional capacitance provided by FG sidewalls. The better immunity to the cell-to-cell interference by reducing the FG height is at the expense of the dwindling CR, and as a consequence even higher P/E voltage is required. P/E voltages are projected still at 15 V for NAND Flash until the end of roadmap in 2018 [5]. Higher P/E voltage leads to higher power dissipation and adversely affects the parallel writing process. It also adds tremendous overhead on power consumption and area of the peripheral circuit for both stand-alone and embedded memory [6]. Even more importantly, the endurance under many P/E cycles is deteriorated by the high field in the thin tunnel oxide. The resulting threshold voltage V_{th} shifting and SILC in short-retention bits are the key reliability concerns. Therefore, a fundamentally new approach to scale cell size without compromising memory performance is of great importance in Flash memory technology.

Meanwhile, there has been very active research on alternative nonvolatile memories that do not employ charge storage in FG. Among the most mature are ferroelectric random access memory (FRAM) [7], magnetoresistive random access memory (MRAM) [8], and phase-change random access memory (PRAM) [9]. Although enormous progress has been made, none of them have stood up to completely address the strict requirements for low-cost, high-density, and high-speed nonvolatile storage. FRAM relying on the charge polarization in small capacitors has limited scaling potential. Its destructive read is also undesirable. MRAM and PRAM are still under active investigation to realize an efficient P/E scheme compatible with the current drive capacity of scaled access transistors in the one-transistor-one-magnetic-tunnel-junction (1T1MTJ) and one-transistor-one-resistor (1T1R) cells. In addition, any emerging technology has to be a cost-effective replacement, a daunting

challenge considering the maturity of today's Flash technology as well as the prevalent implementation of multiple bits per cell. Therefore, it is safe to project that Flash memory will still be the main workhorse of the portable nonvolatile storage for many years to come [10]-[12]. The question is how we are able to extend its longevity by overcoming aforementioned scaling challenges before any viable alternative becoming a reality.

In an attempt to address this, in this chapter we highlight the importance of smart and heterogeneous integration of materials throughout the entire device structure, including charge storage medium, tunnel oxide, control oxide, control gate, and sensing channel. In addition to maintaining retention, in the order of importance, we need to continuously make functional density (bits/cm²) higher, cycling endurance longer, P/E voltage lower (negated by the read disturbance, multi-level possibility and noise margin), and P/E time faster (helped by inserting SRAM buffer at system interface). From both theory and experiments, we will compare the advantages and disadvantages in various material choices in view of three-dimensional (3D) electrostatics, quantum transport and CMOS process compatibility. We will limit our discussion mainly on NAND-type Flash memory owing to its better scalability and dominate role in portable massive storage. However, many viewpoints presented here may apply to NOR-type Flash as well.

1.2 Material Selection in Flash Memory

1.2.1 Charge Storage Medium

Flash memory relies on the static-charge storage in an isolated FG. The conventional choice of material for FG has been n-doped poly-Si because of its process compatibility in the Si process. However, many aforementioned scaling challenges stem from the continuous poly-Si FG. Non-scaleable thickness of the

tunnel oxide due to the poor immunity against SILC and significant cell-to-cell interference are two main inherent disadvantages. Although the industry has every reason to push the continuous FG to its limit, with recent demonstration in the 43nm-node technology [13], it is of little doubt that at some point discrete charge storage, which consists of multiple discrete FGs instead of a continuous one, has to be utilized in order to fundamentally resolve these issues [1], [2], [14]. The discreteness among FGs prevents complete loss of memory states through localized SILC, and greatly suppresses the FG-to-FG coupling. This enables both tunnel oxide and cell size scaling. Proposed device implementation is basically divided into two major categories, silicon-oxide-nitride-oxide-silicon (SONOS) or SONOS-type memories [15]-[25] and nanocrystal (NC) memories [26]-[42]. SONOS-type memories utilize natural traps in dielectrics while NC memories utilize semiconductor or metal NCs embedded inside dielectrics for charge storage. Here we are interested in the best option available to address the remaining roadblock, the high P/E voltage.

In the present Flash memory, the ratio between retention time t_R and P/E time t_{PE} is about 10^{12} - 10^{14} . In order to realize this tremendous ratio, field-asymmetric tunneling processes in the tunnel barrier have to be deliberately engineered between retention and P/E. The asymmetry in conventional Flash is most often provided by external P/E voltage. For example, in NAND Flash, the asymmetry between the Fowler-Nordheim (F-N) tunneling under P/E and the direct tunneling (DT) during retention is exploited as illustrated in Fig. 1.1(a). However, this approach also limits the scalability of P/E voltage. Metal NC memory [32]-[37] has been proposed to enhance the tunneling asymmetry at low P/E voltage. The material-dependent FG work function of metal NCs provides additional band offset to the Si band edges of the channel. During retention, only a small portion of thermally excited charge in metal NCs is able to directly tunnel back to Si channel due to the Si forbidden bandgap. This

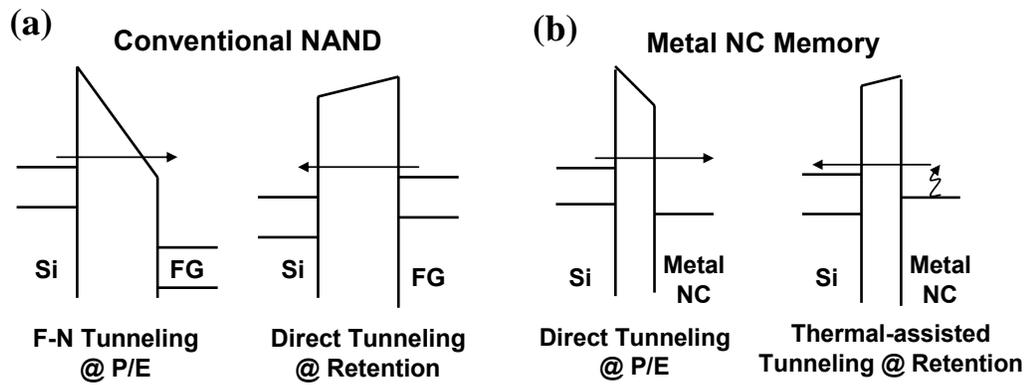


Figure 1.1 Energy band diagram representation at P/E and retention in the nonvolatile memory cells with (a) thick tunnel oxide and poly-Si FG and (b) thinner tunnel oxide and metal NC FG.

greatly improves memory retention even with a thinner tunnel oxide. Meanwhile, the thinner tunnel oxide allows fast P/E operation through DT at low P/E voltage. The asymmetry between the DT under P/E and the thermal-assisted tunneling during retention as illustrated in Fig. 1.1(b) is fundamentally different from that in the conventional NAND. On the contrary, semiconductor NCs, such as Si, Ge, and SiGe NCs [26]-[30], provides little or none band offset to the channel. The quantum-size effect of semiconductor NCs further broadens bandgap larger than that in bulk Si. In metal NCs, this bandgap broadening is suppressed by the large density of states in metal for the size of metal NCs we are generally interested in [43]. Furthermore, previous studies suggested that charge retention in semiconductor NC memories is governed by interface traps surrounding NCs with deep energy level inside the Si bandgap [44], [45]. However, this mechanism is less reliable because the interface traps are subjected to many process variations such as the backend forming gas annealing, and there is no known method to reliably engineer deep-level traps. So are true for SONOS-type memories relying on bulk traps in dielectrics. Many studies have

shown that retention at high temperature is problematic for SONOS with shallow-level traps [14], [20], [46]. In brief summary, the metal NC memory is a unique approach to further scale down the tunnel oxide without compromising retention. Therefore, the cell size scaling, low P/E voltage, and robust memory reliability may be realized simultaneously.

Electrostatics is another important consideration to achieve low P/E voltage. For better P/E efficiency, any potential drop on FGs has to be minimized especially with aggressive scaling on the thickness of tunnel and control oxide. In the conventional n-dope poly-Si FG, the poly depletion is present. In SONOS memories, the voltage drop on nitride is substantial because the nitride permittivity is only two times larger than SiO_2 and the thickness of nitride is comparable to oxide barriers. That is one of the reasons why higher- κ trap layers, such as Al_2O_3 [20], HfAlO [21], HfSiO [22], AlN [23], HfO_2 [24], and Ta_2O_5 [25], are more desirable. With the relatively higher Si permittivity and the small NC size, semiconductor NC memories seem to mitigate this concern. Nevertheless, as discussed in the later sections, the integration of high- κ dielectrics such as HfO_2 with $\kappa = 20$ as both the tunnel and control oxide makes the voltage drop on semiconductor NCs unavoidable. Therefore, metal NCs are the best option to eliminate the voltage drop with the orders of magnitude higher free electron concentration than the semiconductor counterparts. In addition, the above analysis is solely based on one-dimensional (1D) electrostatic approximation and too simplistic for NC memories because of the nature of the 3D spherical NCs and their two-dimensional (2D) placement. This is highlighted in the cross-sectional STEM and plain-view SEM in Fig. 1.2 [34]. We will discuss in greater detail about the significance of the 3D electrostatics in NC memories in Chapter 2 and Chapter 3. For instance, the field-enhancement effects around NCs [47], [48] are very substantial. For a typical design of metal NC memory, the potential drop in the tunnel

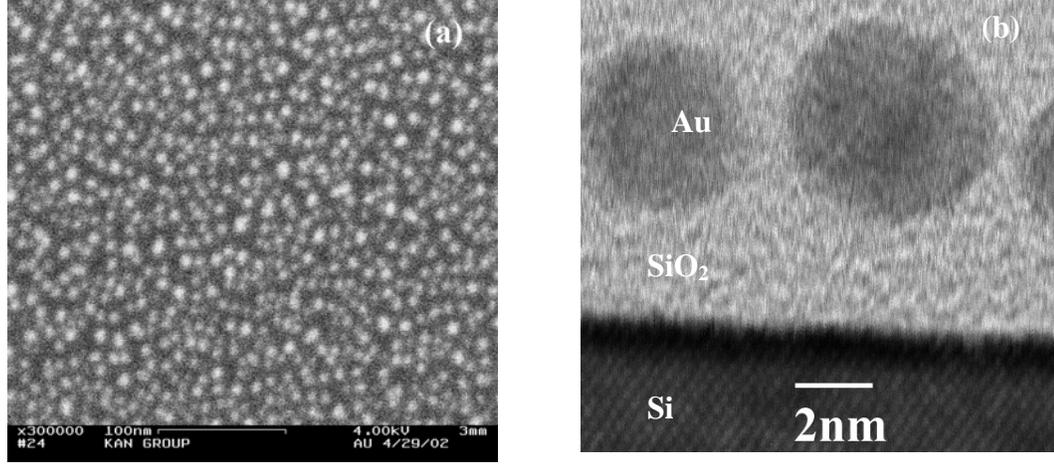


Figure 1.2 (a) A SEM plane-view image of Au NCs with area density of $4 \times 10^{11}/\text{cm}^2$, and (b) A STEM cross-sectional image of Au nanocrystals embedded in SiO_2 [34].

oxide can be more than 40% higher than that in the continuous FG memory, resulting in great improvement on the P/E efficiency. This field enhancement is subject to not only geometrical parameters, many times being able to be solved only by numerical simulation, but also the choice of materials of NC and surrounding dielectric. Considering a simplified case when the top gate, the sensing channel, and other NCs are relatively far away and a NC with charge amount of Q stored is placed in a uniform field E_0 , the analytical solution of the electric field intensity exists and can be expressed as [47], [48]:

$$E_r = E_0 \left(1 + \frac{2a^3}{r^3} \left(\frac{\epsilon_{NC} - \epsilon}{\epsilon_{NC} + 2\epsilon} \right) \right) \cos\theta + \frac{\sum_i Q_i}{4\pi\epsilon r^2} \quad (1.1)$$

$$E_\theta = -E_0 \left(1 - \frac{a^3}{r^3} \left(\frac{\epsilon_{NC} - \epsilon}{\epsilon_{NC} + 2\epsilon} \right) \right) \sin\theta \quad (1.2)$$

where the origin of the spherical coordinate (r, θ) is at the center of the NC, a is the NC radius, ϵ_{NC} is the NC permittivity, ϵ is the dielectric permittivity, and θ is the angle between r and E_0 . For a metal NC with infinite ϵ_{NC} , the field-enhancement term is

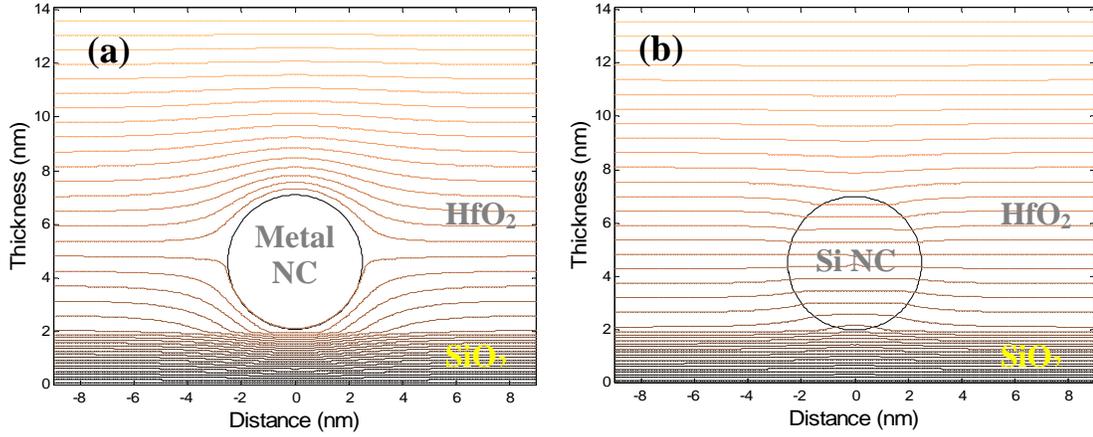


Figure 1.3 Cross-sectional view of the 3D electrostatic potential contours in NC memory unit cells with (a) a metal NC and (b) a Si NC. The NC diameter, the thickness of SiO₂ tunnel oxide, and the thickness of HfO₂ control oxide are 5 nm, 2 nm, and 7 nm, respectively. $V_G = 8$ V and no charge is stored in the NCs. The potential is monotonic from top to bottom, and the contour spacing is 0.2 V.

reduced to $(a/r)^3$ even with high- κ dielectrics. On the other hand, for a Si NC with $\epsilon_{NC} = 11.7$ embedded in SiO₂, the field-enhancement is merely $0.4 \times (a/r)^3$ and it gets less or even becomes negative when embedded in a high- κ matrix. The numerical simulation of 3D potential contours in unit cells of metal and semiconductor NC memories is shown in Fig. 1.3 with a high- κ HfO₂ control oxide. The potential drop inside the Si NC and the electric field decrease around it are in strong contrast with the unit cell of metal NC memory. Therefore, metal NCs are preferable choice over semiconductor NCs as the integration with high- κ dielectrics is inevitable for future scaled memory devices [49]. Meanwhile, due to the infinitesimal physical size of traps, the $(a/r)^3$ term vanishes. Therefore, SONOS memories remain similar to the conventional continuous FG memory without additional field-enhancement from 3D electrostatics.

Despite aforementioned advantages, the discreteness of FGs also poses

fundamental challenges in maintaining P/E efficiency. First, the charging energy arising from shrinking capacitance in the discrete FGs becomes substantial. Single-electron charging energy E_{CH} is the electrostatic energy required to store an additional electron in a small capacitor due to the Coulomb repulsion energy. It can be expressed as e^2/C_{FG} where e is the elemental charge and C_{FG} is the self-capacitance of the charge storage node from the 3D electrostatic calculation. In NC memories, C_{FG} is a strong function of the NC size, and E_{CH} increases dramatically with the NC size scaling [50]. Therefore, in a typical design with a NC diameter of 5 nm, the maximum number of charges each NC can stably hold is around 10. In the SONOS-type memory, because of the infinitesimal size of traps, each trap can hold at most one charge. As a result, to warrant sufficient memory window and P/E efficiency, the NC or trap density has to be deliberately engineered to provide charge storage capability comparable to the conventional continuous FG. One interesting example is by stacking multiple layers of NC vertically to provide additional storage capacity [33], [39], [40]. Detailed theoretical treatment on this novel configuration will be given in Chapter 4. Furthermore, small physical size of discrete FGs, also true for the extreme cases of traps, associates with small capture cross-sectional area σ during charge injection. This may adversely affect P/E efficiency. Lastly, the partial coverage of NCs over the surface of the Si channel results in less control on the channel potential, *i.e.* less memory window. A channel-control factor R between 0 and 1 is usually adopted in comparison with a continuous FG cell with R equal to 1 [48]. A smart design to boost R without increasing NC density will be further discussed later by utilizing high- κ control oxide or small sensing channel. Overall NC with its moderate size provides low E_{CH} , large σ , and sufficient R , which greatly suppress adverse effects on P/E efficiency.

From the aspect of manufacturability, controlling tight V_{th} distribution at P/E

states in a large memory array is very critical. It is a major drawback of the scaled NC memory cell as the fluctuation in the NC size and the NC number in each cell becomes substantial. However, it was projected that NC memory technology still has strong potential to scale beyond 65-nm node with current NC self assembly methods [14], [51], [52]. Recent efforts on ordered placement of NCs with controllable spacing of 3-15 nm [41], [42] may push the scaling limit even further. On the other hand, SONOS may provide better immunity to device variations owing to the large number density of traps. A heterogeneous NC/nitride stack may improve both V_{th} distribution and P/E efficiency for superior scalability [53], [54]. Both the semiconductor NC memory and the SONOS memory are fully compatible with the embedded Flash technology. They have been demonstrated for embedded applications to be fully compatible with CMOS, using even less masking steps compared with the conventional FG memory [16], [27]. High- κ trap layers and metal NCs are less compatible due to the concern of thermal stability and contamination. However, as high- κ dielectrics and metal gates become inevitable in the future Si technology, this may be less critical with better understanding and control on new material integration.

Molecules with versatile and tunable properties may find many applications in integration with traditional Si technology. In NC memory, the finite and undesired size-dispersion in NC formation could degrade the device parametric yield and impede the implementation of the multi-level cell, where different memory states are represented by different amount of charge stored. In Chapter 6, we will report replacing NCs with C_{60} molecules as the FG in a nonvolatile memory cell. The monodisperse nature of C_{60} with its molecular size results in substantial and precise step charging into molecular orbitals (MO), and hence could potentially achieve reliable multi-level charge storage utilizing different C_{60} redox states.

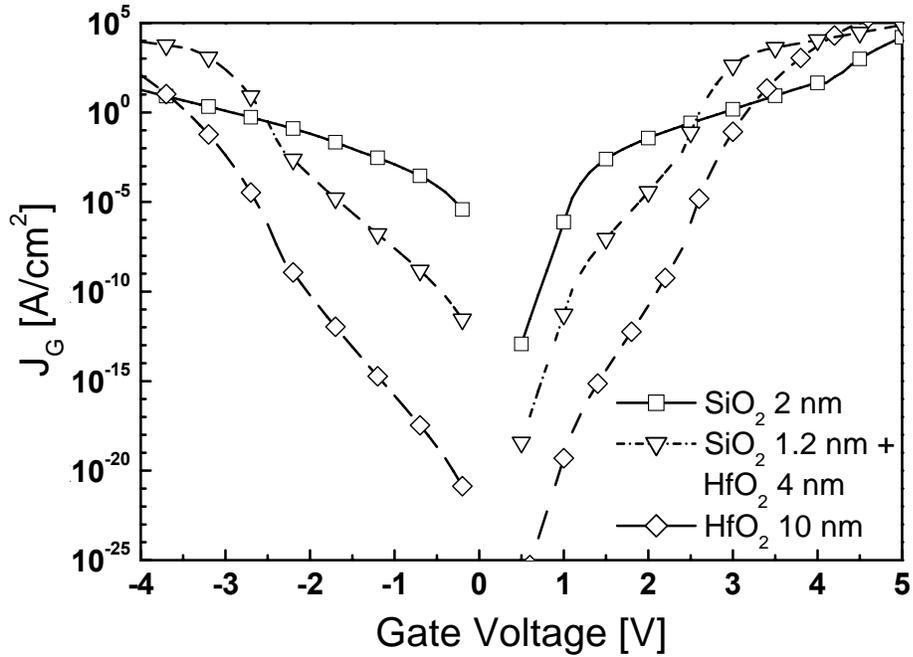


Figure 1.4 Calculated tunneling current from the WKB approximation for three tunnel oxide with the same 2-nm EOT.

1.2.2 Tunnel Oxide

Tailoring the band structure of the tunnel barrier is another effective way to achieve significant tunneling asymmetry. High- κ dielectrics with lower electron / hole barriers are better field-sensitive tunnel barriers than SiO_2 [35]. In Fig. 1.4, tunneling current calculation based on Wentzel-Kramer-Brillouin (WKB) approximation [55] is shown for SiO_2 and HfO_2 with the same 2-nm equivalent oxide thickness (EOT). The WKB approximation of transmission probability T_{WKB} at the DT regime is expressed as:

$$T_{WKB} = \exp\left(-2 \int_0^{t_{ox}} \sqrt{\frac{2m^*}{\hbar^2} (\Delta E - qF_{ox} \cdot x)} dx\right) \quad (1.3)$$

where ΔE , F_{ox} , and t_{ox} are the dielectric / Si band offset, oxide electric field and oxide thickness, respectively. dT_{WKB}/dF_{ox} suggests T_{WKB} has stronger field dependence with

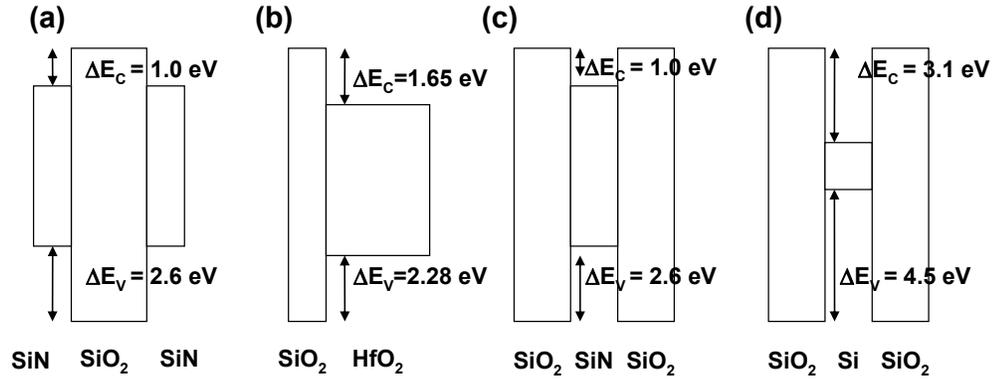


Figure 1.5 Energy band diagram representation of (a) the crested barrier, (b) the asymmetric layered barrier, (c) the bandgap-engineered ONO, and (d) the double tunnel junction.

smaller ΔE . Therefore, the current of HfO_2 has much stronger field dependence than that of SiO_2 at the DT regime. The lower transition voltage from DT to F-N due to the lower ΔE further enhances the overall asymmetry. Composite tunnel barriers with multiple layers of dielectrics such as crested tunnel barriers [29], [56] and asymmetric layered barriers [49], [57], [58] as illustrated in Fig. 1.5 are designed by the same principle. In Fig. 1.4, a 1.2-nm SiO_2 + 4-nm HfO_2 with EOT of 2 nm exhibits similar field-sensitivity as a pure HfO_2 dielectric. The interfacial SiO_2 between high- κ dielectrics and the Si channel exists at many high- κ deposition processes, and also desirable to ease the severe mobility degradation caused by the remote phonon scattering [59] and reduce the interface traps that can affect cycling endurance. More thorough investigation on the improvement of memory performance will be presented in Chapter 3.

The other class of field-sensitive tunnel barriers such as bandgap-engineered Oxide-Nitride-Oxide (ONO) [17], and double tunnel junction [19], [39] is also illustrated in Fig. 1.5. The structure consists of a small bandgap dielectric layer (SBL) sandwiched between two large bandgap dielectric layers (LBL). Resonant tunneling

through the bound states at SBL is utilized to enhance the transmission probability at high fields. However, this process is quenched at low fields with bound state energy at SBL higher than the energy of injecting carriers. The only remaining transport is the DT current through the composite LBL/SBL/LBL, which is very low. Therefore, superior t_R / t_{PE} ratio at low P/E voltage has been demonstrated at highly scaled memory cells [39]. We will further present a very simple implementation of double tunnel junction in Chapter 7 based on the monodisperse C_{60} molecules.

The employment of high- κ tunnel dielectrics is hampered by other disadvantages, such as mobility degradation in the channel and more importantly insufficient reliability caused by interface states D_{it} and dielectric traps. Transport mechanism of many high- κ dielectrics at low fields is governed by the trap-assisted tunneling or interface-state assisted tunneling. Hence the large field-asymmetry estimated from an ideal high- κ dielectric is over optimistic. Furthermore, both natural and stress-induced traps in high- κ may degrade the cycling endurance and V_{th} distribution. However, through the advance of process technology, high- κ gate dielectrics have met strict reliability requirements for future CMOS [60]. Continuous P/E voltage scaling of Flash memory may eventually make reliable high- κ tunnel oxide feasible.

1.2.3 Control Oxide and Control Gate

Under P/E, the electric field at the control oxide increases significantly with the charge build-up at the FG, and so is the inter-poly leakage current. P/E saturation occurs when the inter-poly leakage current is comparable to the injecting current from the channel. As shown in Fig. 1.4, except under very high bias, high- κ dielectrics have substantially less leakage current than SiO_2 at the same EOT due to the physical thickness. Therefore, high- κ control oxide may be exploited to reduce the inter-poly

leakage current and to increase CR simultaneously. This enables large memory window at lower P/E voltage or at faster P/E speed [61]. Combining with a metal electrode of high work function, the inter-poly current can be even further suppressed during erase [62].

In addition, spherical NCs are discretely placed on top of a 2D channel in NC memories. The coupling between NCs and the channel is subjected to 3D electrostatics. The detail of this coupling is important to determine the NC self capacitance, *i.e.* E_{CH} . It is also important to determine the channel-control factor R . Smaller E_{CH} allows more charges being stably stored in NCs, and larger R provides wider memory window where $R = 1$ representing the upper limit of a continuous FG. Both are critical in optimizing memory P/E and retention characteristics. The cross-sections of the 3D potential contours in the NC unit cell with SiO₂ and HfO₂ control oxide are plotted in Fig. 1.6. The EOT remains the same for both stacks. It is obvious that the fringing fields through HfO₂ to the Si channel are much stronger due to the higher permittivity of HfO₂. As a result, E_{CH} with HfO₂ is only a half of that with SiO₂, and R increases from 0.55 to 0.85 dramatically. In Chapter 3, we will show this leads to the increase of t_R / t_{PE} ratio by more than four orders of magnitude [49].

Fermi-level pinning is known to shift the effective gate work functions of metal/high- κ and polysilicon/high- κ gate stacks substantially [63]. Similar effects have been found critical in NC memories integrated with high- κ control oxide [64]. The effective NC work function is not only a bulk property of the NC, but also governed by the interface with the surrounding dielectric due to the formation of interface dipoles. It has to be taken into account in engineering NC work function and design optimization of NC memories. The complete Fermi-level pinning model and experimental proof will be detailed in Chapter 5.

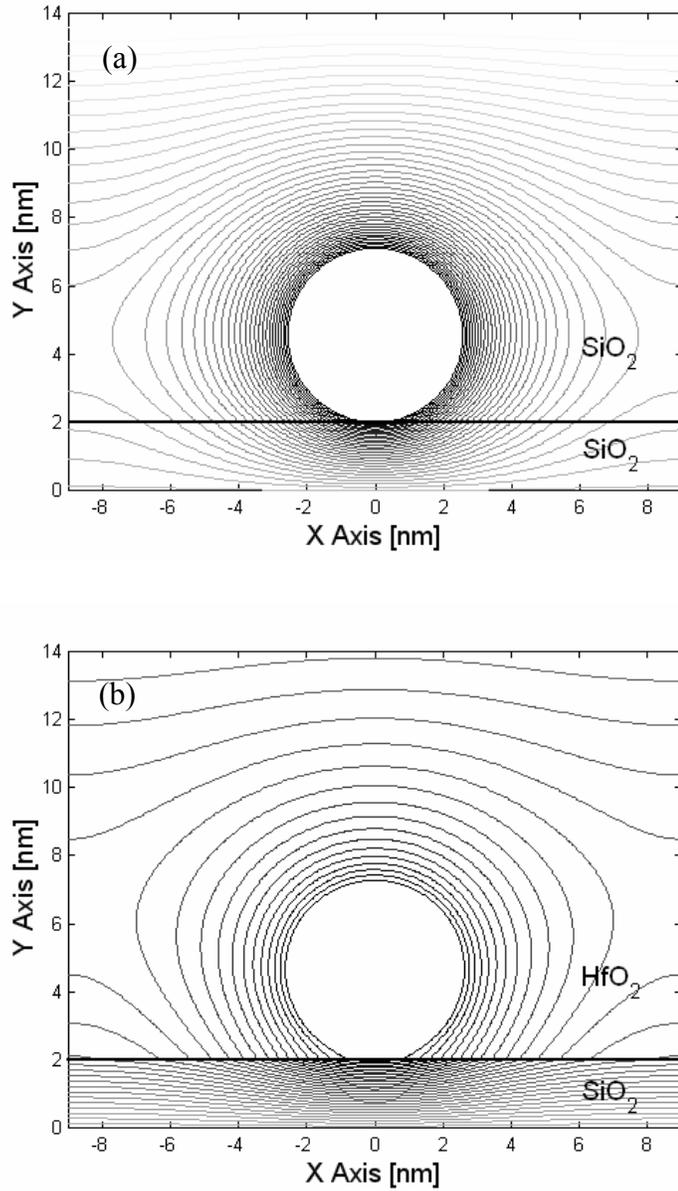


Figure 1.6 Cross-sectional view of the 3D electrostatic potential contours in the NC memory unit cell with (a) 7 nm SiO₂, and (b) 35 nm HfO₂ as the control oxide. Only part of the HfO₂ is shown in (b). The NC potential is set as -1 V while $V_G = 0$ V. The potential increases monotonically from the NC with the contour spacing of 25 mV.

1.2.4 Sensing Channel

Utilizing the field-effect transistor (FET) as a sensing channel sets Flash memory apart from other competitive nonvolatile technologies with very high sensitivity. However, the conventional planar FET on the bulk Si substrate is facing tremendous scaling challenges on its own. The short channel effects, line edge roughness, and substrate dopant fluctuation all make device variations intolerable in high-density memory array with aggressively scaled FETs. Flash memory based on novel FET structures aiming for better scalability has been under investigation [65]-[67]. In particular, FinFETs exhibits outstanding electrostatics from the 3D geometry of the control gate and the ultra-thin Si channel. Moreover, the fully-depleted ultra-thin channel warrants excellent immunity against the dopant fluctuation because V_{th} is controlled mainly by the gate work function instead of the substrate doping. The charge retention is also improved because of the floating body potential during retention. However, the same reason that prohibits FinFETs from replacing planar FETs in logic applications, high manufacturing cost, has to be first resolved.

Instead of the straight cell-size scaling, the other feasible way to increase the bit density per area is by stacking memory devices vertically [68]-[70]. This 3D staking approach has to be a cost-effective and low thermal-budge process without perturbing the characteristics of the underlying devices. The low-temperature thin-film transistor (TFT) is a promising candidate for 3D stacking, owing to its high maturity and low cost after active development for decades. High-performance control circuitry can be implemented on the Si substrate while the memory array is stacked multiple times vertically to reduce the chip size and bit-cost. Nevertheless, non-ideal subthreshold characteristics governed by high trap density at grain boundaries of TFTs remain as the major obstacles. Recently, the ultra-thin poly-Si TFT has been explored to significantly sharpen subthreshold slope and tighten V_{th} distribution by

reducing the total number of traps [70].

Another important structural variation investigated has been the 1D ultra-narrow sensing channel, where sub-10 nm channel width have been fabricated in silicon-on-insulator (SOI) structures and by nano-scale carbon nanotubes (CNT) [71], [72]. One major difference between 1D and 2D channels with discrete charge storage is that the 1D current is mainly controlled by the maximum barrier in the channel, and can be modulated by a single charge-storage node near the 1D channel, known as the “bottleneck” effect [71]. In the 2D channel case, the least resistive path controls the current in the percolation process. The bottleneck effect enhances the memory window and the single-electron sensibility of the device. Another important advantage of 1D channel is its larger CR when combined with a 2D control gate. This is essentially due to the fringing fields and highly 3D electrostatics evident in the potential contour of a back-gated CNT FET in Fig. 1.7(a) [73]. The sharp potential gradient around the CNT indicates sharp potential drop between the CNT and the NC while the much gentle potential gradient between the gate and the NC. The potential outline in Fig. 1.7(b) further reveals that a very high CR can be realized even with a thickness ratio between the control oxide and the tunnel oxide over 7. The superior electrostatic coupling together with the bottleneck effect allows very efficient P/E [72], [73]. Therefore, the narrow-channel configuration deserves serious consideration for Flash memory operated at ultra-low P/E voltage.

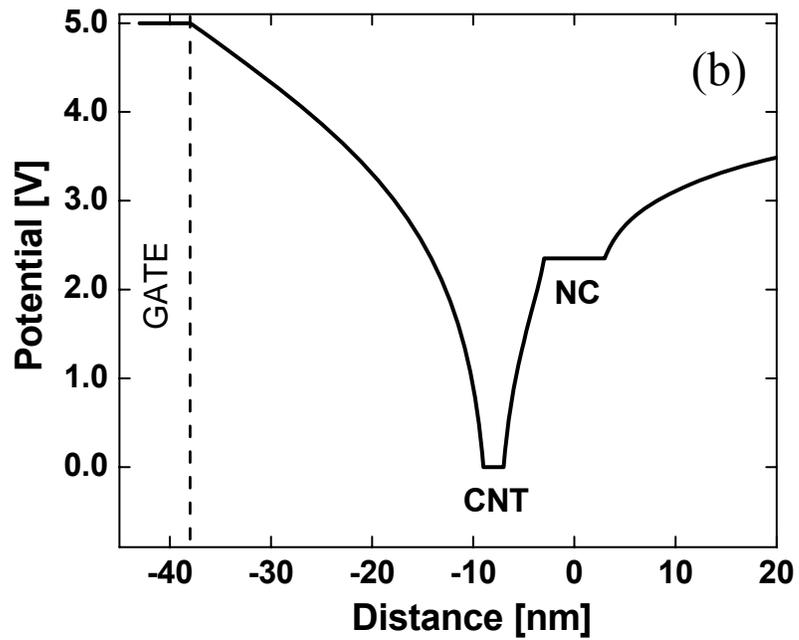
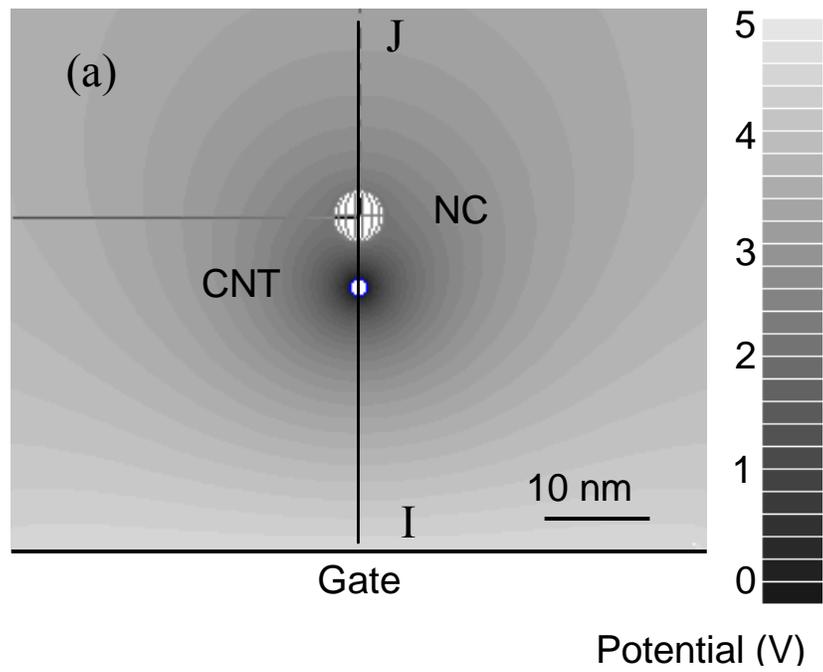


Figure 1.7 (a) Potential contour plot for a bottom-gate CNT FET with gate bias at 5V and CNT grounded, and (b) potential profile along the I-J cutline [73].

