MODELLING AND CHARACTERIZATION OF THE QUANTUM DOT FLOATIING GATE FLASH MEMORY

ZHOU KAI HONG

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2004

Name: Zhou Kai Hong
Degree: M.Eng
Department: Electrical & Computer Engineering
Thesis Title: Modeling and Characterization of the Quantum Dot Floating Gate Flash Memory

Abstract:

This thesis discusses the physics, modeling and design issues of the nanoscale quantum dot flash memory. The characteristics of the flash memory device with one quantum dot floating gate are predicted successfully for the purpose of design. The advantages and applicability of emerging dielectric and quantum dot materials are demonstrated and quantified using simulation for the first time.

The characterization of the quantum dot floating gate flash memory is investigated by a self-consistent solution of Schrödinger- Poisson equation. The tunneling current of the flash memory is calculated by a semi-classical WKB approximation. The programming and retention times are evaluated to the scalability of the tunnel oxide. Studies are further extended to the applicability and advantages of high-k dielectrics, including HfO_2 and HfAlO. The impact of Ge and SiGe quantum dot on the retention time of the flash memory is also studied. This research work gives a comprehensive and detailed simulation of the quantum dot flash memory device with emerging materials. Based on this quantum modelling, ideal quantum dot flash memory device is finally proposed.

Keywords: Quantum Dot, Flash Memory, Self-consistent Solution, High-k Dielectric

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof Ganesh S Samudra, not only for the insightful and valuable guidance and support to this project, which has led me get into the gate of research, but also for the encouragement and positive comments, which have always given me confidence in the last two years. I am also very grateful to Dr Bai Ping, who gives the expert advice, valuable discussion which has helped me a lot in completing the project successfully. I must also thank Dr Rajendra Patrikar, for helping me a lot in understanding basic knowledge of nanoelectronics. The award of a research scholarship by the Institute of High Performance Computing is also gratefully acknowledged.

I am further indebted to Dr Chong Chee Ching for his continuous help during the last half year. I also thank him for his prompt reading and careful critique of my thesis. I wish to thank Dr Yeo Yee Chia, Mr. Hou Yong Tian and Prof Yoo Won Jong for freely sharing their expertise in this project. It was enjoyable working with my fellow students in IHPC and SNDL group. I really want to express my thanks and best wishes to them for their kind help and discussion throughout the project.

I must take this opportunity to express my deep gratitude to my family for their love, care, understanding, support and encouragement during these two years.

Table of Contents

List of Acronyms

QD	Quantum Dot
NC	Nanocrystal
FG	Floating Gate
FET	Field Effect Transistor
CMOS	Complementary Metal-Oxide-Semiconductor
EOT	Equivalent Oxide Thickness
ONO	Oxide-Nitride-Oxide Layer
NFM	Nonvolatile Flash Memory
CVD	Thermal Chemical Vapor Deposition
WKB	Wentzel-Kramers-Brilliouin method
WFM	Wave-Function-Match Method
XPS	X-ray Photoelectron Spectroscopy
ITRS	International Technological Road map for Semiconductor
V_T	Threshold Voltage
Et	Trap Energy
Ec	Conduction Band Shift
V_g	Control Gate Voltage
D	Diameter of the Quantum Dot
T	Temperature (^{0}C)
V_{ox}	Oxide voltage drop

ϕ_B	Conduction band offset
t_ox	Tunnel oxide thickness
χ_c	Electron affinity
V_d	Drain current
Sub	Substrate
E_{ox}	Electric field
ΔV_T	Threshold voltage shift
V_T	Threshold voltage

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Overview

In the late 60's, solid-state nonvolatile memory devices were first introduced and their commercial development followed quickly. As a nonvolatile memory device, the flash memory has many ideal memory characteristics and is consequently considered as a driver for the semiconductor industry in the next decade. The statistic shows that the worldwide market for semiconductor memory was valued at nearly $47 billion in 2002, and expected to cross $86 billion by 2007[1]. Although there is a huge commercial success，conventional floating gate flash memory devices are facing their scaling limitation, that is, it is becoming increasingly difficult to shrink flash memory chips. Indeed, electrons begin to leak out of an ultra thin tunnel oxide weak spot, leading to data corruption or loss.

In order to overcome the scaling problem to improve the memory characteristics, nanocrystal-based memories have been proposed [5]. It is believed that they could potentially become an evolutionary replacement of conventional polycrystalline floating gate flash memories. These new memory devices have been experimentally demonstrated and shown excellent memory performance and high scalability. In many laboratories, the quantum dot or nanocrystal based flash memory is rapidly

approaching length scales of less than 10 nm, in order to yield a higher packing intensity and a faster circuit speed.

As the size of the quantum dot flash memory is continually scaled down to nanometer regime, many important physical phenomena, especially quantum mechanical effects, play important role and become significant[2,4]. For example, the quantum effects become significant as the confinement of electrons becomes stronger within a nanoscale device [4]. Furthermore, in order to optimize the memory characteristics at low voltage, in recent years, high-k dielectrics and metal quantum dot were proposed to replace SiO_2 and Si, respectively[10,20]. Hence, their performance in the flash memory needs to be explored and studied carefully as well with such new materials.

In this context, fundamental physics poses stringent challenges and difficulties on the traditional theoretical simulation. In the traditional simulator, it becomes difficult to describe and analyse these quantum phenomena which occur in small nanoscale dimensions, such as quantum effects, single electron effects and F-N/direct tunneling in high-k dielectrics.

For this reason, in this thesis, a device simulation model using new theory and approaches is proposed to allow a comprehensive understanding of the memory characteristics of the flash memory with various new materials. In this research work, we developed a new TCAD (technology computer aided design) tool to accomplish

the task of understanding the device physics, designing flash memory devices, and predicting their performance limits. The results of modeling and characterization of the single Si/SiGe/Ge quantum dot floating gate flash memory device with SiO_2, HfO_2 and HfAlO as dielectrics will make up this thesis.

1.2 Objectives

The aims of this research work are to develop a simulation tool to study the quantum dot flash memory device and implement the appropriate physical methodologies in device modeling. The simulation tool developed investigates characteristics of programming and retention phenomena of flash memories and explores the effect of new dielectric materials on the memory performance. The impact of the dot size of Si and Ge quantum dot on the retention characteristic of memory device is studied and discussed.

1.3 Scope

This work mainly focuses on developing a simulation tool to construct an optimized quantum dot flash memory structure, including the study of electrons charging phenomena of the quantum dot, addressing programming/retention properties, and investigating various alternative high-k dielectrics, such as HfO_2 and HfAlO. The Coulomb Blockade is considered using an approximate method. Both the Si and Ge

quantum dots with different dot size are considered and their performance, in particular, programming and retention are examined. The self-consistent solution of the Poisson-Schrödinger equation and a modified WKB approximation are adopted in developing the simulation tool.

Chapter 2 gives a brief review about the current research progress and development in the study of quantum dot flash memories. It serves as a background introduction to this work, in which essential concepts and the vital methodology are elaborated. The new proposed materials and their applications in the flash memory device are also introduced.

Chapter 3 is devoted to the theory and methodology implemented in this simulation model. The main physical model, theory and methodology are described and explained. The method of solving Schrödinger and Poisson equations self-consistently is described. Various numerical techniques used in developing this simulator, such as Poisson equation boundary conditions and mode-space method, are explained. The semi-classical analytical WKB approximation used in the calculation of gate current is also described. Finally, the way to estimate the programming and retention times is presented.

In chapter 4, the verification of simulation results is presented in order to verify our device model. The results are compared and contrasted with published theoretical and

experimental data. Good agreement of our results with the reported data is demonstrated. The differences between them are discussed and explained as well.

In Chapter 5, we consider the silicon quantum dot flash memory with SiO_2 as the tunnel dielectric. The developed simulator nanoFM-1.0 is described. Using this simulation tool, we examine the performance related characterization of the quantum dot flash memory, considering single electron charging effect approximately. The tunneling current and the impact of the tunnel oxide thickness on the tunneling current are also investigated carefully. The programming and retention characteristics are estimated and are used to explore the scalability of the quantum dot flash memory.

In Chapter 6, we model the flash memory with the quantum dot embedded in high-k dielectrics and its characteristics are compared to the SiO_2 flash memory device. The model explores the effect of alternative high-k tunnel dielectrics on the memory performance. The advantages of high-k materials, including HfO_2 and HfAlO, are analysed and their potential of replacing the SiO_2 is demonstrated. The efficient programming and good retention of the flash memory with high-k dielectrics are shown by simulated results.

In Chapter 7, germinum nanocrystal is studied with special attention to the effect of trap energy on the retention time. SiGe nanocrystal is considered and the basic properties and characteristics are explored. The impact of the trap energy on the

retention time is examined using germinum nanocrystal. The effect of dot size on the characteristics of flash memory is also discussed briefly. Finally, based on our current detailed physical model and through analyzing the results, we propose an optimum memory structure which shows close to ideal memory characteristics, if perfect materials and interfaces are used.

In Chapter 8, we conclude the work presented in this thesis, and reinforce some of its results. Also, some potential directions for future work are suggested.

Chapter 2
Literature Review

2.1 Introduction

This chapter reviews the recent research progress of nonvolatile flash memories, including the traditional and innovative memories. The development and applications of new materials that can optimize the performance of flash memories are presented. An overview to the physical theory and methodology which are implemented for describing the new quantum phenomena in innovative memories is given.

Section 2.2 introduces the development of flash memories. Section 2.3 provides a brief review of new materials applied in flash memories, including high-k dielectrics and Ge nanocrystals. Furthermore, section 2.4 reviews the main physical concepts and methodology used in the study of various characteristics of nanocrystal memories. Section 2.5 summaries the content of this chapter.

2.2 Nonvolatile Flash Memory

2.2.1 Conventional flash memory

Solid-state memory devices that retain information once the power supply is switched

off are called “nonvolatile” memories. There are two most common solutions used to store the information in nonvolatile memories:

(1) in traps which are present in the insulator or at the interface between two dielectric and other materials. The most commonly used interface is the silicon oxide/nitride interface.

(2) in a conductive material layer between the gate and the channel, and completely surrounded by the insulator. This is called the “floating gate”(FG) device.

The nonvolatile memories based on charge trapping are a very low fraction of the total nonvolatile memory production. On the contrary, floating gate flash memories form the basis of every modern nonvolatile memory, and are used in particular for flash application. The single cell of floating gate memories can be electrically programmed, and a large number of cells, called a block, sector or page, are electrically erasable at the same time [4]. The word “flash” means that the whole memory can be erased at once and the erase time can be very short.