1.2.5 Summary of Material Selection in Flash Memory Cell

We have comprehensively reviewed the advantages and disadvantages in various material choices throughout the entire Flash device structure. A brief summary is provided in Table I. Despite tremendous technological challenges ahead for future FG nonvolatile memories, many viable solutions are also in sight. The heterogenous integration of new materials, such as metal NCs, high- κ trap layers, high- κ tunnel and control oxides, and metal control gates, enables new design space to increase the bit density and to optimize the memory performance at low P/E voltage. Moreover, the cost-effective 3D stacking and the innovative narrow-width channel provide new opportunities to go beyond the limitation imposed by the current planer memory architecture.

1.3 Chapter Organization

This dissertation intends to expand several critical considerations aforementioned in greater detail. Chapter 2 will present a 3D electrostatic model of the NC memory explicitly. Furthermore, the optimization strategies of NC array design including NC size and number density and gate stack design including high- κ tunnel oxide and control oxide will be highlighted in Chapter 2 and Chapter 3, respectively. Chapter 4 will introduce a generalized model for both SL and DL NC memories. Chapter 5 deals with the non-ideal Fermi-level pinning effect at the NC/high- κ interface. Moreover, in the efforts of realizing hybrid molecular / Si electronics, an interesting demonstration of utilizing C_{60} molecules as charge storage nodes will be given in Chapter 6. C_{60} molecules can be further employed in a double-junction resonant tunnel barrier as shown in Chapter 7. Finally, a summary and suggestions for future work will be discussed in Chapter 8.

Table 1.1 Material selections in flash memory

	Material	Main Advantages	Main Disadvantages	Note
Control gate	Poly-Si	– Conventional process	– Erase saturation	
	Metal	– Less erase saturation		TaN [62]
Control oxide	CVD SiO ₂	– Conventional process	– EOT ↑ & coupling ratio ↓ – Channel control of NC ↓ – Coulomb energy of NC ↑ – P/E saturation	
	High-κ	– EOT ↓ & coupling ratio ↑ – Channel control of NC ↑ – Coulomb energy of NC ↓ – Less P/E saturation		Al ₂ O ₃ [61-62], HfO ₂ [35], HfAlO [23],
Charge storage	Poly-Si film	– Conventional process	– Floating gate crosstalk ↑ – SILC immunity ↓ – ψ drop	
	Dielectric traps	– Compatible process for nitride – Number density ↑	– EOT ↑ – Capture X-section ↓ – ψ drop with high-κ barriers – Compatibility of high-κ	Nitride [15-19], Al ₂ O ₃ [20], HfAlO [21], HfSiO [22], AlN [23], HfO ₂ [24], Ta ₂ O ₅ [25]
	Semi. NC	– Compatible process	– ψ drop with high-κ barriers – Field diverge with high-κ – Small memory window – Number fluctuation	Si [26-29], Ge [30], C ₆₀ [31]
	Metal NC	– Tunable workfunction – No ψ drop – Field converge with any dielectric	– Compatibility – Number fluctuation	Au, Pt, Ag [32-34], Ni [35], NiSi ₂ [36], TiSi ₂ [37], W, Co [40], C-tube [38]
Tunnel oxide	Thermal SiO ₂	– Good endurance – Low D_{it}	– Nonscalable with retention time requirements	
	High-κ	– Improve t_R / t_{PE} ratio	– Endurance ↓	Nitride [17, 29], Al ₂ O ₃ , HfO ₂ [35], HfAlO [30]
Channel	Bulk silicon	– Conventional process	– Short-channel effect	
	SOI/Fin	– Short-channel effect ↓ – V_{th} variation ↓ – Floating at retention	– Manufacturing cost	[65-67]
	TFT	– Stackable – Floating at retention	– V_{th} variation and poor S.S. in low-T poly and a-Si TFT	[68-70]
	Nanowire/nanotube	– Coupling ratio ↑ – Memory window ↑ – Floating at retention	– Device yield ↓	Si NW [71], CNT [72]

REFERENCES

- [1] K. Kinam, J.-H. Choi, J. Choi, and H.-S. Jeong, "The future prospect of nonvolatile memory", in *Proc. VLSI Technology (VLSI-TSA-Tech)*, 2005, pp. 88-94.
- [2] R. Bez, and P. Cappelletti, "Flash memory and beyond", in *Proc. VLSI Technology (VLSI-TSA-Tech)*, 2005, pp. 84-97.
- [3] G. Atwood, "Future directions and challenges for ETox Flash memory scaling", *IEEE Trans. Device and Materials Reliability*, vol. 4, pp. 301-305, 2004.
- [4] International Technology Roadmap for Semiconductors, 2005 edition, [Online]. Available: <http://public.itrs.net/>
- [5] H. Pon, "Technology scaling impact on NOR and NAND Flash memories and their applications", in *Proc. Int. Solid-State and Integrated Circuit Technology*, 2006, pp. 697-700.
- [6] J.-H. Kim *et al.*, "Manufacturing technologies for a highly reliable 0.34 μm^2 -cell, 64 Mb, and 1T1C FRAM, in *IEDM Tech. Dig.*, 2006, pp. 45-48.
- [7] S. Tehrani, "Status and outlook of MRAM memory technology", in *IEDM Tech. Dig.*, 2006, pp. 585-588.
- [8] J. H. Oh *et al.*, "Full integration of highly manufacturable 512 PRAM based on 90nm technology," in *IEDM Tech. Dig.*, 2006, pp. 49-52.
- [9] Y. Shin, "Non-volatile memory technologies for beyond 2010," in *Proc. VLSI Circ. Dig.*, 2005, pp. 156-159.
- [10] L. Geppert, "The new indelible memories," *IEEE Spectrum*, vol. 40, pp. 48-54, 2003.
- [11] N. Flaherty, "Not in a flash," *IEE Review*, vol. 49, pp. 50-53, 2003.
- [12] M. Noguchi *et al.*, "A high-performance multi-level NAND Flash memory with 43nm-node floating-gate technology", in *IEDM Tech. Dig.*, 2007, pp. 445-448.

- [13] B. DeSalvo *et al.*, “Performance and reliability features of advanced nonvolatile memories based on discrete traps (silicon nanocrystals, SONOS)”, *IEEE Trans. Device Mater. Rel.*, vol. 4, pp. 377-389, 2004.
- [14] H. A. R. Wegener *et al.*, “The variable threshold transistor, a new electrically alterable, non destructive read-only-storage device,” in *IEDM Tech. Dig.*, 1967, p. 70.
- [15] C. T. Swift *et al.*, “An embedded 90 nm SONOS nonvolatile memory utilizing hot electron programming and uniform tunnel erase”, in *IEDM Tech. Dig.*, 2002, pp. 927- 930.
- [16] H.-T. Lue *et al.*, “BE-SONOS: A bandgap engineered SONOS with excellent performance and reliability,” in *IEDM Tech. Dig.*, 2005, pp. 547-600.
- [17] Y. Park *et al.*, “Highly manufacturable 32Gb multi-level NAND Flash memory with $0.0098 \mu\text{m}^2$ cell size using TANOS(Si-Oxide- Al_2O_3 -TaN) cell technology,” in *IEDM Tech. Dig.*, 2006, pp. 29-32.
- [18] R. Ohba, Y. Mitani, N. Sugiyama, and S. Fujita, “15 nm planar bulk SONOS-type memory with double junction tunnel layers using sub-threshold slope control, ” in *IEDM Tech. Dig.*, 2007, pp. 75-78.
- [19] T. Sugizaki, *et al.*, “Novel multi-bit SONOS type flash memory using a high- κ charge trapping layer,” in *Symp. VLSI Tech.*, 2003, pp. 27-28.
- [20] Y. Tan, W. Chim, W. Choi, M. Joo, T. Ng, and B. Cho, “High- κ HfAlO charge trapping layer in SONOS-type nonvolatile memory device for high speed operation,” in *IEDM Tech. Dig.*, 2004, pp. 889-892.
- [21] Y.-H. Lin, C.-H Chien, C.-T. Lin, C.-W. Chen, C.-Y. Chang, and T.-F. Lei, “High performance multi-bit nonvolatile HfO_2 nanocrystal memory using spinodal phase separation of hafnium silicate, in *IEDM Tech. Dig.*, 2004, pp. 1080-1082.

- [22] C. H. Lai, *et al.*, “Novel SiO₂/AlN/HfAlO/IrO₂ memory with fast erase, large ΔV_{th} and good retention,” in *Symp. VLSI Tech.*, 2005, pp. 210-211.
- [23] Y. Q. Wang *et al.*, “Fast erasing and highly reliable MONOS type memory with HfO₂ high- κ trapping layer and Si₃N₄/SiO₂ tunneling stack,” in *IEDM Tech. Dig.*, 2006, pp. 971-974.
- [24] X. Wang, J. Liu, W. Bai, and D.-L. Kwong, “A novel MONOS-type nonvolatile memory using high- κ dielectrics for improved data retention and programming speed,” *IEEE Trans. Electron Devices*, vol. 51, pp. 597-602, 2005.
- [25] S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, and D. Buchanan, “Volatile and non-volatile memories in silicon with nanocrystal storage,” in *IEDM Tech. Dig.*, 1995, pp. 521-524.
- [26] R. Muralidhar *et al.*, “A 6V embedded 90nm silicon nanocrystal nonvolatile memory,” in *IEDM Tech. Dig.*, 2003, pp. 601-604.
- [27] B. De Salvo *et al.*, “How far will silicon nanocrystals push the scaling limits of NVMs,” in *IEDM Tech. Dig.*, 2003, pp. 597–600.
- [28] S. Baik, S. Choi, U. I. Chung, and J. T. Moon, “High speed and non volatile Si nanocrystal memory for scaled flash technology using highly field-sensitive tunnel barrier,” in *IEDM Tech. Dig.*, 2003, pp. 545-548.
- [29] J. H. Chen *et al.*, “Nonvolatile flash memory device using Ge nanocrystals embedded in HfAlO high- κ tunneling and control oxides: Device fabrication and electrical performance,” *IEEE Trans. Electron Devices*, vol. 51, pp. 1840-1848, 2004.
- [30] T. H. Hou, U. Ganguly, and E. C. Kan, “Programmable molecular orbital states of C₆₀ from integrated circuits,” *Appl. Phys. Lett.*, vol. 89, 253113, 2006.

- [31] Z. Liu, C. Lee, V. Narayanan, G. Pei, and E. C. Kan, "Metal nanocrystal memories—Part I: Device design and fabrication," *IEEE Trans. Electron Devices*, vol. 49, pp. 1606–1613, 2002.
- [32] C. Lee, A. Gorur-Seetharam, and Edwin C. Kan, "Operational and reliability comparison of discrete-storage nonvolatile memories: Advantages of single- and double-layer metal nanocrystals," in *IEDM Tech. Dig.*, 2003, pp. 557-560.
- [33] C. Lee, J. Meter, V. Narayanan, and E. C. Kan, "Self-assembly of metal nanocrystal on ultra-thin oxide for nonvolatile memory applications," *J. Electronic Materials*, vol. 34, pp. 1-11, 2005.
- [34] J. J. Lee, and D.-L. Kwong, "Metal nanocrystal memory with high- κ tunneling barrier for improved data retention," *IEEE Trans. Electron Devices*, vol. 52, pp. 507-511, 2005.
- [35] P. H. Yeh *et al.*, "Fabrication of NiSi₂ nanocrystals embedded in SiO₂ with memory effect by oxidation of the amorphous Si/Ni/SiO₂ structure," *J. Vac. Sci. Technol. A*, vol. 23, pp. 851-855, 2005.
- [36] Y. Zhu, D. Zhao, R. Li, and J. Liu "Self-aligned TiSi₂/Si heteronanocrystal nonvolatile memory," *Appl. Phys. Lett.*, vol. 88, 103507, 2006.
- [37] X. B. Lu and J. Y. Dai, "Memory effects of carbon nanotubes as charge storage nodes for floating gate memory applications," *Appl. Phys. Lett.*, vol. 88, 113104, 2006.
- [38] R. Ohba, N. Sugiyama, K. Uchida, J. Koga, and A. Toriumi, "Nonvolatile Si quantum memory with self-aligned double-stacked dots," *IEEE Trans. Electron Devices*, vol. 49, pp. 1392-1398, 2002.
- [39] M. Takata *et al.*, "New non-volatile memory with extremely high density metal nano-dots," in *IEDM Tech. Dig.*, 2003, pp. 553-556.

- [40] S. Tang, C. Mao, Y. Liu, D. Q. Kelly, and S. K. Banerjee, "Nanocrystal flash memory fabricated with protein-mediated assembly," in *IEDM Tech. Dig.*, 2005, pp. 181-184.
- [41] K. W. Guarini, C. T. Black, Y. Zhang, I. V. Babich, E. M. Sikorski, and L. M. Gignac, "Low voltage, scalable nanocrystal flash memory fabricated by templated self assembly," in *IEDM Tech. Dig.*, 2003, pp. 541-544.
- [42] C.N.R. Rao, G. U. Kulkarni, P. J. Thomas, and P. P. Edwards, "Size-dependent chemistry: properties of nanocrystals," *Chem. Eur. J.*, vol. 8, pp. 29-35, 2002.
- [43] M. She, and T.-J. King, "Impact of crystal size and tunnel dielectric on semiconductor nanocrystal memory performance," *IEEE Trans. Electron Devices*, vol. 50, pp. 1934-1940, 2003.
- [44] Y. Liu, S. Tang, and S. K. Banerjee, "Tunnel oxide thickness dependence of activation energy for retention time in SiGe quantum dot flash memory," *Appl. Phys. Lett.*, vol. 88, 213504, 2006.
- [45] Y. Yang, and M. H. White, "Charge retention of scaled SONOS nonvolatile memory devices at elevated temperatures," *Solid-state Electron.*, vol. 44, pp. 949-958, 2000.
- [46] C. Lee, U. Ganguly, V. Narayanan, T.-H. Hou, and E. C. Kan, "Asymmetric electric field enhancement in nanocrystal memories," *IEEE Electron Device Lett.*, vol. 26, pp. 879-881, 2005.
- [47] T.-H Hou, C. Lee, V. Narayanan, U. Ganguly, and E. C. Kan, "Design optimization of metal nanocrystal memory—part I: nanocrystal array engineering," *IEEE Trans. Electron Devices*, vol. 53, pp. 3095-3102, 2006.
- [48] 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 Trans. Electron Devices*, vol. 53, pp. 3103-3109, 2006.

- [49] T. H. Hou, C. Lee, and E. C. Kan, "Modeling of multi-layer nanocrystal memory," in *Device Research Conf. Dig.* 2007, pp.221-222.
- [50] L. Perniola *et al.*, "Modeling of the programming window distribution in multi-nanocrystals memory," *IEEE Trans. Nanotechnol.*, vol. 2, pp. 277-284, 2003.
- [51] R. Gusmeroli, C. M. Compagnonia and A. S. Spinellia "Statistical constraints in nanocrystal memory scaling," *Microelectronic Eng.*, vol. 84, pp. 2869–2874, 2007.
- [52] S. Huang, K. Arai, K. Usami, and S. Oda, "Toward long-term retention-time single-electron-memory devices based on nitrated nanocrystalline silicon dots," *IEEE Trans. Nanotech.*, vol. 3, pp. 210-214, 2005.
- [53] C. Lee, T-H Hou, and E. Kan, "Nonvolatile memory with a metal nanocrystal/nitride heterogeneous floating-gate," *IEEE Trans. Electron Devices*, vol. 52, pp. 2697-2702, 2005.
- [54] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices," *IEEE Trans. Electron Devices*, vol. 46, pp. 1464-1471, 1999.
- [55] K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices," *Appl. Phys. Lett.* vol. 73, pp. 2137-2139, 1998.
- [56] P. Blomme, B. Govoreanu, M. Rosmeulen, J. Van Houdt, and K. DeMeyer, "VARIOT: a novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices," *IEEE Electron Device Lett.*, vol. 24, pp. 99-101, 2003.
- [57] C. M. Compagnoni, D. Ielmini, A. S. Spinelli, and A. L. Lacaita, "Optimization of threshold voltage window under tunneling program/erase in nanocrystal memories," *IEEE Trans. Electron Devices*, vol. 52, pp. 2473-2481, 2005.

- [58] M. V. Fischetti, D. A. Neumayer, and E. A. Cartier, "Effective electron mobility in Si inversion layers in metal–oxide–semiconductor systems with a high- κ insulator: The role of remote phonon scattering," *J. Appl. Phys.*, vol. 90, pp. 4587-4608, 2001.
- [59] K. Mistry *et. al*, "A 45nm logic technology with high- κ + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging," in *IEDM Tech. Dig.*, 2007, pp. 247-250.
- [60] W.-H. Lee, J. T. Clemens, R. C. Keller, and L. Manchanda, "A novel high- κ inter-poly dielectric (IPD), Al₂O₃ for low voltage/high-speed Flash memories: erasing in msec at 3.3V," in *Symp. VLSI Tech. Dig.*, 1997, pp. 117-118.
- [61] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, K. C. Park, and K. Kim, "A novel SONOS structure of SiO₂/SiN/Al₂O₃ with TaN metal gate for multi-giga bit Flash memories," in *IEDM Tech. Dig.*, 2003, pp. 613-616.
- [62] J. Robertson, "Band offsets of wide-band-gap oxides and implications for future electronic devices," *J. Vac. Sci. Technol. B*, vol. 18, pp. 1785–1791, 2000.
- [63] T. H. Hou, U. Ganguly, and E. C. Kan, "Fermi-level pinning in nanocrystal memories," *IEEE Electron Device Letters*, vol. 28, pp.103-106, 2007.
- [64] D. Burnett, D. Shum, and K. Baker "An advanced flash memory technology on SOI," in *IEDM Tech. Dig.*, 1998, pp. 983-985.
- [65] P. Xuan, M. She, B. Harteneck, A. Liddle, J. Bokor, and T.-J. King, "FinFET SONOS flash memory for embedded applications," in *IEDM Tech. Dig.*, 2003, pp. 609-613.
- [66] H. Silva, M.K. Kim, A. Kumar, U. Avci, S. Tiwari, "Few electron memories: finding the compromise between performance, variability and manufacturability at the nano-scale," in *IEDM Tech. Dig.*, 2003, pp. 271-274.

- [67] S.-M. Jung *et al.*, “Three dimensionally stacked NAND Flash memory technology using stacking single crystal Si layers on ILD and TANOS structure for beyond 30nm node,” in *IEDM Tech. Dig.*, 2006, pp. 37-40.
- [68] E.-K. Lai *et al.*, “A multi-layer stackable thin-film transistor (TFT) NAND-type Flash memory,” in *IEDM Tech. Dig.*, 2006, pp. 41-44.
- [69] Y. Fukuzumi *et al.*, “Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable Flash memory”, in *IEDM Tech. Dig.*, 2007, pp. 449-452.
- [70] M. Saitoh, E. Nagata and T. Hiramoto, "Large memory window and long charge-retention time in ultranarrow-channel silicon floating-dot memory," *Appl. Phys. Lett.*, vol. 82, pp. 1787-1789, 2003.
- [71] U. Ganguly, E. C. Kan and Y. Zhang, "Carbon nanotube-based nonvolatile memory with charge storage in metal nanocrystals," *Appl. Phys. Lett.*, vol. 87, 43108, 2005.
- [72] U. Ganguly, C. Lee, T. H. Hou, and E. C. Kan, “Enhanced electrostatics for low-voltage operations in nanocrystal based nanotube/nanowire memories,” *IEEE Trans. Nanotech.*, vol. 6, pp. 22-28, 2007.

CHAPTER 2
DESIGN OPTIMIZATION OF METAL NANOCRYSTAL MEMORY —
PART I: NANOCRYSTAL ARRAY ENGINEERING

2.1 Abstract

The three-dimensional (3D) electrostatics together with the modified WKB tunneling model has been implemented to simulate the program and retention characteristics of the metal nanocrystal (NC) memories. Good agreements with experimental data are first demonstrated to calibrate the transport parameters. In contrast to previous works, the 3D electrostatic effects investigated in this model are proven very significant in the memory operations. Therefore, new design criteria of metal NC memories are investigated. This chapter presents the physical model and the NC array design optimization. A sparse and large-size NC array, which is suitable for the 1D narrow-channel memories, provides higher program/erase tunneling current density due to the field enhancement effect, and lower charging energy due to the large NC capacitance. On the other hand, to achieve a sufficient memory window, fast program speed, and long retention time in the typical 2D channel memories, a dense and large-size NC array is favorable while taking the trade-off with the NC number density into account. Based on the same theoretical model, we continue in Chapter 3 to consider the design optimization when high- κ dielectrics are employed.

2.2 Introduction

While the conventional Flash memories become the fastest-growing segment of the semiconductor industry in recent years, severe scaling challenges are in sight [1]. Most noticeably, the program/erase (P/E) voltages do not scale with the technology

nodes because of the non-scalable thickness of tunnel oxide from data retention requirements. P/E voltages are projected still at 15V for NAND Flash until the end of roadmap in 2018 [2]. As a result, low-power operations in portable devices and embedded nonvolatile memories for system-on-a-chip (SoC) integration are very difficult to implement. Although many alternative nonvolatile technologies were heavily investigated, the nanocrystal (NC) memory remains among the most promising because of its compatibility to the current Flash technology [3]. In particular, metal NC was proposed above its semiconductor counterpart due to the selectable work function and large density of states [4]-[8].

Although theoretical models of NC memories have been investigated by several groups [9]-[12], only one-dimensional (1D) electrostatic models were employed to approximate the potential profile in the memory cell, despite the very nature of the three-dimensional (3D) spherical NCs and their two-dimensional (2D) arrayed distribution. Recently, we have reported asymmetric electric field enhancement effects between the NCs and the sensing channel as a result of the 3D electrostatics [13]. It is particularly significant in the metal NC memories and results in an improved P/E operation with low voltages while maintaining satisfactory retention properties. In addition, the NC charging energy and the channel-control factor [10], [12], which can only be accurately determined from the 3D electrostatic model, have not yet been quantitatively addressed. Without including these 3D electrostatic effects, the scaling models of metal and semiconductor NC memories can be quite misleading.

We developed a physical model utilizing the 3D numerical solution to electrostatics and the modified 1D Wentzel-Kramers-Brillouin (WKB) approximation to calculate the time-dependent tunneling current density in the metal NC memories. Moreover, with the channel-control factor extracted from the electrostatic solution, an

improved expression of the flatband voltage shift ΔV_{FB} is introduced for the memory window approximation. Then the transport model parameters such as the tunneling capture cross section and oxide effective mass are calibrated with experimental data. Finally, because of the geometrical dependence of the 3D electrostatics, the design criteria of the metal NC array in terms of the NC spacing and diameter are comprehensively examined based on the program and retention characteristics. The results suggest that a sufficient memory window, low P/E voltage, high speed, and long retention time can be simultaneously achieved by an appropriate design of NC array in aggressively scaled nonvolatile memories. Chapter 3 will then elaborate further design optimization with high- κ dielectrics.

2.3 *Device Modeling*

2.3.1 *Simulation Structures*

Figure 2.1 illustrates the schematics of the NC cells considered in the simulation. A p-type (100) Si substrate with a doping level of $2 \times 10^{17} \text{ cm}^{-3}$ and an initial flatband voltage $V_{FB} = 0$ before program are assumed. Spherical metal NCs [14] are embedded in a high-quality trap-free SiO_2 . The NC work function is set as 5.1 eV, which can be tuned by selecting different metal species [4], [14]. Both the stand-alone NC and the hexagonal-packed NC array were investigated to study the NC distribution effects. Recent efforts on ordered placement of NCs with controllable spacing of 3-15 nm have enabled this engineering design space [8], [15], and even the spontaneous NC formation on predefined 30-90 nm cell areas can yield reasonable control and acceptable number fluctuations [16]. The NC diameter D_{NC} is set larger than 4 nm so that the NC quantum confinement energy is minimized. The NC spacing S_{NC} is set larger than 5 nm so that lateral tunneling between NCs is negligible. The NC unit cell diameter C_{NC} is the sum of the NC diameter D_{NC} and the NC spacing S_{NC} . The NC

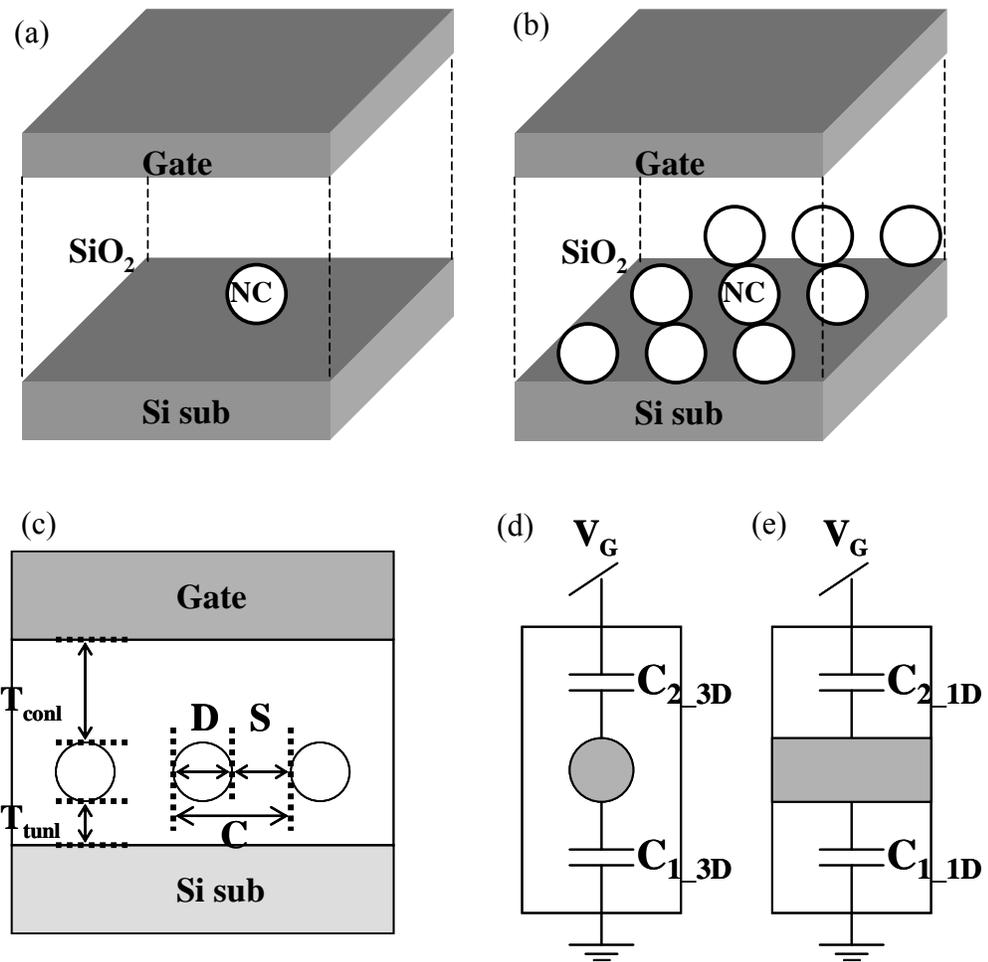


Figure 2.1 Schematics of the metal NC cell. (a) 3D geometry with the stand-alone NC. (b) 3D geometry with the hexagonal-packed NC array. (c) The corresponding cross-section view of (b). (d) The capacitor model in the unit cell of NC memories. (e) The capacitor model in the conventional floating gate memories.

unit cell area A_{NC} is inversely proportional to the NC number density N_{NC} . In addition, the thickness of control oxide T_{conl} between the NCs and the control gate is much thicker than the thickness of tunnel oxide T_{tunl} to ensure that NCs mainly interact with the channel carriers.

2.3.2 3D Electrostatics and 1D Tunneling Calculation

Numerical solution of the 3D electrostatic potential profile inside the gate stack is developed by solving the 3D Poisson equation with the finite element method (FEM). Nevertheless, the electrostatics in the Si substrate is still treated by the 1D approximation to save the computational effort, given the fact that the 3D electric field enhancement diminishes quickly with the distance from metal NCs, and Si has higher permittivity than SiO₂ [13]. The electron energy quantization in the inversion layer is treated by the triangular well approximation [17]. The energy levels and carrier concentrations in all 2D subbands are calculated. For the tunneling current model, because of the 3D nature of our simulated structure, it is more appropriate to use a 3D tunneling formalism, which is however not established in the present literature with realistic geometries. In this model, we opt to employ a 1D tunneling model considering only the least-action tunneling path along the NC central axis where the tunneling current density is the largest [11], [12]. Figure 2.2 presents the calculated band diagrams along the least-action path under the P/E and retention operations, respectively. All tunneling components considered in this work, including the NC coupling with the channel and the control gate, are also indicated. The electron tunneling current $J_{E,SN}$ from the discrete 2D subbands of the Si inversion layer into NCs can be expressed as [17], [18]:

$$J_{E,SN} = \sum_{ij} \frac{qk_B T}{\pi^2 \hbar^3} \left[\frac{g_i m_{di} E_{ij}}{j} \right] T_{WKB}(E_{ij}) \ln \left[1 + \exp \left(\frac{E_{f,Si} - E_{ij}}{k_B T} \right) \right] \quad (2.1)$$

where $i=1,2$ $j=1,2,3,\dots$, g_i is the degeneracy of the i^{th} valley, m_{di} is the electron density-of-states effective mass in the i^{th} valley, $E_{f,Si}$ is the Fermi level in the Si substrate, and E_{ij} is the energy level of the j^{th} subband in the i^{th} valley. $T_{WKB}(E_{ij})$ is the modified WKB transmission probability of an incident electron with energy E_{ij} by

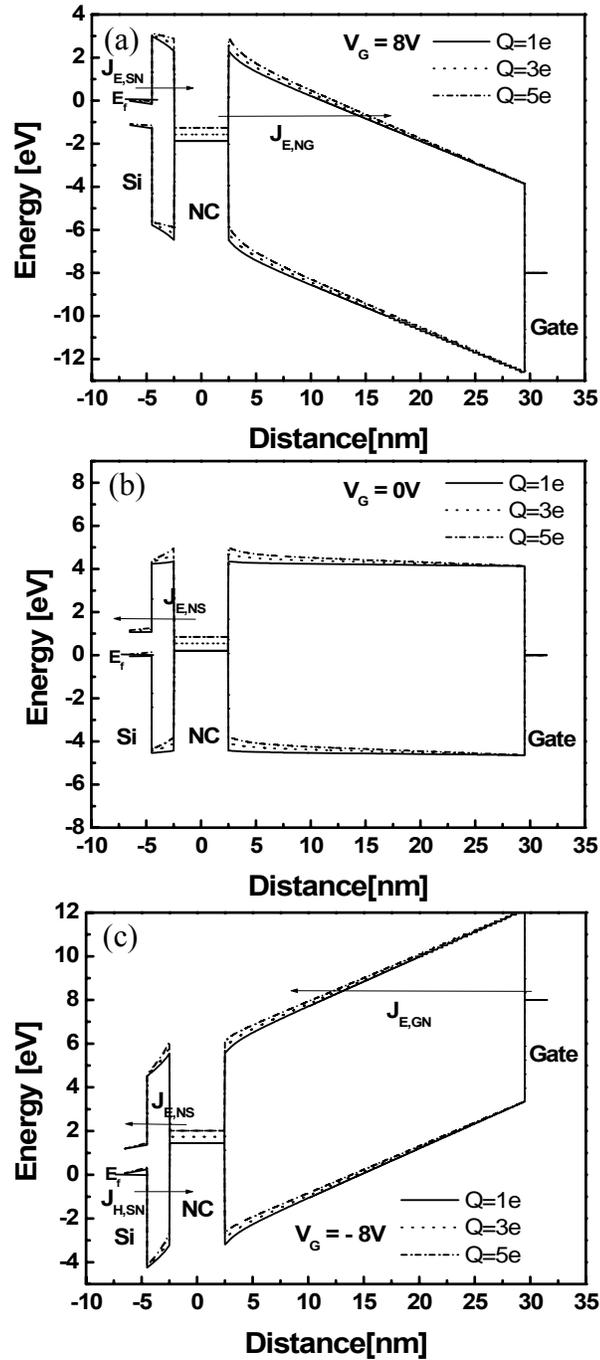


Figure 2.2 Calculated band diagrams as a function of stored electrons under (a) program ($V_G = 8V$), (b) retention ($V_G = 0V$), and (c) erase ($V_G = -8V$) conditions. All tunneling components considered are indicated. The Fermi-level of the Si substrate is referred at 0 eV. Parameters for simulation: $T_{tunl} = 2$ nm, $T_{cont} = 27$ nm, $D_{NC} = 5$ nm, and $S_{NC} = 13$ nm.

taking into account the boundary reflections. Similar expression is used for the hole tunneling current $J_{H,SN}$ from the Si accumulation layer into NCs. Other tunneling components are treated by the classical tunneling model. For example, the electron tunneling current $J_{E,NS}$ from NCs back to the Si substrate is given by [19]:

$$J_{E,NS} = \frac{qm^*k_B T}{2\pi^2\hbar^3} \int_{E_{Si_C}}^{\infty} T(E) \ln \left(\frac{1 + \exp\left(\frac{E_{f,NC} - E}{k_B T}\right)}{1 + \exp\left(\frac{E_{f,Si} - E}{k_B T}\right)} \right) dE \quad (2.2)$$

where E_{Si_C} and $E_{f,NC}$ are the Si conduction band edge and the NC Fermi level, respectively. m^* is the electron density-of-states effective mass in NCs. Other non-ideal transport mechanisms, such as the Frenkel-Poole emission and the interface-state assisted tunneling [20] are not considered in this study.

2.3.3 P/E and Retention Dynamics

Figure 2.2 also shows the dependence of electrostatic potential profiles on the number of charges stored in NCs due to the Coulomb blockade effect. Hence, the tunneling current has to be evaluated as a function of the stored charge in NCs. The evolution of the number of charges is the consequence of the net current flow between injection and emission. Therefore, the time required for the Q^{th} electron being stored in NCs, t_Q , can be written as:

$$t_Q = t_{(Q-1)} + \frac{e}{\sigma_{NC} J_{net}(Q-1)} \quad (2.3)$$

where e is the elementary charge, $J_{net}(Q-1)$ is the net electron current density flowing into NCs with $(Q-1)$ electrons stored, and σ_{NC} is the NC tunneling capture cross section, a fitting parameter from experimental calibration. The uncertainty from the

tunneling statistics is supposed to be negligible, given by the energy-time uncertainty principle:

$$\Delta E \cdot \Delta t \geq \frac{\hbar}{2} \quad (2.4)$$

For uncertainty in energy of milli-eV in typical tunneling current calculation, the uncertainty in time for a tunneling process is less than 10^{-12} sec, well below the time scale we consider in the memory operation. Two additional features of NC memories as a result of the Coulomb blockade effect are worth mentioning. First, the charge build-up during program will reach self-saturation as the Si channel departing from strong inversion, when any practical program process ceases due to insufficient electron concentration in the channel. Second, the charges cannot be effectively stored in NCs once the NC potential is raised higher than the Si conduction band edge, owing to the large direct tunneling current. This also suggests the metal NC work function is relevant to the retention characteristics [14].

Knowing the time-dependent charge state in NCs, the P/E and retention characteristics can be readily derived with an appropriate flatband or threshold voltage model. In the NC memories, due to the discrete nature of the NC array, rigorous calculation of ΔV_{FB} under the influence of charges in NCs can only be obtained by evaluating the entire 3D potential profile in the Si substrate, which is very computationally intensive. To simplify the problem without losing the essential 3D electrostatics, a semi-empirical model is adapted to describe ΔV_{FB} [10], [12] as

$$\Delta V_{FB} = R \cdot \frac{e \cdot Q}{C_{2_3D}} \quad (2.5)$$

where Q is the number of electrons stored in the NC, C_{2_3D} is the 3D NC-gate coupling capacitance in the unit cell [see Fig. 2.1 (d)], and R is the channel-control factor considering the partial coverage of NCs over the surface of the Si channel. In other

words, the charges in NCs can only perturb the channel potential in a smaller effective coverage area less than the NC unit cell area A_{NC} , but significantly larger than the NC cross-section area due to the 3D fringing effect. Instead of treating R completely as a fitting parameter [10], [12], which makes the simulation with various geometrical setups questionable, the effective coverage area is estimated using the effective capacitive coupling between NCs and the substrate. Thus, R can be expressed as

$$R = k \cdot \frac{C_{1_3D}}{C_{1_1D}} \quad (2.6)$$

where C_{1_3D} is the 3D substrate-NC coupling capacitance, C_{1_1D} is the 1D substrate-NC coupling capacitance using the parallel-plate approximation [see Fig. 2.1 (e)], and k is a correction constant translating the capacitance ratio to ΔV_{FB} , which is very close to unity in our simulation. Therefore, ΔV_{FB} can be rewritten as

$$\Delta V_{FB} = k \cdot \frac{C_{1_3D}}{C_{1_1D}} \frac{e \cdot Q}{C_{2_3D}} = k \cdot \left(\frac{C_{1_3D}}{C_{1_1D}} / \frac{C_{2_3D}}{C_{2_1D}} \right) \cdot \frac{e \cdot Q}{C_{2_1D}} = R_{3D} \cdot \frac{e \cdot Q}{C_{2_1D}} \quad (2.7)$$

$$C_{2_1D} = \frac{\epsilon_{conl}}{T_{conl}} \cdot A_{NC} = \frac{\epsilon_{conl}}{T_{conl}} \cdot \frac{1}{N_{NC}} \quad (2.8)$$

where R_{3D} is the 3D channel-control factor and ϵ_{conl} is the permittivity of the control oxide. The 3D coupling capacitances, C_{1_3D} and C_{2_3D} , are solved numerically from our 3D model. Therefore, the conventional 1D continuous floating gate model can be used to describe ΔV_{FB} with the correction of R_{3D} summing all 3D electrostatic effects from the discrete NCs.

2.3.4 Model Validation

In Fig. 2.3, good agreements can be obtained between the simulated and experimental program transients in an Au metal NC memory by using a single value

of capture cross section σ_{NC} . Details of the device fabrication were similar to those in [5], [14]. The simulation parameters, $T_{tunnel} = 2$ nm, $T_{cont} = 27$ nm, $D_{NC} = 5$ nm, $S_{NC} = 13$ nm, $C_{NC} = 18$ nm, $N_{NC} = 4 \times 10^{11}$ cm $^{-2}$, and $\sigma_{NC} = 5.3 \times 10^{-14}$ cm 2 per NC, about a quarter of the NC cross-section area, are fairly close to estimation from various types of physical characterization in the given sample [14], which validates the accuracy of our formalism. Unless otherwise mentioned, the same set of parameters is used throughout this chapter to be consistent.

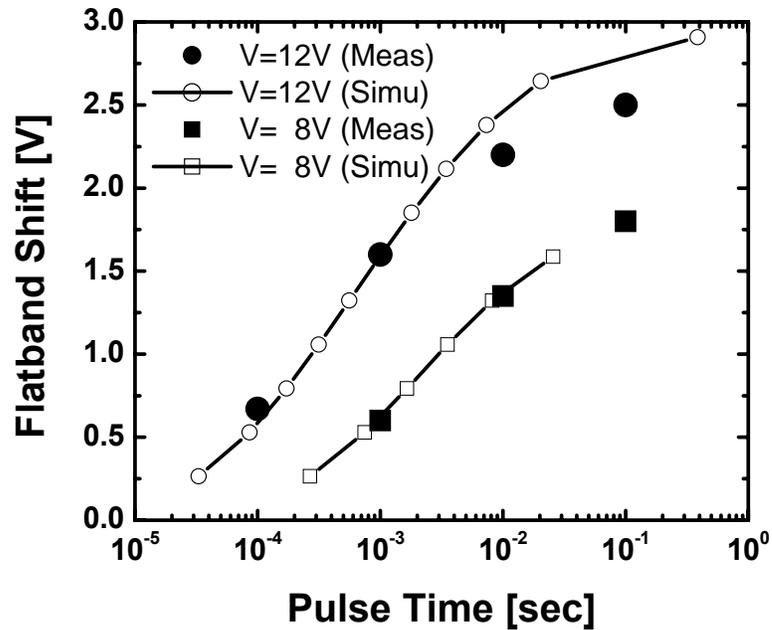


Figure 2.3 Flatband voltage shift versus program pulse time in a metal NC memory. Both simulated and experimental data are shown for $V_G = 8$ V and $V_G = 12$ V with a single set of parameters.

2.4 Results and Discussion

2.4.1 NC Spacing Effect

The cross-sectional view of the 3D electrostatic potential contours in the NC memory cells with the stand-alone NC and the hexagonal-packed NC array are shown in Fig. 2.4, respectively. The potential contours between the NCs and the Si channel are less crowded in Fig 2.4 (b) because of the influence from the adjacent NCs. The

horizontal electric field is forced to be zero at the midpoint of two NCs as a result of the symmetric requirement in the geometry. This additional boundary condition contributes to the less electric field enhancement. Figure 2.5 plots the electrostatic potential profile inside the tunnel oxide along the NC central axis, which is the dominant least-action tunneling path. A case of homogenous stack without NCs is also plotted for comparison, where the equivalent oxide thickness (EOT) remains the same as $T_{cont} + T_{tunl}$ in the homogenous stack. Therefore, the larger potential drop in the tunnel oxide of metal NC memory cells solely comes from the field enhancement effect [13]. Moreover, the influence of adjacent NCs is also observed in Fig. 2.5.

The oxide electric field at the SiO_2/Si interface and the tunneling current density into NCs are shown in Fig. 2.6 as a function of the NC spacing while the NC diameter is fixed. The simulation was done with no charge stored in the NCs, which represents the beginning of the program operation. Both the electric field and the tunneling current density exhibit a similar trend. The field enhancement becomes larger with the NC spacing, but gradually saturates when the adjacent NCs are too far away to significantly disturb the potential near the NC under consideration. Accordingly, it is rational to expect that only an adequate range of NC spacing can best exploit this enhancement for a faster program speed. Besides the higher tunneling current density, another advantage of a sparser NC array is that the single-electron charging energy due to the Coulomb blockade effect is smaller as shown in Fig. 2.7. This can be explained by the larger C_{I_3D} and C_{2_3D} due to the larger NC unit cell area. The increase of the NC self capacitance, $C_{FG} = C_{I_3D} + C_{2_3D}$, reduces the single-electron charging energy, e^2/C_{FG} . As a consequence, more electrons can be injected during a fixed program time. This also helps the retention characteristics by maintaining a higher band offset between the NC Fermi level and the Si conduction band with the same number of electrons stored.

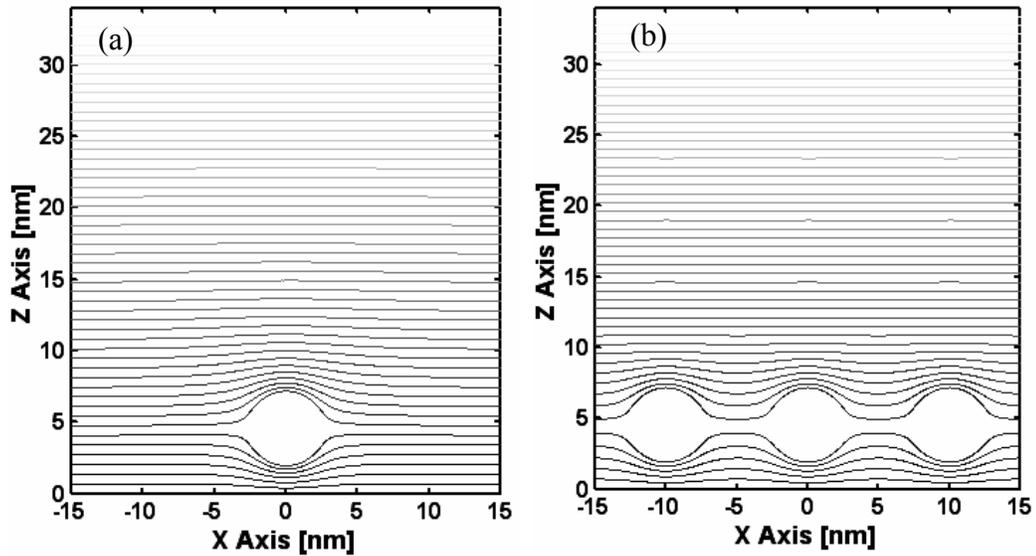


Figure 2.4 Cross-sectional views of the 3D electrostatic potential contours in the SiO_2 of NC memory cells with (a) the stand-alone NC, and (b) the hexagonal close-packed NC array. The NC diameter is 5 nm in both cases, and the NC spacing is 5 nm in the array. $V_G = 8$ V and no charge stored in the NCs. The potential is monotonic from top to bottom, and the contour spacing is 0.14 V.

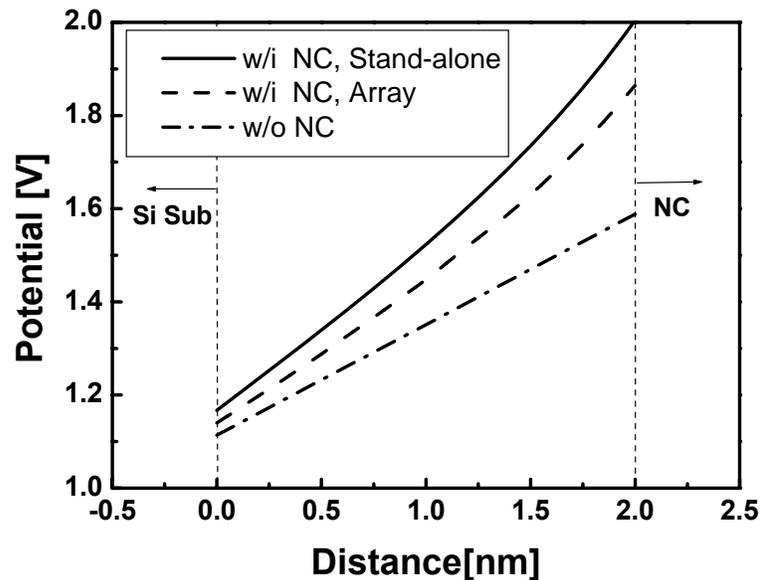


Figure 2.5 Electrostatic potential profiles inside the tunnel oxide along the cut line of the least-action tunneling path in Fig. 2.4. A reference of a homogeneous stack without NC at the same EOT is also shown for comparison.

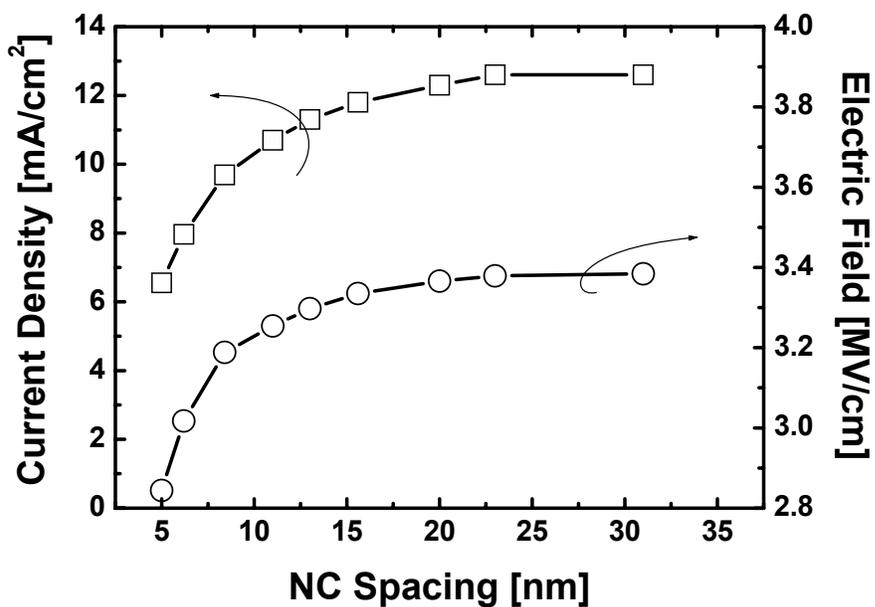


Figure 2.6 Oxide electric field at the SiO₂/Si interface and tunneling current density into NCs as a function of the NC spacing. The NC diameter is fixed at 5 nm, $V_G = 8$ V, and no charge stored in the NCs.

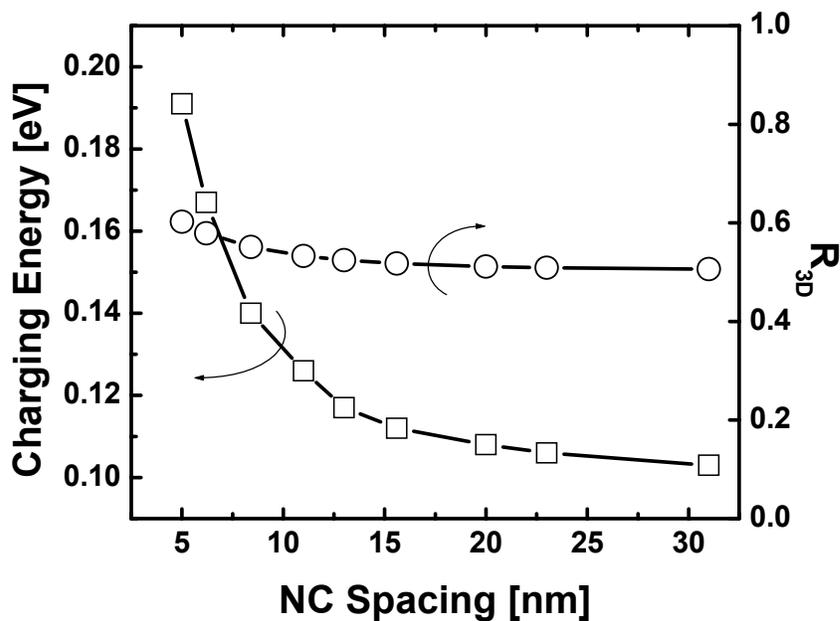


Figure 2.7 Single-electron charging energy and 3D channel-control factor R_{3D} as a function of the NC spacing. The NC diameter is fixed at 5 nm.

In Fig. 2.7, R_{3D} is a relatively weak function of the NC spacing because both $C_{1_{3D}} / C_{1_{1D}}$ and $C_{2_{3D}} / C_{2_{1D}}$ ratios vary simultaneously. According to Eqs. (2.7) and (2.8), ΔV_{FB} is then mainly determined by N_{NC} . Although the advantages of a sparser NC array are evident in terms of the tunneling efficiency, in the typical NC memories of 2D channels we study here, they are significantly set back by the cost of small memory window as shown in Fig. 2.8, where ΔV_{FB} is calculated as a function of program pulse time. In contrast, for a memory where the current flowing in the sensing channel is mainly controlled by the local maximum barrier (the bottle-neck effect), such as those in 1D narrow-channel memories [21], [22], the adverse effect of low NC number density on the memory window is much reduced. Therefore, a sparse NC array design is preferable to improve the program speed.

In Fig. 2.9, the retention characteristics are also investigated as a function of the NC spacing. To ensure program saturation, 1s program pulse at $V_G = 8$ V is assumed prior to the retention period with $V_G = 0$ V. Owing to the larger ΔV_{FB} after program, the device with a denser NC array still maintains a larger memory window after the desired 10-year lifetime, even with an apparently steeper degradation slope. The steeper slope is because the large NC number density amplifies any change of electrons stored per NC to ΔV_{FB} . To obtain a more stable memory window through the 10-year lifetime, one can choose a shorter program time, *i.e.* achieving a smaller but still appreciable memory window with fewer electrons stored per NC, which leads to larger band offset between the NC Fermi level and the Si conduction band. Therefore, both fast program and long/stable retention are possible with a dense NC array in the 2D channel NC memories. This design also assures less fluctuation of the memory window across devices when going toward smaller and smaller cell areas [23]. However, the lower bound of NC spacing should be set such that enough reliability margins are guaranteed by stopping the lateral tunneling between NCs.

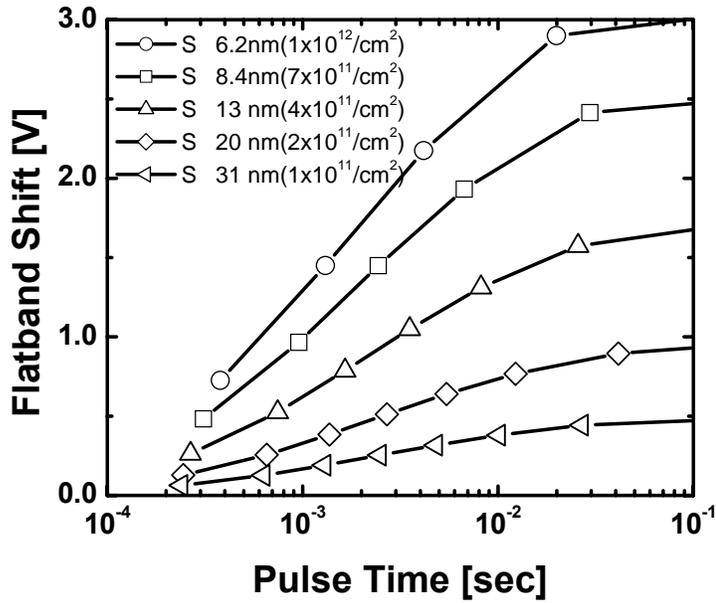


Figure 2.8 Flatband voltage shift versus program pulse time as a function of the NC spacing S_{NC} at $V_G = 8$ V and $D_{NC} = 5$ nm. Each point in the plot represents an additional electron stored per NC.

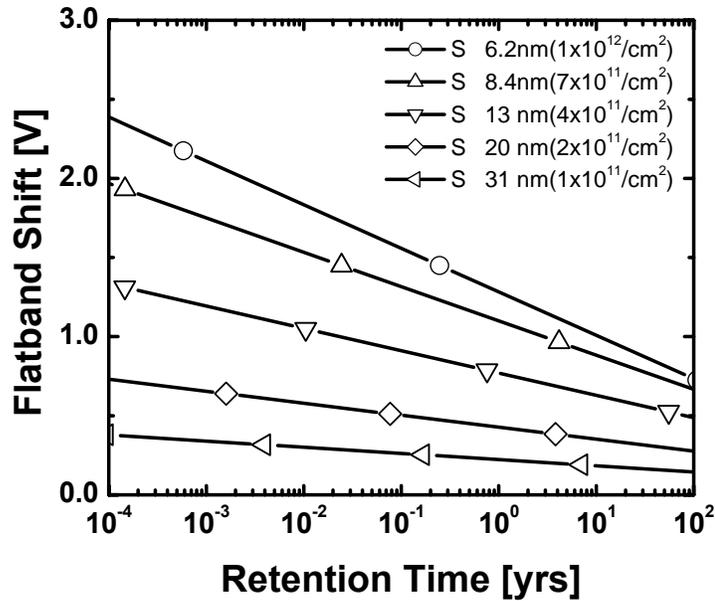


Figure 2.9 Room-temperature retention characteristics at $V_G = 0$ V and $D_{NC} = 5$ nm as a function of the NC spacing S_{NC} . 1s program pulse at $V_G = 8$ V is applied prior to the retention period to ensure program saturation. Each point in the plot represents an additional electron loss per NC.

2.4.2 NC Diameter Effect

A larger NC diameter was suggested beneficial for the program and retention characteristics [11]. Nevertheless, it was solely attributed to the smaller single-electron charging energy due to the larger NC capacitance. The significance of the 3D electrostatic effect was not fully discussed. From the analytical electrostatic solution of placing a NC in a uniform field E_0 , which represents the case when the top gate, the sensing channel, and other NCs are relatively far away, the disturbed electrostatic potential Φ and fields E with charge amount of Q stored in the NC can be expressed as [13], [24]:

$$\Phi = -E_0 \left(r - \frac{a^3}{r^2} \left(\frac{\epsilon_{NC} - \epsilon}{\epsilon_{NC} + 2\epsilon} \right) \right) \cos\theta + \frac{\sum_i Q_i}{4\pi\epsilon r} \quad (2.9)$$

$$E_r = E_0 \left(1 + \frac{2a^3}{r^3} \left(\frac{\epsilon_{NC} - \epsilon}{\epsilon_{NC} + 2\epsilon} \right) \right) \cos\theta + \frac{\sum_i Q_i}{4\pi\epsilon r^2} \quad (2.10)$$

$$E_\theta = -E_0 \left(1 - \frac{a^3}{r^3} \left(\frac{\epsilon_{NC} - \epsilon}{\epsilon_{NC} + 2\epsilon} \right) \right) \sin\theta \quad (2.11)$$

where the origin of the spherical coordinate (r, θ) is at the center of the NC, a is the NC radius, ϵ_{NC} is the NC permittivity, ϵ is the dielectric permittivity, and θ is the angle between r and E_0 . Equations (2.10) and (2.11) suggest that the field enhancement be a cubic function of the a/r ratio. This field focusing effect has been widely used in many other applications such as the optical tweezers [25]. Even with all boundary conditions presented in our structure, the field enhancement with the increasing NC diameter in Fig. 2.10 is still significant, but gradually saturates when the NC diameter close to the NC unit cell diameter. The saturation is explained by the influence from adjacent NCs mentioned above. Notice that total EOT along the central axis of NCs and NC number density remain the same while the NC diameter varies in this analysis. The single-electron charging energy and R_{3D} are plotted in Fig. 2.11. R_{3D} is not sensitive to the

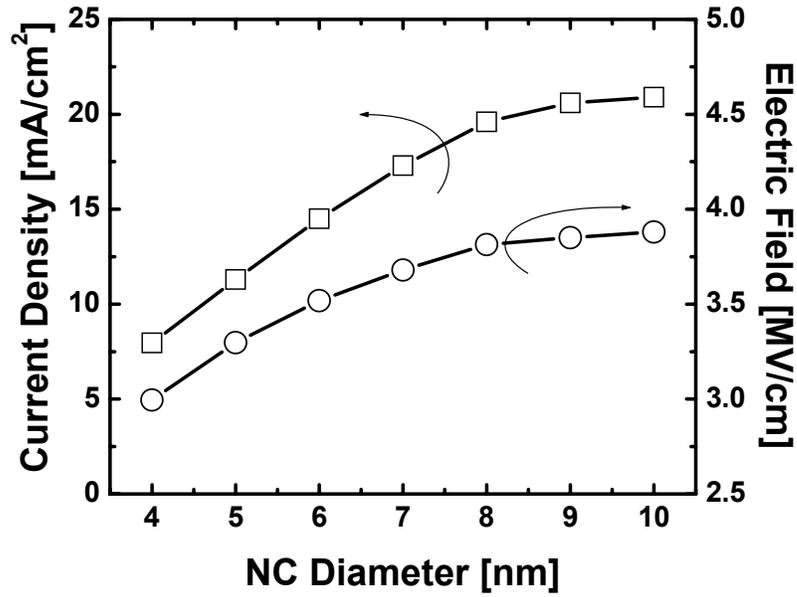


Figure 2.10 Oxide electric field at the SiO₂/Si interface and tunneling current density into NCs as a function of the NC diameter D_{NC} . The NC unit cell diameter C_{NC} is fixed at 18 nm, $V_G = 8$ V, and no charge stored in the NCs.

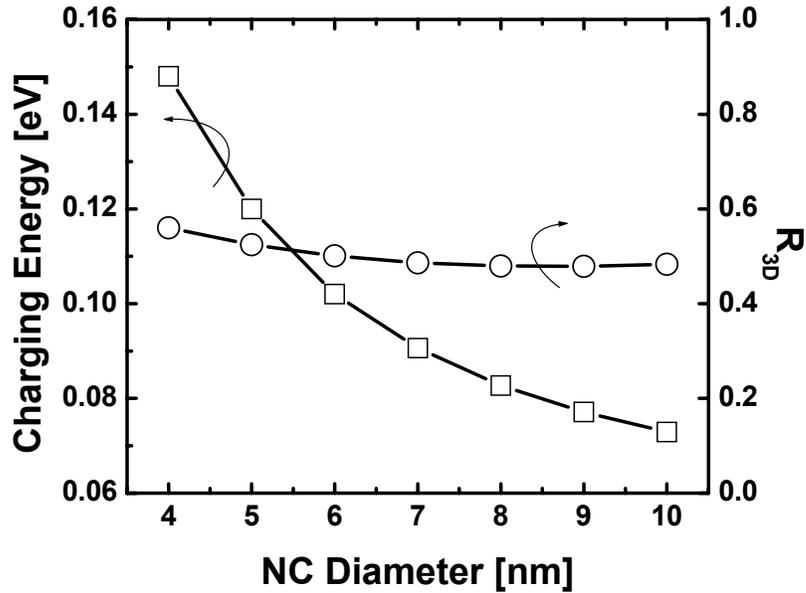


Figure 2.11 Single-electron charging energy and 3D channel-control factor R_{3D} as a function of the NC diameter. The NC unit cell diameter C_{NC} is fixed at 18 nm.

NC diameter because both C_{1_3D} / C_{1_1D} and C_{2_3D} / C_{2_1D} ratios vary simultaneously, while the decreasing charging energy with the increase of NC diameter is expected. Finally, in combination with the benefits of electric field enhancement and larger tunneling capture cross-sectional area, which is assumed to scale with the NC diameter, larger NCs achieve much-improved program speed in Fig. 2.12.

Similar to the previous NC spacing analysis, larger NCs also provide larger memory window after the desired 10-year lifetime as shown in Fig. 2.13, thanks to the smaller charging energy that allows more charges stably stored in NCs. Therefore, the larger NC diameter benefits both program and retention characteristics, as long as the NC number density remains the same and the lateral tunneling between NCs is negligible. This is valid for both 2D channel and 1D narrow-channel NC memories.

2.4.3 *Figure of Merit for NC array design*

As a figure of merit, program time versus retention time for various NC array designs is presented in Fig. 2.14. The program time is defined at a realistic 1.5 V memory window with $V_G = 8$ V. The retention time is chosen to guarantee less than 20 % memory window variation. It can be seen that both decreasing the NC spacing while keeping the NC diameter constant and increasing the NC diameter while keeping the unit cell diameter constant present substantial improvement at the retention time to program time ratio. The significance of this becomes even more obvious when we compare with the thickness scaling of tunnel oxide. In contrast to the NC array optimization, the fast program speed with a thinner thickness of tunnel oxide can only be obtained at the expense of the degraded retention characteristic. Given the unit cell diameter C_{NC} ranges from 11 nm to 18 nm, the proposed NC array design in Fig. 2.14 is realistic even for an aggressively scaled memory cell of sub-50nm feature size. However, to ensure acceptable parametric yield [23],

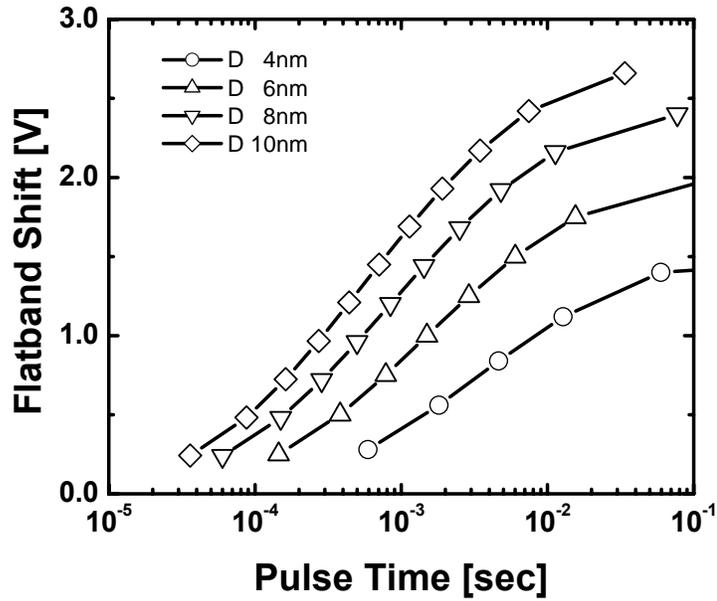


Figure 2.12 Flatband voltage shift versus program pulse time as a function of the NC diameter D_{NC} at $V_G = 8$ V and $C_{NC} = 18$ nm. Each point in the plot represents an additional electron stored per NC.

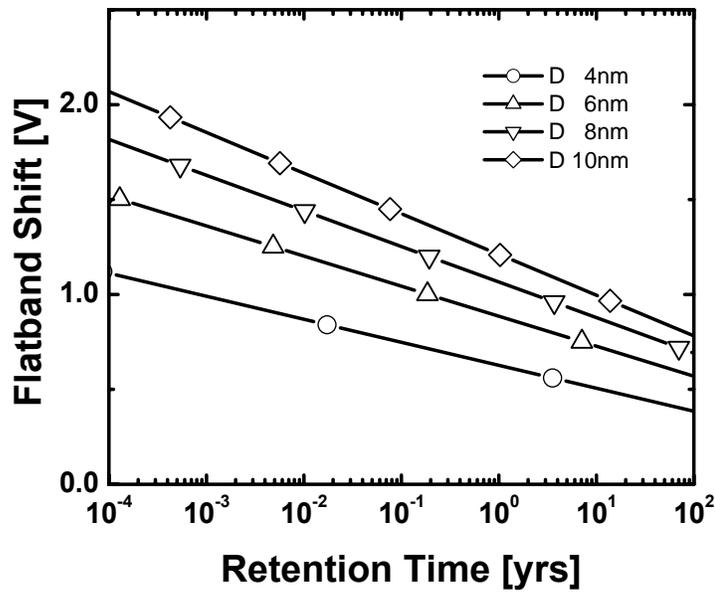


Figure 2.13 Room-temperature retention characteristics at $V_G = 0$ V and $C_{NC} = 18$ nm as a function of the NC diameter D_{NC} . 1s program pulse time at $V_G = 8$ V is applied prior to the retention period to ensure program saturation. Each point in the plot represents an additional electron loss per NC.

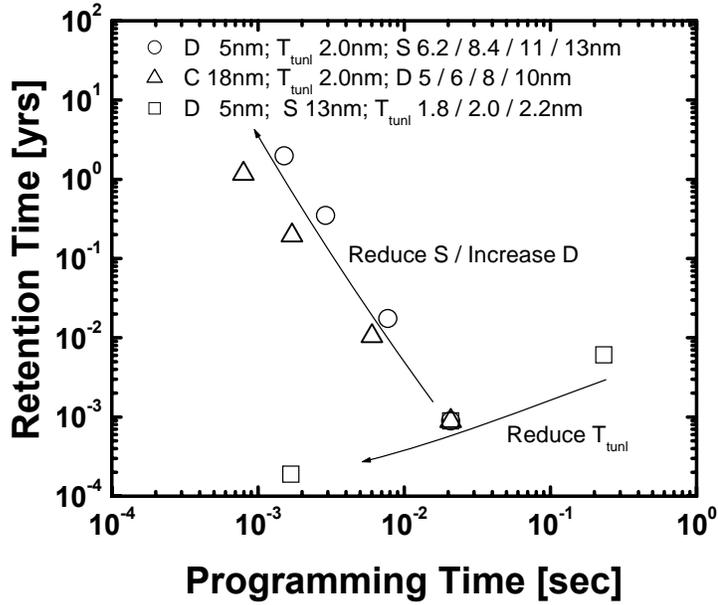


Figure 2.14 Program time versus retention time for various designs of the NC array. The program time is defined at a realistic 1.5V memory window with $V_G = 8$ V. The retention time is chosen to guarantee less than 20 % memory window variation.

certain ordered NC placement techniques [8], [15] may be required to suppress the fluctuation of the number of NC per cell. The unit cell size can be further driven down with use of high- κ control dielectric due to the reduction in Coulomb charging energy, which will be discussed in Chapter 3.

2.5 Conclusion

We present a physical model based on 3D electrostatics and 1D tunneling to simulate the characteristics of the metal NC memories. The unique 3D electrostatic effects are proven very important to optimize the NC array design, such as the NC spacing and the NC diameter. In the typical 2D channel metal NC memories, the desired large memory window, fast program speed and long retention can be obtained with a high-density and large-size NC array. The different design consideration in the 1D narrow-channel memories is also addressed.

REFERENCES

- [1] G. Atwood, "Future directions and challenges of ETox flash memory scaling," *IEEE Trans. Device and Materials Reliability*, vol. 4, no. 3, pp. 301-305, Sept. 2004.
- [2] International Technology Roadmap for Semiconductors, 2005 edition, [Online]. Available: <http://public.itrs.net/>
- [3] S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, and D. Buchanan, "Volatile and non-volatile memories in silicon with nano-crystal storage," in *IEDM Tech. Dig.*, 1995, pp. 521-524.
- [4] Z. Liu, C. Lee, V. Narayanan, G. Pei, and E. C. Kan, "Metal nanocrystal memories, part I: Device design and fabrication," *IEEE Trans. Electron Devices*, vol. 49, pp. 1614-1622, Sept. 2002.
- [5] C. Lee, A. Gorur-Seetharam, and Edwin C. Kan, "Operational and reliability comparison of discrete-storage nonvolatile memories: Advantages of single- and double-layer metal nanocrystals," in *IEDM Tech. Dig.*, 2003, pp. 557-560.
- [6] M. Takata, S. Kondoh, T. Sakaguchi, H. Choi, J.-C. Shim, H. Kurino, and M. Koyanagi, "New non-volatile memory with extremely high density metal nanodots," *IEDM Tech. Dig.*, 2003, pp. 553-556.
- [7] J. J. Lee, and D.-L. Kwong, "Metal nanocrystal memory with high- κ tunneling barrier for improved data retention," *IEEE Trans. Electron Devices*, vol. 52, no. 4, pp. 507-511, Apr. 2005.
- [8] S. Tang, C. Mao, Y. Liu, D. Q. Kelly, and S. K. Banerjee, "Nanocrystal flash memory fabricated with protein-mediated assembly," in *IEDM Tech. Dig.*, 2005, pp. 181-184.