During the early growth stage of the flash memory device industry, a dominant design emerged, the so-called continuous floating gate flash memory. In this conventional flash memory, the information is stored in a continuous polysilicon layer, called floating gate (FG). The floating gate is located between the channel and the conventional gate of the FET, surrounded completely by dielectrics. The charge stored in the floating gate can be sensed easily because it is directly proportional to the

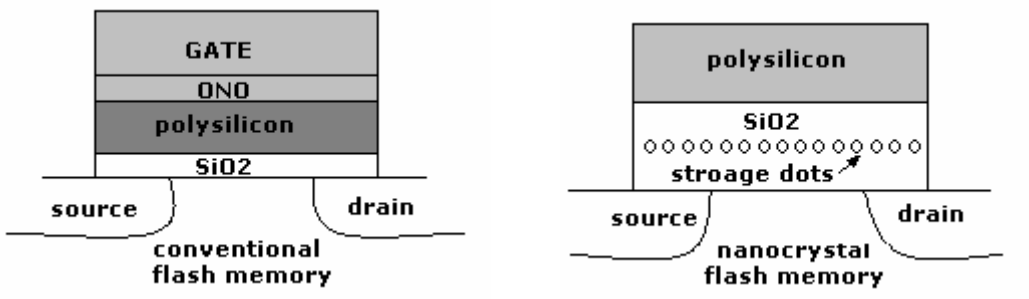
threshold voltage of the FET (Fig.2.1 (a)). When electrons are on the FG, they modify the electric filed in the gate region, which modifies the threshold voltage of the memory device. Hence, when the memory is read by placing a specific voltage on the control, electric field will either flow or not flow, depending on the threshold voltage of the memory. This presence or absence of current is sensed and translated into 1 or 0s, reproducing the stored data. Therefore, the charge stored in the floating gate can be sensed easily. The traditional dielectric used in flash memory is silicon dioxide. The writing and erasing operations are done by increasing or decreasing the control gate voltage. Two standards are used to describe how “good” and reliable is a nonvolatile memory. They are: (1) endurance: the capability of maintaining the stored information after erase, program, or read cycling, (2) retention: the capability of keeping the stored information ever ling time.

Although a huge commercial success, conventional flash memories are confronted with challenges. They are: (1) multilevel cell development, (2) cell scaling and scaling limitations, (3) low-voltage compatibility, (4) product diversification. (1) and (4) mainly come from industry manufacturability consideration. The most prominent one today is the limited potential for continued scaling of the device structure and low voltage operation.

The scaling limitation primarily stems from the future application requirements in terms of densities and performances, in particular the extreme requirements imposed

on the tunnel oxide between the FG layer and the silicon substrate. The tunnel oxide needs to provide fast, low voltage write/erase operations. In other words, it requires an ultrathin tunnel oxide to provide quick and efficient charge transfer to and from the floating gate. On the other hand, the tunnel oxide has to allow superior isolation under retention and disturbance conditions in order to ensure ten years maintenance of stored information (the industry standard). This retention mainly depends on the thickness of the tunnel oxide. Due to above conflicting requirements, the conventional flash memory has only marginally improved with device scaling, with the compromise tunnel oxide thickness of the conventional flash memory ranging from 9nm-11nm. Although theoretically use of thin oxide is possible, a single weak spot in the oxide can adversely affect the retention as all FG charge can leak through spot.

In order to alleviate the scaling limitation of the conventional floating gate flash memory, quantum dots flash memory that is not susceptible to weak dielectric spot, is proposed as a candidate and aims to replace the conventional flash memory in recent years.



(a) (b)

Fig.2.1 *Schematic representation (a) a conventional floating gate nonvolatile memory cell (b) a nanocrystal nonvolatile flash memory cell. ONO=oxide-nitride-oxide layer.*

2.2.2 Nanocrystal nonvolatile flash memory

The first nanocrystals flash memory was introduced in the 1995[5] (see Fig.2.1 (b) for a schematic representation). In a nanocrystals flash memory, the conventional floating gate is replaced by a layer of discrete, isolated, nanocrystals or dots, normally made of semiconductor materials. The memory is programmed by applying to the gate a positive voltage of a few volts that lowers the thin oxide conduction band and enhances tunneling of electrons from the substrate to the quantum dot. Electrons get trapped in the quantum dot, since further tunneling to the gate is inhibited by the thicker top oxide. The information stored in the memory is then simply read by measuring the device current using to a gate voltage significantly smaller than that used for programming. The memory is erased by applying a negative gate voltage that ejects electrons from the nanocrystals into the channel. The V_T shift between the programmed and erased states is denoted by a quantity known as the “memory window”.

Electrons (charges) are confined in discrete 3-D dots instead of the continuous polysilicon floating gate. The distributed dots or nanocrystals make the stored charge more robust and thus the memory device shows the potential of affording a thinner tunnel oxide [6, 7] without sacrificing the retention time. Hence, the quantum dot flash memory provides advantages of shorter write-erase times, lower operation voltage and longer retention time compared to the conventional flash memory.

Also, in conventional flash memories, one weak spot will create a fatal discharge path and lose the information stored in the floating gate. The novel discrete and isolated dots floating gate layer will not make the memory device prone to failure just because of one weak spot. Due to the distributed nature of the charge storage in the nanocrystal layer, the nanocrystal flash memory shows good immunity to stress induced leakage current and oxide defects. On the other hand, the Coulomb Blockade effect in the quantum dot flash memory can enable both the single and multi bits storage [22, 23]. Coulomb blockade is based on the charging energy of a small capacitor and allows the transport of single electrons. If one electron is stored in the nanocrystal, the system will be raised by the electrostatic charging energy $e^2/2C$ [56]. When electrons are to tunnel into the QD through tunnel oxide, the capacitor must be charged. When applied a voltage larger than the threshold voltage, electrons can tunnel through tunneling oxide and to the other reservoir. In this case only single electron transport occurs. The suppression of the current due to modified field is called "Coulomb blockade" [56].

There are some other advantages for the use of the nanocrystal flash memory. From the fabrication process viewpoint, the nanocrystal flash memory devices process adds only a few steps to the conventional complementary metal-oxide-semiconductor (CMOS) technology, offering a reduced number of masks compared with the conventional FG flash memory process. Therefore, it leads to a corresponding reduction in cost for system-on-a-chip application employing such devices [8]. The nanocrystal memory also allows the use of a shorter channel length and therefore a

smaller cell area.

There are several shortcomings as well to these nanocrystal flash memories. An important one is the low capacitive coupling between the external control gate and nanocrystal floating gate. This weakness results in a somewhat higher voltage operation, thus offsetting the benefits of the thinner tunnel oxide thickness. It degrades the important parameter, coupling ratio, which is used to optimize the performance/reliability tradeoff.

However, generally, the nanocrystal floating gate flash memory device is still a promising candidate for replacing the conventional flash memory in future. Both the experimental and theoretical studies of the nanocrystal flash memory have been explored to demonstrate their advantages in recent years.

2.3 Scaling Limitation of Nanocrystal Memory

2.3.1 Alternative high-k dielectrics

The primary driver behind flash memories is the potential to scale down the tunnel oxide thickness, which results in lower operation voltage and fast programming speed. However, most recent results, for instance, find reasonable programming efficiency with 2.3nm SiO_2 tunnel oxide, but lose 25% of its stored charges in several tens of

seconds[9]. Because the continuous scaling of the tunnel oxide results in a significant degradation in the retention performance, in ITRS 2004, the tunnel oxide thickness of 4.6nm is considered as a practical limit. As a result, it becomes difficult to improve the programming speed(voltage and /or time) and data retention simultaneously, because they both rely on the tunneling current through an ultra thin tunnel oxide between the floating gate and silicon substrate.

Based on the above discussions, the nanocrystal flash memory device with SiO_2 dielectric is rapidly approaching a point where device fabrication can no longer be progressively scaled to a smaller size [10]. In order to overcome this problem, alternative materials with dielectric constants ranging from 10-80 are proposed to replace the traditional SiO_2. It implies that the physical thickness of the dielectric, which possesses thinner equivalent oxide thickness(EOT) to maintain electrical properties, can be increased.

Using the high-k dielectrics, both lower programming/erasing voltage and better retention performance can be achieved. This is due to the smaller conduction band offset between Si substrate and high-k dielectrics, and the larger physical thickness of high-k dielectrics [11, 12], respectively. When high-k dielectrics are used as a control oxide, the control gate coupling ratio can be increased because of smaller EOT of the control oxide. Hence, the control gate voltage couples to the tunneling oxide more effectively, which provides lower programming voltage and enlarged memory

window. Recent experiments and simulations have identified these advantages offered by high-k dielectrics in nanocrystal flash memory [11-13].

Many materials have been explored as potential alternative gate dielectric candidates for flash memory devices. The most commonly studied high-k gate dielectric candidates are $SrTiO_3$, Ta_2O_5, Al_2O_3 and HfO_2.

It is important to distinguish between the requirements for memory and transistor applications so that optimization strategies become clear. For flash memory, it requires extremely low leakage currents and very high capacitance density for charge storage, while the interface quality is not as critical. Since the main requirement of flash memory is that the floating gate capacitor stores the charge, current transport along the dielectric interface is not that important. However, the stability of the interface is still critical in the reading process. Therefore, all of the requirements amount to the important distinction that the bottom dielectric interface quality is not as critical to capacitor performance.

In contrast, a key requirement of a Field Effect Transistor is that the electric field should induce a channel in Si to modulate carrier transport, and that the dielectric-channel interface be of a very high quality. The channel must be of course Si, so any potential high-k dielectric must be compatible with Si. Transistors have more lenient leakage requirement for high-performance processors, although high

capacitance densities are still needed.

The most critical distinction between high-k materials requirements for capacitors versus gate dielectrics is the interface and materials compatibility: gate dielectric must form an extremely high-quality interface with Si, and also be able to withstand CMOS processing conditions especially source-drain annealing while in contact (or near contact) with Si.

2.3.2 Considerations of high-k dielectrics properties

All high-k dielectrics must meet the following requirements [14] in order to be a successful gate dielectrics. The several criteria are summarized in this section.

(1) Permittivity and barrier height: It is essential to select a gate dielectric with a higher permittivity than that of SiO_2. However, the required permittivity must be balanced by corresponding change in the barrier height for the tunneling process. It is more appropriate to find a dielectric which provides a moderate increase in k value and also has a tunneling barrier preventing the large leakage current in the retention.

(2) Thermodynamic stability on silicon: The dielectric should be thermodynamically stable on Si substrate with respect to formation of uncontrolled SiO_2 or silicates at the Si/high-K interface during the deposition or post deposition annealing (PDA). Most of

the high-k dielectrics require an interfacial reaction barrier to ensure the thermodynamic stability on Si substrate.

(3) Interface quality: For potential high-k dielectrics, it is crucial to attain a sufficiently high-quality optimal high-k-Si interface. Therefore, the origin of the interface properties of high-k dielectrics should be understood clearly in order to create a good interface as that of SiO_2.

(4) Film morphology: A high-k dielectric with an amorphous film structure is an ideal gate dielectric in flash memory device. It will be helpful to prevent the effects of mass or electrical transport along grain boundaries and overcome the extent of crystallization.

(5) Gate compatibility: One significant issue for high-k dielectrics is that they should be compatible with Si-based gates which could create the desired threshold voltage V_T by tuning the dopant implant.

(6) Process compatibility: The deposition process for the dielectric must be compatible with current or expected FG flash memory processing, cost, and throughput.

(7) Reliability: The electrical reliability of a new gate dielectric must also be considered critical for application in flash memory technology. It requires a well-characterized materials system for high-k dielectrics.

2.3.3 Interface between silicon substrate and high-k dielectrics

Except for Al_2O_3, many high-k dielectrics are not thermodynamically stable in direct contact with silicon. As an attempt to prevent/minimize reaction with the underlying silicon, and to maintain high channel carrier mobility, interface engineering schemes form oxynitrides and oxide/nitride reaction barriers between these high-k dielectrics and silicon [15, 16, 17] which have been tried. Recently, investigation of amorphous ZrO_2-SiO_2 and HfO_2-SiO_2 alloys have been studied extensively [18].

2.3.4 Ge nanocrystal flash memory

In nanocrystal memories, electrons are stored in the traps or the conduction band of nanocrystals. Experiments demonstrate nanocrystal memories with electrons stored in interface states or bulk traps, rather than the conduction band can provide good retention performance [19]. Since narrower band gaps can provide lower conduction band edge, better confinement of electrons and longer retention time, nanocrystals with narrow band gap materials are good candidates to replace silicon nanocrystals in flash memories. For example, compared with silicon, Ge nanocrystals have narrower

band gap and similar electron affinity.

Therefore, Ge nanocrystals are expected to provide both a higher confinement barrier for retention time and a smaller barrier for program and erase mode [11]. Since the fabrication of Ge dot on the insulator is much more difficult than Si dot, an alternative technique is implemented to form $Si_{1-x}Ge_x$ directly on the insulator using thermal chemical vapor deposition (CVD)[20].

2.4 Quantum Dots Flash Memory Modeling

2.4.1 Device modeling

As discussed previously, the detailed simulation tool is necessary, and helpful in understanding new physical phenomena occurring in nanocrystal flash memories. As flash memories are scaled to the nanometer regime, quantum effect plays an important role. Therefore quantum mechanical model is required to explore the new characterization of flash memories. Many quantum models have been employed in simulating nanocrystals flash memory devices.

The most commonly used model is a self-consistent simulation of Schrödinger's and Poisson's equation, in which the potential and electrons distribution of the device system are solved self-consistently [21, 22]. Farhan Rana et al also use a quantum

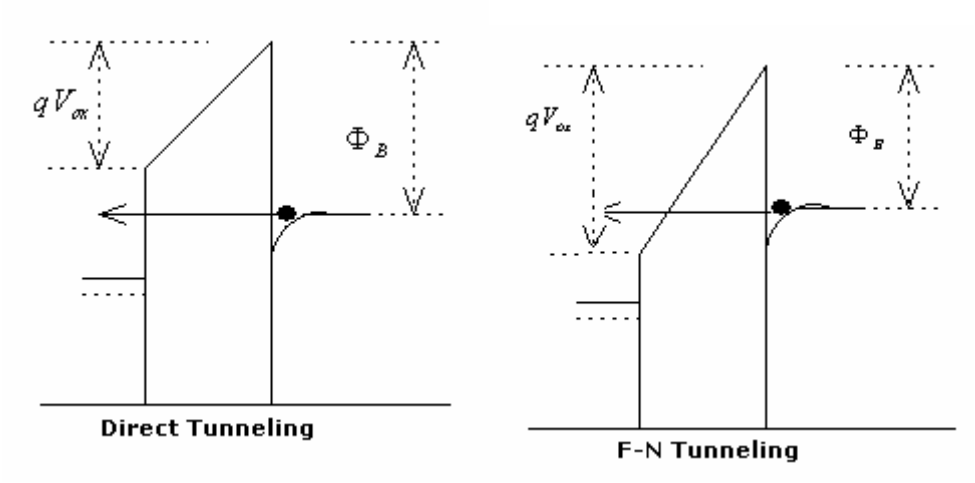
kinetic approach, based on a master equation for modeling the injection and ejection of electrons into and from the quantum dot, in which Heisenberg representation is employed [23] and Coulomb Blockade effect is simulated. J.S.de Sousa et al use Kohn-Sham-Poisson self-consistent scheme to obtain electronic spectrum of a silicon nanocrystal [24]. H.G.Yang et al apply Bardeen's transfer Hamiltonian formalism for flash memory systems modeling, in which the tunneling process of electrons could be considered as that of the transition between the two eigenstates of H_1(silicon substrate) and H_2(floating gate)[25]. Among these quantum simulation methods, the self-consistent solution of Schrödinger-Poisson using a computational mesh is the most popular and mature method. It is proved to be sufficient to provide good agreement with experimental results [21, 22, 26]

2.4.2 Tunneling models

The operations referred as writing and erasing the memory cell, require either the increasing or reducing the amount of charge stored on the FG, with electrons tunneling between the floating gate and silicon substrate. The tunneling mechanism in flash memories includes the direct tunneling and the Fowler-Nordheim (FN) tunneling.

The F-N tunneling is a quantum mechanical process in which electrons tunnel through a thin dielectric from (or to) a floating gate to (or from) a conducting channel [27]. The direct tunneling happens when the oxide voltage drop is less than the conduction band

offset of insulator(tunnel oxide) and silicon substrate, the electrons can tunnel directly through the forbidden energy gap of the insulator(tunnel oxide). The illustrations of direct tunneling and F-N tunneling are shown in the Fig.2.2.



(a) (b)

Fig.2.2 *Illustrations of (a) direct tunneling and F-N tunneling (b)* Φ_B *is the conduction band offset and* V_{ox} *is the oxide voltage drop.*

2.4.3 Various tunneling current calculation models

The tunneling phenomenon through a forbidden energy barrier has been studied for a long time and its basic mechanism has been known [28]. Many approaches are proposed to study the tunneling phenomenon. Some typical models are reviewed in this section.

(1) *Classical Tunneling Model*: Classical tunnelling current model focuses on the carriers in the extended states(3-D). In this 3-D model, the transmission probability is well-defined as the ratio of transmission and incident flux [29, 30]. The tunnelling

current is decided by weighting the electron distribution function by the carrier transmission probability.

(2) *Transverse Resonance Method* : In flash memory devices, the carriers are indeed of 2-D nature and distributed in the discrete subbands, while, the classical model does not consider the 2-D quantum effects and its transmission probability is not accurate enough for describing the confined carriers in the potential well. As a result, a full quantum mechanical model, named transverse resonant method is proposed. It uses the life-time τ of these quasi-bound states to evaluate the tunnelling current [31-33]

$$J = \sum_n N_n / \tau_n(E_n) \tag{1.17}$$

where N_n is the carrier density of nth subband.

(3) Wentzel-Kramers-Brillioun(WKB) Approximation: WKB approximation is a simple and a well-known method for the calculation of tunnelling probability [34]. The transmission probability can be expressed as

$$T_{WKB}(E) = e^{-2\int k(z)dz}$$

where $k(z)$ is the imaginary part of wave number of the carrier.

(4) Semi-classical tunneling model: In transverse resonant method, though the life-time of quasi-bound states can be evaluated by the width of the quasi-bound states resonance, tremendous numerical effort is required. Hence, a semi-classical tunnel

model was developed to give an efficient evaluation of the life-time of quasi-bound states [31]. Recently, a first-principle approach has been used to produce a modified WKB tunnelling expression, which, for the trapezoidal barrier, is similar but not identical in form to that of WFM, and for low to moderate voltages results are similar to those which are from numerical analysis [35-37]. They present comparison of the quasi-classical model to the full quantum numerical calculation [38-39]. They show good agreement and demonstrate the applicability of the semi-classical model.

2.5 Summary

In this chapter, an overview of current research progress of the floating gate flash memory and focus on the quantum dot flash memory with various dielectrics is given. The advantages of the quantum dot flash memory are presented and the important concepts used in flash memories are explained. The application and properties of high-k dielectrics are presented. The importance of the simulation of the quantum dot flash memory with high-k dielectric is established. The main simulation models are reviewed and their advantages and shortcomings are compared and contrasted. The main shortcoming of these models is the lack of ability of simulating the quantum phenomena occurring in the new device. Through the comparison, a new simulator will be implemented in this thesis in order to include the main quantum effects which occur in nanoscale regime. Therefore, a new simulation tool focusing on quantum effects is adopted in this thesis, in which the F-N/direct tunnelling is calculated. The

detailed explanation of the simulation tool, verification by comparing to published data and new nanocrystal flash memory simulation results with high-k dielectrics will be given in the following chapters.

Chapter 3

PHYSICAL THEORY, MODEL AND METHODOLOGY

3.1 Introduction

In nanocrystal embedded flash memories, electrons are confined in 3-D nanocrystals and therefore a more aggressive scaling of the tunnel oxide is enabled. With the rapid progress of the nanocrystal flash memory, the device dimension today has been scaled into the nanometer regime. Under this situation, new phenomena, especially quantum effects, play important roles, and therefore a quantum computation model to explore new characteristics of the flash memory device is required. Moreover, it facilitates the design and optimization of the memory device. In this thesis, a two-dimensional simulator, called nanoFM-1.0, is developed to perform the analysis of the nanocrystal flash memory. The physical model and methodology used in this thesis are presented in this chapter.

A self-consistent solution of the Poisson-Schrödinger equation simulation method is used to evaluate the charging process of nanocrystal memories. The potential profile and electrons distribution of the device system are obtained by solving the Poisson and Schrödinger equation, self-consistently. The tunneling characteristics of the thin dielectric between the floating gate and silicon substrate are calculated by using an analytic modified semi-classical WKB approximation. Two methods are used to

evaluate the programming time. One is to find the distribution function of the relationship between the stored charge and tunneling current by fitting a theoretical curve. Another one is calculated from the time-dependent tunneling current density. The retention time is evaluated by calculating the probability of an electron escaping from the quantum dot. There are 17 routines in nanoFM-1.0 and the code is implemented by using Matlab 6.1. The layout of the simulator that comprises of 17 routines is illustrated in Fig.3.1 in which the main routines are presented. Except for accessorial routines, the rest were developed during the course of this project.

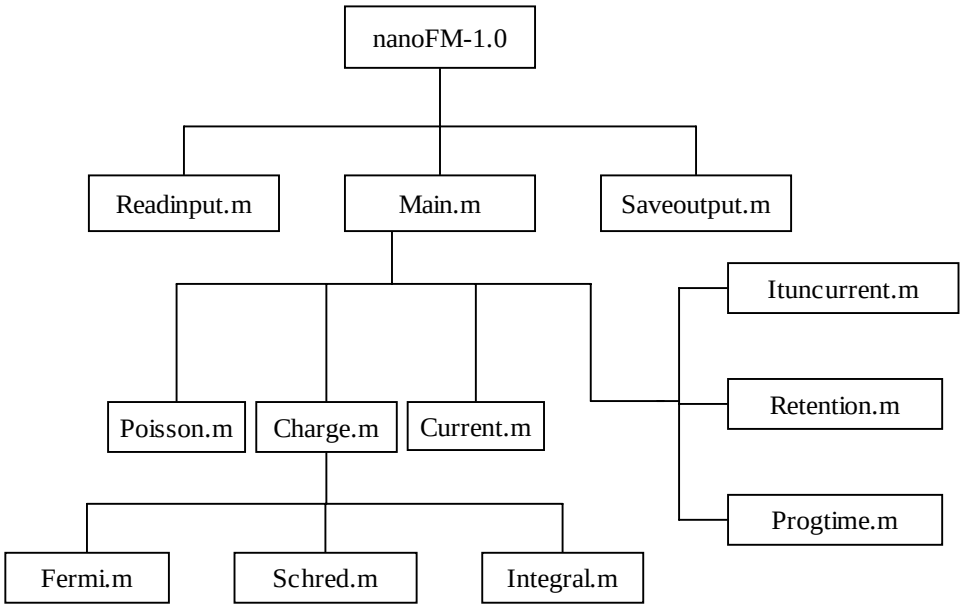


Fig.3.1 *Main routines of 2-D simulator nanoFM-1.0.*

Among these routines, readinput and saveoutput routines are used to input the parameters and output/plot the simulation results. Poisson and Schred routines are implemented to solve the Poisson equation and Schrödinger equation, respectively. The tunneling current is calculated in the Ituncurrent routine. Retention and progtime

routines evaluate the retention and programming times. Current routine simulates the drain current of the flash memory device. NanoFM-1.0 and main routines are the main functions to call other routines.