- [9] F. Rana, S. Tiwari, and J. J. Welser, "Kinetic modeling of electron tunneling processes in quantum dots coupled to field-effect transistors," *Supperlatt. Microstruct.*, vol. 23, pp. 757-770, 1998.
- [10] B. D. Salvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Baron, N. Buffet, A. Fernandes, and B. Guillaumot, "Experimental and theoretical investigation of nano-crystal and nitride-trap memory devices," *IEEE Trans. Electron Devices*, vol. 48, no. 8, pp. 1789-1799, Aug. 2001.
- [11] M. She, and T.-J. King, "Impact of crystal size and tunnel dielectric on semiconductor nanocrystal memory performance," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1934-1940, Sept. 2003.
- [12] C. M. Compagnoni, D. Ielmini, A. S. Spinelli, and A. L. Lacaita, "Modeling of tunneling P/E for nanocrystal memories," *IEEE Trans. Electron Devices*, vol. 52, no. 4, pp. 569-576, Apr. 2005.
- [13] C. Lee, U. Ganguly, V. Narayanan, T.-H. Hou, and E. C. Kan, "Asymmetric electric field enhancement in nanocrystal memories," *IEEE Electron Device Lett.*, vol. 26, no. 12, pp. 879-881, Dec. 2005.
- [14] C. Lee, J. Meter, V. Narayanan, and E. C. Kan, "Self-assembly of metal nanocrystal on ultra-thin oxide for nonvolatile memory applications," *J. Electronic Materials*, vol. 34, no. 1, pp. 1-11, Jan. 2005.
- [15] K. W. Guarini, C. T. Black, Y. Zhang, I. V. Babich, E. M. Sikorski, and L. M. Gignac, "Low voltage, scalable nanocrystal flash memory fabricated by templated self assembly," in *IEDM Tech. Dig.*, 2003, pp. 541-544.
- [16] C. Lee, U. Ganguly and E. C. Kan, "Characterization of number fluctuations in gate-last metal nanocrystal nonvolatile memory array beyond 90nm CMOS technology," *Material Research Symposium (MRS)*, Boston, MA, Nov. 29 – Dec. 3, 2004.

- [17] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices," *IEEE Trans. Electron Devices*, vol. 46, no. 7, pp. 1464-1471, July 1999.
- [18] Y. T. Hou, M. F. Li, Y. Jin, and W. H. Lai, "Direct tunneling hole currents through ultrathin gate oxides in metal-oxide-semiconductor devices," *J. Appl. Phys.*, vol. 91, pp. 258-264, 2002.
- [19] A. Shanware, J. Shiely, H. Massoud, E. Vogel, K. Henson, A. Srivastava, C. Osburn, J. Hauser, and J. Wortman "Extraction of the gate oxide thickness of N- and P-channel MOSFETs below 20Å from the substrate current resulting from valence-band electron tunneling," in *IEDM Tech. Dig.*, 1999, pp. 815-818.
- [20] A. Ghetti, E. Sangiorgi, J. Bude, T. W. Sorsch, and G. Weber, "Tunneling into interface states as reliability monitor for ultrathin oxide," *IEEE Trans. Electron Devices*, vol. 47, no. 12, pp. 2358-2364, July 2000.
- [21] M. Saitoh, E. Nagata, and T. Hiramoto, "Large memory window and long charge-retention time in ultranarrow-channel silicon floating-dot memory," *Appl. Phys. Lett.*, vol. 82, pp. 1787-1789, 2003.
- [22] U. Ganguly, E. C. Kan, and Y. Zhang, "Carbon nanotube-based nonvolatile memory with charge storage in metal nanocrystals", *Appl. Phys. Lett.*, vol. 87, 043108, 2005.
- [23] L. Perniola, B. D. Salvo, G. Ghibaud, A. F. Para, G. Pananakakis, V. Vidal, T. Baron, and S. A. Lombardo, "Modeling of the programming window distribution in multinanocrystals memory," *IEEE Trans. Nanotechnol.*, vol. 2, no. 4, pp. 277-284, Dec. 2003.
- [24] J. M. Crowley, *Fundamentals of Applied Electrostatics*, John Wiley, 1986.

- [25] L. Novotny, R. X. Bian, and X. S. Xie, “Theory of nanometric optical tweezers,”
Phys. Rev. Lett. 79, pp. 645–648, July 1997.

CHAPTER 3
DESIGN OPTIMIZATION OF METAL NANOCRYSTAL MEMORY —
PART II: GATE STACK ENGINEERING

3.1 Abstract

Based on the physical model of nanocrystal memories [1] described in Chapter 2, we present a systematic investigation of gate stack engineering, including high- κ control oxide and tunnel oxide. The high- κ control oxide enables the EOT scaling without compromising the memory performance, owing to the low charging energy and large channel-control factor from the three-dimensional electrostatics. The high- κ tunnel oxide, on the other hand, improves the retention characteristics utilizing the asymmetric tunnel barrier more effectively away from the direct tunneling regime. Finally, with the optimization strategies introduced in both Chapter 2 and 3, a metal nanocrystal memory design with 1.0 V memory window, 13 μ s program, 2.5 μ s erase and over 10-year retention time has been demonstrated at ± 4 V operation, which highlights the potential of nanocrystal memories as the next-generation nonvolatile memory.

3.2 Introduction

Many recent efforts on the high- κ dielectric development [2], [3], driven by the continuous miniaturization of MOSFETs, enable much better control on the gate leakage current, the high- κ crystalline structure, and undesirable defects such as the interface states and traps. This also inspired investigation for their use in nonvolatile memories. The high- κ dielectric integration has been experimentally demonstrated in the continuous floating gate memory [4], the silicon-oxide-nitride-oxide-silicon (SONOS) memory [5], and the nanocrystal (NC) memory [6]. However, a

comprehensive physical-based model to predict the memory performance did not gain as much attention. Although a few existing works simulated the high- κ dielectric NC memories using the one-dimensional (1D) electrostatic model with somewhat high program/erase (P/E) voltage [7], [8], the three-dimensional (3D) electrostatic nature and the advantage of low P/E voltage operations in the NC memory have not yet been fully explored.

In Chapter 2, the 3D electrostatics together with 1D tunneling model has been established and verified with the experimental characteristics of metal NC memories. The benefits of applying this model in the memory design have also been presented in the case of NC array optimization. In this chapter, to further improve the performance, the strategy of the gate stack optimization in the memory cell, including high- κ control oxide and tunnel oxide, is explicitly examined based on the same model with an emphasis on the low-voltage operation. The significance of the 3D electrostatics is also highlighted. Finally, relying on the optimization strategies introduced in Chapter 2 and here, we report a scaled metal NC memory design allowing ± 4 V operation and excellent memory characteristics.

3.3 *Device Modeling*

To be consistent with the simulation done in Chapter 2, the same set of parameters extracted from the experimental results is applied, NC diameter $D_{NC} = 5$ nm, NC number density $N_{NC} = 4 \times 10^{11}$ cm⁻², capture cross-section $\sigma_{NC} = 5.3 \times 10^{-14}$ cm² per NC, and NC work function (WF) 5.1 eV. A p-type (100) Si substrate with a doping level of 2×10^{17} cm⁻³ and an initial flatband voltage $V_{FB} = 0$ before program are assumed. The relevant HfO₂ material parameters used in the tunneling current calculation are dielectric constant ϵ_{HfO_2} 20, HfO₂/Si conduction band offset ΔE_c 1.5 eV, HfO₂/Si valence band offset ΔE_v 2.5 eV, electron effective mass 0.22 m_0 , and hole

effective mass $0.15 m_0$, where m_0 is the free electron mass. These material parameters can vary from the deposition techniques, gate stack compositions, and even the extraction methods [9]-[11]. The values we choose here are close to those reported in [10]. All dielectrics in the stack are assumed to be ideal without considering the second-order effects, such as grain boundaries, interface states, traps, and fixed charges, which may affect the device performance in a negative or positive [12] manner. Our choice here to ignore these effects is mainly to establish the ideal design space without ambiguous process variations.

3.4 Results and Discussion

3.4.1 High- κ Dielectric as Control Oxide

The benefits of the control oxide equivalent oxide thickness (EOT) scaling are twofold. First, it enables the memory cell size to scale aggressively by enhanced control of the short-channel effect. Second, it increases the coupling ratio, which improves the P/E efficiency in the conventional continuous floating gate devices. In the NC memories, however, due to the Coulomb blockade effect, unlike their continuous floating gate counterparts, the maximum number of stored charges is self-saturated as a function of the NC charging energy and the bias condition. High coupling ratio by scaling the control oxide may allow more charges stored in NCs at self saturation, but does not guarantee larger flatband shift ΔV_{FB} , which is influenced by the combined effect of the number of stored charges, the control oxide EOT, and the 3D channel-control ratio R_{3D} for a given NC density, according to Eqs. (2.7) and (2.8) described in Chapter 2. Even more importantly, if the NC charging energy does not scale much with the control oxide EOT, more charges stored in NCs may adversely affect retention characteristics. Therefore, the trade-off between the thickness of control oxide and NC memory characteristics is a fundamental problem to

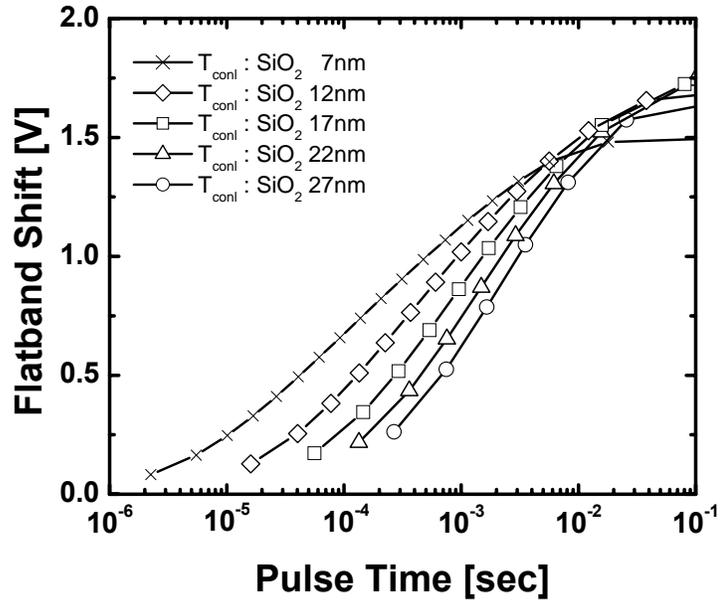


Figure 3.1 Flatband voltage shift versus program pulse time at $V_G = 8\text{V}$ using SiO_2 as the control oxide with thickness of 7 nm to 27 nm. The tunnel oxide is fixed as 2 nm SiO_2 . Each point in the plot represents an additional electron stored per NC.

consider in the NC memory cell design.

Figure 3.1 shows the program transient curve at $V_G = 8\text{V}$ using SiO_2 as the control oxide with thickness T_{cont} of 7nm to 27nm. It is clear to see the high coupling ratio benefits the program efficiency at the initial stage of charge build-up, and then gradually yields to the fact of less flatband shift due to the thinner control oxide toward the self-saturated region. For a moderate 1.5V ΔV_{FB} , there is no clear edge of thinner control oxide in terms of the program speed. Notice that this statement is only applicable to the given NC array design. For example, if the NC number density is increased, thinner control oxide may improve the program time referred to the 1.5V ΔV_{FB} although the inherent trade-off still exists. In general, given that maximum NC number density is limited by the available process technology, the choice of thickness of control oxide is a matter of the design optimization. Another consideration of the

thickness of control oxide is to suppress the electron tunneling current from the NCs to the control gate ($J_{E,NG}$) during program. In all simulated conditions except $T_{cont} = 7\text{nm}$, $J_{E,NG}$ through the Fowler-Nordheim (F-N) tunneling is negligible compared with the direct tunneling current from the Si-channel to the NCs ($J_{E,SN}$). Thus the charge build-up in NCs is solely determined by $J_{E,SN}$. Nevertheless, for $T_{cont} = 7\text{nm}$ of SiO_2 , before it approaches the self-saturated condition, $J_{E,NG}$ reaches the level of $J_{E,SN}$ given the high electric field on the control oxide, and the charge build-up in NCs saturates earlier due to the cancellation of two currents. This is another constrain to be taken into account when scaling the control oxide.

While thinner EOT by scaling the control oxide is inevitable to reduce the NC memory cell size, the trade-off mentioned above makes the optimization of P/E efficiency very challenging. Here high- κ dielectrics as the control oxide are proposed as a solution. High- κ dielectrics were first introduced to the floating gate devices to suppress the inter-poly leakage [4], analogous to $J_{E,NG}$ in the NC memories. However, the true significance of high- κ control oxide in NC memories lies in the unique 3D electrostatic nature, which is usually overlooked, but proven very important in our model. Figure 3.2 plots the calculated single-electron charging energy and the 3D channel-control factor R_{3D} through the 3D electrostatic simulation as a function of the control oxide dielectric constant and EOT. The charging energy is relatively insensitive to the control oxide EOT because the NC self-capacitance is mainly determined by the capacitive coupling between the NC and the Si-channel, where the tunnel oxide EOT is usually several times smaller than the control oxide EOT. On the other hand, it is a strong function of the dielectric constant. The charging energy with HfO_2 is 2 times less than that with SiO_2 . R_{3D} also strongly depends on the dielectric constant and is much higher with HfO_2 . To better understand this, the cross-sections of the 3D potential contours in the NC unit cell are plotted in Fig. 3.3. The NC

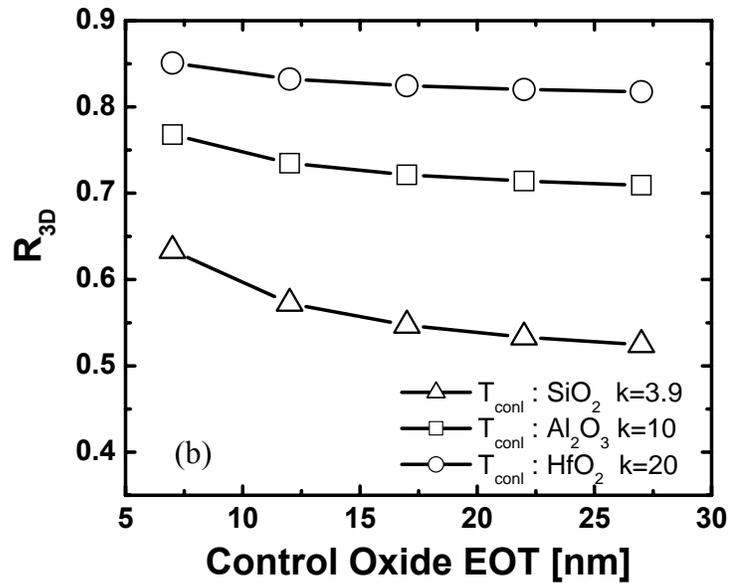
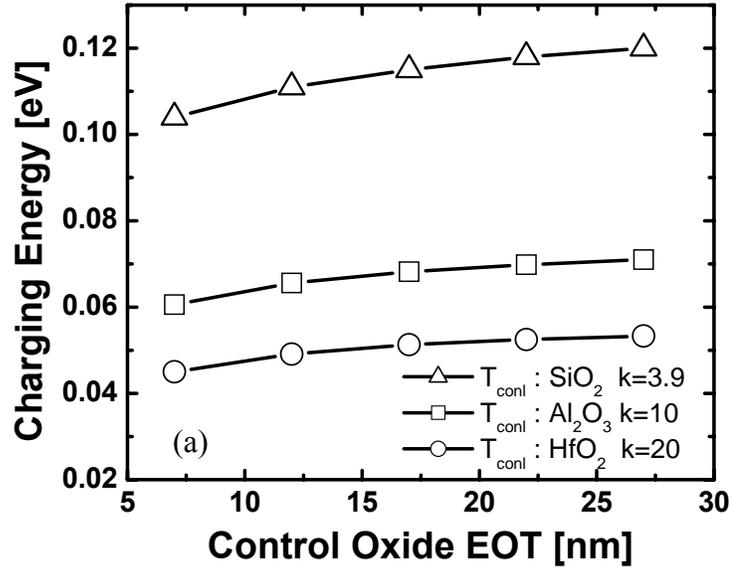


Figure 3.2 (a) Single-electron charging energy, and (b) 3D channel-control factor R_{3D} as a function of the control oxide EOT and dielectric constant through the 3D electrostatic simulation.

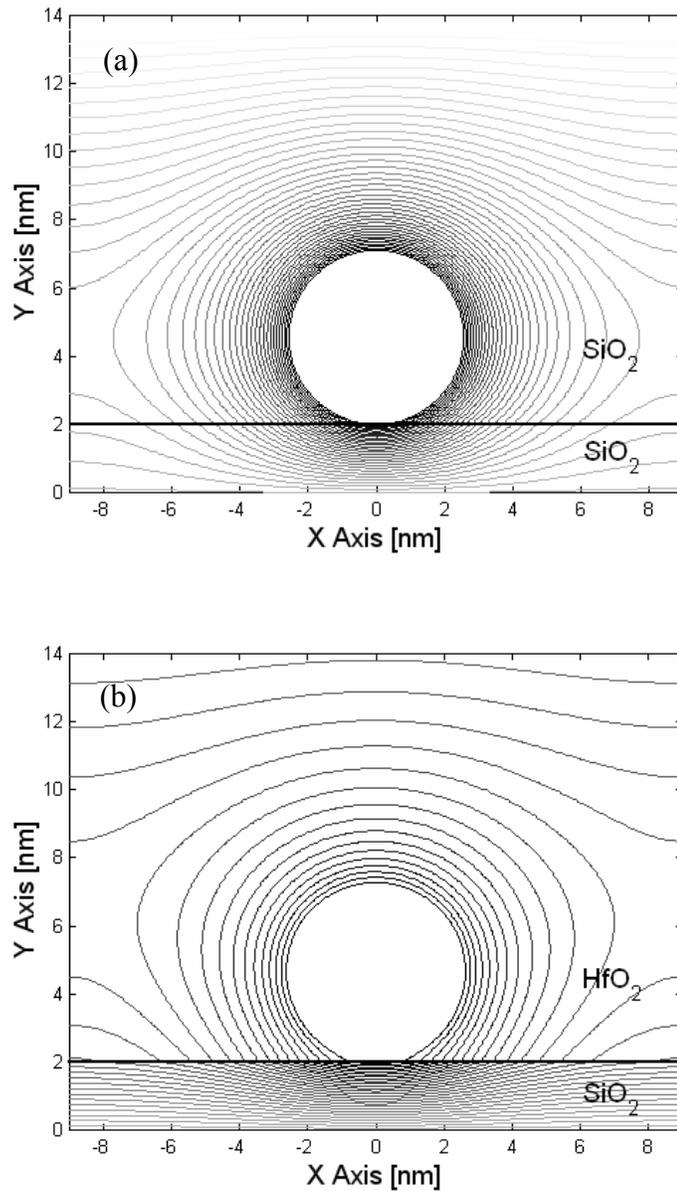


Figure 3.3 Cross-sectional views of the 3D electrostatic potential contours in the NC memory unit cell with (a) 7 nm SiO₂, and (b) 35 nm HfO₂ as the control oxide. Only part of the HfO₂ is shown in (b). The NC potential is set as -1 V while $V_G = 0$ V. The potential increases monotonically from the NC with the contour spacing of 25 mV.

potential is set as -1 V while $V_G = 0$ V, and the EOT remains the same for both stacks. It is obvious that the fringing fields through HfO₂ to the Si channel are much stronger due to the higher dielectric constant of HfO₂. This substantially increases the substrate-NC coupling capacitance and the effective channel coverage area under the influence of the charges stored in NCs.

The phenomena lead to significant consequences in the NC memory operations. Large R_{3D} results in larger ΔV_{FB} when the same number of charges stores in NCs. Meanwhile, low charging energy allows even more charges stored per NC at the given bias condition and program time. These compensate the insufficient memory window due to the trade-off in the control oxide EOT scaling. As a result, with HfO₂ as the control oxide, much better program efficiency is shown in Fig. 3.4. Finally, the retention characteristics are examined. To guarantee long retention time in the NC memories, the NC Fermi level has to remain substantially lower than the Si conduction band edge at the retention condition to suppress the direct tunneling current. In fact, any excess charge stored in NCs during the program pulse can tunnel back to the Si channel in a short time. Only lowering the charging energy can allow more charge storage without scarifying the retention characteristics, which leads to the appreciable improvement shown in Fig. 3.5.

As a short summary, we demonstrate, by replacing SiO₂ with HfO₂ as the control oxide, both the program efficiency and the retention characteristics are improved as a result of low charging energy and large channel-control factor from the 3D fringing effect. These results confirm 3D electrostatics instead of 1D should be considered in NC memory modeling.

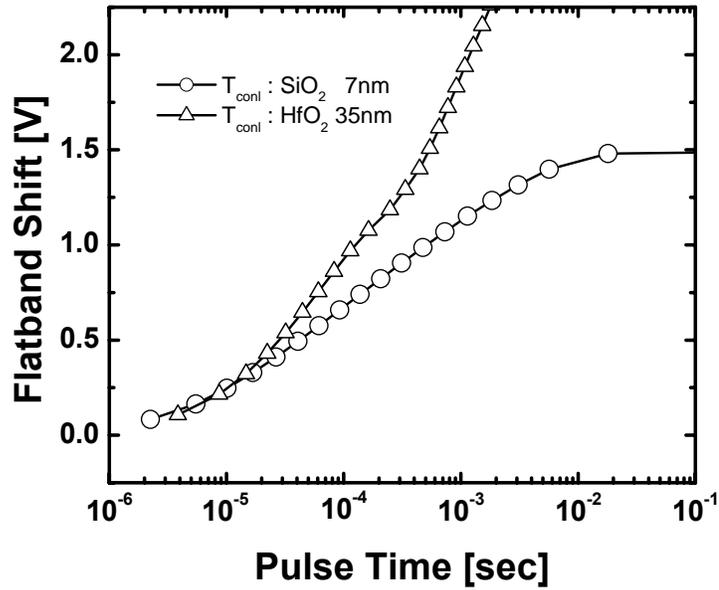


Figure 3.4 Flatband voltage shift versus program pulse time at $V_G = 8\text{V}$ with 7 nm SiO₂ and 35 nm HfO₂ as the control oxide. The tunnel oxide is fixed as 2 nm SiO₂.

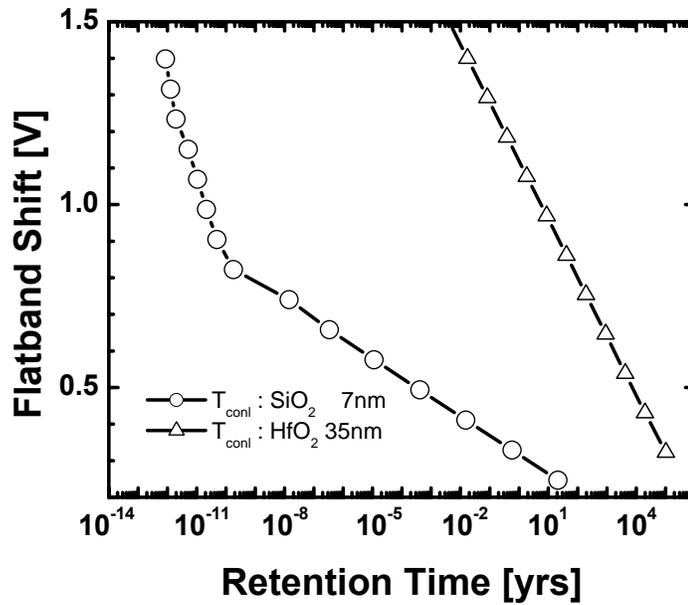


Figure 3.5 Room-temperature retention characteristics at $V_G = 0\text{V}$ with 7 nm SiO₂ and 35 nm HfO₂ as the control oxide. The tunnel oxide is fixed as 2 nm SiO₂. The NC memories are pre-charged to 1.5 V ΔV_{FB} . Each point in the plot represents an additional electron loss per NC.

3.4.2 High- κ Dielectric as Tunnel Oxide

Even with the benefits of the high- κ control oxide, the retention time shown in Fig. 3.5 still cannot meet the 10-year lifetime requirement, owing to the very thin tunnel oxide. This is another trade-off to overcome in the NC memory design where both program and retention characteristics rely on the same physical barrier. A SiO₂/HfO₂ bi-layer stack aimed to break the trade-off is exploited as the field-sensitive tunnel barrier [8], [13], maximizing asymmetry between the effective barriers seen at the program and the retention operations. In fact, the SiO₂/HfO₂ bi-layer not only improves the tunnel barrier properties, but is also a realistic consideration while integrating high- κ dielectrics. The interfacial SiO₂ growth between high- κ dielectrics and the Si channel is difficult to avoid in many high- κ deposition methods [14], and even desirable to ease the severe mobility degradation caused by the remote phonon scattering [15].

With the same EOT, the 1.2 nm SiO₂ / 4 nm HfO₂ bi-layer stack shows compatible program speed as the case of the 2.0 nm SiO₂ in Fig. 3.6 in spite of much thicker physical thickness, thanks to the low ΔE_c of HfO₂ and the appropriate barrier engineering by the deliberate design of the SiO₂/HfO₂ thickness ratio. Figure 3.7 shows the calculated band diagrams of two stacks at $\Delta V_{FB} = 0$ V and at $\Delta V_{FB} = 2.0$ V under the program operation. When there is no charge stored in NCs, the subband electrons injecting into NCs only have to tunnel through a portion of the HfO₂ barrier, resulting in the faster program speed in the low ΔV_{FB} regime. Note that the curvature near the NC, which also contributes to the thinner HfO₂ barrier, is the signature of the 3D field enhancement effect discussed in Chapter 2. As the charge accumulation in NCs gradually raises the NC potential, electrons tunneling through HfO₂ can no longer take advantage of the F-N tunneling. This creates the turning point in the program transient in Fig. 3.6 and lowers the program speed toward the larger ΔV_{FB} regime.

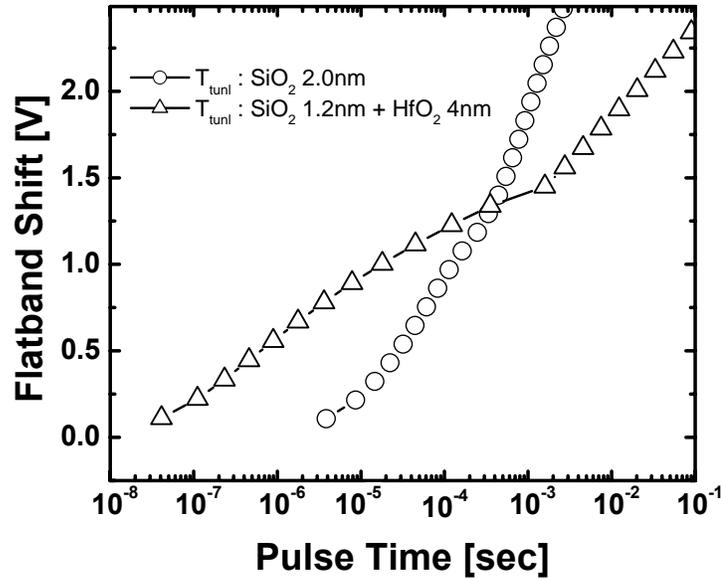


Figure 3.6 Flatband voltage shift versus program pulse time at $V_G = 8\text{V}$ with 2 nm SiO₂ and 1.2 nm SiO₂ / 4 nm HfO₂ bi-layer as the tunnel oxide. The control oxide is fixed as 35 nm HfO₂.

Meanwhile, due to the much thicker physical thickness of the tunnel oxide, 5 orders of magnitude longer retention time with the bi-layer stack is achieved in Fig. 3.8. The calculated band diagrams of two stacks at $\Delta V_{FB} = 2.0\text{ V}$ under the retention condition are plotted in Fig. 3.9. By combining both the high NC WF (5.1 eV) and the field-sensitive barrier, the asymmetry of the effective tunnel barrier between Figs. 3.7 and 3.9 is explicitly shown. In this section, we demonstrate the SiO₂ / HfO₂ bi-layer stack as the tunnel oxide improves the retention characteristics of the NC memories without compromising the P/E efficiency.

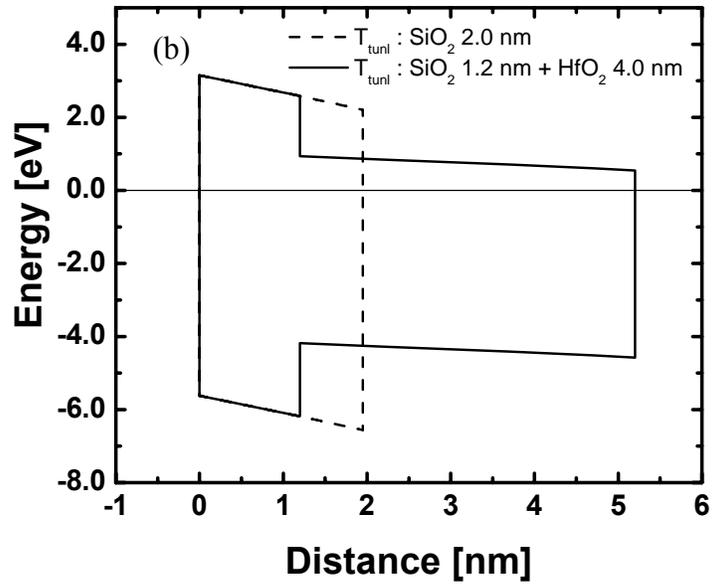
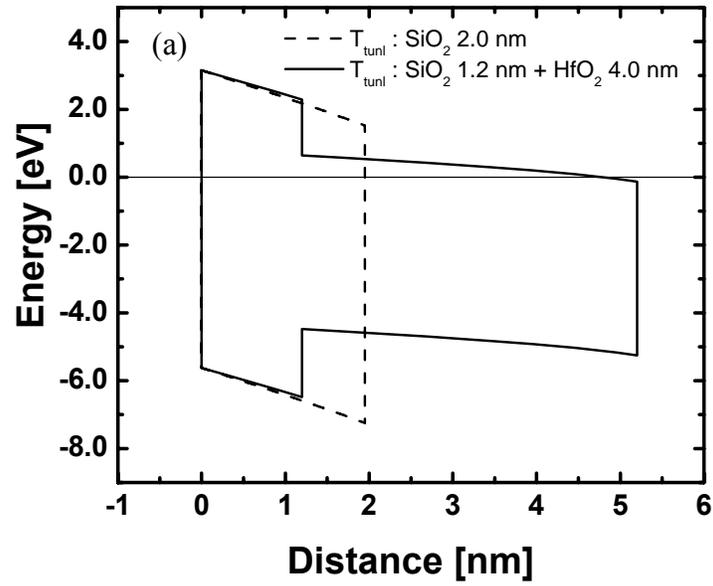


Figure 3.7 Calculated band diagram of 2 nm SiO₂ and 1.2 nm SiO₂ / 4 nm HfO₂ bi-layer tunnel oxide at $V_G = 8$ V while (a) $\Delta V_{FB} = 0$ V, and (b) $\Delta V_{FB} = 2$ V. The control oxide is fixed as 35 nm HfO₂. The conduction band edge of the Si channel is referred at energy 0 eV, and the Si / SiO₂ interface at distance 0 nm.

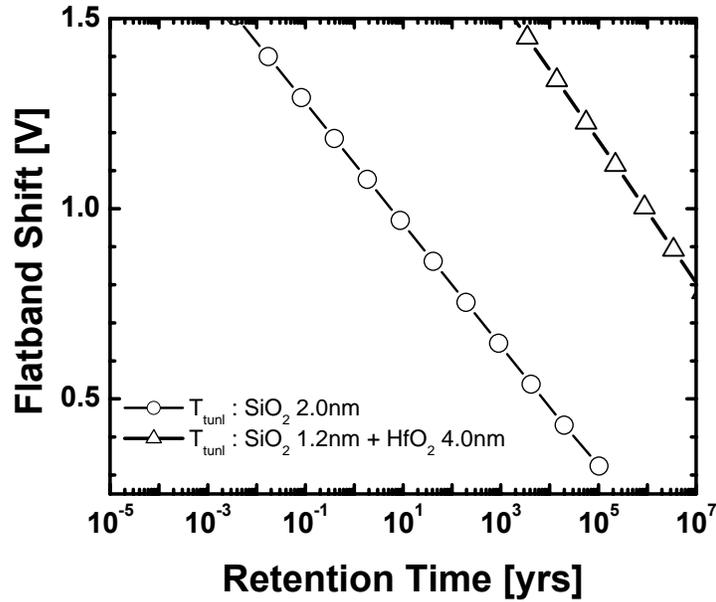


Figure 3.8 Room-temperature retention characteristics at $V_G = 0$ V with 2 nm SiO₂ and 1.2 nm SiO₂ / 4 nm HfO₂ bi-layer as the tunnel oxide. The control oxide is fixed as 35 nm HfO₂. The NC memories are pre-charged to 1.5 V ΔV_{FB} .

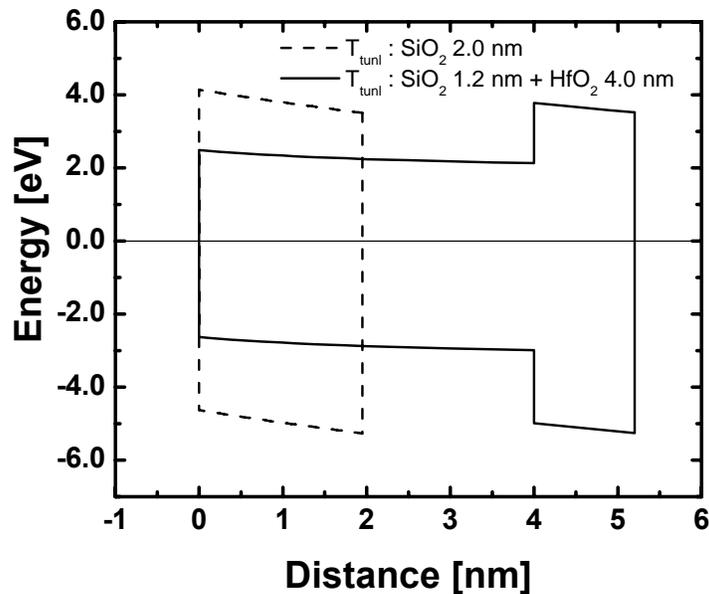


Figure 3.9 Calculated band diagram of 2 nm SiO₂ and 1.2 nm SiO₂ / 4 nm HfO₂ bi-layer tunnel oxide at $V_G = 0$ V while $\Delta V_{FB} = 2$ V. The control oxide is fixed as 35 nm HfO₂. The NC Fermi level is referred at energy 0 eV, and the tunnel oxide / NC interface at distance 0 nm.