Since the quantum dot flash memory uses nano-scale island to store charge, Coulomb blockade becomes more prominent. In this work, Coulomb blockade is emulated simply by simulating single electron charging effect approximately in Fig.5.6 and Fig.6.7.

This chapter is organized as follows. Section 3.2 describes the scheme of self-consistent solution of the Schrödinger-Poisson equation, including the description of 2-D Poisson equation and 1-D Schrödinger equation. Section 3.3 introduces the semi-classical WKB method which is used to calculate the tunneling current. Section 3.4 explains how to estimate the programming time and retention time. Section 3.5 gives a summary.

3.2 Self-consistent Solution of Schrödinger-Poisson equation

The self-consistent simulation tool, called nanoFM-1.0, is modified from nanoMOS-2.0[40], which is originally used in the simulation of double-gate MOS devices. The main programming methodology of nanoFM-1.0 is illustrated in Fig. 3.2.

The developed code consists of two iterative loops: the Poisson equation which is solved to obtain the device potential profile, and the transport equation (1-D Boltzmann and 1-D Schrödinger equations) which is solved to obtain the electrons distribution in the device system. In order to solve Schrödinger and Poisson equations, the finite difference discretization scheme is used.

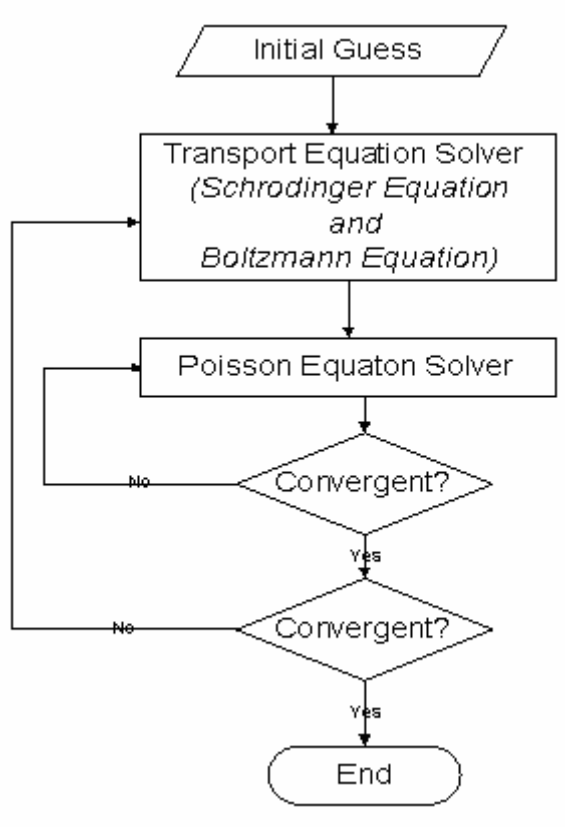


Fig.3.2 *A illustration of self-consistent solution of the Schrödinger-Poisson equation.*

3.2.1 Computational Scheme

In this work, the simulation domain of the entire flash memory device is partitioned into mesh grids as showed in Fig. 3.3.

As N_x and N_y are defined as the number of grids in the x and y direction, the solution domain consists of $N_{total} = N_x \times N_y$ nodes. In order to obtain N_{total} value of the potential and electron density, the Poisson equation is discretized into N_{total} equations and the Schrödinger equation is solved slice by slice at every x location by using the mode-space method.

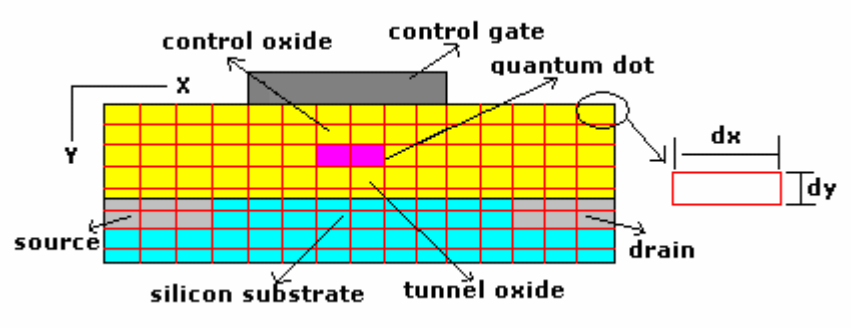


Fig.3.3 *The cross-section of the quantum dot flash memory device with uniformly spaced grids in X and Y direction. The width and height of a grid are dx and dy, respectively.*

Normally, for an ultra-thin body, the spacing dy should be chosen smaller than the spacing dx in order to obtain an accurate simulation result and efficient convergence. In order to ensure an efficient and faster convergence, values of dx and dy should be selected appropriately.

3.2.2 Poisson Equation

The potential profile of the device is obtained from the solution of a 2-D Poisson's equation which is obtained by including Gauss's law,

$$\oiint[\varepsilon \vec{E}(x,z).d\vec{S}] = \int_{\Omega} q[p - n + N_D - N_A]d\Omega, \quad (3.1)$$

where $\vec{E}$ is the electric field, p is the hole concentration (neglected in this n type floating gate flash memory), n is the electron concentration, N_D and N_A are donor and acceptor concentrations, respectively, q is the elementary charge, and ε is the position dependent dielectric constant.

As the simulation domain is divided into grids, the Poisson equation is discretizated into discrete equations, in which N_{total} potential values at each node are obtained. Applying Eq. 3.1 at all internal nodes, the linearized finite difference form of Eq.3.1 is given as [40]

$$\frac{a}{b}V_{m-1,n} + \frac{b}{a}V_{m,n-1} - 2(\frac{a}{b} + \frac{b}{a})V_{m,n} + \frac{b}{a}V_{m,n+1} + \frac{a}{b}V_{m+1,n} = -\frac{ab}{\varepsilon}q(N_D - N_A - n_e)_{m,n} \quad (3.2)$$

where a and b are the spacings in the x and y directions, V the vacuum potential. m and n denote row and column. n_e is the electron concentration (but for other sections n is used to denote the electron concentration). $\varepsilon = \varepsilon_{ox}$ and $\varepsilon = \varepsilon_{si}$ are the dielectric constants in the oxide region and silicon regions, respectively. In this work, all the interface nodes between two different materials are considered as internal nodes, therefore there is no need to form the discrete Poisson equations for the interface nodes.

The boundary condition is applied at different regions in memory device. In source and drain region, the Neumann boundary condition $\vec{n} \cdot \vec{\nabla} V = 0$ is employed. It

means that potentials at the contact can float to any values which are necessary for ensuring charge neutrality of the contact. The Dirichlet boundary condition $V_{m,n} = V_G$ is implemented for nodes under the control gate. V_G is the gate potential, which is determined by the gate bias voltage and the workfunction of gate contact material.

Zero electric field conditions are imposed on other boundary nodes. For the left and right edges, the boundary condition is $V_{m,n} - V_{m\pm1,n} = 0$. For the top and bottom edges, the boundary condition is set as $V_{m,n} - V_{m,n\pm1} = 0$. For the two corner nodes along the left edge, the boundary condition is expressed as $2V_{m,n} - V_{m+1,n} - V_{m,n\pm1} = 0$. As for the two corner nodes along the right edge, $2V_{m,n} - V_{m-1,n} - V_{m,n\pm1} = 0$ is assumed.

Using Eq. 3.2 and above boundary conditions, the N_{total} discrete nonlinear Poisson equations are obtained, which are solved by the Netwon-Raphson method. These nonlinear Poisson equations are denoted by $F_\alpha(V) = 0$, where the index α means the number from 1 to N_{total}. The Jacobian matrix is obtained as

$$F_{\alpha,\beta}(V) \equiv \frac{\partial F_\alpha(V)}{\partial V_\beta} \tag{3.3}$$

Given an initial guess of the previous solution V_{old}, the projected solution is $V_{new} = V_{old} + \Delta V$. By using a Taylor expansion of the first order, we have

$$F_\alpha(V_{new}) \approx F_\alpha(V_{old}) + F_{\alpha,\beta}(V_{old}) \cdot [\Delta V]_\beta = 0 \tag{3.5}$$

Therefore, we obtain

$$[\Delta V]_\beta = -F_{\alpha,\beta}(V_{old}) / F_\alpha(V_{old}) \tag{3.6}$$

Above process is repeated until the residual of $F_\alpha(V)$ is less than the specified

convergence norm. Since the Netwon-Raphson approach provides a quadratic convergence, the number of iterations is small. Because the size of the Jacobian is $(N_x \times N_y)^2$, the memory and time to carry out Gaussian eliminations can be excessive.

3.2.3 1D Transport Equation

The 1D ballistic transport is modeled at a semi-classical level, that is, using the Boltzmann transport equation (BTE) and Schrödinger equation. The solution of Boltzmann and Schrödinger equations assumes that the vertical potential profile variations along the channel direction are negligible. This assumption is reasonable because for the flash memory device, the length of the channel and floating gate is always larger than 10nm. Hence, the quantum effect in the transverse direction can be neglected. As a result, the 1D BTE is solved in the transverse direction slice by slice, and the solution is a charge-sheet description. About details on the solution of 1D Boltzmann transport equation, please refer to the reference [40].

The Schrödinger equation is solved by using mode-space method [40]. The main principle of mode-space method is to transform an original 2D Schrödinger equation to a 1D partial differential Schrödinger equation. The detailed explanation can be found in Ren Zhibin's work [40]. As a result, the size of original 2D problem can be reduced greatly. In the flash memory device, the envelop wavefunctions of electrons

are assumed to penetrate into tunnel oxide region. The zero boundary condition is applied to the dielectric layer between the floating gate and control oxide.

The 2D electron density can be obtained by a semi-classical BTE method. The 3D electron density can be calculated by multiplying the corresponding distribution function to the 2D density matrix at each longitudinal lattice mode. The 3D electron density is then fed back to the Poisson equation solver for self-consistent solutions.

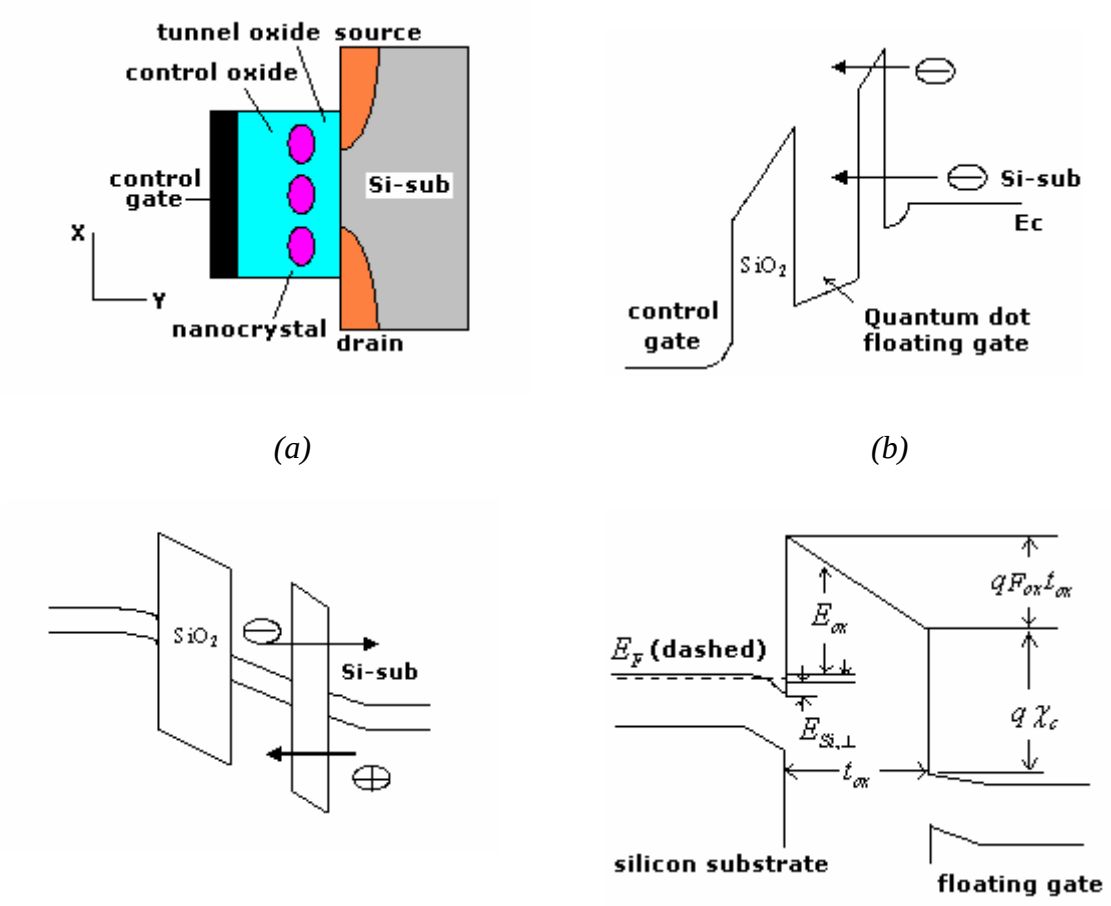
Finally, the self-consistent simulation of the Schrödinger and Poisson equations are achieved through an interactive scheme as shown in Fig.3.2. The Poisson equation is solved for the potential profile, and the transport equation (1D BTE and Schrödinger equation) is solved for electron density and wavefunction distribution. Iterations continue till both solutions are consistent.

3.3 The Calculation of the Tunneling Current

3.3.1 Tunneling mechanism in the flash memory

It is well known that as SiO_2 is scaled below 3.5 nm, the direct and/or F-N tunneling mechanisms dominate. A large amount of current can pass through the tunnel oxide at a low voltage [41]. In this situation, a thin tunnel oxide is highly desirable for achieving fast programming/erase time. In the programming model, a positive gate voltage is applied to form channel inversion-layer. The oxide field thus increases and

the potential barrier seen by electrons near the band bottom changes from trapezoidal to triangular. The flow of electrons through the trapezoidal barriers is referred as direct tunneling current, as electrons tunnel directly into the floating gate. In F-N tunneling case, electrons are injected into the floating gate through a triangular barrier.



(a) *(b)*

(c) *(d)*

Fig. 3.4 *(a) Quantum dot floating gate flash memory device structure (b) Illustration of programming state (c) Illustration of retention state (d) Band diagram for WKB approximation.*

Eventually, an electron will tunnel from silicon substrate into the floating gate or from the floating gate to silicon substrate. In this process, the direct tunneling and F-N tunneling contribute to the tunneling current, as shown in Fig. 3.1(b). During the

retention mode, ideally no charges should be lost. However, because the conduction band edge inside the quantum dot is higher than that of silicon substrate, so electrons in the floating gate can tunnel back to the silicon substrate, as shown in Fig.3.4.(c), thereby leaking to charge loss and finite retention time.

3.3.2 Semi-classical WKB approximation

In this research work, a modified WKB approximation [36, 37] is implemented to study the tunneling current through the tunnel oxide. The band diagram in WKB calculation is showed in Fig. 3.4(d) [36]. In order to simplify the theoretical model, we assume that the control oxide thickness is large enough to prevent electrons from tunneling into the control gate. The tunneling probability calculated by a modified WKB approach [42, 43] is expressed as

$$T = T_{WKB} T_{R1} T_{R2} \tag{3.7}$$

where T_{WKB} is the usual WKB tunneling probability valid for smoothly varying potentials T_{WKB} is defined as

$$T_{WKB} = \exp\left[\frac{E_g \sqrt{2m_{ox}}}{4\eta q F_{ox}} (2\gamma' \sqrt{\gamma} + \sqrt{E_g} \sin^{-1} \gamma') \right]_{E_{ox}=q\phi_{in}}^{E_{ox}=q\phi_{cat}} \tag{3.8}$$

where

$$q\phi_{cat} = q\chi_c - (E_{si,\perp} + E_{si,\parallel}) \tag{3.9}$$

$$q\phi_{in} = q\chi_c - (E_{si,\perp} + E_{si,\parallel}) - qF_{ox}t_{ox} \tag{3.10}$$

are net barrier heights for electrons at the cathode and anode interfaces, respectively.

t_{ox} is the thickness of the tunnel oxide and F_{ox} is the electrical field in terms of the control gate voltage (Fig.3.4(d)). χ_c=3.15 eV is the silicon substrate-tunnel oxide conduction band discontinuity. $E_g = 9$ eV is the band gap of Silicon oxide. In Eq.(3.8), γ and γ' are defined as

$$r = \frac{\eta^2 k_{ox}^2}{2m_{ox}} = E_{ox}(1 - \frac{E_{ox}}{E_g}) \tag{3.11}$$

$$\gamma' = \frac{d\gamma}{dE_{ox}} = (1 - \frac{2E_{ox}}{E_g}) \tag{3.12}$$

T_{R1}, T_{R2} in Eq. (3.7) are corrections for reflections from potential discontinuities. They are obtained by considering reflections from the material interface and the band structure diagram is showed in Fig. 3.4. (d). The correction factor is defined as

$$T_R = T_{R1} T_{R2} \tag{3.13}$$

T_{R1} and T_{R2} depend on the group velocity of electrons through,

$$T_{R1} = \frac{4v(\mathrm{E_{Si,\perp}})\mathrm{v_{ox}}(\mathrm{q}\phi_{\mathrm{cat}})}{v_{si,\perp}{}^2(\mathrm{E_{Si,\perp}}) + v_{ox}{}^2(\mathrm{q}\phi_{\mathrm{cat}})} \tag{3.14}$$

$$T_{R2} = \frac{4v_{si,\perp}(\mathrm{E_{Si,\perp}} + qF_{OX}\,\mathrm{t_{OX}})v_{\mathrm{ox}}(\mathrm{q}\phi_{\mathrm{an}})}{v_{si,\perp}{}^2(\mathrm{E_{Si,\perp}} + qF_{OX}\,\mathrm{t_{OX}}) + v_{ox}{}^2(\mathrm{q}\phi_{\mathrm{an}})} \tag{3.15}$$

where v_{ox} is the group velocity of electrons and is defined as

$$v_{ox}(E) = \frac{1}{\gamma'}\sqrt{\frac{2\gamma}{m_{ox}}} \tag{3.16}$$

$m_{si,\perp} = 0.98m_e$, and $m_{si,\parallel} = 0.19m_e$, where m_e is the free space electron rest mass. E_{ox} is the magnitude of the electron energy referenced to the oxide conduction band edge. $E_{si,II}$ and $E_{si,\perp}$ are the subband quantization energies along direction parallel and perpendicular to the interface, respectively, and written as

$$E_{Si,\perp} = 0.6 \times \frac{\left(3\pi\eta q m_{Si,\perp}\right)^{2/3}}{2m_{Si,\perp}} \left(\frac{\varepsilon_{OX} F_{OX}}{\varepsilon_{Si}} \right)^{2/3} \tag{3.17}$$

$$E_{si,II} = \frac{1}{2}(E_f - E_{si,\perp}) \tag{3.18}$$

With the relationship between accumulated charge, impact frequency, and tunneling probability, gate current can be evaluated analytically by

$$J_g = \frac{nqm_{si,\parallel} f}{\pi\eta^2} \int_0^{E_F - E_{si,\perp}} T dE_{si,\parallel} \cong QfT \Big|_{E_{si,\parallel}=1/2(E_f = E_{si,\perp})} \tag{3. 19}$$

where a midpoint approximation to the integral is used for simplicity to obtain the mean tunneling probability while introducing very little error. T is the tunneling probability calculated by a modified WKB approach and is given by Eq.3.7. One limit of this WKB approximation is that it uses semi-classical transport to simulate the quantum transport, such as coulomb blockade. However, as a simplified model, it is acceptable as Min She's work also uses WKB approximation to calculate the tunneling current density [44]. On the other hand, because Coulomb blockade effect should be reduced by using big nanocrystals, hence for a quantum dot with dimensions about $6nm \times 10nm \times 10nm$ used in this thesis, using semi-classical tunneling model is reasonable.

3.4 Programming and Retention Times

3.4.1 Programming time

Two methods are used to evaluate the programming time. One method is to estimate the programming time by finding an expression of the programming time as a

function of number of electrons in the quantum dot from theoretical data [23]. Another method is to calculate the programming time by calculating the time-dependent tunneling current density.

This part will introduce how to find the distribution expression of the programming time as a function of stored charge. With the help of the tunneling current obtained above, the programming time will be estimated. As we know, the tunneling current is given by

$$I = dQ / dt \tag{3.20}$$

where Q is the charge in the quantum dot and t is the tunneling time which tells us how long the charge needs to inject into the floating gate. Therefore, once the tunneling current and the charge in the quantum dot are obtained, t can be estimated. In order to obtain an accurate programming time, reported data [23], showing the relationship between the charge in the quantum dot and programming time, is used to derive a distribution function $Q_{QD}=f(t)$[23].

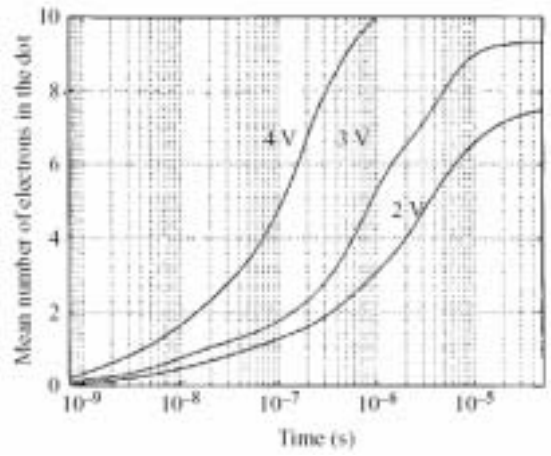


Fig.3.5 *Finding the expression of tunneling current as a function of number of electrons in the quantum dot*[23].