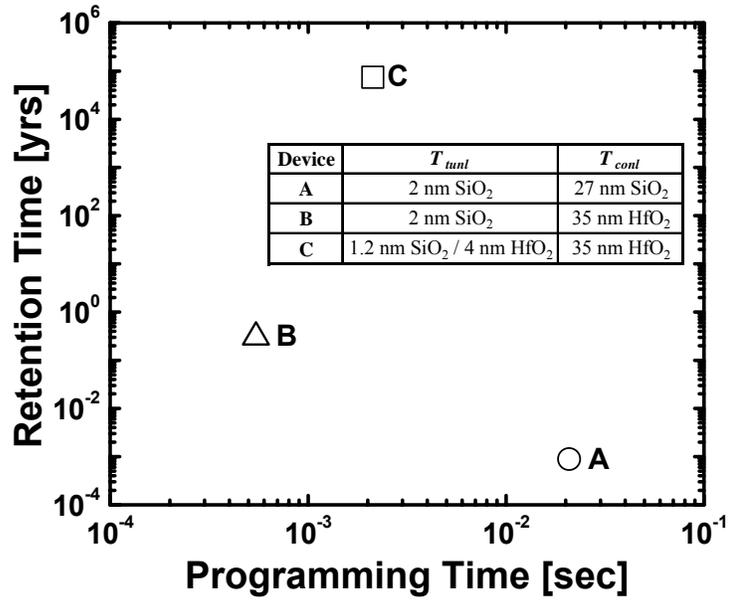


Figure 3.10 Program time versus retention time for various gate stack designs in this study. The program time is defined at 1.5V memory window with $V_G = 8$ V. The retention time is chosen to guarantee less than 20 % memory window variation.

3.4.3 Figure of Merit

To summarize the trend of design optimization, program versus retention time for various gate stack designs in this study is presented in Fig. 3.10. The program time is defined at a realistic 1.5 V memory window with $V_G = 8$ V. The retention time is chosen to guarantee less than 20 % memory window variation. With the high- κ control oxide, EOT of the entire stack is scaled down substantially with improved memory performance, which is difficult to achieve by directly scaling down the SiO₂ control oxide. Moreover, the integration of the high- κ tunnel oxide enables sufficient retention time without compromising the P/E efficiency, which is also impossible to obtain with the SiO₂ tunnel oxide alone. Most noticeably, the benefits of these two approaches are accumulative without contradicting each other. Compared with the conventional device A, the device C exhibits much thinner EOT, 10 times shorter program time and

8 orders of magnitude longer retention time. In addition, the NC memory performance can be further enhanced if the optimization strategy of the NC array presented in Chapter 2 is exploited. This proves the significance of the deliberate NC memory design through accurate modeling.

3.4.4 NC Memory Design for Low P/E Voltage Operation

To investigate the potential of scalability, we have further explored a scaled NC memory configuration allowing ± 4 V P/E operations, by optimizing both the gate stack and the NC array. The simulation parameters are: NC number density 5×10^{11} cm^{-2} , NC diameter 8 nm, NC WF 5.1 eV, 0.7 nm SiO_2 / 2.3 nm HfO_2 as the tunnel oxide, and 25 nm HfO_2 as the control oxide. Please note these are very realistic parameters achievable by either present metal NC [16], [17] or high- κ deposition methods [2], [3]. Figure 3.11 demonstrates 13 μs program and 2.5 μs erase time with 1.0 V ΔV_{FB} at ± 4 V P/E operations. Over 10-year retention time is projected with less than 20% memory window variation. A recent experiment of the metal NC memory using a thicker 5.1 nm HfO_2 as the tunnel oxide also reported 1 ms program time at ± 4 V P/E operation with promising retention characteristics [17]. Both of our model and the experimental evidence suggest that the metal NC memory is very promising for the low-power, low-voltage, long-retention and high-speed memory applications.

In addition, the scaling effect of P/E voltage is shown in Fig. 3.12. The nanosecond P/E time can be obtained with ± 8 V operations while maintaining the same memory window and retention characteristics, which presents the potential of metal NC memories to replace SRAM and DRAM as the cache memory and the primary memory in the portable electronics with the advantage of nonvolatile data storage.

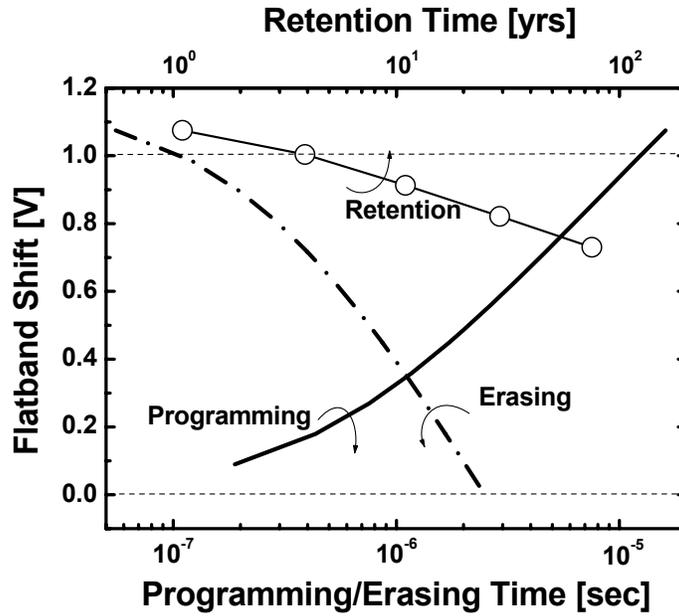


Figure 3.11 P/E and retention characteristics of a scaled NC memory design operated at ± 4 V. The simulation parameters are: NC number density $5 \times 10^{11} \text{ cm}^{-2}$, NC diameter 8 nm, NC WF 5.1 eV, 0.7 nm SiO_2 / 2.3 nm HfO_2 as the tunnel oxide, and 25 nm HfO_2 as the control oxide.

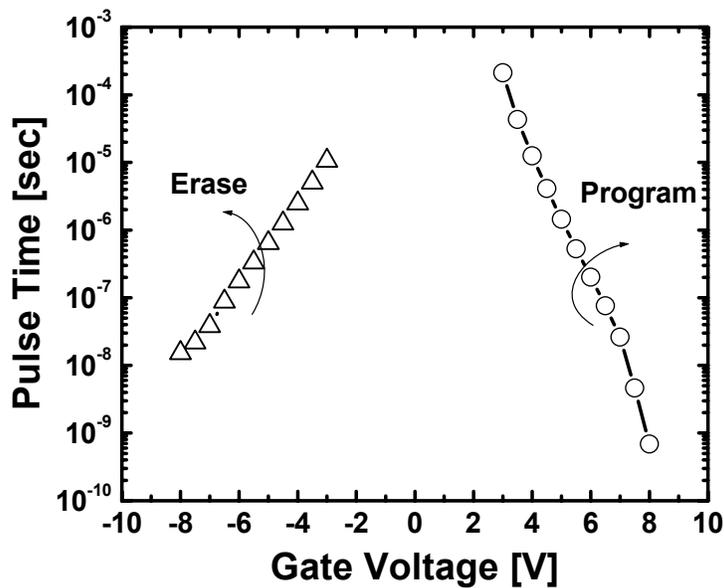


Figure 3.12 P/E characteristics of the scaled NC memory design as a function of P/E voltage. The P/E pulse time and voltage are chosen to assure 1.0 V memory window and over 10-year retention time.

In the prior art of Flash memory cell scaling based on 1D electrostatics and similar tunneling models [18], it had been concluded that even in the optimal design of control and tunnel dielectrics, the P/E voltage scaling can only be down to 4 V with a 5 ms program time, given the fixed retention time at 10 years. The ratio between the retention time and P/E time t_R / t_{PE} is around 10^{11} , which comes directly from the asymmetry between the tunneling currents in P/E and retention conditions. By inclusion of the 3D electrostatics and the bi-layer design of tunnel barrier, our calculations show much more design freedom. For designs in Figs. 3.10 and 3.11, the t_R / t_{PE} ratio is over 10^{14} for a reasonable memory window with the P/E voltages at ± 8 V and ± 4 V, respectively. This ratio can be further enlarged with material optimization of control-gate work function [6], NC work function [16], and injection effective mass [19], as well as heterogeneous floating-gate layers [12]. Notice also that within the channel length only two NCs are necessary to take advantage of most of the effects discussed in Chapter 2 and Chapter 3, and the two-bit-per-cell virtual ground architecture [20], [21]. Therefore, we can safely conjecture that there is still much room of Flash cell scaling beyond the 50-nm technology node.

3.5 Conclusion

We present a systematic study on the gate stack engineering of NC memories based on our physical model. The integration of high- κ dielectrics is proven very desirable. Together with the NC array optimization described in Chapter 2, the results serve as useful guidelines in the NC memory design with the demonstration of the extremely low-voltage NC memory configuration.

REFERENCES

- [1] T.-H. Hou, C. Lee, V. Narayanan, U. Ganguly, and E. C. Kan, "Design optimization of metal nanocrystal memory — Part I: nanocrystal array engineering," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1934-1940, Sept. 2003.
- [2] Y. T. Hou, F. Y. Yen, P. F. Hsu, V. S. Chang, P. S. Lim, C. L. Hung, L. G. Yao, J. C. Jiang, H. J. Lin, Y. Jin, S. M. Jang, H. J. Tao, S. C. Chen, and M. S. Liang, "High performance tantalum carbide metal gate stacks for nMOSFET application," in *IEDM Tech. Dig.*, 2005, pp. 35-38.
- [3] S. Inumiya, Y. Akasaka, T. Matsuki, F. Ootsuka, K. Torii, and Y. Nara, "A thermally-stable sub-0.9nm EOT TaSi_x/HfSiON gate stack with high electron mobility, suitable for gate-first fabrication of hp45 LOP devices," in *IEDM Tech. Dig.*, 2005, pp. 27-30.
- [4] W.-H. Lee, J. T. Clemens, R. C. Keller, and L. Manchanda, "A novel high-κ inter-poly dielectric (IPD), Al₂O₃ for low voltage/high-speed Flash memories: erasing in msec at 3.3V," in *Symp. VLSI Tech. Dig.*, 1997, pp. 117-118.
- [5] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, K. C. Park, and K. Kim, "A novel SONOS structure of SiO₂/SiN/Al₂O₃ with TaN metal gate for multi-giga bit Flash memories," in *IEDM Tech. Dig.*, 2003, pp. 613-616.
- [6] J. J. Lee, X. Wang, W. Bai, N. Lu, J. Liu, and D. L. Kwong, "Theoretical and experimental investigation of Si NC memory device with HfO₂ high-κ tunneling dielectric," in *Symp. VLSI Tech. Dig.*, 2003, pp. 33-34.
- [7] M. She, and T.-J. King, "Impact of crystal size and tunnel dielectric on semiconductor nanocrystal memory performance," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1934-1940, Sept. 2003.

- [8] C. M. Compagnoni, D. Ielmini, A. S. Spinelli, and A. L. Lacaita, "Optimization of threshold voltage window under tunneling program/erase in nanocrystal memories," *IEEE Trans. Electron Devices*, vol. 52, no. 11, pp. 2473-2481, Nov. 2005.
- [9] W. J. Zhu, T.-P. Ma, T. Tamagawa, J. Kim, and Y. Di, "Current transport in metal/hafnium oxide/silicon structure," *IEEE Electron Device Lett.*, vol. 23, no. 2, pp. 97-99, Feb. 2002.
- [10] Y. T. Hou, M. F. Li, H. Y. Yu, Y. Jin, and D.-L. Kwong, "Quantum tunneling and scalability of HfO₂ and HfAlO gate stacks," in *IEDM Tech. Dig.*, 2002, pp. 731-734.
- [11] S. Zafar, V. Narayanan, A. Callegari, F. R. McFeely, P. Jamison, E. Gusev, C. Cabral, and R. Jammy, "HfO₂ / metal stacks: determination of energy level diagram, work functions & their dependence on metal deposition," in *Symp. VLSI Tech. Dig.*, 2005, pp. 44-45.
- [12] C. Lee, T-H Hou, and E. Kan, "Nonvolatile memory with a metal nanocrystal/nitride heterogeneous floating-gate," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2697-2702, Dec. 2005.
- [13] S. J. Baik, S. Choi, U. -I. Chung, and J. T. Moon, "High speed and non-volatile Si nanocrystal memory for scaled Flash technology using field sensitive tunnel barrier," in *IEDM Tech. Dig.*, 2003, pp. 545-548.
- [14] W. Tsai, L. Ragnarsson, P. J. Chen, B. Onsia, R. J. Carter, E. Cartier, E. Young, M. Green, M. Caymax, S. De Gendt, and M. Heyns, "Comparison of sub 1 nm TiN/HfO₂ with poly-Si/HfO₂ gate stacks using scaled chemical oxide interface," in *Symp. VLSI Tech. Dig.*, 2003, pp. 21-22.
- [15] M. V. Fischetti, D. A. Neumayer, and E. A. Cartier, "Effective electron mobility in Si inversion layers in metal-oxide-semiconductor systems with a high-κ

- insulator: The role of remote phonon scattering,” *J. Appl. Phys.*, vol. 90, pp. 4587-4608, 2001.
- [16] C. Lee, J. Meter, V. Narayanan, and E. C. Kan, “Self-assembly of metal nanocrystal on ultra-thin oxide for nonvolatile memory applications,” *J. Electronic Materials*, vol. 34, no. 1, pp. 1-11, Jan. 2005.
- [17] J. J. Lee, and D.-L. Kwong, “Metal nanocrystal memory with high- κ tunneling barrier for improved data retention,” *IEEE Trans. Electron Devices*, vol. 52, no. 4, pp. 507-511, Apr. 2005.
- [18] V. Zhirnov, “NVM limits,” presented at the SRC Forum on Non-Volatile Memories, Nov.18-19, 2004, Stanford, CA.
- [19] C. B. Duke, *Tunneling in solids*. New York: Academic Press, 1969.
- [20] B. Eitan, P Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, “NRROM: A novel localized trapping, 2-bit nonvolatile memory cell,” *IEEE Electron Device Lett.*, vol. 21, no. 11, pp. 543-545, Nov. 2000.
- [21] Z. Liu, C. Lee, V. Narayanan, G. Pei, and E. C. Kan, “A novel quad source/drain metal nanocrystal memory device for multi-per-cell storage,” *IEEE Electron Device Lett.*, vol. 24, no. 5, pp. 345-347, May 2003.

CHAPTER 4

MODELING OF MULTI-LAYER NANOCRYSTAL MEMORY

4.1 Abstract

This chapter presents a physical model applicable to both single- and multi-layer nanocrystal (NC) memory. With the ability to redistribute charges to the upper-layer NCs, our model shows that the double-layer NC memory enlarges the retention time to program time ratio as well as the memory window, in good agreement with experimental data. Among several optimization approaches investigated, the retention time may be further prolonged by engineering sufficient Coulomb and quantum blockade energy with the lower-layer NC diameter around 1 nm. Meanwhile, extended from the double-layer configuration, a multi-layer scheme with heterogeneous composition warrants both good scalability and retention, and thus is a very promising candidate for future charge-based nonvolatile memory.

4.2 Introduction

Although there has been enormous progress in developing alternative nonvolatile memories such as ferroelectric random access memory (RAM) [1], magnetoresistive RAM [2], and phase-change RAM [3], Flash memory based on electrostatic charge storage in the floating gate will remain as the mainstream mass-storage technology for many years to come [4]-[6]. Flash technology is well established and cost-effective with a small cell size desirable for high-density integration. More importantly, it keeps renovating itself with new developments to stay ahead of the competition, notably the discrete charge storage for better reliability and scalability [7], and the implementation of the multiple bits per cell [8], [9]. As a result, a 32Gb TANOS (TaN-Al₂O₃-nitride-oxide-silicon) multi-level memory with a

cell size as small as $0.0098 \mu\text{m}^2$ will be in mass production shortly [10].

Sharing similar benefits of the TANOS memory with discrete charge storage in isolated traps, the nanocrystal (NC) memory is promising for realizing high-density nonvolatile storage [11]. Moreover, its inherent advantage of low-voltage operation could potentially further increase the longevity of the current Flash technology [12], with the recent demonstration of sub-4V operation [13]. Like all other nonvolatile memories, the main figures of merit for performance evaluation are the ratio between retention time t_R and program time t_P and magnitude of the memory window. The large memory window is particularly important for the multi-level implementation. The double-layer (DL) NC memory utilizing two layers of NCs for charge storage was proposed and shown superior on both figures over the conventional single-layer (SL) NC memory [14]-[17]. On the other hand, toward better understanding in theory and more quantitative design strategy, we recently reported a physical model based on the three-dimensional (3D) electrostatics and the one-dimensional (1D) Wentzel-Kramers-Brillouin (WKB) tunneling current calculation [18], [19], which can accurately simulate the program/erase (P/E) and retention characteristics of the SL NC memory.

In this chapter, we show that the same physical model can be extended to the DL NC memory when appropriate time evolution of charge states is employed. Device operations are then investigated in detail to explain the improved t_R / t_P time ratio and memory window in the DL NC memory. To further explore the entire design space, design criteria including the variations on the inter-NC oxide, the work-function offset and the size ratio between two NC layers are explicitly examined. Other optimization parameters sharing with the SL NC memory, such as the NC array engineering, the control oxide scaling, the high- κ tunnel and control oxide have been reported in previous two chapters and will not be included here to avoid duplications. Finally, a configuration with multiple layers of NCs as the floating gate and its implications on

the device scalability and retention time are presented.

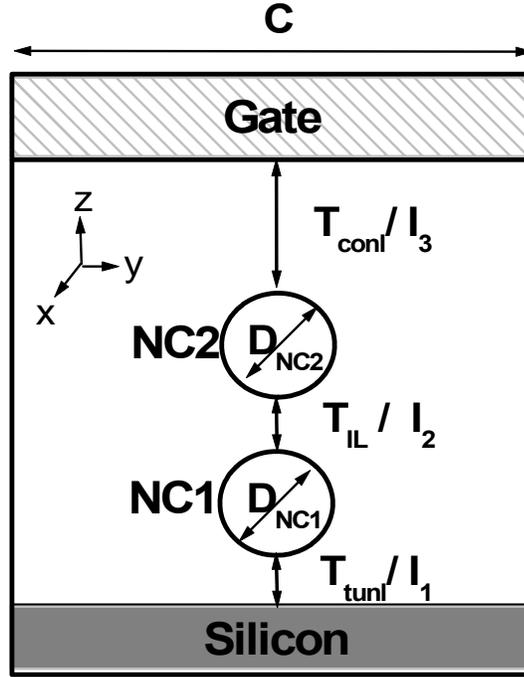


Figure 4.1 Schematic of the unit cell of the DL NC memory, assuming azimuthal symmetry around the central axis in the z direction.

4.3 Device Modeling

Figure 4.1 illustrates the unit cell of the DL NC memory, assuming azimuthal symmetry around the central axis in the z direction. Two layers of spherical metal NCs [16], [20], the lower-layer NC1 with diameter D_{NC1} and the upper-layer NC2 with diameter D_{NC2} , are embedded in a trap-free dielectric, which consists of tunnel oxide, inter-NC oxide, and control oxide with thickness of T_{tunl} , T_{IL} , and T_{conl} , respectively. The NC unit cell diameter C_{NC} is derived from the NC number density N_{NC} . To exploit the simplicity of azimuthal symmetry, we limit ourselves to the case that NC2 is perfectly aligned on top of NC1, which is an approximation in our structure but can be achieved by deliberate process designs [14]. Although only metal NCs are discussed

here, the model is applicable to other semiconductor NCs as well by considering their finite dielectric constant [12], [18]. The currents through the tunnel oxide I_1 , the inter-NC oxide I_2 , and the control oxide I_3 depend on the charge states in NC1 and NC2. They are calculated by solving the 3D electrostatic potential profile inside the gate stack, the 1D WKB tunneling approximation along the least-action path, and appropriate capture cross sections as fitting parameters [18]. Notice that the Coulomb charging energy of NCs for the given geometric configuration is included in the simulation without ambiguity. When D_{NC1} and D_{NC2} are smaller than 5 nm, quantum size effects will become prominent and are also accounted for. Furthermore, the flatband shift ΔV_{FB} under the influence of the charges in the discrete NCs is also directly derived from the 3D electrostatic solution without assuming a coupling ratio [18].

To analyze P/E and retention dynamics, the evolution of charge states in NCs and the corresponding time stepping have to be evaluated. In the SL NC memory, it is intuitive to apply the current continuity equation, which yields monotonically increasing charge evolution under program and monotonically decreasing charge evolution under erase and retention by an integer step [18]. However, satisfying the continuity equation does not necessarily force the system to the steady or thermal equilibrium state after successive evolution. Particularly in the DL NC memory, it is prone to trap in loops due to the additional charge transfer between NC1 and NC2. Here we developed a semi-empirical method to fast evaluate the evolution of charge state based on the minimization of the charge flux. At the thermal equilibrium after long retention time, the charge flux in the system is exactly zero. At the steady state after long P/E time, the charge flux is not only balanced but also reaches its minimum because the increase of electrostatic energy from the stored charge in the gate stack retards the flux driven by the fixed program voltage. In the structure given by Fig. 4.1,

minimization of I_{rms} , the root mean square of all current components including I_1 , I_2 , and I_3 , should guide the system to its steady or thermal-equilibrium states. For instance, in a program process with an initial state $(Q_{NC1}, Q_{NC2}) = (m, n)$, where Q_{NC1} and Q_{NC2} are the numbers of charges in NC1 and NC2 respectively, there are two possible nearest states, $(m+1, n)$ and $(m-1, n+1)$. The former represents an additional electron injected into NC1 from the substrate, while the later represents an electron transferring from NC1 to NC2 without substrate injection. For simplicity of representation, the charge transfer between NC2 and the control gate is assumed disabled with the thicker T_{conl} . The minimum of $I_{rms}(m+1, n)$ and $I_{rms}(m-1, n+1)$ is then chosen to determine which state should be evolve from (m, n) . Even though in some cases I_{rms} of the next state could be higher than that of its previous state, the general trend of I_{rms} minimization still retains through the entire evolution of charge states. This is a common minimization problem with both local and global minima, where the true steady or equilibrium state resides at the global minimum. This method is applicable to the SL case yielding the intuitive monotonic sequence, and is essential for multiple layers of NCs where the inter-NC charge transfer becomes crucial in determining the memory operations.

4.4 Results and Discussion

4.4.1 Program Characteristics

The fabrication process of the SL and DL Au metal NC memory was similar to that described before [16], [20], [21]. The DL sample differs from the SL one by only the insertion of the inter-NC oxide and the upper-layer NCs. Figure 4.2 demonstrates good agreement between simulated and experimental program transients in SL and DL samples. The same simulation parameters from various types of physical characterization were applied in both SL and DL cases, including a p-type

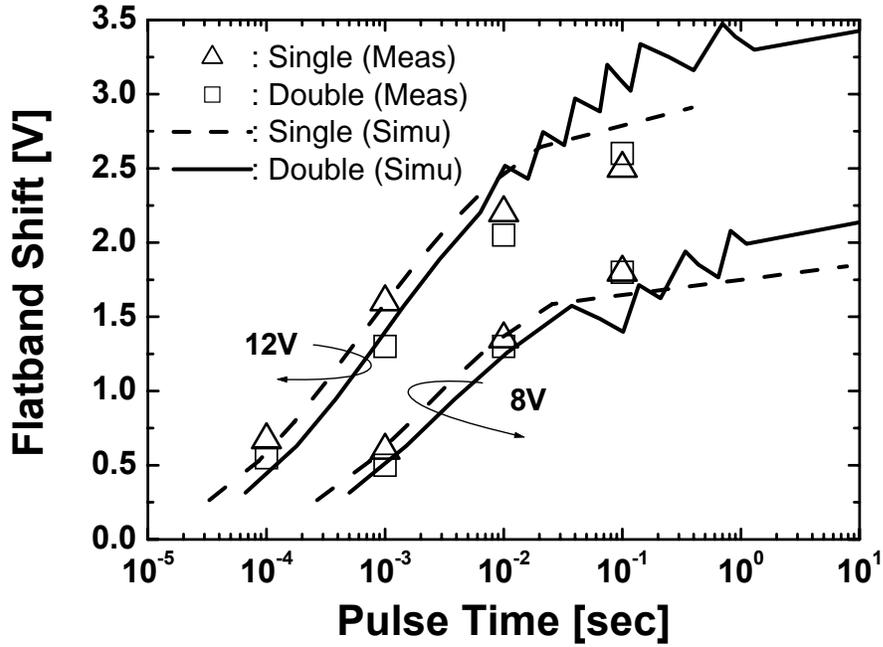


Figure 4.2 Measured and calculated flatband voltage shift versus program pulse time for both SL and DL devices at $V_G = 8$ V and $V_G = 12$ V. A consistent set of parameters is applied in the simulation.

(100) Si substrate with a doping level of $2 \times 10^{17} \text{ cm}^{-3}$, an initial flatband voltage $V_{FB} = 0$ before program, NC work function $\Phi_{NC} = -5.1 \text{ eV}$, $T_{tunl} = 2 \text{ nm}$, $T_{IL} = 2.5 \text{ nm}$, $T_{cont} = 27 \text{ nm}$, $D_{NC1} = D_{NC2} = 5 \text{ nm}$, $C_{NC} = 18 \text{ nm}$, and $N_{NC} = 4 \times 10^{11} \text{ cm}^{-2}$. The nominal T_{IL} is close to 5 nm by plasma enhance chemical vapor deposition (PECVD). The better fitting from a thinner T_{IL} in the simulation probably results from the non-ideal step coverage over NC1 and electrical properties of the PECVD oxide. The pure fitting parameters in our model are σ_{NC1} and σ_{NC2} , the capture cross sections of NC1 and NC2. $\sigma_{NC1} = \sigma_{NC2} = 5.3 \times 10^{-14} \text{ cm}^2$ per NC was used in the simulation, although they do not necessarily have to be equal due to the 3D nature of the tunneling process. Better fitting can be expected by tuning σ_{NC1} and σ_{NC2} independently, but that should not affect the general results of this study. Unless otherwise mentioned, the above parameters are consistently used throughout this chapter.

In Fig. 4.2, the program speed for a moderate ΔV_{FB} in the DL sample is slightly slower as a result of the thicker oxide thickness seen from the control gate with the addition of inter-NC oxide. However, the maximum memory window is slightly larger because of the trade-off between the memory window and control oxide thickness [19] and larger charge storage capability provided by NC2. The evident oscillations in the program transients occur when the charges stored in NC1 transfer into NC2, which is farther away from the channel and has less influence on ΔV_{FB} . Notice that, instead of oscillation, a statistical average is more likely to be observed experimentally, considering time stepping in each unit cell of the NC array is not synchronous due to the non-ideality of NC size and spatial distribution. With very little computational effort, the current model based on a single unit cell presents satisfactory estimation of the memory performance.

The calculated band diagrams along the least-action path and the corresponding evolution of charge state are plotted in Fig. 4.3. In the beginning of the program step, electrons accumulate in NC1 due to the larger I_1 through the thinner T_{unt} in comparison with the smaller I_2 through the thicker T_{IL} . This also represents the regime with smooth increase of ΔV_{FB} in Fig. 4.2. It is worth mentioning that the largest electric field in Fig. 4.3(a) occurs at the inter-NC oxide as a consequence of the field enhancement [12], [18] from both NC1 and NC2. The charge and potential build-up in NC1 finally increase I_2 to a comparable level with I_1 , and prompt the charge transfer from NC1 to NC2. In the presented configuration, the memory window is self-saturated as the Si channel departing from strong inversion. In Fig. 4.3(b), the DL sample manages to store more charges by distributing a part of them to NC2 while keeping the NC1 potential steady and the Si channel under inversion, resulting in the larger saturated memory window.

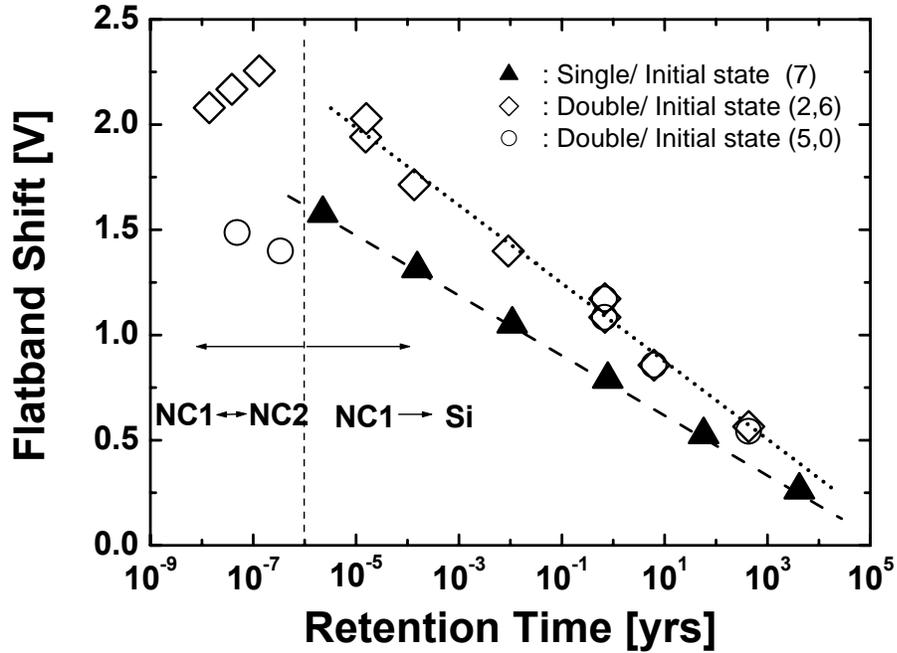


Figure 4.4 Calculated room-temperature retention time at $V_G = 0$ V for both SL and DL devices. The DL devices with two different initial states after program exhibit similar retention characteristics.

4.4.2 Retention Characteristics

In Fig. 4.4, the retention characteristic is calculated. The DL sample exhibits 1 to 2 orders of magnitude longer retention time than the SL sample, consistent with the experimental results reported previously [14], [16], [20]. It also shows two distinct regimes in the curve of the DL sample. Depending on the initial charge state, in a time scale of seconds, the redistribution of charges through the inter-NC transfer results in increasing or decreasing ΔV_{FB} . The long-term retention is still set by the charge loss to the substrate, which is determined by the band offset between Φ_{NC} and the Si conduction band edge, independent of the initial charge state. Notice that the charge redistribution effect still needs to be carefully considered to prevent possible read errors especially for the multi-level memory, although it does not significantly affect retention.

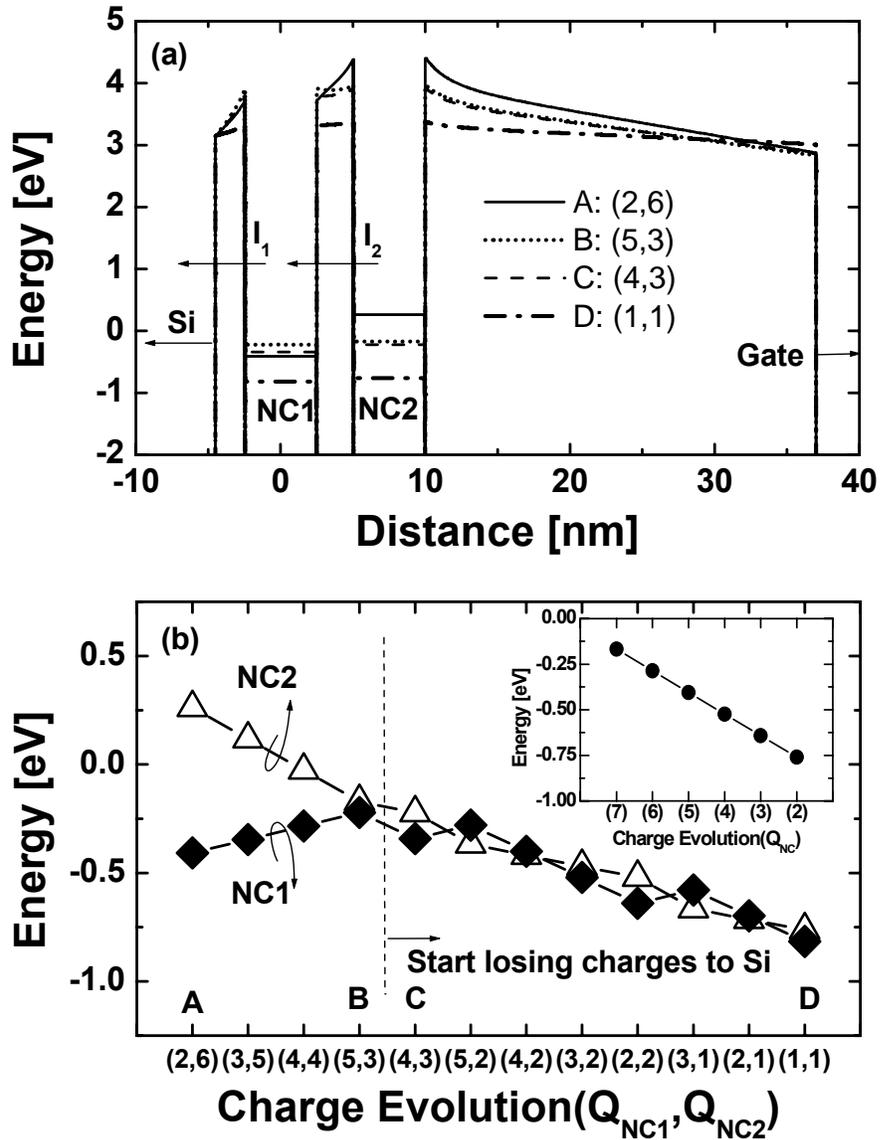


Figure 4.5 (a) Calculated energy band diagrams of the DL NC device along the least-action path as a function of the charge state (Q_{NC1} , Q_{NC2}) at $V_G = 0V$. The Si conduction band edge is referred at 0 eV. (b) Evolution of charge states during retention and the corresponding NC potential energies, with the inset showing a comparison to the SL device.

Figure 4.5 illustrates the calculated band diagrams along the least-action path and corresponding evolution of charge state during retention. When the external bias is removed after program, large imbalance between the NC1 and NC2 potentials may

appear. With available states for transition by tunneling through the inter-NC oxide, the potential balance are restored shortly as shown in Fig. 4.5 (b), followed by the long-term charge loss to the Si substrate. With the same number of charges stored, the NC potentials in the DL sample are consistently lower than that in the SL sample during retention. In other words, the average Coulomb charging energy for each electron in the DL sample is smaller, thanks to the even charge distribution in a larger storage volume. Lower NC potentials provide larger band offset to the Si conduction band edge, and thus grant better retention.

It was long believed that the larger Coulomb and quantum confinement energy in NC1 prevent the charges in NC2 from back tunneling to NC1 [14], [16]. Therefore, better retention is probable in the DL NC memory. Nevertheless, in most experiments reported [14]-[17], [20], the size of NC1 is still very substantial, and multiple charges can be accommodated in NC1. Once the charges in NC1 back tunnel to the substrate during retention, those in NC2 experience little blockade to backfill NC1 as already illustrated in Fig. 4.5. Therefore, strictly speaking, the physical explanation based on the blockade effect [14] is somewhat misleading. The aforementioned larger charge storage capability presents a clearer physical picture. For the blockade effect to be dominant, which will be later investigated in Sec. 4.4.5, the NC1 size has to be significantly scaled [22] so that NC1 serves mainly as a potential barrier to the back-tunneling current rather than a charge storage node.

With only a fractional degradation in the program speed and orders of magnitude improvement in retention, the t_R / t_P ratio of the DL NC memory outperforms the SL counterpart, in addition to the larger memory window which is critical for the multi-level memory. We continue in the following sections to further elaborate the optimization strategies of the DL NC memory.

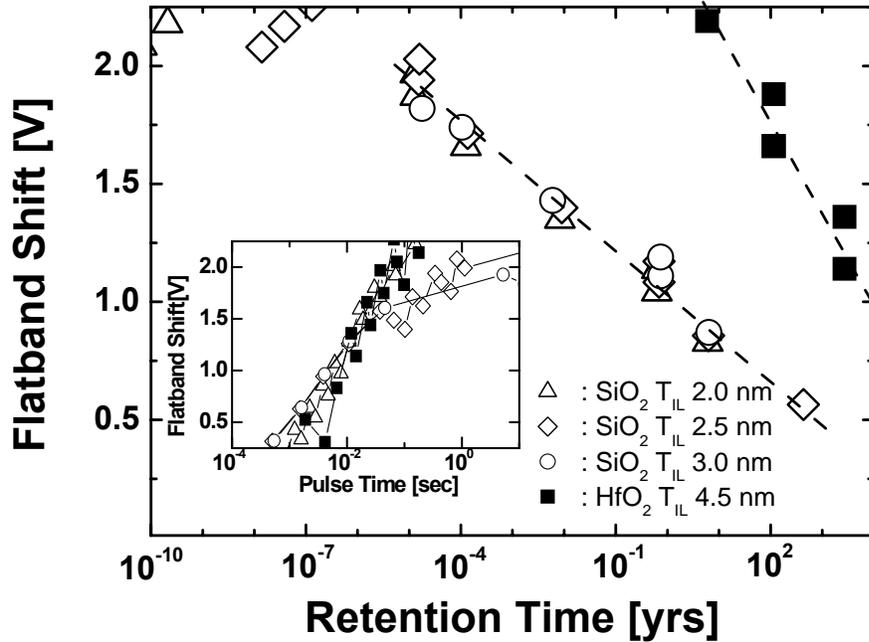


Figure 4.6 Calculated room-temperature retention time at $V_G = 0$ V for DL devices with variations of inter-NC oxide. The inset shows the calculated program transients at $V_G = 8$ V.

4.4.3 Inter-NC Oxide

If the inter-NC oxide T_{IL} is too thick to allow efficient inter-NC charge transfer in the time scale of interest for program and retention, the advantages of larger memory window and longer retention time due to the additional storage capacity provided by NC2 diminish dramatically. On the other hand, for a moderate T_{IL} ranging from 2 nm to 3 nm, variations on T_{IL} may affect the time constant of the charge transfer, but hardly disturb the overall memory performance. As shown in Fig. 4.6 and its inset, the calculated program and retention characteristics are relatively independent of T_{IL} , except that larger ΔV_{FB} is obtained with thinner T_{IL} for very long program pulses. Moreover, the dielectric constant of the inter-NC oxide influences device operations significantly through 3D electrostatics. Similar to high- κ dielectrics as the control oxide [19], by replacing SiO₂ with HfO₂ as the inter-NC oxide in Fig.

4.6, remarkable improvement in memory performance is not surprising due to its lower Coulomb charging energy and stronger coupling to the channel. The thinner equivalent oxide thickness of HfO_2 is chosen to allow inter-NC charge transfer. Otherwise, the device essentially operates as a SL NC memory with the high- κ control oxide.

4.4.4 *Work-Function Offset of Nanocrystals*

Tuning metal NC work function by selecting appropriate metal species was regarded as an important step to realize larger retention time and memory window [16]. In the DL NC memory, another degree of freedom by engineering work function separately for NC1 and NC2 can potentially bring about other intriguing effects. For instance, the design with the NC1 work function higher than the NC2 seems to be able to reduce back-tunneling from NC2 to NC1 during retention, as a result of the potential barrier built by the work-function offset. This in principle can lead to better retention characteristics. On the contrary, the NC2 work function higher than the NC1 blockades charge injection to NC2 during low-voltage program. Therefore, multiple memory states may be controlled more precisely by variations of program voltage instead of pulse time. However, while the charge transfer between the substrate and NCs during retention is prohibited by the Si band gap except for very few energetic carriers, the time constant of the inter-NC charge transfer, set by merely the barrier of the inter-NC oxide, is relatively short even in the absence of external bias. For a NC1 work function of -5 eV and a NC2 work function of -4 eV, the difference of work functions vanishes almost immediately in a matter of seconds. Hence, we conclude that designing the work-function offset between two layers of NCs will not be a significant effect.

4.4.5 Nanocrystal Size

In our previous study on the SL NC memory in Chapter 2, the large NC size was found beneficial due to smaller charging energy, stronger field enhancement, and larger capture cross-section. Therefore, in the DL NC memory, it is also possible to explore the advantages of the NC size engineering, especially the size ratio of NC1 to NC2. The calculated program and retention characteristics are presented in Fig. 4.7 with respect to the NC2 size while keeping the NC1 size and the NC number density constant. Here we assume the capture cross-section to be proportional to its physical dimension. The program transients appear to be insensitive to the NC2 size, which agrees with the previous result that the program process is mostly governed by NC1 prior to the inter-NC charge transfer taking place. Meanwhile, enlarging the NC2 size lowers its Coulomb charging energy and allows more charges being distributed in NC2 and less in NC1. Thus better retention is enabled by the larger band offset between NC1 and the Si conduction band edge. However, given the fact that the program speed is slower in the DL sample, a SL implementation with larger NCs seems more straightforward considering its faster speed, easier process and comparable retention. But it is worth mentioning that, as the continuous memory cell scaling now beyond the sub-50nm node, increasing the NC number density is essential to sustain a reasonable parametric yield [23], which makes enlargement of the NC size in either SL or DL NC memory impractical.

Another design perspective is to exploit the blockade effect to achieve better retention by shrinking the NC1 size, so that during the retention NC1 actually acts as a potential barrier to the charges in NC2 instead of a charge storage node. Besides the Coulomb blockade effect already included in our 3D electrostatic model, quantum-mechanical effects become significant in NCs with diameter smaller than 5 nm. In contrast to the significant bandgap increase in the semiconductor NCs of nanometer

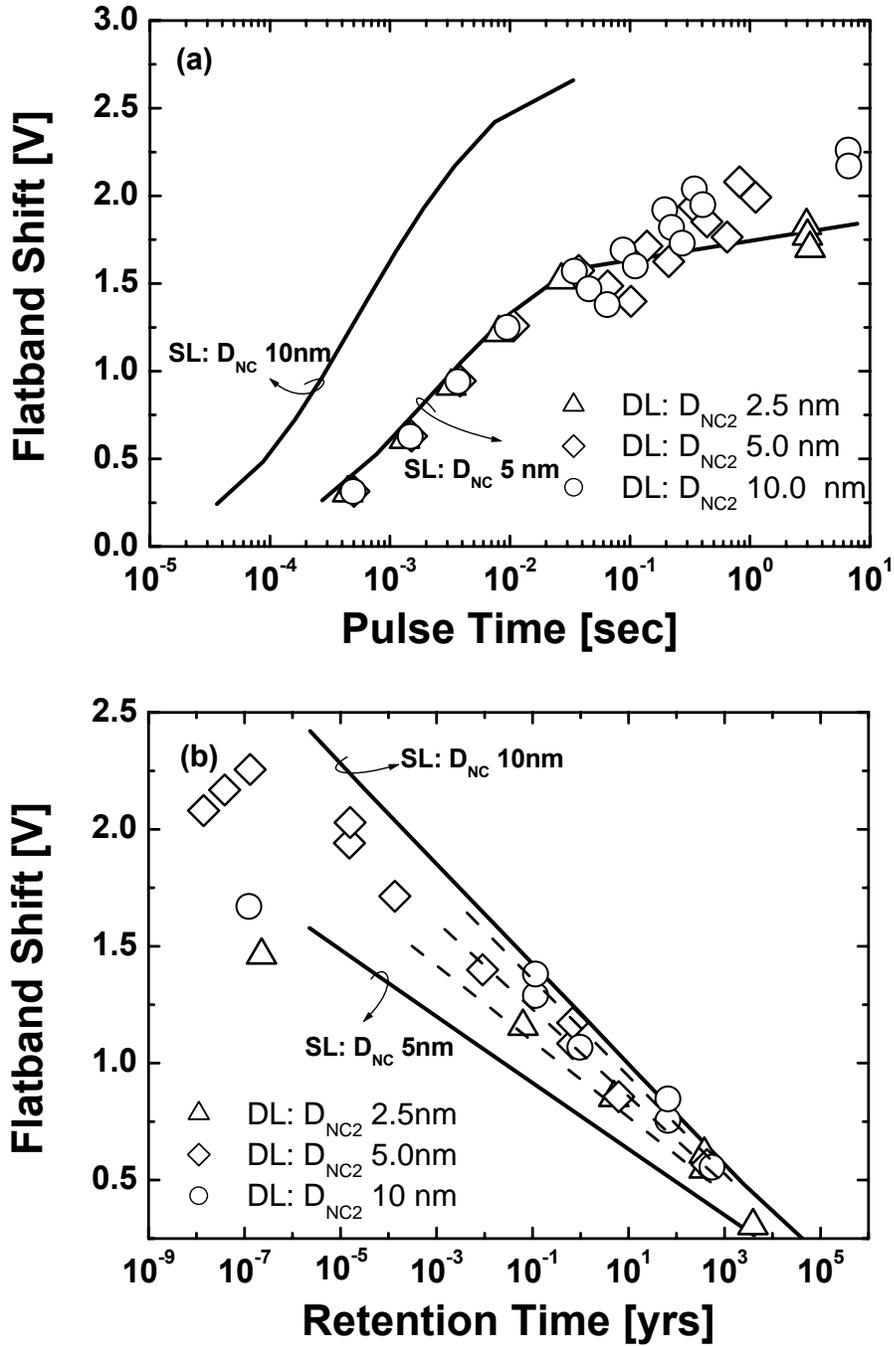


Figure 4.7 (a) Calculated program transients at $V_G = 8$ V and (b) room-temperature retention time at $V_G = 0$ V for DL devices with NC2 diameter ranging from 2.5 nm to 10 nm. The NC1 diameter of 5 nm and the NC number density of $4 \times 10^{11} \text{ cm}^{-2}$ are fixed. The results from SL devices with the same configuration are also plotted for comparison.

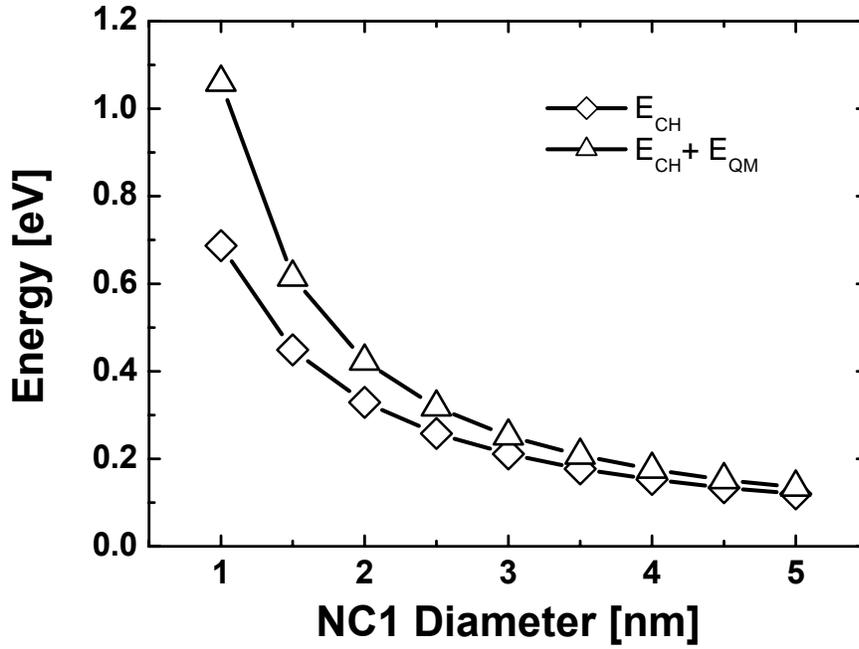


Figure 4.8 Calculated Coulomb charging energy E_{CH} and total blockade energy including quantum confinement energy E_{QM} as a function of the NC1 diameter. The NC2 diameter of 5 nm and the NC number density of $4 \times 10^{11} \text{ cm}^{-2}$ are fixed.

dimension, the electronic structures of most metal NCs are relatively undisturbed from bulk materials even down to a 1-nm diameter. The average energy-level spacing of successive quantum levels, known as the Kubo gap, is often close to or smaller than the thermal energy at room temperature [24]. The other more pronounced quantum-size effect originates from the tiny quantum potential well. Instead of solving the Schrödinger equation exactly, the first-order approximation of the quantum confinement energy E_{QM} was carried out by the well-known ground-state eigenenergy E_1 of the infinite square well, which can be expressed as [25]:

$$E_1 = \frac{\pi^2 \hbar^2}{2m^* D_{NC}^2} \quad (4.1)$$

where m^* is the electron effective mass in NCs. Figure 4.8 depicts the Coulomb charging energy E_{CH} and the total blockade energy $E_{CH} + E_{QM}$ as a function of the

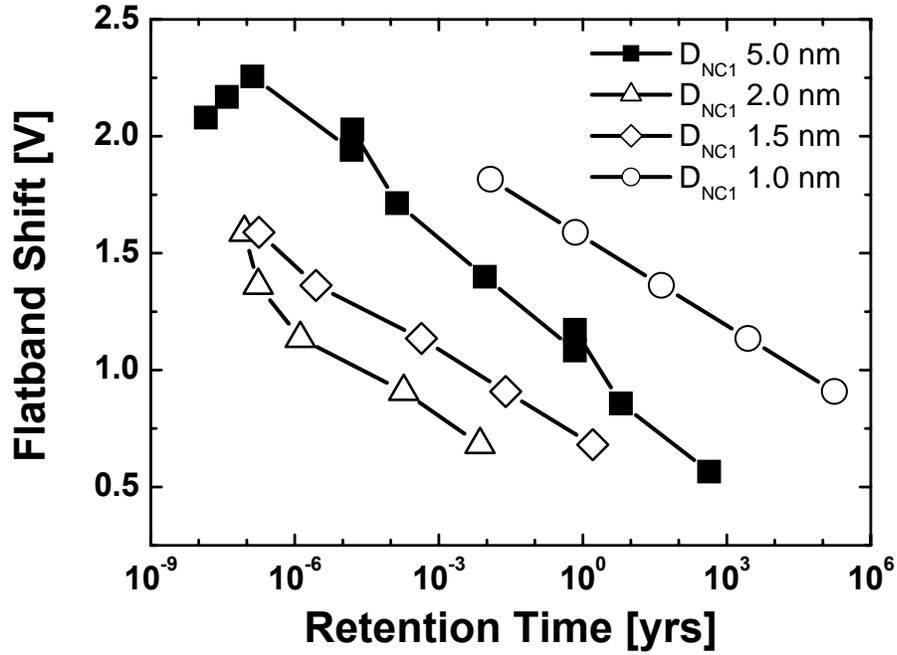


Figure 4.9 Calculated room-temperature retention time at $V_G = 0$ V for DL devices with NC1 diameter ranging from 1 nm to 5 nm. The NC2 diameter of 5 nm and the NC number density of $4 \times 10^{11} \text{ cm}^{-2}$ are fixed. Notice that the trend is not monotonic with respect to the NC2 size, with the worst retention at 2 nm and the best retention at 1 nm.

NC1 diameter. It was found that the Coulomb charging energy remains dominant through most of the interested range of the metal NC diameter.

Given the charging energy of NC2 at hundreds of milli-eV, sizeable blockade energy in a range of eV is required to retain multiple charges in NC2 for an appreciable retention time and a realistic memory window. The calculated retention characteristics with NC1 diameters ranging from 1 nm to 5 nm are illustrated in Fig. 4.9. The improvement over the nominal device can only be obtained for a NC diameter of 1 nm and the blockade energy up to 1 eV. For other slightly larger NCs, the sharply decreased blockade energy of NC1 is no longer sufficient to prevent back-tunneling from NC2. Because of its own large E_{CH} and E_{QM} , NC1 is ineffective to hold

charges either. Besides, with the much smaller NC1, most charges are located at NC2 with less influence on the channel V_{FB} . All these effects contribute to the abrupt degradation of retention time with slightly larger NCs. By replacing metal NCs with semiconductor NCs, the large bandgap increase may relax the strict requirement of the NC size [22]. On the other hand, the program efficiency may be compromised due to the same blockade energy of NC1, which reduces charge injection from the channel to NC1. However, this blockade effect is less significant when high program voltage is allowed [14]. In summary, depending on its size, NC can act as a charge storage node or an effective tunnel barrier. The tradeoff has significant impacts on the retention characteristics.

4.4.6 Multiple layer of NCs

As discussed previously, enlarging the NC2 size is favorable for better retention but seems an impractical solution for aggressively scaled devices. By replacing one layer of oversize NC2, multiple upper layers of NCs with smaller size each can retain similar advantages. Figure 4.10 exhibits further improved t_R / t_P ratio in a quad-layer NC memory with the same NC diameter of 5 nm and the NC number density as the DL NC memory. In order to integrate more layers of NCs in a reasonable fabrication process, an ultimate multiple-layer implementation is by replacing all upper-layer NCs and inter-NC oxides with a dielectric with a large number of intrinsic traps such as nitride. Although only a single charge is able to be stored in each trap site, the large number of traps (in both the planar number density and the number of vertical layers) by engineering the trap density and layer thickness may still provide sufficient storage capability. This class of heterogeneous NC/nitride memory has been experimentally shown to be superior to the SONOS or SL NC memory [26], [27]. Unlike the conventional SONOS memory, the lower-layer NC

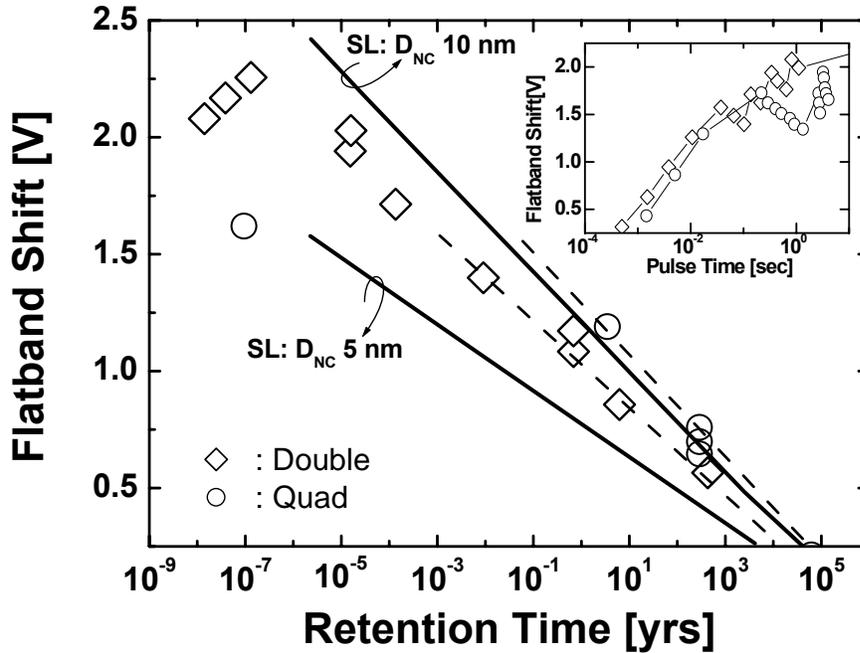


Figure 4.10 Calculated room-temperature retention time at $V_G = 0$ V for DL and quad-layer devices. The inter-NC oxide is 2.5 nm, the NC diameter is 5 nm and the NC number density is $4 \times 10^{11} \text{ cm}^{-2}$ in all layers. The inset shows the calculated program transients at $V_G = 8$ V.

ensures fast program in low voltage, while the charge redistribution to traps helps retention. More recently, by utilizing the blockade effect of small Si NCs with a diameter of 1 nm, better retention over SONOS devices with little program efficiency sacrifice has been demonstrated in a 25 nm heterogeneous device [22].

4.5 Conclusion

A physical model based on the 3D electrostatics and the 1D WKB tunneling calculation is presented and found to be applicable to both SL and DL NC memories. The superior retention characteristics and memory window in the DL NC memory are explained by the larger storage capacity due to the charge redistribution to the upper-layer NCs. By scaling NC1 size down to 1 nm or smaller, the back-tunneling current

from NC2 is suppressed by the Coulomb and quantum confinement energy of NC1. Hence, further improvement on retention time can be feasible. Finally, the multiple-layer configuration such as the heterogeneous NC/nitride layer is shown promising for further aggressive scaling.

REFERENCES

- [1] J.-H. Kim, D. J. Jung, S. K. Kang, Y. M. Kang, H. H. Kim, J. Y. Kang, E. S. Lee, W. W. Jung, H. J. Joo, J. Y. Jung, J. H. Park, H. Kim, D. Y. Choi, S. Y. Lee, H. S. Jeong, and K. Kim, "Manufacturing technologies for a highly reliable 0.34 μm^2 -cell, 64 Mb, and 1T1C FRAM, in *IEDM Tech. Dig.*, 2006, pp. 45-48.
- [2] S. Tehrani, "Status and outlook of MRAM memory technology", in *IEDM Tech. Dig.*, 2006, pp. 585-588.
- [3] J. H. Oh, J. H. Park, Y. S. Lim, H. S. Lim, Y. T. Oh, J. S. Kim, J. M. Shin, J. H. Park, Y. J. Song, K. C. Ryoo, D. W. Lim, S. S. Park, J. I. Kim, J. Yu, F. Yeung, C. W. Jeong, J. H. Kong, D. H. Kang, G. H. Koh, G. T. Jeong, H. S. Jeong, and K. Kim, "Full integration of highly manufacturable 512 PRAM based on 90nm technology," in *IEDM Tech. Dig.*, 2006, pp. 49-52.
- [4] Y. Shin, "Non-volatile memory technologies for beyond 2010," in *Proc. VLSI Circ. Dig.*, 2005, pp. 156-159.
- [5] L. Geppert, "The new indelible memories," *IEEE Spectrum*, vol. 40, no. 3, pp. 48-54, Mar. 2003.
- [6] N. Flaherty, "Not in a flash," *IEE Review*, vol. 49, no. 11, pp. 50-53, Dec. 2003.
- [7] B. DeSalvo, C. Gerardi, R. van Schaijk, S. A. Lombardo, D. Corso, C. Plantamura, S. Serafino, G. Ammendola, M. van Duuren, P. Goarin, W. Y. Mei, K. van der Jeugd, T. Baron, M. Gely, P. Mur, and S. Deleonibus, "Performance and reliability features of advanced nonvolatile memories based on discrete traps (silicon nanocrystals, SONOS)", *IEEE Trans. Device Mater. Rel.*, vol. 4, no. 3, pp. 377-389, Sep. 2004.
- [8] B. Eitan, R. Kazerounian, A. Roy, G. Crisenza, P. Cappelletti, and A. Modelli, "Multi-level flash cells and their trade-offs", in *IEDM Tech. Dig.*, 1996, pp. 169-172.

- [9] B. Eitan, P Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "NRROM: A novel localized trapping, 2-bit nonvolatile memory cell," *IEEE Electron Device Lett.*, vol. 21, no. 11, pp. 543-545, Nov. 2000.
- [10] Y. Park, J. Choi, C. Kang, C. Lee, Y. Shin, B. Choi, J. Kim, S. Jeon, J. Sel, J. Park, K. Choi, T. Yoo, J. Sim, and K. Kim, "Highly manufacturable 32Gb multi-level NAND Flash memory with $0.0098 \mu\text{m}^2$ cell size using TANOS(Si-Oxide- Al_2O_3 -TaN) cell technology," in *IEDM Tech. Dig.*, 2006, pp. 29-32.
- [11] S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, and D. Buchanan, "Volatile and non-volatile memories in silicon with nanocrystal storage," in *IEDM Tech. Dig.*, 1995, pp. 521-524.
- [12] C. Lee, U. Ganguly, V. Narayanan, T.-H. Hou, and E. C. Kan, "Asymmetric electric field enhancement in nanocrystal memories," *IEEE Electron Device Lett.*, vol. 26, no. 12, pp. 879-881, Dec. 2005.
- [13] J. J. Lee, and D.-L. Kwong, "Metal nanocrystal memory with high- κ tunneling barrier for improved data retention," *IEEE Trans. Electron Devices*, vol. 52, no. 4, pp. 507-511, Apr. 2005.
- [14] R. Ohba, N. Sugiyama, K. Uchida, J. Koga, and A. Toriumi, "Nonvolatile Si quantum memory with self-aligned double-stacked dots," *IEEE Trans. Electron Devices*, vol. 49, no. 8, pp. 1392-1398, Aug. 2002.
- [15] M. Takata, S. Kondoh, T. Sakaguchi, H. Chou, J-C. Shim, H. Kurino, and M. Koyanagi, "New non-volatile memory with extremely high density metal nanodots," in *IEDM Tech. Dig.*, 2003, pp. 553-556.
- [16] C. Lee, A. Gorur-Seetharam, and Edwin C. Kan, "Operational and reliability comparison of discrete-storage nonvolatile memories: Advantages of single- and double-layer metal nanocrystals," in *IEDM Tech. Dig.*, 2003, pp. 557-560.

- [17] T. Z. Lu, M. Alexe, R. Scholz, V. Talelaev, and M. Zacharias, "Multilevel charge storage in silicon nanocrystal multilayers," *Appl. Phys. Lett.*, vol. 87, p. 202110, 2005.
- [18] T.-H. Hou, C. Lee, V. Narayanan, U. Ganguly, and E. C. Kan, "Design optimization of metal nanocrystal memory—part I: nanocrystal array engineering," *IEEE Trans. Electron Devices*, vol. 53, no. 12, pp. 3095-3102, Dec. 2006.
- [19] 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 Trans. Electron Devices*, vol. 53, no. 12, pp. 3103-3109, Dec. 2006.
- [20] C. Lee, "Charge storage in nanocrystals," Ph.D. thesis, Cornell University, 2004.
- [21] C. Lee, J. Meter, V. Narayanan, and E. C. Kan, "Self-assembly of metal nanocrystal on ultra-thin oxide for nonvolatile memory applications," *J. Electronic Materials*, vol. 34, no. 1, pp. 1-11, Jan. 2005.
- [22] R. Ohba, Y. Mitani, N. Sugiyama, and S. Fujita, "25 nm planar bulk SONOS-type memory with double tunnel junction," in *IEDM Tech. Dig.*, 2006, pp. 959-962.
- [23] L. Perniola, B. D. Salvo, G. Ghibaudo, A. F. Para, G. Pananakakis, V. Vidal, T. Baron, and S. A. Lombardo, "Modeling of the programming window distribution in multi-nanocrystals memory," *IEEE Trans. Nanotechnol.*, vol. 2, no. 4, pp. 277-284, Dec. 2003.
- [24] C. N. R. Rao, G. U. Kulkarni, P. J. Thomos, and P. P. Edwards, "Size-dependent chemistry: properties of nanocrystals," *Chem. Eur. J.*, vol. 8, no. 1, pp. 29-35, Jan. 2002.
- [25] D. J. Griffiths, *Introduction to quantum mechanics*, 2nd ed, NJ: Prentice-Hall, 2005, pp.30-40.

- [26] S. Huang, K. Arai, K. Usami, and S. Oda, "Toward long-term retention-time single-electron-memory devices based on nitrated nanocrystalline silicon dots," *IEEE Trans. Nanotech.*, vol. 3, no. 1, pp. 210-214, Mar. 2005.
- [27] C. Lee, T-H Hou, and E. Kan, "Nonvolatile memory with a metal nanocrystal/nitride heterogeneous floating-gate," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2697-2702, Dec. 2005.

CHAPTER 5

FERMI-LEVEL PINNING IN NANOCRYSTAL MEMORIES

5.1 Abstract

Nanocrystal (NC) work function engineering, which plays an important role on the NC memory characteristics such as memory window and retention time, was long regarded as a matter of choice on NC materials. In this chapter, we report opposite polarities of charge storage in Au NC memories with different control oxides. The effective NC work function is found to be not only a bulk property of the NC, but also governed by the interface with surrounding dielectric, as a result of the Fermi-level pinning. By replacing Au NCs with C_{60} molecules, we also show the pinning effect generally exists at quantum-dot-based devices with high density of interface states. This fundamental interface property should be taken into account in the selection of NC and dielectric materials for the NC memory optimization.

5.2 Introduction

Nanocrystal (NC) memories are considered as one of the promising candidates for future nonvolatile, high-density and low-voltage memory applications, owing to their inherent immunity to the local oxide defects by discrete charge storage, which allows more aggressive scaling of the tunnel oxide thickness [1]. Among the efforts to further improve the performance of NC memories, metal NC was proposed above its semiconductor counterpart due to large density of states, three-dimensional electric field enhancement, and selectable work function [2], [3]. The work function engineering provides another degree of freedom to increase the charge storage capacity and retention time with a deeper potential well, whereas the typical Si NC has no band offset to the Si channel. A wide range of metal NC materials, such as Au, Ag,

Pt, Ni, and W, have been successfully implemented [3]-[7]. In addition, high- κ dielectrics were introduced in NC memories to facilitate the equivalent oxide thickness scaling, the inter-poly leakage current reduction, and the field-sensitive tunneling [8]. Despite the obvious benefits of integrating both metal NCs and high- κ dielectrics into NC memories, the potential Fermi-level pinning at the metal NC/high- κ dielectric interface has not yet been fully investigated. Fermi-level pinning is known to shift the effective gate work functions of metal/high- κ and polysilicon/high- κ gate stacks substantially [9]-[11]. Similar effects may be proven critical in terms of work function engineering on metal NCs.

In this chapter, a systematic set of experiments with material variations on the floating gate and control oxide were carried out. Two types of floating gates, Au NCs and C_{60} molecules, were studied. The purposes of implementing C_{60} molecules as the floating gate are twofold. First, the monodisperse nature of C_{60} results in substantial and precise step-charging into its molecular orbitals (MO), and hence could potentially achieve reliable multi-level charge storage in a single memory cell [12]. More details about the C_{60} experiment will be highlighted in Chapter 6. Second, unlike some metal NCs such as Pt [13] and Ni, use of C_{60} can eliminate the concern of chemical reaction with surrounding dielectric that makes the definition of NC work function ambiguous. The superior chemical stability of C_{60} due to its entirely satisfied chemical bonding and the known MO are thus very feasible to study the Fermi-level pinning effect. The floating gate work function is found to be strongly affected by the surrounding dielectric, which is well explained by the Fermi-level pinning theory. This interfacial nonideality should be considered carefully in the selection of NC and control oxide materials for optimizing the memory characteristics.

5.3 *Device Fabrication*

The metal-oxide-semiconductor (MOS) capacitor structure with conventional isolation, similar to that used in the Flash memory devices [13], was fabricated on a p-type (100) Si substrate of 1-5 $\Omega\cdot\text{cm}$ resistivity. After 1.8-nm dry thermal oxide was grown on the Si substrate, spherical Au nanocrystals were self-assembled on the oxide by the electron-beam evaporation of 1.2-nm Au without annealing. In another alternative structure, the fullerenes C_{60} were thermally evaporated to a thickness of 0.4 to 0.6 nm as measured by the quartz crystal monitor [14]. Then, control oxide was implemented by atomic layer deposition (ALD) of 30-nm Al_2O_3 or by evaporation of 4-nm SiO_2 together with plasma-enhanced chemical vapor deposition (PECVD) of 26-nm SiO_2 . ALD Al_2O_3 deposition was done by trimethylaluminum (TMA) and H_2O at 300°C. The evaporated oxide served as a barrier layer to prevent possible plasma damage to C_{60} during the PECVD process. Finally a top Cr gate was deposited and patterned, followed by 400°C forming gas annealing for 30 min.

5.4 *Fermi-Level Pinning Theory in NC Memory*

High frequency capacitance-voltage (CV) measurements (Fig. 5.1) were carried out by applying stress voltages on the control gate for 3 s, followed by sweeping CV from positive to negative for substrate electron injection and from negative to positive for substrate hole injection, respectively. There is a remarkably larger memory window swing by substrate electron injection in the Au NC capacitor with Al_2O_3 control oxide. On the contrary, the injected holes easily tunnel back to the Si substrate as the CV gate bias sweeps from negative to positive, which leads to the CV stretch-out and small memory windows. The above observation suggests electron storage be favored in the Au NC capacitor with Al_2O_3 . However, opposite hole storage is found preferable with evaporated SiO_2 . There are a few possible explanations. We

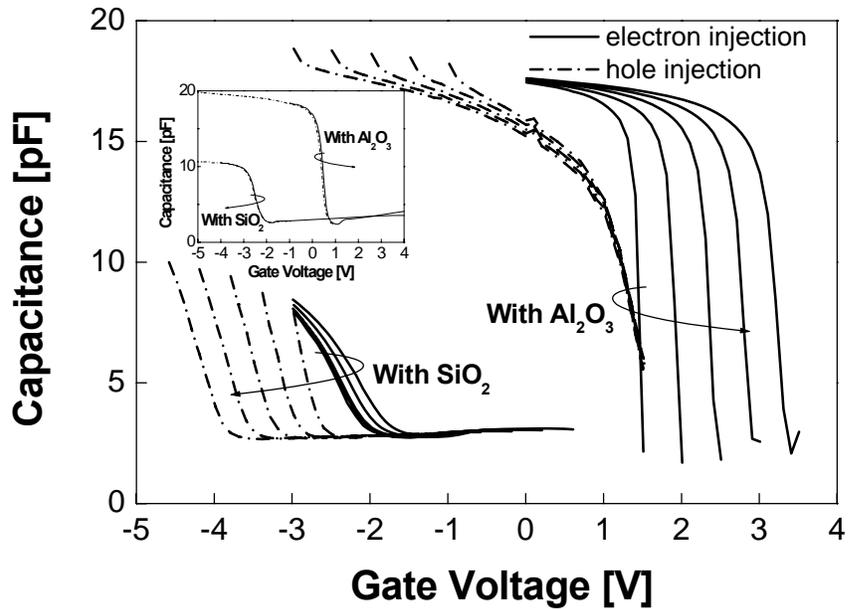


Figure 5.1 High-frequency CV measurements of Au NC capacitors with Al₂O₃ control oxide and evaporated/PECVD SiO₂ control oxide. The capacitor area is 10⁻⁴ cm⁻². The stress voltages for electron injection are +1.5 V to +3.5 V in 0.5 V step with Al₂O₃ and -1 V to +0.6 V in 0.4 V step with SiO₂, respectively. The stress voltages for hole injection are -1 V to -3 V in -0.5 V step with Al₂O₃ and -3 V to -4.6 V in -0.4 V step with SiO₂, respectively. Negligible flatband shift in the control samples with stress voltages of ±5V is shown in the inset.

first eliminate that charges are injected directly into the bulk traps of the control oxide. Negligible flatband shift is found in the control samples without NCs at similar program conditions, as shown in the inset of Fig. 5.1. Furthermore, CV measurements at cryogenic temperature of 10K (not shown) exhibit very similar asymmetric memory window and control oxide dependence as at room temperature. Because of the weak temperature dependence and the low oxide field we used, it is unlikely that charges are first injected into NCs and then hop to farther bulk traps toward the control gate through the Frenkel-Poole (F-P) emission, which is believed to be the injection mechanism in other trap-based nonvolatile memories [15]. Therefore, we conclude

that charges are most likely stored in NCs. If there is charge transfer between NCs and traps in the control oxide, those traps or so-called interface states must be located in a short distance from NCs where the direct tunneling current is significant.

In typical CV measurements, the slow ramp rate of DC voltage together with the initial voltage setup time ensures the program operation reaches the steady state at each sampling point. Hence, the charge storage capability is insensitive to the differences of electron / hole tunnel barrier heights and transport effective mass, but mainly governed by the metal NC work function [13]. Theoretically, electron storage in metal NCs with high work functions is preferable because of the deeper electron potential well formed by the NC-Si band offset. Likewise, metal NCs with low work function facilitate hole storage, while metal NCs with mid-gap work function exhibit symmetric charge storage capacity. However, better understanding on the influence of the NC–dielectric interface is required to interpret the contrary results obtained from ALD Al_2O_3 and evaporated SiO_2 that surround the same Au NC with the work function of -4.94 eV.