The theoretical result of programming time as a function of the charge in QD with control gate voltage of 2 V, 3 V and 4 V is given in Fig. 3.5[23], in which the curve at 2 V is chosen for fitting purpose. Fig.3.5 shows the number of electrons in a quantum dot as a function of time for three different values of the magnitude of the applied voltage pulse. A general feature is that larger the number of electrons already inside the quantum dot the longer it takes to add one more electron to the quantum dot. The detailed explanation for the reason will be given in section 5.5.

Using the relation of $I = (Q_1 - Q_2)/t$, function expression is given by

$$I = A\exp(-Q/Q_0) \tag{3.21}$$

Since

$$I = \frac{dQ}{dt} \tag{3.22}$$

Hence the differential equation is given as

$$\frac{dQ}{dt} = A\exp(-Q/Q_0) \tag{3.23}$$

Finally we have

$$Q = Q_0 \ln\left[\frac{A}{Q_0}(t - t_0)\right] \tag{3.24}$$

In Eq. (3.24), there are two unknown constant parameters. Therefore, in order to obtain these two unknown parameters, we have to get two sets of values of tunneling current with two different stored charges. In this work, one is obtained from MOS device and another one is obtained from flash memory device.

At first, we calculate the tunneling current J_{MOS} of the MOS device with the tunnel oxide thickness *t_ox,* and the stored charge Q_{MOS} is assumed to be 0. Secondly, we

calculate the tunneling current J_{FM} and the stored charge Q_{FM} in flash memory device, in which the *t_ox* is similar to the case of the MOS device. Using these two sets of values, Q_{MOS} J_{MOS} and Q_{FM} J_{FM}, the constant parameters, Q_0 and A, can be calculated using Eqn (3.24). Finally, the function of describing the relationship between the tunneling current and the electron charge is obtained. The detailed process to evaluate the programming time is described in the following flowchart.

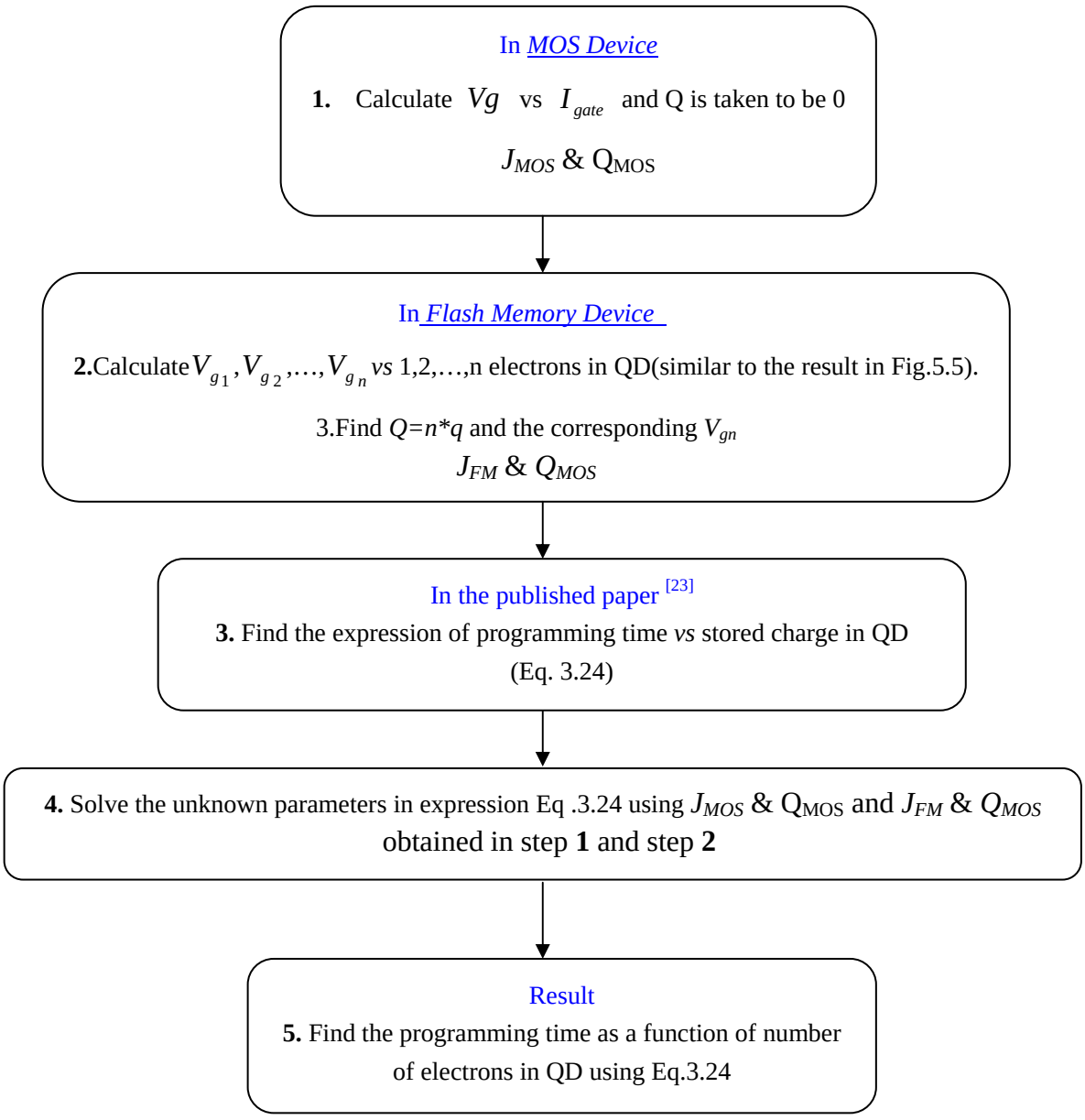


Fig.3.6 *Calculation method of programming time.*

The second method used to calculate the programming time is obtained from the time-depended tunneling current. The time-dependent tunneling current is [44]

$$J(t) = e\sum_{i,j} g_i \int_{E \geq E_{cn}} P(E) f_j(E) \rho_i(E) f(E) dE \tag{3.25}$$

where $P(E)$ is the transmission probability across the tunnel oxide calculated by the WKB approximation mentioned in section 3.3, *i* is the index for the two degenerate valleys, *j* is the index of subband for each conduction band valley, $\rho_i(E)$ is the density of states for each valley, $f(E)$ is the Fermi distribution, g_i is the degeneracy for these two degenerate valleys, E_{cn} is the conduction band edge in the nanocrystal, $f_i(E)$ is the impact frequency of the electrons impinging on the tunnel layer/silicon substrate interface and is expressed as [44]

$$f_j(E) = \frac{eE_{si}}{4\varepsilon_{si}} (m_z E_j / 3)^{-1/2} \tag{3.26}$$

where E_{si} is the silicon surface electrical field, ε_{si} is the dielectric constant, m_z is the silicon electron effective mass. E_j is the *j-th* subband bottom energy. The electric field and electron density are calculated by the simulator nanoFM-1.0. Hence, the total charge in the nanocrystal is defined as

$$Q = \int_0^{t_p} J(t) A dt \tag{3.27}$$

where t_p is the programming time and A is the quantum dot capture cross section area. Since the electric field across the tunnel oxide depends on the charge in the quantum dot, the tunneling current is time- dependent.

3.4.2 Retention time

During the retention mode, electrons will be thermally de-trapped to the conduction band and then tunnel back to the channel. The retention time in this work [44] is defined as the time when 20% of the charge leaks at zero gate bias from the quantum dot.

The probability of an electron escaping from the deep trap states back to the channel is given by [44]:

$$P(t) = \int_{E>E_{cn}} \alpha P(E) f_{imp}(E) \exp(-\frac{E+E_t}{kT}) \rho(E) dE \qquad (3.28)$$

E_{cn} : conduction band edge

α : a fitting parameter for nanocrystal shape

$P(E)$: transmission probability across the tunnel oxide calculated with WKB approximation

$f_{imp}(E)$: Weinberg impact frequency $(E+E_s)/h$ which describes the escape frequency of the electron from the conduction band

E_s : quantum confinement energy that is equal to the conduction band shift

E_t : relative trap energy level below the conduction band

$\rho(E)$: density of states for each valley $4\pi(\frac{2m^*}{h^2})^{3/2} E^{1/2}$

In Eq. 3.28, the transmission probability across the tunnel oxide is given by

$$T(E) = \left[1 + \frac{\sinh^2 \kappa a}{\frac{E}{V_0}\left(1 - \frac{E}{V_0}\right)}\right]^{-1} \tag{3.29}$$

where

$$\kappa = \frac{1}{\eta}\sqrt{2m(V_0 - E)} \tag{3.30}$$

Though Eq.(3.29) is for a square barrier, under a low programming voltage that is less than 2 V, it is acceptable in order to simplify the model. And also if the barrier drop is close to the potential of quantum dot, this simplification is reasonable. E_s varies with the size of nanocrystals, and for 5 nm, 3 nm, 2 nm nanocrystals it is taken as 0.15 eV, 0.5 eV and 1 eV, respectively. E_t is the trap energy in Ge quantum dot. For Ge quantum dot, E_t is about 0.51eV and the geometry factor α is 9.08×10^{-3} which are extracted from the experiment retention data [49]. For Si nanocrystal, in this thesis, the trap energy is assumed to be 0 eV. Finally, the charge in the nanocrystal is expressed as

$$dQ(t)/dt = -P(t)Q(t) \tag{3.31}$$

$$Q(t) = Q(0)e^{-\int P(t)dt} \tag{3.32}$$

As a result, the remaining charge on the nanocrystal can be calculated in terms of time and temperature.

3.5 Summary

This chapter focuses on the theory, models and methodology used in the simulation. Firstly, a self-consistent solver of the Schrödinger-Poisson equation is described. The

Poisson equation solver, 1D Boltzmann Transport equation and Schrödinger equation solver are presented and explained. The semi-classical WKB approximation that is used to calculate the tunneling current through the tunnel oxide is discussed. Finally, the way to estimate the programming and retention times is introduced. This chapter is important in providing theoretical background to the discussion of the simulation results in the following chapters.

Chapter 4
Verification of Simulation Framework

4.1 Introduction

In order to ensure the validity of simulation results in this research work, this chapter concentrates on the verification and evaluation of our simulation results. The simulation results are compared with the reported data. And if there are any discrepancies between them, reasonable explanations are given. Through the verification, evaluation and explanation exercises, we can check our simulation results and provide a better understanding for our physically detailed model. After the verification, it ensures that our simulation results are reliable and can be accepted widely.

In this chapter, important results of our research work are selected and compared with published results. Firstly, the electron density in the floating gate is verified by comparing reported theoretical data using the same device structure and size. This quantity is a basic and essential parameter for calculating other results. Secondly, due to the lack of experimental and theoretical data in the case of flash memory device, our simulation model has to be applied for the MOS device and the results are compared with reported experimental data coming from the MOS device. Finally, the programming and retention times are compared with the published results. Once

credibility of the software is established, we will apply it for new and novel devices.