Similar phenomena have been observed previously without thorough understanding and detailed explanation [3]-[7]. Table 5.1 summarizes the preferable polarity of charge storage from similar CV measurements as those used in this work. Notice that the Au NC in Table 5.1 is in direct contact with PECVD SiO_2 [3], instead of the evaporated SiO_2 used in Fig. 5.1. The deviations from the theoretical predictions are clearly seen. For example, with a control oxide of PECVD SiO_2 where the composition stoichiometry is not as perfect as thermally grown oxide, Pt and Au behave as if mid-gap metals rather than high work-function metals. Ag, on the other hand, prompts hole storage like a low work-function metal. There is a trend of shifting all metal NC work functions upward with this PECVD SiO_2 . From the independent works on Ni and W NCs, inconsistent results are also reported depending on the

Table 5.1 Preferable polarity of charge storage from the multiple CV measurements of metal NC memories available in the literature

Nanocrystal	Work function (eV)	Control oxide	Preferable polarity of charge storage
Ag ^[3]	4.46	PECVD SiO ₂	Hole
Au ^[3]	4.94	PECVD SiO ₂	Symmetric electron/hole
Pt/Si ^[3]	4.95	PECVD SiO ₂	Symmetric electron/hole
Ni ^[4]	5.15	CVD HfO ₂	Symmetric electron/hole
Ni ^[5]	5.15	CVD HfO ₂	Electron
W ^[6]	4.50	Oxidized a-Si	Symmetric electron/hole
W ^[7]	4.50	ALD HfAlO	Electron

process and material of the control oxide. Therefore both our experimental results and previous findings suggest the “effective” metal NC work function is not only a bulk material property, but also strongly influenced by the interface with surrounding dielectric. Furthermore, this strong interfacial effect is not restricted to metal NCs. Very similar results are shown in Fig. 5.2 by replacing Au NCs with C₆₀ molecules.

The above observations are best explained by the Fermi-level pinning theory, first introduced to explain the insensitivity of metal work function to the Schottky barrier height [16], and later was proved applicable at semiconductor–dielectric and metal–dielectric interfaces [9]-[11]. For a dielectric in contact with a metal, there are dangling bond states at the interface dispersed across the band gap of the dielectric. Those states are known as metal-induced gap states (MIGS) [17]. A charge neutrality level (CNL) is defined where states with energy smaller than the CNL are occupied, and states with energy larger than the CNL are empty for a neutral surface [9]. It can be thought of as a local Fermi level of the interface states. For an inert metal with negligible chemical interaction at the interface, the density of interface states and the

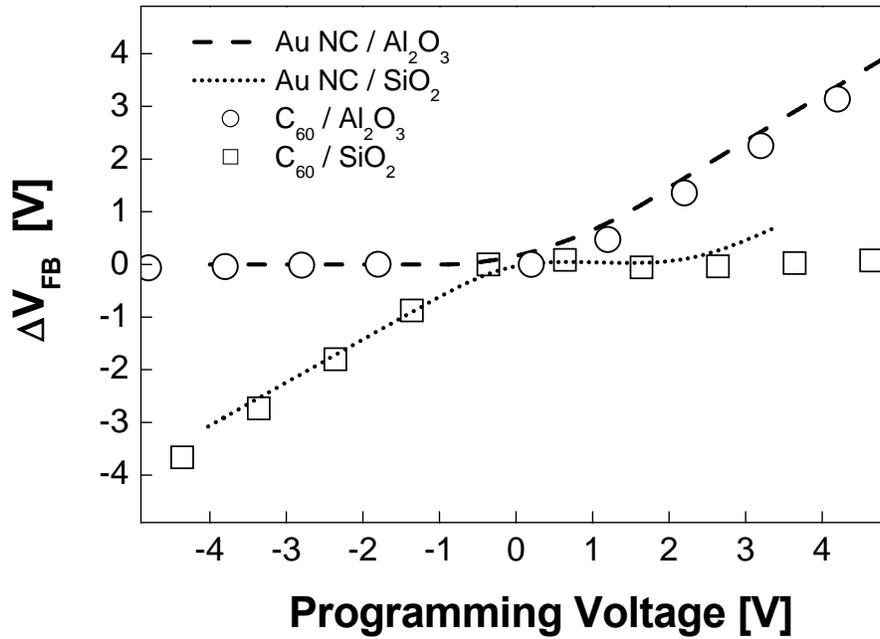


Figure 5.2 Flatband voltage shift ΔV_{FB} as a function of program voltage in Au NC and C₆₀ capacitors. The program voltage is defined as the difference between the applied stress voltage and initial flatband voltage without stress.

CNL can be calculated directly from the dielectric band structure [9]-[11]. High- κ dielectrics are generally known to have substantially higher density of interface states with metals than thermally grown SiO₂. Figure 5.3 illustrates the dipole formation at the NC–dielectric interface and the corresponding 1D energy-band diagram. Because the CNL is close to the Si conduction band edge in this example, interface dipoles are formed by transferring electrons from interface states into metal NCs at equilibrium. The effective metal NC work function is shifted toward the CNL accordingly. Please note the narrow triangular barrier shown in Fig. 5.3 is transparent to charge transport due to the proximity of paired dipoles. If the density of interface states is high enough, metal work functions can be pinned near the CNL regardless of the choice of metal NC materials. The CNL of ALD Al₂O₃ is known at -5.2 eV [11], which gives the effective Au work function even higher than -4.94 eV and thus suppresses hole

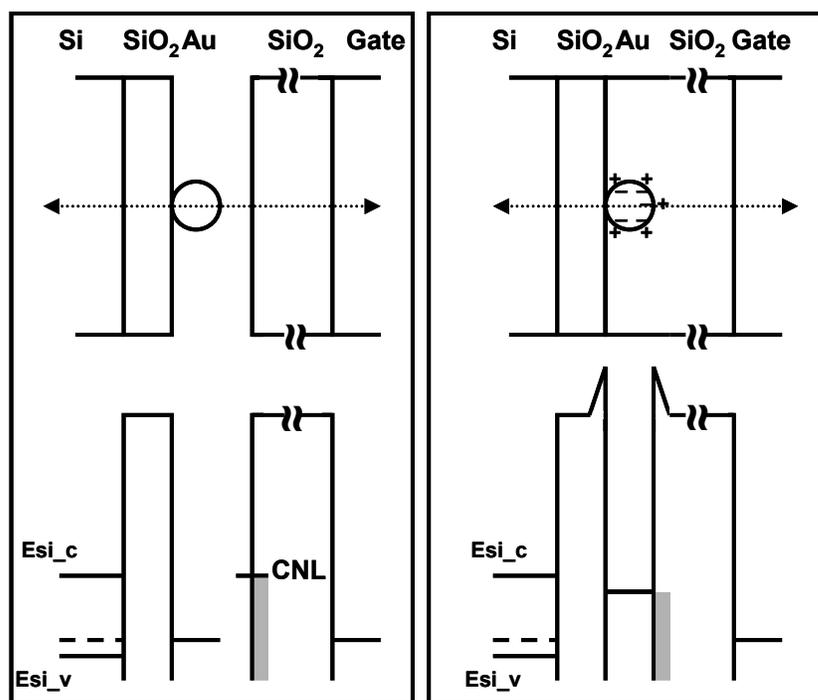


Figure 5.3 Dipole formation at the NC–dielectric interface and corresponding 1D energy-band diagrams of Au NC capacitors with high density of interface states. The left and right figures represent the thermal equilibriums before and after NC in contact with the control oxide, respectively. The charge neutrality level is close to the Si conduction band edge. The electron-filled interface states are indicated as the shaded regions.

storage in Au NCs. On the other hand, although the CNL of evaporated SiO₂ is not known, it is expected to be close to the Si conduction band edge in favor of hole storage. Because the CNL and the density of interface states are determined by the atomic configuration of dielectrics, they are sensitive to composition stoichiometries and deposition processes, which explain the inconsistency in Table 5.1. It is worthwhile to mention that the control oxide has more profound effect on metal NC work function than the tunnel oxide because of the larger contact surface and higher number of interface states with spherical metal NCs. The same model is found

applicable to the pinning effect at the C_{60} -dielectric interface. However, the impacts of the C_{60} molecular-orbital structure and the single electron charging energy should be evaluated at the molecular scale [18]. A more detailed discussion is given in Chapter 6.

5.5 *Conclusion*

We show that the work function engineering is not merely a choice of metal NC materials, but also the surrounding dielectric, owing to the Fermi-level pinning effect appeared at the NC-dielectric interface. The formation of interface dipoles, which depends on the density and the energy distribution of interface states, shifts the NC work function apart from its bulk value. This has important implications on the design optimization of NC memories. For instance, to maximize the electron potential well for better charge storage capacity and retention time in a typical n-channel memory, a dielectric with high interface states and low CNL should be avoided when integrating with a high work-function metal NC.

REFERENCES

- [1] S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, and D. Buchanan, "Volatile and non-volatile memories in silicon with nano-crystal storage," in *IEDM Tech. Dig.*, 1995, pp. 521-524.
- [2] C. Lee, U. Ganguly, V. Narayanan, T.-H. Hou, and E. C. Kan, "Asymmetric electric field enhancement in nanocrystal memories," *IEEE Electron Device Lett.*, vol. 26, no. 12, pp. 879-881, Dec. 2005.
- [3] C. Lee, A. Gorur-Seetharam, and E. C. Kan, "Operational and reliability comparison of discrete-storage nonvolatile memories: Advantages of single- and double-layer metal nanocrystals," in *IEDM Tech. Dig.*, 2003, pp. 557-560.
- [4] Z. Tan, S. Samanta, W. Yoo, and S. Lee, "Self-assembly of Ni nanocrystals on HfO₂ and N-assisted Ni confinement for nonvolatile memory application", *Appl. Phys. Lett.*, vol. 86, 013107, 2005.
- [5] J. Lee, Y. Harada, J. Pyun, and D.-L. Kwong, "Nickel nanocrystal formation on HfO₂ dielectric for nonvolatile memory device application," *Appl. Phys. Lett.*, vol. 86, 103505, 2005.
- [6] T. C. Chang, P. T. Liu, S. T. Yan, and S. M. Sze, "Electron charging and discharging effects of tungsten nanocrystals embedded in silicon dioxide for low-voltage nonvolatile memory technology," *Electrochem. Solid-State Lett.*, vol. 8, no. 3, G71-G73, 2005.
- [7] S. Samanta, Z. Tan, W. Yoo, G. Samudra, S. Lee, L. Bera, and N. Balasubramanian, "Self-assembled tungsten nanocrystals in high- κ dielectric for nonvolatile memory application," *J. Vac. Sci. Technol. B*, vol. 23, no. 6, pp. 2278-2283, 2005.

- [8] J. J. Lee, X. Wang, W. Bai, N. Lu, J. Liu, and D. L. Kwong, "Theoretical and experimental investigation of Si NC memory device with HfO₂ high-κ tunneling dielectric," in *Symp. VLSI Tech. Dig.*, 2003, pp. 33-34.
- [9] J. Robertson, "Band offsets of wide-band-gap oxides and implications for future electronic devices," *J. Vac. Sci. Technol. B*, vol. 18, no. 3, pp. 1785–1791, 2000.
- [10] C. Hobbs, L. Fonseca, V. Dhandapani, S. Samavedam, B. Taylor, J. Grant, L. Dip, D. Triyoso, R. Hegde, D. Gilmer, R. Garcia, D. Roan, L. Lovejoy, R. Rai, L. Hebert, H. Tseng, B. White, and P. Tobin, "Fermi-level pinning at the polySi / metal oxide interface," in *Symp. VLSI Tech. Dig.*, 2003, pp. 9-10.
- [11] S. Samavedam, L. La, P. Tobin, B. White, C. Hobbs, L. Fonseca, A. Demkov, J. Schaeffer, E. Luckowski, A. Martinez, M. Raymond, D. Triyoso, D. Roan, V. Dhandapani, R. Garcia, S. Anderson, K. Moore, H. Tseng, C. Capasso, O. Adetutu, D. Gilmer, W. Taylor, R. Hegde, and J. Grant, "Fermi level pinning with sub-monolayer MeOx and metal gates," in *IEDM Tech. Dig.*, 2004, pp. 307-310.
- [12] C. Li, W. Fan, B. Lei, D. Zhang, S. Han, T. Tang, X. Liu, Z. Liu, S. Asano, M. Meyyappan, J. Han, and C. Zhou, "Multilevel memory based on molecular devices," *Appl. Phys. Lett.*, vol. 84, pp. 1949-1951, 2004.
- [13] C. Lee, J. Meter, V. Narayanan, and E. C. Kan, "Self-assembly of metal nanocrystal on ultra-thin oxide for nonvolatile memory applications," *J. Electronic Materials*, vol. 34, no. 1, pp. 1-11, Jan. 2005.
- [14] U. Ganguly, C. Lee, and E. C. Kan, "Experimental observation of non-volatile charge injection and molecular redox in fullerenes C₆₀ and C₇₀ in an EEPROM type device," in *Proc. MRS. Boston, MA, Nov. 29-Dec. 3 2004*, pp.D7.5.

- [15] H. Bachhofer, H. Reisinger, E. Bertagnolli, and H. Philipsborn, "Transient conduction in multielectric silicon-oxide-nitride-oxide semiconductor structures," *J. Appl. Phys.*, vol. 89, pp.2791-2800, 2001.
- [16] J. Bardeen, "Surface states and rectification at a metal semi-conductor contact," *Phys. Rev.*, vol. 71, pp.717-727, 1947.
- [17] J. Tersoff, "Theory of semiconductor heterojunctions: the role of quantum dipoles," *Phys. Rev. B*, vol. 30, pp. 4874-4877, 1984.
- [18] T.-H. Hou, U. Ganguly, and E. C. Kan, "Programmable molecular orbital states of C₆₀ from integrated circuits," *Appl. Phys. Lett.*, vol. 89, 253113, 2006.

CHAPTER 6

PROGRAMMABLE MOLECULAR ORBITAL STATES OF C₆₀ FROM INTEGRATED CIRCUITS

6.1 Abstract

We have reported experimental demonstration of four C₆₀ programmable molecular orbital states, C₆₀⁰, C₆₀¹⁻, C₆₀²⁻, and C₆₀³⁻, through a charge-based nonvolatile memory cell. Owing to the monodisperse nature and molecular size of C₆₀, very sharp Coulomb staircase is observed at room temperature. A physical model based on the molecular orbit structure, the charging energy solved by an electrostatic method, and the Fermi-level pinning theory yields good quantitative agreement with experiments. These not only lead to better understanding of the C₆₀ molecular orbital structure and corresponding chemical redox levels, but also potentially pave the way for realizing reliable multi-level molecular memories.

6.2 Introduction

There has been a great deal of interest in carbon nanotubes and fullerenes because of their unique electronic, mechanical, and chemical properties as well as the low dimensional nanoscale features from bottom-up assembly. They received particular attention in nanoelectronics, where device miniaturization is typically driven and limited by the top-down photolithography. Taking advantage of the maturity built on the top-down technology, a hybrid of bottom-up and top-down approach could be an effective way toward the ultimate scalability and functionality in nanoelectronics. Several interesting examples have been reported such as carbon nanotube decoder circuits [1] and memories [2].

Nanocrystal memories are considered promising candidates for nonvolatile

storage with low power consumption [3]. However, the finite and undesired size-dispersion in the nanocrystal formation could degrade the device parametric yield. Meanwhile, to increase the storage density, the multi-level cell [4], where different memory states are represented by different amount of charge stored, is proposed for multiple bits per cell. However, due to lack of self-convergence in programming, it usually requires complicate peripheral circuits and operations to precisely control the memory states. In theory, Coulomb islands such as nanocrystals are ideal to realize the multi-level cell concept through the self-convergent Coulomb blockade effect. Nevertheless, their size of several nanometers in diameter together with the size-dispersion prohibits any appreciable step charging at room temperature. In this study, we report using C_{60} molecules to replace nanocrystals as the floating gate in a nonvolatile memory cell. The monodisperse nature of C_{60} with its molecular size results in substantial and precise step charging into molecular orbitals (MO), and hence could potentially achieve reliable multi-level charge storage in a single memory cell. Different from previous works that applied various molecules as charge storage nodes [5], [6], the discrete C_{60} redox states in the molecular Coulomb islands are first reported at room temperature. Besides the potential application as memory devices, this study also provides valuable understanding of the C_{60} molecular orbital structure in the electrochemical environment, which is essential for future realization of interface between molecules and integrated circuits.

6.3 Device Fabrication

The metal-oxide-semiconductor (MOS) capacitor structure with conventional isolation, similar to that used in the Flash memory devices [7], [8], was fabricated. After 1.8-nm dry thermal oxide was grown on the silicon substrate, the fullerenes C_{60} were thermally evaporated to a thickness of 0.4 to 0.6 nm as measured by the quartz

crystal monitor. The C_{60} source used in this study was obtained commercially (MER Corporation 99.9 %). Then, control oxide is implemented by atomic layer deposition (ALD) of 30 nm Al_2O_3 or by evaporation of 4-nm SiO_2 together with plasma-enhanced chemical vapor deposition (PECVD) of 26 nm SiO_2 . ALD Al_2O_3 deposition was done by trimethylaluminum (TMA) and H_2O at $300^\circ C$. The evaporated oxide served as a barrier layer to prevent possible plasma damage to C_{60} during the PECVD process. Finally a top Cr gate was deposited and patterned, followed by $400^\circ C$ forming gas annealing for 30 min.

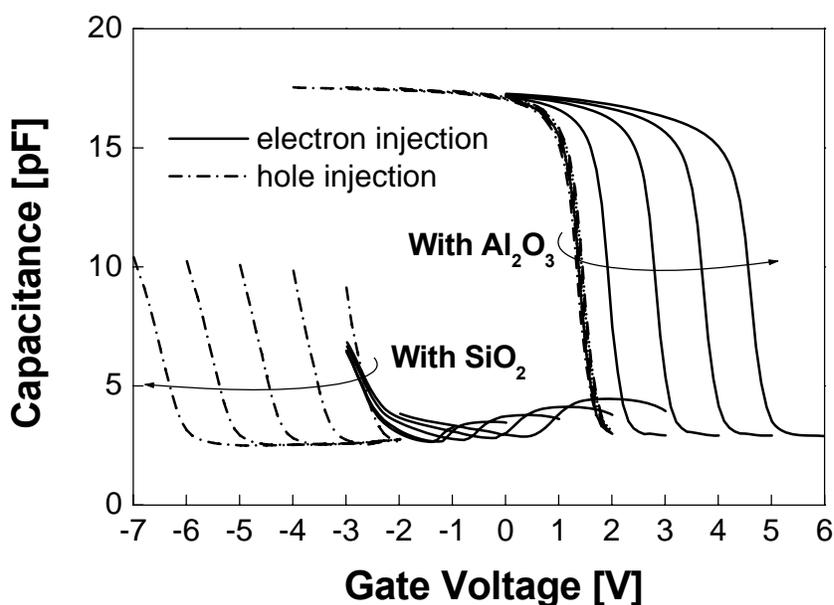


Figure 6.1 High frequency CV measurements of C_{60} capacitors with Al_2O_3 control oxide and SiO_2 control oxide. The capacitor area is $1 \times 10^{-4} \text{ cm}^2$. Substrate electron and hole injections are carried out by applying stress voltages on the control gate for 3 s before each CV sweep.

6.4 Redox States of C_{60} Molecules

From high frequency capacitance-voltage (CV) measurements at low program voltages (Fig. 6.1), only electron injection is allowed in C_{60} capacitors with Al_2O_3

control oxide, whereas only hole injection is allowed with SiO₂. Negligible CV shift in control samples without C₆₀ confirms C₆₀ is responsible for the observed charge storage. This asymmetric injection phenomenon apparently depends on the material of control oxide. It is similar to the Fermi-level pinning at the interface of metal gate and high-κ dielectric [9], [10], and has been reported at another organic-dielectric interface as well [11]. More extensive discussion of the Fermi-level pinning theory has been given in Chapter 5. Hole injection with Al₂O₃ control oxide is enabled only after overcoming a negative blockade voltage as shown in Fig. 6.2(a). There is no obvious difference of the blockade voltages measured at 300K and 10K, implying this effect likely originates from the intrinsic properties of C₆₀. It is worthwhile to note that the temperature-dependent Frenkel-Poole conduction [12] in Al₂O₃, responsible for flatband voltage shift ΔV_{FB} saturation at high positive gate bias as well as the turning point at high negative gate bias at 300K, is significantly suppressed at 10K. Meanwhile, electron injection with SiO₂ control oxide is allowed only after overcoming a positive blockage voltage as shown in Fig. 6.2(b). In contrast to the continuous ΔV_{FB} in the control sample without C₆₀, which comes from electron injection into the traps of control oxide through the Fowler-Nordheim and Frenkel-Poole conduction, the Coulomb staircase due to the C₆₀ MO can be clearly observed.

In this experimental setup, the voltage sweep at the control gate results in the potential change of C₆₀. It is analog to scanning C₆₀ energy levels relative to the Si substrate. The C₆₀ energy change ΔE_{C60} can be calculated by using the series capacitor model of conventional floating gate devices. For example, the blockade plateau shown in Fig. 6.2(a) corresponds to 2.2 eV ΔE_{C60} seen from the Si substrate. To consistently and quantitatively explain the above results, a model based on the Fermi-level pinning theory and the C₆₀ MO structure is established. Figure 6.3 illustrates the energy band/level diagrams. The highest occupied molecular orbital (HOMO) and lowest

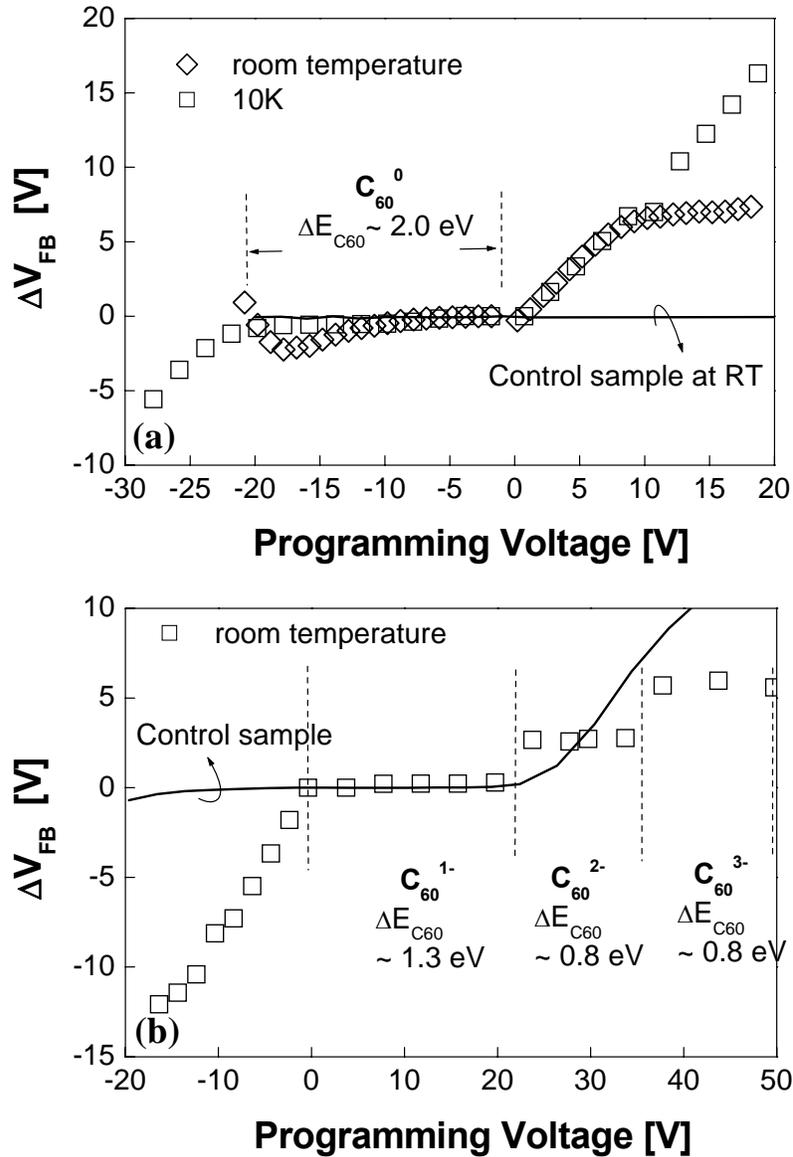


Figure 6.2 Flatband voltage shift ΔV_{FB} as a function of program voltage measured at (a) both room temperature and 10K for C_{60} capacitors with Al_2O_3 control oxide (b) room temperature for C_{60} capacitors with SiO_2 control oxide. Measurements from control samples without C_{60} are also shown for comparison. The program voltage is defined as the difference between applied stress voltage and initial flatband voltage without stress. It can be correlated to the C_{60} energy given the gate stack configuration. For example, the plateau in (a) corresponds to about 2.2 eV change experienced by C_{60} . The C_{60} anion state in each plateau is also indicated.

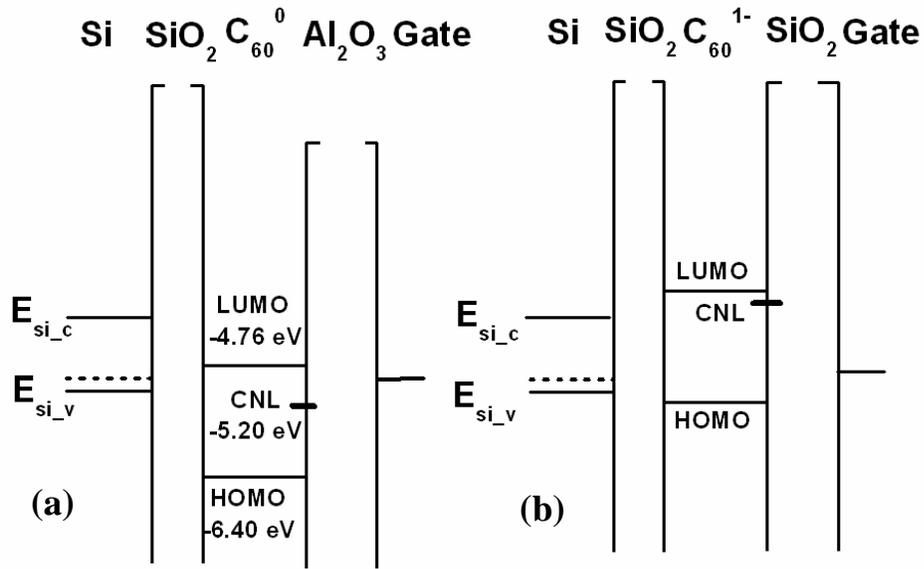


Figure 6.3 Energy band/level diagram representation of C₆₀ capacitors with (a) Al₂O₃ control oxide and (b) SiO₂ control oxide. The thermal equilibrium states of C₆₀ are neutral and monoanion for Al₂O₃ and SiO₂, respectively.

unoccupied molecular orbital (LUMO) energy for neutral C₆₀ (C₆₀⁰) are -6.40 eV and -4.76 eV, respectively [13], and the charge neutrality level (CNL) of ALD Al₂O₃ is -5.2 eV [10]. The CNL can be thought of as a local Fermi level of the interface states, where states with energy smaller than the CNL are occupied, and states with energy larger than the CNL are empty for a neutral surface [9], [10]. Those states are called metal-induced gap states (MIGS) [14], and can be thought of as the dangling bond states at the interface dispersed across the band gap of the dielectric. Therefore, according to the C₆₀ HOMO-LUMO energy and the CNL of Al₂O₃, C₆₀ remains neutral at thermal equilibrium before program. Electron injection into the C₆₀⁰ LUMO is enabled by electrons with energy higher than -4.76 eV plus the C₆₀ single electron charging energy E_{CH} . The injected electron is then quickly relaxed into surrounding interface states in favor of the lower CNL energy. Due to the wide spatial and energy dispersion of interface states, the distinct Coulomb staircase is not expected and was

not experimentally observed even at 10K. On the other hand, at low negative bias, hole injection into the C_{60}^0 HOMO is forbidden due to the C_{60} HOMO-LUMO gap, and is only allowed after overcoming the negative blockade voltage, which moves the HOMO energy E_{C60_HOMO} higher than the Si valence band edge E_{Si_V} by one E_{CH} . Hence, the C_{60} energy change required to start hole injection is $\Delta E_{C60} = E_{Si_V} - E_{C60_HOMO} + E_{CH} = 1.23 \text{ eV} + E_{CH}$. E_{CH} can be approximated using the conventional three-dimensional (3D) electrostatic method described in Chapter 2 by assuming C_{60} as a metal sphere with a radius of 4 Å (the outer radius of C_{60}) embedded in two parallel plates. For a C_{60} density of $2 \times 10^{12} \text{ cm}^{-2}$, 1.8 nm SiO_2 as the tunnel oxide, and 30 nm Al_2O_3 as the control oxide, E_{CH} is 0.52 eV. It gives the calculated ΔE_{C60} around 1.75 eV, in reasonable agreement with that found in Fig. 6.2(a). For the case with SiO_2 control oxide, although the CNL of evaporated SiO_2 is not known, it is expected to be close to the silicon conduction band edge to favor hole injection at low bias. The interface dipole formation at the C_{60} and SiO_2 interface and the C_{60} charging energy lead to a different thermal equilibrium state before program, likely being monoanion (C_{60}^{1-}), established by electron transferring from the interface state into C_{60}^0 . As the energy band/level diagram shown in Fig. 6.3(b), while hole injection from the Si valence band into the partially filled LUMO can happen at low bias (the injected hole is then quickly relaxed into the surrounding interface states to restore the stable C_{60}^{1-} state), electron injection into the C_{60}^{1-} LUMO to become dianion (C_{60}^{2-}) is not favorable until the LUMO is one E_{CH} lower than the Si conduction band edge, which is responsible for the 1.3 eV blockade energy in Fig. 6.2(b). Likewise, other high energetic states such as trianion (C_{60}^{3-}) are also possible by overcoming necessary charging energy. As several distinct blockade regions under positive gate bias shown in Fig. 6.2(b), C_{60}^{2-} and C_{60}^{3-} are stable without losing electrons to interface states because of the proximity of the LUMO and CNL levels. The observation of C_{60}^{4-} or

other higher anions was prohibited by severe CV distortion, possibly due to oxide damage. The 0.8 eV blockade energy for both C_{60}^{2-} and C_{60}^{3-} states corresponds to exactly one E_{CH} , in excellent agreement with 0.86 eV approximated from the previous 3D electrostatic analysis but with 30 nm SiO_2 as the control oxide. The C_{60} density calculated from the amount of the flat band shift is $2 \times 10^{12} \text{ cm}^{-2}$, also agreed well with the density used for the E_{CH} simulation.

6.5 Conclusion

In summary, we have demonstrated four programmable molecular orbital states of C_{60} , including C_{60}^0 , C_{60}^{1-} , C_{60}^{2-} , and C_{60}^{3-} , from a MOS-based nonvolatile memory cell. The C_{60} programmable states strongly depending on the material of control oxide are best explained by the Fermi-level pinning theory due to the interface dipole formation. The room-temperature Coulomb staircase, originating from the blockade effects through the C_{60} HOMO-LUMO gap and charging energy, is quantitatively interpreted, and shown good agreement with theoretical predictions.

REFERENCES

- [1] Y. Tseng, P. Xuan, A. Javey, R. Malloy, Q. Wang, J. Boker, and H. Dai, "Monolithic Integration of Carbon Nanotube Devices with Silicon MOS Technology," *Nano Lett.*, vol. 4, pp. 123-127, 2004.
- [2] T. Rueckes, K. Kim, E. Joselevich, G. Tseng, C. Cheung, and C. M. Lieber, "Carbon nanotube-based nonvolatile random access memory for molecular computing," *Science*, vol. 289, pp. 94-97, 2000.
- [3] S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, and D. Buchanan, "Volatile and non-volatile memories in silicon with nanocrystal storage," in *IEDM Tech. Dig.*, 1995, pp. 521-524.
- [4] B. Eitan, R. Kazerounian, A. Roy, G. Crisenza, P. Cappelletti, and A. Modelli, "Multilevel flash cells and their trade-offs," in *IEDM Tech. Dig.*, 1996, pp. 169-172, 1996.
- [5] X. Duan, Y. Huang, and C. M. Lieber, "Nonvolatile Memory and Programmable Logic from Molecule-Gated Nanowires," *Nano Lett.*, vol. 2, pp. 487-490, 2002.
- [6] C. Li, W. Fan, B. Lei, D. Zhang, S. Han, T. Tang, X. Liu, Z. Liu, S. Asano, M. Meyyappan, J. Han, and C. Zhou, "Multilevel memory based on molecular devices," *Appl. Phys. Lett.*, vol. 84, pp. 1949-1951, 2004.
- [7] C. Lee, J. Meter, V. Narayanan, and E. C. Kan, "Self-assembly of metal nanocrystal on ultra-thin oxide for nonvolatile memory applications," *J. Electronic Materials*, vol. 34, pp. 1-11, 2005.
- [8] U. Ganguly, C. Lee, and E. C. Kan, "Experimental observation of non-volatile charge injection and molecular redox in fullerenes C₆₀ and C₇₀ in an EEPROM type device," *Mater. Res. Soc. Symp. Proc.*, vol. 830, D7.5., 2005.
- [9] J. Robertson, "Band offsets of wide-band-gap oxides and implications for future electronic devices," *J. Vac. Sci. Technol. B*, vol. 18, pp. 1785-1791, 2000.

- [10] C. Hobbs, L. Fonseca, V. Dhandapani, S. Samavedam, B. Taylor, J. Grant, L. Dip, D. Triyoso, R. Hegde, D. Gilmer, R. Garcia, D. Roan, L. Lovejoy, R. Rai, L. Hebert, H. Tseng, B. White, and P. Tobin, "Fermi-level pinning at the polySi / metal oxide interface," in *Symp. VLSI Tech. Dig.*, 2003, pp. 9-10.
- [11] L. Chen, R. Ludeke, X. Cui, A. Schrott, C. Kagan, and L. Brus, "Electrostatic Field and Partial Fermi Level Pinning at the Pentacene-SiO₂ Interface," *J. Phys. Chem. B*, vol. 109, pp. 1834-1938, 2005.
- [12] M. Specht, M. Stadele, S. Jakschik, and U. Schroder, "Transport mechanisms in atomic-layer-deposited Al₂O₃ dielectrics," *Appl. Phys. Lett.*, vol. 84, pp. 3076-3079, 2004.
- [13] W. Green, S. Gorun, G. Fitzgerald, P. Fowler, A. Ceulemans, and B. Titeca, "Electronic structures and geometries of C₆₀ anions via density functional calculations," *J. Phys. Chem.*, vol. 100, 14892-14898, 1996.
- [14] J. Tersoff, "Theory of semiconductor heterojunctions: the role of quantum dipoles," *Phys. Rev. B*, vol. 30, pp. 4874-4877, 1984.

CHAPTER 7
NONVOLATILE MEMORY WITH
MOLECULE-ENGINEERED TUNNEL BARRIERS

7.1 Abstract

We report a novel field-sensitive tunnel barrier by embedding C_{60} in SiO_2 for nonvolatile memory applications. C_{60} is a better choice than ultra-small nanocrystals due to its monodispersion. Moreover, C_{60} provides accessible energy levels to prompt resonant tunneling through SiO_2 at high fields. However, this process is quenched at low fields due to HOMO-LUMO gap and large charging energy of C_{60} . Furthermore, we demonstrate an improvement of more than an order of magnitude in retention to program/erase time ratio for a metal nanocrystal memory. This shows promise of engineering tunnel dielectrics by integrating molecules in the future hybrid molecular-silicon electronics.

7.2 Introduction

In the present charge-based nonvolatile Flash memory technology, the ratio between retention time t_R and program/erase (P/E) time t_{PE} is about 10^{12} - 10^{14} . In order to realize this tremendous ratio, field-asymmetric tunneling processes in the tunnel barrier have to be deliberately engineered between retention and P/E. The asymmetry in conventional Flash cells is most often provided by the external P/E voltage. For example, in the NAND Flash, the asymmetry between the Fowler-Nordheim tunneling under P/E and the direct tunneling during retention is exploited. However, this approach limits the scalability of the P/E voltage, which is quickly becoming the major scaling roadblock, considering power dissipation, cycling endurance, and peripheral circuitry design [1], [2].

In order to address these issues, the metal nanocrystal (NC) memory has been proposed [3]. Metal NCs enhance the tunneling asymmetry due to the additional band offset between the material-dependent floating-gate work function and Si band edges of the channel. Along with the 3D electrostatic advantages [4], extremely low P/E voltages can be realized [5], [6]. Furthermore, tailoring the band structure of the tunnel barrier is another effective way to achieve significant tunneling asymmetry. Various theoretical and experimental approaches based on crested tunnel barriers [7], [8], asymmetric layered barriers [5], [9], [10], bandgap-engineered Oxide-Nitride-Oxide (ONO) [11], and double tunnel junction [12], [13] have been undertaken. Among them, the double tunnel junction proposed by R. Ohba *et. al.* [12], [13] is of particular interest with its superior t_R / t_{PE} ratio at low P/E voltage and the demonstration of excellent memory scalability. This structure consists of a layer of about 1-nm Si NCs sandwiched between two SiO₂ layers. These infinitesimal Si NCs are crystallized by annealing a SiO₂ / a-Si / SiO₂ structure. The size of NCs plays an important role in the memory performance [12]. But its precise control determined by the thin Si layer thickness and total thermal history remains complicated. Therefore, device variation within large memory array and reproducibility from run to run are potential issues.

Molecules with versatile and tunable properties may find many applications in integration with traditional Si technology. We have discussed the redox states of C₆₀ molecules for multi-level charge storage [14] in Chapter 6. In this chapter, we present a simpler implementation of the double tunnel junction by utilizing the monodisperse nature of these nanoscale entities. C₆₀ molecules instead of Si NCs are embedded inside the oxide barrier to overcome the aforementioned limitation on the NC size control. To our best knowledge, it is the first demonstration of molecule-engineered tunnel barrier in Si devices. We will further show improved t_R / t_{PE} ratio in a metal NC memory integrated with this barrier.

7.3 *Device Fabrication*

The metal-oxide-semiconductor (MOS) capacitors with conventional local oxidation of Si (LOCOS) isolation on p-type substrates were fabricated. After 2.5-nm dry thermal oxidation, C₆₀ molecules were thermally evaporated to a thickness of 0.4 to 0.6 nm as measured by the quartz crystal monitor, followed by SiO₂ evaporation of 3 nm to complete the tunnel barrier formation. The C₆₀ molecules used in this study were obtained commercially (MER Corporation 99.9 %). The area density of C₆₀ molecules estimated from the electrical measurement [14] presented in Chapter 6 is around $2 \times 10^{12}/\text{cm}^2$. As for the metal NC memory cells, after the tunnel oxide formation, spherical Au NCs were self-assembled on the oxide by the electron-beam evaporation of 1.2-nm Au without annealing. SiO₂ control oxide was deposited by plasma-enhanced chemical vapor deposition (PECVD) to a thickness of 30 nm. Finally, a top Cr gate was patterned, followed by 400°C forming gas annealing for 30 min. The schematics of various heterogenous gate stacks (S1-S5) investigated in this study are illustrated in Fig. 7.1 (a).

7.4 *Resonant Tunnel Barrier*

We first examine the gate current through the proposed C₆₀-embedded tunnel barrier in Fig. 7.2(a). S1 consists of tunnel oxide (2.5-nm thermal SiO₂ + C₆₀ + 3-nm evaporated SiO₂) but not top layers of Au NC and PECVD SiO₂. A control sample S2 without the C₆₀ layer is also shown for comparison. Identical equivalent oxide thickness (EOT) of S1 and S2 is confirmed by the capacitance-voltage (CV) in Fig. 7.2(b). The strong field-dependence and temperature-dependence of the gate current in Fig. 7.3(a) suggest that the current transport in S2 is governed by the Frenkel-Poole emission through the shallow traps inside the evaporated SiO₂. S1 shows exponential gate current increase by four orders of magnitude due to resonant tunneling through

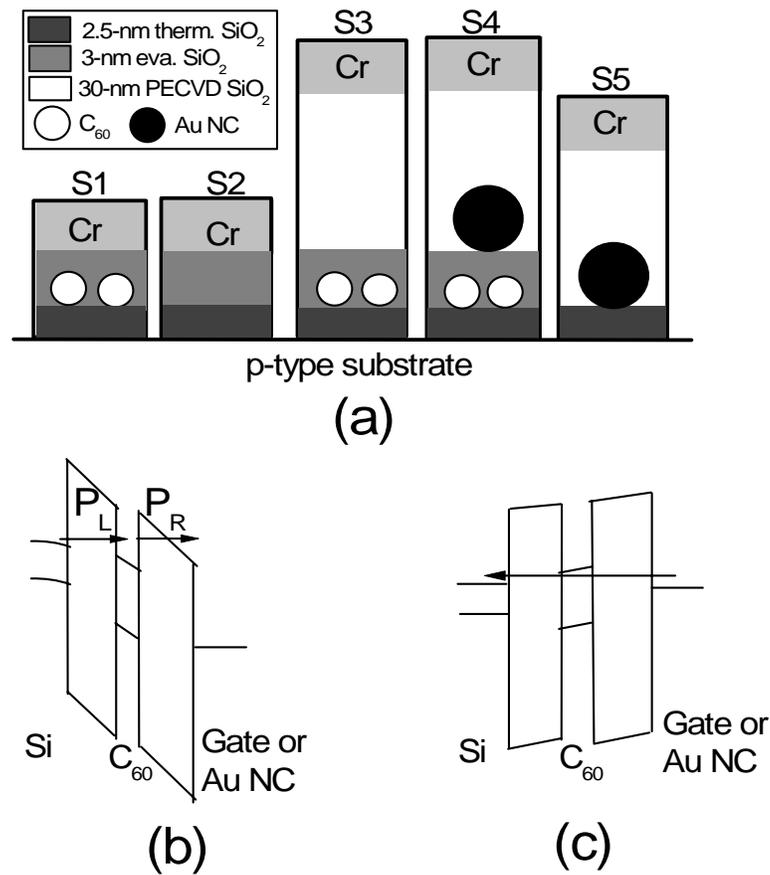


Figure 7.1 (a) Schematics of heterogeneous gate stacks (S1-S5) examined in this work. Energy band/level diagram representation of tunnel barriers with (b) resonant tunneling through C₆₀ under high electric field, and (c) direct tunneling through C₆₀ under low electric field.

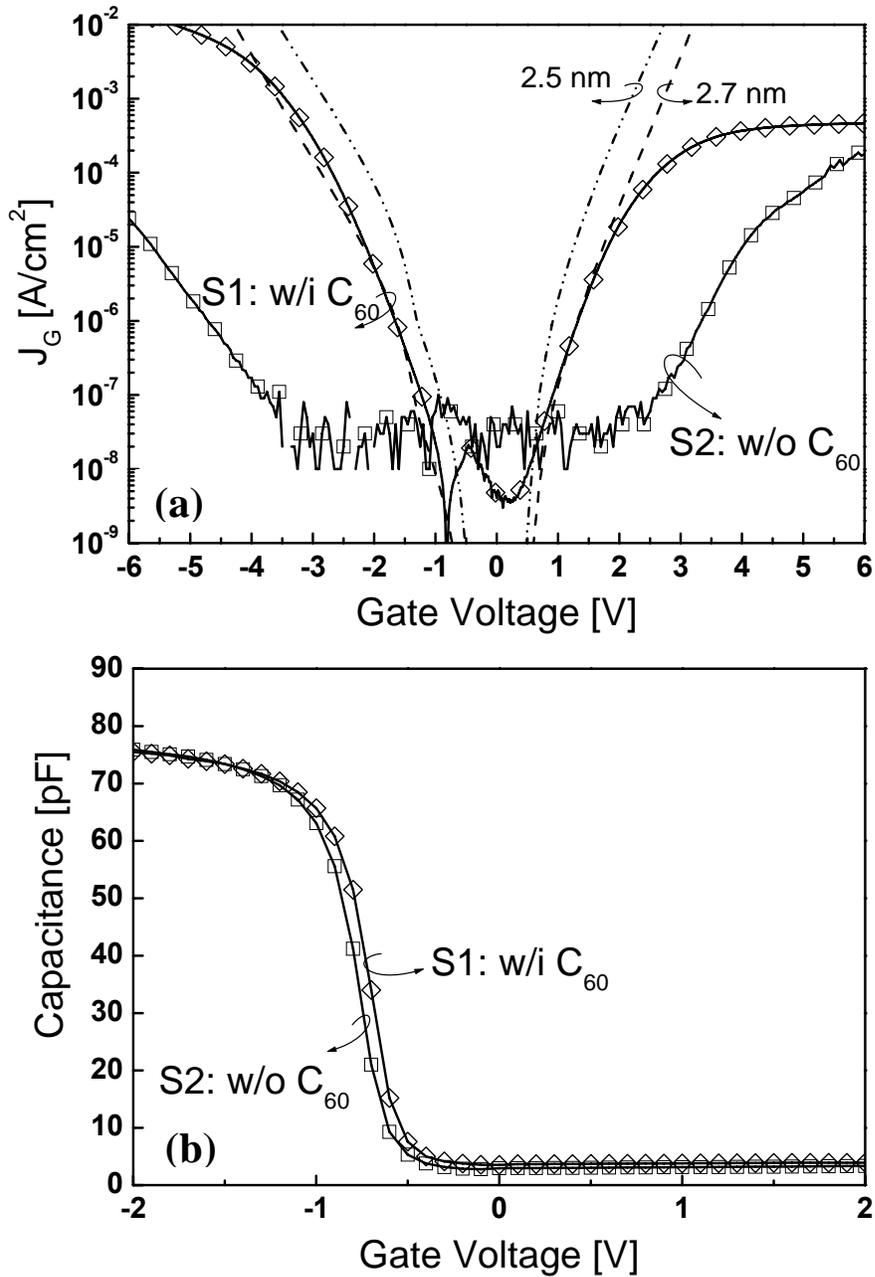


Figure 7.2 (a) Tunneling current and (b) CV measurement of S1 with a 2.5-nm SiO_2 + C_{60} + 3-nm evaporated SiO_2 barrier and S2 with a 2.5-nm SiO_2 + 3-nm evaporated SiO_2 barrier. The dash lines in (a) are calculated from the WKB approximation for 2.5-nm and 2.7-nm SiO_2 .

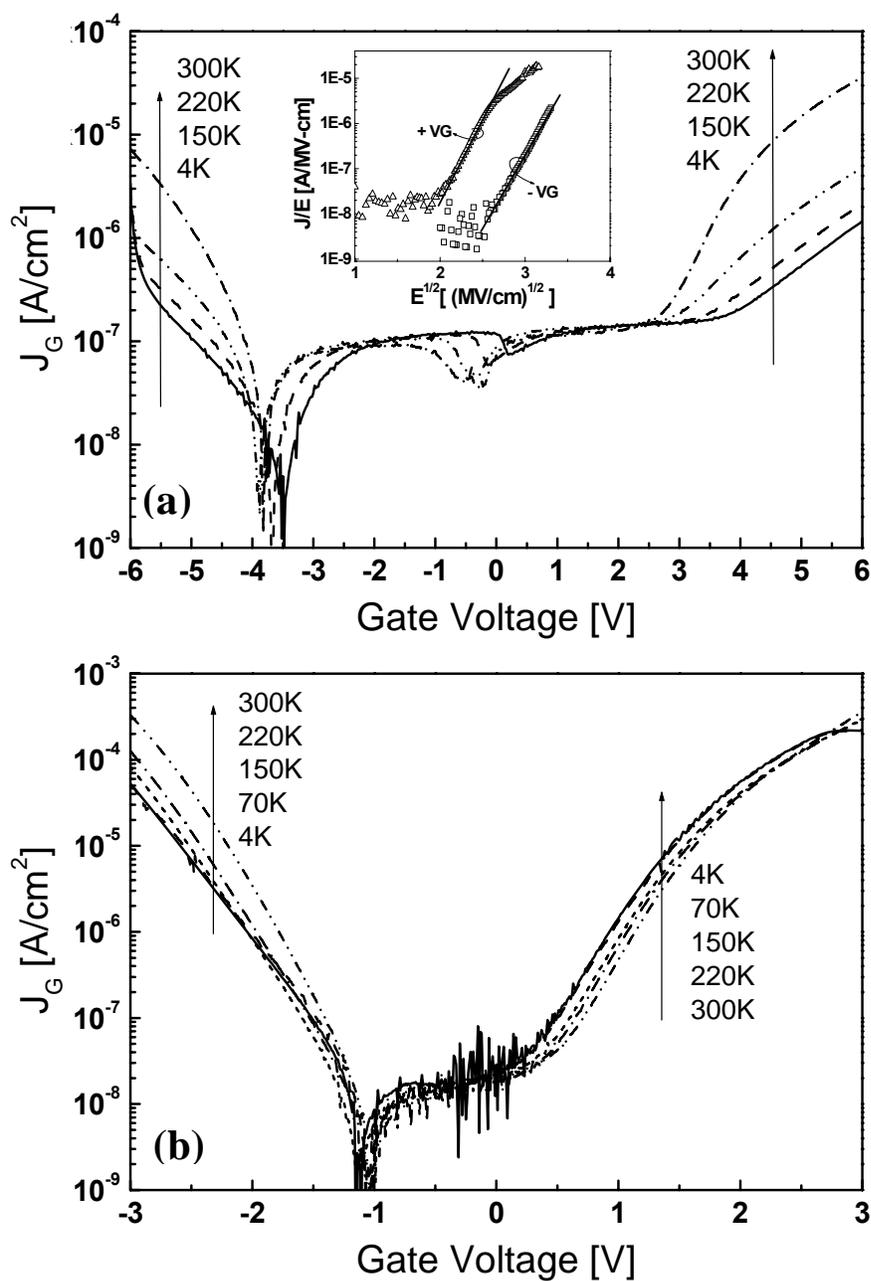


Figure 7.3 Temperature-dependent tunneling current measurement on (a) S2 with a 2.5-nm SiO₂ + 3-nm evaporated SiO₂ barrier and (b) S1 with a 2.5-nm SiO₂ + C₆₀ + 3-nm evaporated SiO₂ barrier. The Frenkel-Poole fitting on the room-temperature J - V of S2 is also shown in the inset of (a).

the molecular levels of C_{60} . C_{60} molecules are closer to the channel and have a larger and more controllable density than the evaporated SiO_2 traps. The gate current saturation above $\pm 3V$ is limited by high substrate resistance and insufficient minority carrier generation under inversion. Because these field conditions are very far away from those in normal memory operations, the present results are satisfactory for the discussion here. Theoretical tunneling current calculation by the Wentzel-Kramer-Brillouin (WKB) approximation [15] is provided to compare with the experimental data in Fig. 7.2(a). The gate current from S1 agrees very well with that from an ideal 2.7-nm SiO_2 barrier. It is only 10 times smaller than the calculated current for a single layer of 2.5-nm SiO_2 , despite much thicker physical thickness provided by the top C_{60} and evaporated SiO_2 layers. The temperature-dependence of gate current measurement in Fig. 7.3(b) is relatively weak, confirming the elastic tunneling process is the dominant transport mechanism.

The energy band/level diagram of the C_{60} -embedded barrier under high-bias conditions, such as the program operation, is illustrated in Fig. 7.1 (b). The HOMO-LUMO gap (highest occupied molecular orbital, lowest unoccupied molecular orbital) of C_{60} is about 1.64 eV [16] with HOMO and LUMO levels being five-fold and three-fold degenerate, respectively. Furthermore, the specific energy level alignment with the bands of the surrounding dielectrics is determined by interface dipole formation and redox states of C_{60} at thermal equilibrium [14] as discussed in Chapter 6. Under sufficient external bias, resonant tunneling through C_{60} energy levels is enabled due to the energy of injected electrons from the Si channel exceeding the C_{60} energy levels and Coulomb charging energy. Although a detailed model using the Coulomb blockade theory of single electron tunneling [17] is more complete, a simple two-step tunneling process [18] is sufficiently intuitive to describe the observed phenomena to the first order. The two-step tunneling current density J in the weak coupling regime

with accessible energy levels provided by the intermediate C_{60} molecules can be expressed as:

$$J = \sum_{E=E_a}^{\infty} q C(E) N_t \sigma_t \frac{P_L(E) P_R(E)}{P_L(E) + P_R(E)} \quad (7.1)$$

where N_t is the density of C_{60} , σ_t is the effective capture cross section of C_{60} , P_L and P_R are the tunneling probabilities through the left and right oxide barriers, C is the electron source function of the channel. The summation takes into account all electrons with energy higher than the first assessable energy-level of C_{60} molecules E_a . Here we assume that the occupancy factors are 1 and 0 for the conduction-band electron states in the channel and gate, respectively. In Fig. 7.2, the similarity of J - V shapes between the 2.5-nm SiO_2 and the C_{60} -embedded barrier implies J is mainly controlled by the left barrier, *i.e.* $P_L \ll P_R$. This is not surprising because bulk traps in the evaporated SiO_2 could greatly enhance P_R through the trap-assisted tunneling process [19]. Very high C_{60} density with a reasonable $\sigma_t = 5 \times 10^{-14} \text{ cm}^2$ can account for the 10 times current reduction in comparison with the single layer of 2.5-nm SiO_2 . On the contrary, under low-bias conditions, the resonant tunneling is forbidden due to both the C_{60} HOMO-LUMO gap and the Coulomb charging energy as shown in Fig. 7.1 (c). The direct tunneling current can be extremely low for a thick barrier and will only be evaluated through the retention measurement in the memory cell discussed below. In reality, the trap-assisted tunneling through the interface states between C_{60} and SiO_2 and the bulk traps in evaporated SiO_2 may lead to higher current. It is worthwhile mentioning that design optimization for a maximum tunneling asymmetry may be possible by engineering the HOMO-LUMO gap and charging energy of different molecules as well as the dielectric thickness.

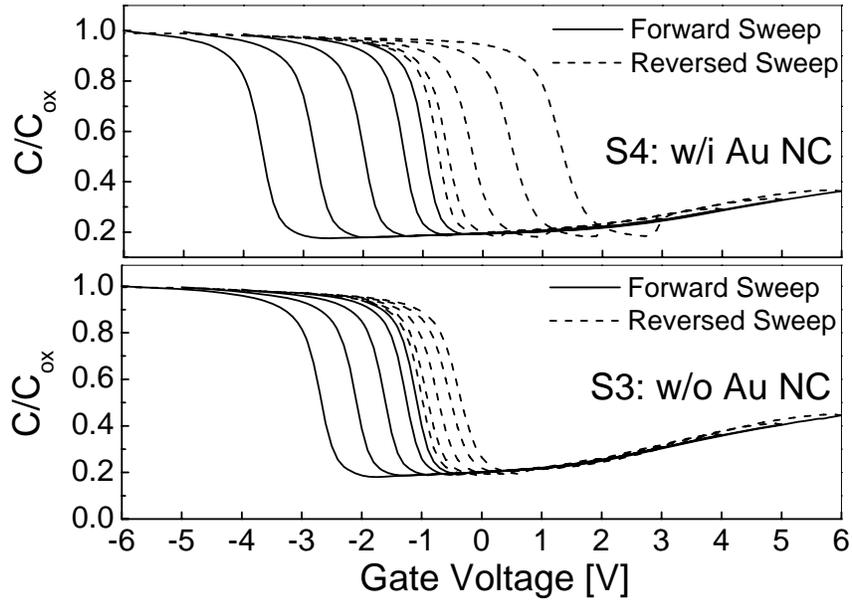


Figure 7.4 High frequency CV sweeps with increasing range from $\pm 2\text{V}$ to $\pm 6\text{V}$ for memory cells without the Au NC layer (S3) and with the Au NC layer (S4). Both S3 and S4 are with the C_{60} -embedded tunnel barrier.

7.5 Improved Tunneling Asymmetry in NC Memory

The high frequency CV sweeps with increasing range from $\pm 2\text{V}$ to $\pm 6\text{V}$ are shown in Fig. 7.4 for memory cells without the Au NC layer (S3) and with the Au NC layer (S4). Both S3 and S4 are with the C_{60} -embedded tunnel barrier. In a separate control sample without both C_{60} and Au NC but with all other dielectric layers, no hysteresis is observed under the same sweep range. In S3, larger negative flatband shifts (ΔV_{FB}) agree with the previous experiment in Chapter 6, indicating the preferable hole storage at monoanion C_{60}^{1-} and the higher charge neutrality level (CNL) of interface states between C_{60} and SiO_2 [14]. In S4, much larger and symmetric ΔV_{FB} clearly indicates that both electron and hole can indeed be injected into the upper Au NCs through the resonant tunneling modes provided by C_{60} .

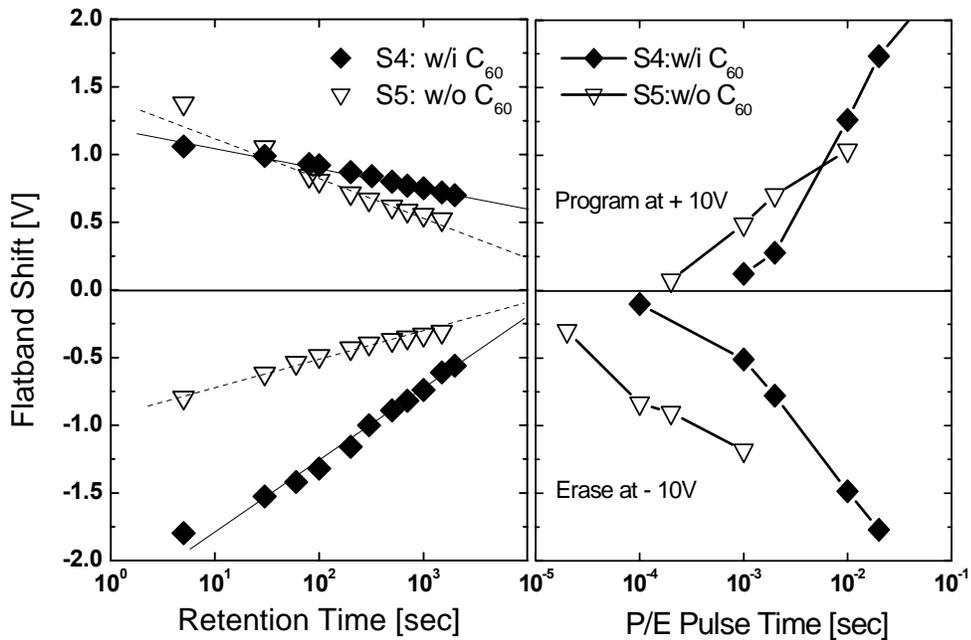


Figure 7.5 Retention and P/E characteristics of metal NC memories, S4 with a composite barrier of 2.5-nm SiO₂ + C₆₀ + 3-nm evaporated SiO₂ and S5 with a single layer of 2.5-nm SiO₂. For consistent initial conditions, the preset bias prior to the retention and P/E measurements is $\pm 5V$ for 3 sec.

Finally, the retention and P/E characteristics of a metal NC memory cell with a single layer of 2.5-nm SiO₂ (S5) are compared with S4 in Fig. 7.5. S4 has longer retention due to the lower escape rate of thermally excited electrons and holes in Au NCs through a physically thicker barrier provided by the additional C₆₀ and top SiO₂ layers. This is more pronounced for electron storage with at least two orders of magnitude improvement in the extrapolated retention time. It is likely due to the suppression of trap-assisted tunneling of electrons with the high CNL at the C₆₀/SiO₂ interface. The P/E speed at $\pm 10V$ in S4 is only about 10 times slower than that in S5, in close agreement with the tunneling current results in Fig. 7.2(a). Even though the P/E voltage has not yet been optimized, which is expected to scale by improving the coupling ratio with either a thinner or a higher- κ control oxide [5], improved t_R / t_{PE}

ratio by at least an order is clearly shown with the field-sensitive C₆₀-embedded tunnel barrier. Further improvement would be possible by reducing the non-ideal effects by the C₆₀/SiO₂ interface states and the bulk traps in the evaporated SiO₂.

7.6 Conclusion

We have demonstrated a field-sensitive asymmetry in tunneling probability through the molecule-embedded dielectric from J-V characteristics and from improved t_R / t_{PE} ratio of the memory cells. By taking advantages of versatile and tunable molecular properties, the integration of molecules in Si-based devices provides a simple and promising way to tailor tunnel dielectric properties. Although only the metal NC memory is investigated here, the proposed tunnel barrier can be applied for other charge-based memories such as conventional NAND flash, Si NC, and SONOS memories as well.

REFERENCES

- [1] G. Atwood, "Future directions and challenges of ETox flash memory scaling," *IEEE Trans. Device and Materials Reliability*, vol. 4, no. 3, pp. 301-305, Sept. 2004.
- [2] H. Pon, "Technology scaling impact on NOR and NAND Flash memories and their applications", in *Proc. Int. Solid-State and Integrated Circuit Technology*, 2006, pp. 697-700.
- [3] C. Lee, A. Gorur-Seetharam, and Edwin C. Kan, "Operational and reliability comparison of discrete-storage nonvolatile memories: Advantages of single- and double-layer metal nanocrystals," in *IEDM Tech. Dig.*, 2003, pp. 557-560.
- [4] T.-H Hou, C. Lee, V. Narayanan, U. Ganguly, and E. C. Kan, "Design optimization of metal nanocrystal memory—part I: nanocrystal array engineering," *IEEE Trans. Electron Devices*, vol. 53, pp. 3095-3102, 2006.
- [5] 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 Trans. Electron Devices*, vol. 53, pp. 3103-3109, 2006.
- [6] J. J. Lee, and D.-L. Kwong, "Metal nanocrystal memory with high- κ tunneling barrier for improved data retention," *IEEE Trans. Electron Devices*, vol. 52, pp. 507-511, 2005.
- [7] K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices," *Appl. Phys. Lett.* vol. 73, pp. 2137-2139, 1998.
- [8] S. Baik, S. Choi, U. I. Chung, and J. T. Moon, "High speed and non volatile Si nanocrystal memory for scaled flash technology using highly field-sensitive tunnel barrier," in *IEDM Tech. Dig.*, 2003, p. 545.

- [9] P. Blomme, B. Govoreanu, M. Rosmeulen, J. Van Houdt, and K. DeMeyer, "VARIOT: a novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices," *IEEE Electron Device Lett.*, vol. 24, pp. 99-101, 2003.
- [10] C. M. Compagnoni, D. Ielmini, A. S. Spinelli, and A. L. Lacaita, "Modeling of tunneling P/E for nanocrystal memories," *IEEE Trans. Electron Devices*, vol. 52, no. 4, pp. 569-576, Apr. 2005.
- [11] H.-T. Lue *et al.*, "BE-SONOS: A bandgap engineered SONOS with excellent performance and reliability," in *IEDM Tech. Dig.*, 2005, pp. 547-600.
- [12] R. Ohba, Y. Mitani, N. Sugiyama, and S. Fujita, "25 nm Planar Bulk SONOS-type Memory with Double Tunnel Junction," in *IEDM Tech. Dig.*, 2006, pp.959-962.
- [13] R. Ohba, Y. Mitani, N. Sugiyama, and S. Fujita, "15 nm planar bulk SONOS-type memory with double junction tunnel layers using sub-threshold slope control," in *IEDM Tech. Dig.*, 2007, pp. 75-78.
- [14] T. H. Hou, U. Ganguly, and E. C. Kan, "Programmable molecular orbital states of C₆₀ from integrated circuits," *Appl. Phys. Lett.*, vol. 89, 253113, 2006.
- [15] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices," *IEEE Trans. Electron Devices*, vol. 46, no. 7, pp. 1464-1471, July 1999.
- [16] W. Green, S. Gorun, G. Fitzgerald, P. Fowler, A. Ceulemans, and B. Titeca, "Electronic structures and geometries of C₆₀ anions via density functional calculations," *J. Phys. Chem.*, vol. 100, 14892-14898, 1996.
- [17] G. W. J. Beenakker, "Theory of Coulomb-blockade oscillations in the conductance of a quantum dot," *Phys. Rev. B*, vol. 44, pp. 1646-1656, 1991.

- [18] E. Suzuki, D. K. Schroder, and Y. Hayashi, "Carrier conduction in ultrathin nitrided oxide films," *J. Appl. Phys.*, vol. 60, pp. 3616-3621, 1986.
- [19] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "Modeling of anomalous SILC in flash memories based on tunneling at multiple defects," *Solid-State Electronics*, vol. 46, pp.1749-1756, 2002.

CHAPTER 8

CONCLUSION

8.1 Summary of Major Contributions

The major contributions of the work described in this dissertation are summarized as follows:

1. A physical model based on 3D electrostatics and 1D WKB tunneling applicable to both single- and multi-layer NC memory has been established.
2. The unique 3D electrostatics plays an important role on the performance of the NC memory. In particular, the field enhancement around NCs and the 3D capacitive coupling are highlighted.
3. Optimization strategies throughout the entire structure of the metal NC memory have been extensively examined, including NC size, NC density, tunnel oxide, control oxide, and multiple-layer NC *etc.*
4. Hybrid molecular integration in the nonvolatile memory cell has been explored. The precise control of the C_{60} redox states at room temperature may potentially realize reliable multi-level molecular memories. C_{60} -embedded tunnel oxide utilizing the resonant tunneling phenomenon improves the tunneling asymmetry substantially.

8.2 Suggestions for Future Work

As mentioned in Chapter 1, the potential deteriorated V_{th} distribution in NC memories stems from the inherent NC registry and size variations. Despite all advantages of the metal NC memory we have explored in great detail, it has to prove acceptable bit error rate (BER) even under aggressive feature-size scaling before been

considered as a legitimate succession of the current Flash technology. The 3D model presented in this work is based on a single unit cell with special geometrical symmetry and not directly applicable for the demanding BER calculation. Moreover, only the charge transport in the gate-stack direction is considered. More quantitative description of charge transport in the channel direction is essential to take into account the percolation effect for accurate estimation of the BER in scaled memory cells. Therefore, developing a more sophisticated model including the full 3D electrostatics in arbitrary cell configurations, and current transport through both the gate-stack and the channel directions is the next important step to advance the current capability.

The SONOS-type memory is a competing technology to the NC memory. The theoretical framework of SONOS has been established back 1970' [1], [2] with limited recent progress. The major controversy in the present model based on the Frenkel-Poole (F-P) emission and the Shockley-Read-Hall (SRH) statistics is the unphysically large capture cross section σ . σ for nitride traps was reported ranging from $5 \times 10^{13}/\text{cm}^2$ [3] to $2.2 \times 10^{12}/\text{cm}^2$ [4], whereas the consensus agrees on an amphoteric trap model [5] with native traps in nitride being neutral. Considering the atomistic nature of traps originated from the unsatisfied dangling bonds, the reported values are apparently very high for neutral traps. Furthermore, since σ is typically treated as a pure fitting parameter in the present model, all detailed physics associated with the elastic and inelastic tunneling processes to zero-dimensional (0D) traps is largely overlooked. As a result, the SONOS memory model has none or little predicting power for guiding the future device design and optimization especially when many high- κ charge trapping layers are under heavy investigation now. Hence an improved physical-based model on the SONOS-type memory is of great scientific and engineering interest. It will also benefit the design of the aforementioned hybrid

NC/nitride memory, which has unique opportunity in realizing both excellent P/E efficiency and scalability.

Regarding the valuable future experiments on the metal NC memory, the performance advantages of employing high- κ tunnel and control oxide has been explicitly shown in this work. However, the systematic experimental demonstration has been hindered by the non-ideal Fermi-level pinning at the high- κ /metal NC interface as discussed in Chapter 5, and trap-assisted tunneling in high- κ tunnel oxide. Hopefully both better experiment design and improved high- κ film quality may finally realize the true potential of the high- κ integration. Furthermore, the narrow-width channel with the 2D control gate is an intriguing configuration with excellent P/E efficiency. Previous demonstration utilizing carbon nanotubes (CNT) [6], [7] is a proof of concept, but far from real applications because of very little control on the CNT growth process. A more controllable implementation possibly based on semiconductor nanowires is worthwhile exploring.

The versatile and tunable properties of molecules developed in chemistry research society may open up new possibilities of unprecedented performance and functionality in Si technology. Our pioneering work of integrating molecules into the nonvolatile memory cell should be further extended in several directions. First, the diverse range of molecules is waiting to be explored. For example, as a charge storage node, negative-U molecules [8], which has total electron-electron correlation energy $U_{ee} < 0$ due to the strong electron-phonon interaction, may enhance charge retention in molecules. Second, the present choice of molecules has been limited to fullerenes owing to their excellent thermal stability and structural robustness. However, many interesting molecules may not be as stable or robust. The general process compatibility with molecules has to be carefully addressed. This may require developments on tailored molecular structures, innovative device designs and low-thermal budget

processes. Lastly, beyond the memory application, the monodisperse nature of molecules and their characteristic redox states provide a viable platform for highly sensitive molecular sensing in the integrated circuit.

REFERENCES

- [1] I. Lundstrom, and C. Svensson, "Tunneling to traps in insulators," *J. Appl. Physics*, vol. 43, pp. 5045-5047, 1972.
- [2] P. C. Arnett, "Transient conduction in insulators at high fields," *J. Appl. Physics*, vol. 46, pp. 5236-5243, 1975.
- [3] K. A. Nasyrov, V. A. Gritsenko, Y. N. Novikov, E.-H. Lee, S. Y. Yoon, and C. W. Kim, "Two-bands charge transport in silicon nitride due to phonon-assisted trap ionization," *J. Appl. Physics*, vol. 96, pp. 4293-4296, 2004.
- [4] A. Paul, C. Sridhar, S. Gedam, and S. Mahapatra, "Comprehensive simulation of program, erase and retention in charge trapping Flash memory," in *IEDM Tech. Dig.*, 2006.
- [5] Y. Yang, and M. H. White, "Charge retention of scaled SONOS nonvolatile memory devices at elevated temperatures," *Solid-state Electron.*, vol. 44, pp. 949-958, 2000.
- [6] U. Ganguly, E. C. Kan and Y. Zhang, "Carbon nanotube-based nonvolatile memory with charge storage in metal nanocrystals," *Appl. Phys. Lett.*, vol. 87, 43108, 2005.
- [7] U. Ganguly, C. Lee, T. H. Hou, and E. C. Kan, "Enhanced electrostatics for low-voltage operations in nanocrystal based nanotube/nanowire memories," *IEEE Trans. Nanotech.*, vol. 6, pp. 22-28, 2007.
- [8] A. S. Alexandrov and A. M. Bratkovsky, "Memory effect in a molecular quantum dot with strong electron-vibron interaction," *Phys. Rev. B*, vol. 67, 235312, 2003.