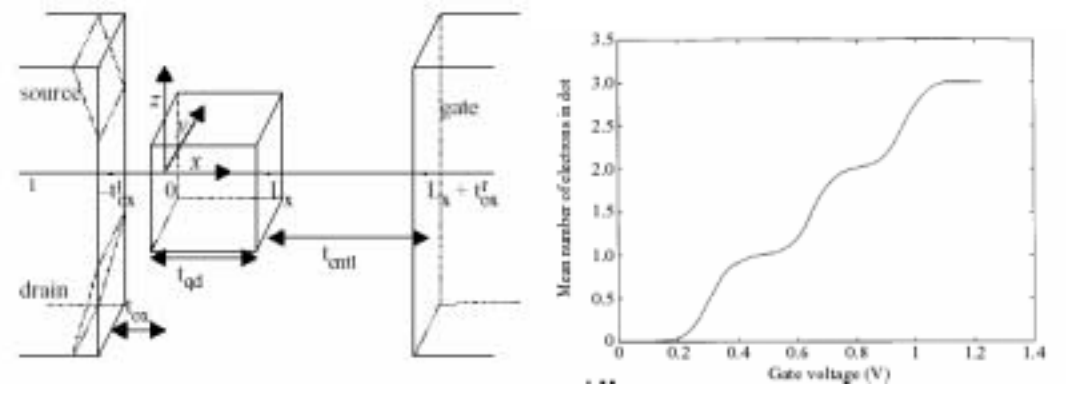
This chapter is organized as follows. In section 4.2, the charging process of the floating gate is compared to the reported data [23]. In section 4.3, the tunneling current in MOS device is calculated and compared with reported results [45]. In section 4.4, results of high-k dielectrics are compared with accepted results [44]. In section 4.5, the computed programming and retention times are verified. In section 4.6, a summary is given.

4.2 Charging Phenomenon of the Floating Gate

The charging phenomenon of the floating gate is a consequence of the quantum effect. The electron distribution is solved through the 1D transport equation, including Schrödinger and Boltzmann equations. It is a very important fundamental parameter for describing other characteristics of the memory device, such as tunneling current and programming/retention times. Therefore, in this section, the electron distribution in terms of number of electrons in the floating gate as a function of control gate voltage is compared with Farahan Rana's work [23]. Though the physical model used in this thesis is different compared with Rana's model, the purpose of comparison is to verify the electron distribution calculated in this work is close to a published result. In this model, a quantum dot is coupled to the silicon substrate as the quantum dot floating gate with dimensions of $6nm \times 10nm \times 10nm$. The calculation assumes a tunnel oxide of thickness 1.5nm, a control oxide of thickness 5nm, substrate doping of

$10^{17}cm^{-3}$ p-type as shown in Fig.4.1 (a).

It is significant that the variance in the electron number shows "staircase" phenomenon and the mean number of electrons is fluctuating rapidly between integer and integer+1 as shown in Fig. 4.1. (b). When control gate voltage is 1.0V, there are about 3 electrons in the quantum dot. The first electron does not appear in the quantum dot until the control gate voltage exceeds the threshold voltage V_{TO} =0.3V. When the quantum dot has one electron, the threshold voltage of the device is shifted up by ΔV_T . For the case of electron transfer, to add second electron into the quantum dot, the control gate voltage needs to be increased by at least ΔV_T beyond the voltage needed to place the first electron. In this device, as shown in Fig.4.1 (b), the threshold voltage shift ΔV_T is about 0.3 V. The second electron can only be trapped at a gate voltage of approximate $V_{TO} + \Delta V_T(v-1)$, where *v* is the number of electrons in the quantum dot.



(a) (b)

Fig.4.1 *(a) Device geometry considered in the model (b) Mean number of electrons in quantum dot as a function of gate voltage* [23].

The similar result from our research work, simulated by self-consistent solution of the

Schrödinger-Poisson equation, is shown in Fig.4.2. It indicates that the threshold voltage V_{TO} is about 0.3 V and when gate voltage is 1.0 V, the mean number of electrons is about 3. This shows good agreement with Rana's work. The threshold voltage shift for a single quantum dot flash memory with *v* electrons in the quantum dot is given by [23]

$$\Delta V_T(v) \approx \frac{ve}{A\varepsilon_{ox}}\left(\frac{t_{qd}\varepsilon_{ox}}{2\varepsilon_{si}} + t_{cntl}\right) \tag{4.1}$$

where t_{cntl} is the thickness of the control oxide, t_{qd} is the height of the quantum dot, and A is the area of cross-section. Thus using our simulation results, threshold voltage shift ΔV_T is calculated as 0.3 V and is same as the Rana's result.

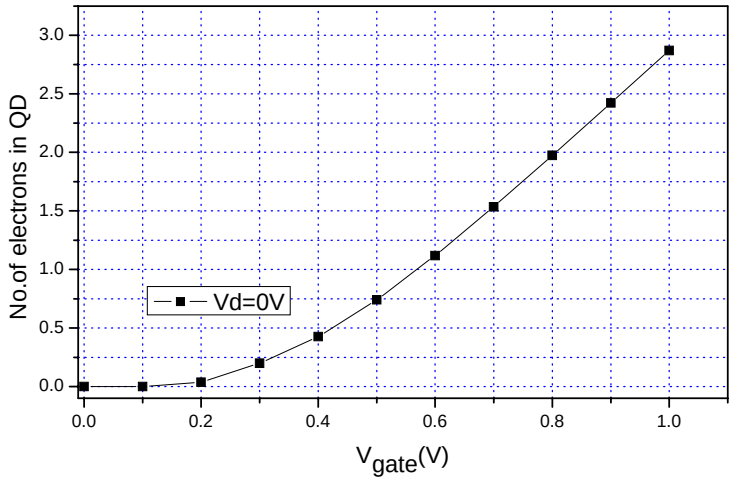


Fig.4.2 *Mean number of electrons in the quantum dot as a function of gate voltage calculated by self-consistent simulation.*

An obvious difference between ours and Rana's result is the "staircase" phenomenon. This discrepancy is due to the different physical models used to calculate the number of electrons (electron distribution). In Rana's work, the quantum kinetic approach based on a master equation is adopted, and it simulate a full quantum confinement effect in the quantum dot. While, in our simulation model, the number of electrons is

evaluated from electron density so that it is a continuous value and doesn't show "staircase" phenomenon.

4.3 Tunneling Current Simulation in MOS Device

After finishing the verification of electron distribution, we proceed to the verification of the tunneling current through the tunnel oxide in the flash memory. However, because of the lack of appropriate theoretical and experimental data in these devices for comparison purpose, our tunneling current calculation model is applied in the MOS device and calculated results are compared with the reported results available for the MOS device.

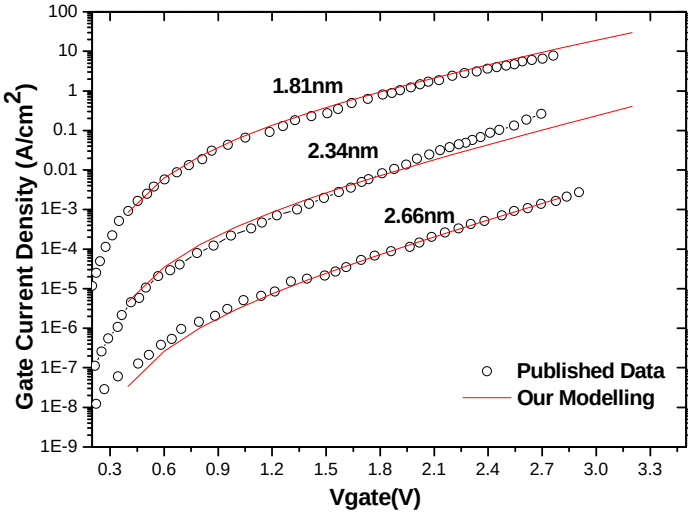


Fig.4.3 *The electron tunneling currents in nMOSFETs with SiO_2 gate dielectric by assuming $m_{ox}=0.61m_0$, compared with published data[45].*

Fig.4.3 gives a comparison between our results and Hou's results [45] for the tunneling current in a MOS device with SiO_2 gate dielectric. Different tunnel oxide of

thicknesses 1.81nm, 2.34nm and 2.66nm are considered. In this example, the conduction band offset is fixed at 3.15eV, and the effective mass of SiO_2 is $0.61m_0$. We assume that the effective mass is a constant value in various tunnel oxide thicknesses. Fig.4.3 shows that our results are in good agreement with the reported data when gate voltage is larger than 0.4V. The slight misfit at a low field is acceptable because the model for tunneling current at a very low field is not as well established as that at high field.

4.4 High-K Dielectrics Flash Memory Simulation

Since the flash memory with high-k dielectrics can optimize the memory characteristics, hence they have been investigated extensively in experiments [11-12]. However, the theoretical exploration of the flash memory with high-k dielectrics is lacking, especially tunneling currents of high k dielectrics when used in non-volatile memories. Therefore, we have to simulate the MOS device with high-k dielectrics to do the comparison. In this section, the tunneling current of the MOS device with three kinds of high-k dielectrics are simulated and compared to the published data. They are Si_3N_4, HfAlO and HfO_2, in which HfAlO and HfO_2 will be further investigated in chapter 6 and chapter 7.

In the early years of studying various high-k materials, Si_3N_4 was the most attractive material, which has a higher dielectric constant value of 7.8 than SiO_2. In our

simulation, the barrier height of 2.1 eV and the effective mass of $0.50m_0$ in Si_3N_4 are considered [45]. The band gap is assumed to be 5.3 eV [45]. The tunneling current in the MOS device with Si_3N_4 dielectric is obtained and confirmed by the published data in Fig.4.4. The simulation result in this work agrees well with published data [45]. As discussed previously, the unsatisfactory fit found at low voltage is due to the insufficient consideration in the numerical model at low field. Except the results at a very low field, our simulation result is acceptable and accurate.

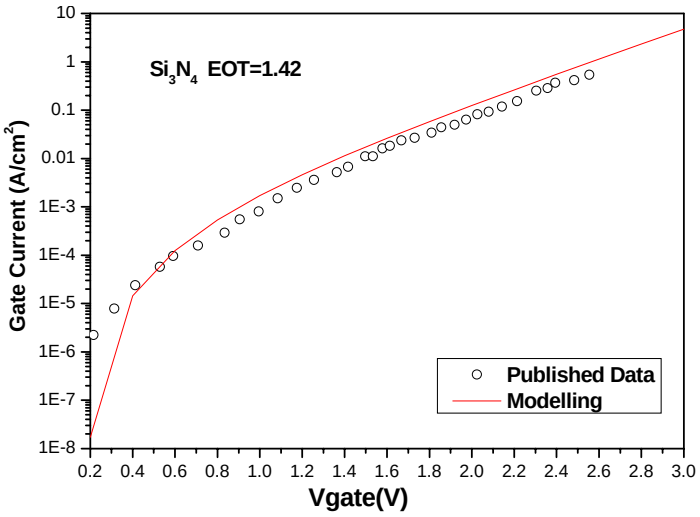


Fig.4.4 *Calculated electron tunneling currents through a Si_3N_4 gate dielectric with EOT of 1.42nm from inversion layer n MOSFET, compared with published data[45].*

Another high-k dielectric of prime importance, HfAlO [53], is investigated in this research work. Based on XPS experiments, HfAlO dependences on the Hf composition are demonstrated to be in a linear relationship. The result of its property is verified by the simulation of the MOS device. In this work, the electron mass and the dielectric constant of HfAlO are assumed to be linear interpolated between those of HfO_2 and Al_2O_3. Fig. 4.5 shows the simulated tunneling current as a function of various Hf compositions with the comparison of published data [45] (red line).

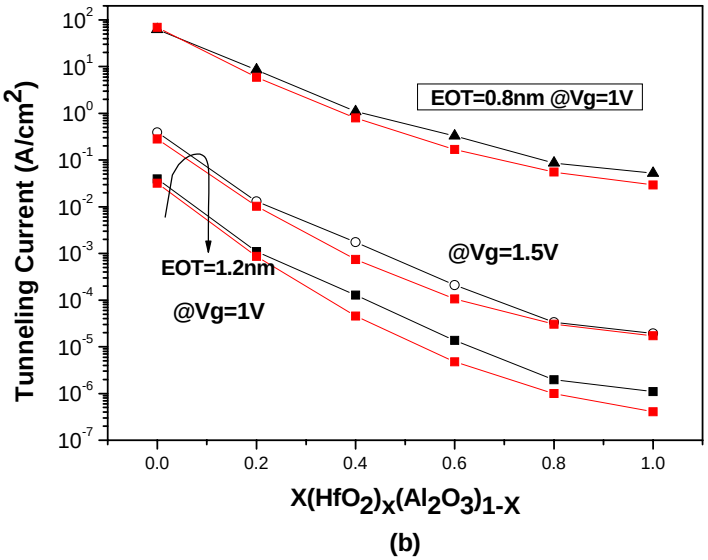


Fig.4.5 *Calculated tunneling currents of HfAlO for various Hf compositions, compared with published data*[45].

In this simulation, the dielectric constants of HfO_2 and Al_2O_3 are taken as 22 and 11 as discussed in Hou's work [45]. The barrier height of HfO_2 and Al_2O_3 are assumed to be 2.0 eV and 2.24 eV [45]. And the effective mass of HfO_2 and Al_2O_3 are set as $0.18m_0$ and $0.28m_0$ [45]. Our simulation result is very close to the reported data in which the higher Al composition results in higher tunneling current. In MOS device, 30% concentration of Al is regarded as the optimized value for HfAlO which is mainly decided by the experimental issue.

Because of the large dielectric constant, small band gap and band offset, HfO_2 has been studied extensively recently and is believed to be a good candidate to replace SiO_2. The verification of our HfO_2 simulation results is done by simulating the tunneling current of HfO_2 versus equivalent effective thickness in the MOS device. The result is shown in Fig.4.6, in which the substrate doping is $10^{18}cm^{-3}$, the dielectric

constant is 22 and the effective mass is $0.18m_0$. There is a good match between the reported data [45] and our simulation results.

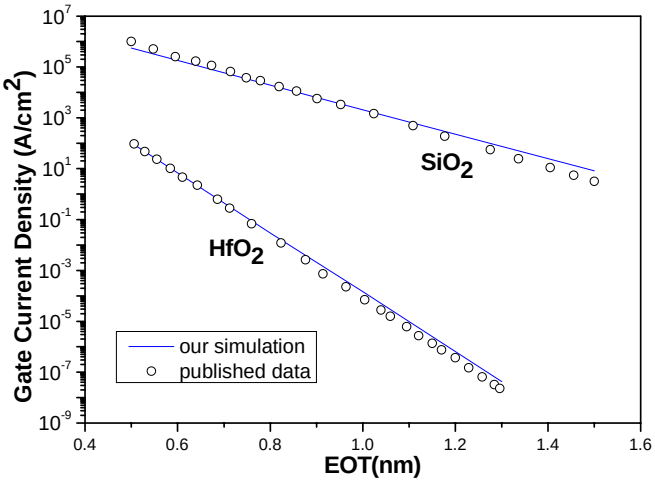


Fig.4.6 *Simulated tunneling current of MOSFET versus EOT for* HfO_2 *and* SiO_2 *gate dielectrics. The substrate doping is* $10^{18}\,cm^{-3}$*, compared with published data*[45].

4.5 Estimation of Programming and Retention Times

4.5.1 Verification of the programming time

In this section, an estimation of programming and retention times is given and verified. Firstly, the number of electrons in the quantum dot as a function of programming time is calculated using the method discussed in chapter3. The programming time of the flash memory with SiO_2 is presented in Fig.4.7 and the theoretical result in Rana's paper is showed in Fig.4.8.

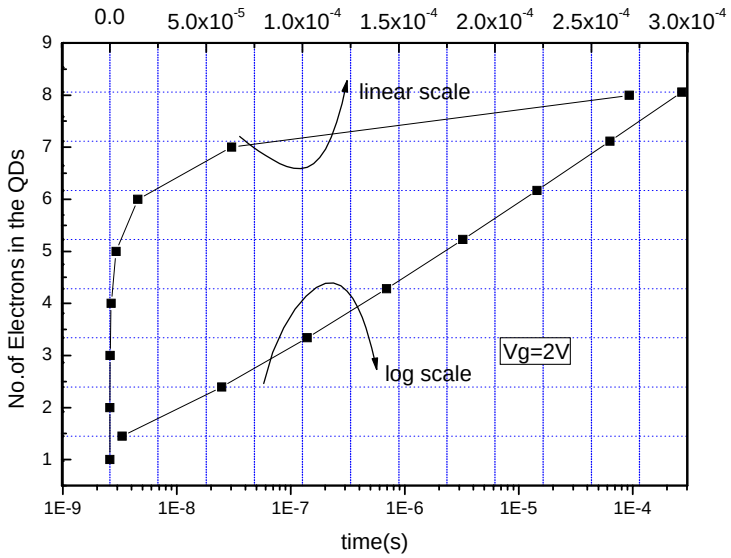


Fig.4.7 *Interpolation result: number of electrons in quantum dot as a function of programming time for V_g=2V.*

Both in Fig.4.7 and in Fig. 4.8, it is easily seen that more the electrons in the quantum dot, the longer it takes to add an extra electron into the quantum dot. It shows a good agreement with Rana's result. For the first 6 electrons, the programming time between two electrons is very close and the reason for this is believed to be due to computational uncertainty when fitting the distribution function. The device parameters used in this device are similar to those in section 4.2.

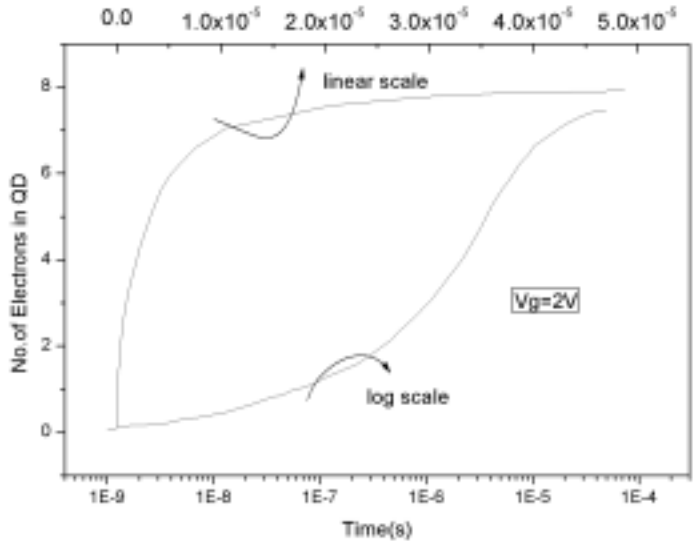


Fig.4.8 *Number of electrons in quantum dot as a function of time at V_g=2V*[23].

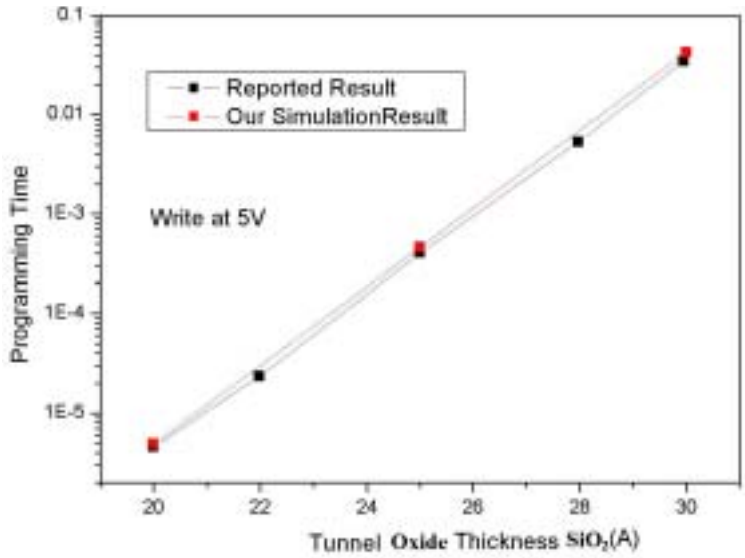


Fig. 4.9 *The programming time at 5V as a function of tunnel oxide thickness, compared with published data*[44].

Fig.4.9 shows the programming time as a function of the tunnel oxide thickness when programming voltage is 5 V. The result is calculated by simulating the time-dependent tunneling current density discussed previously in section 3.4.1 and compared to the reference [44]. In this example, the Ge quantum dot has a diameter 5 nm and control oxide thickness is fixed at 5 nm. Our simulation result shows good agreement with the reported result.

4.5.2 Verification of the retention time

The verification of evaluation of the retention time is done by the comparison with reference [29]. The detail of numerical method used for calculating the retention time was discussed in section 3.4.2.

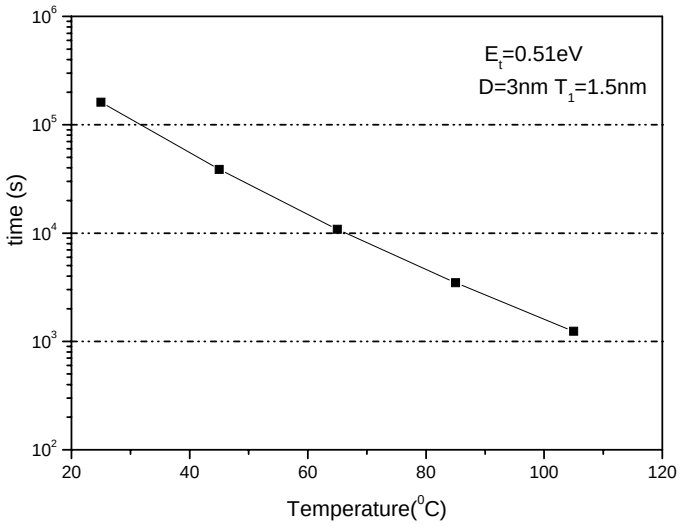


Fig.4.10 *Time as a function of temperature in the retention state.*

In Fig.4.10, the relationship between the retention time and temperature is shown. A Ge nanocrystal flash memory device is considered in this simulation, with the deep trap energy level E_t =0.51 eV and the geometry factor $\alpha = 9.08 \times 10^{-3}$ [44]. Our simulation result matches the result of the reference [44] well. For Ge nanocrystal flash memory, with the increase of temperature, the retention time becomes poor. The simulation model of the retention time calculation is verified.

Fig.4.11 presents a comparison of the retention time with various tunnel oxide thicknesses, considering the dot of diameter 5 nm, 3 nm and 2 nm respectively. The impact of the germanium nanocrystal size on the retention time is observed. As the graph shows, our result is close to the data in the reference [44]. The larger the diameter, the better retention time of the flash memory. A slight difference is considered acceptable due to the use of different values of some parameters in our simulation which are not given in the reported reference.

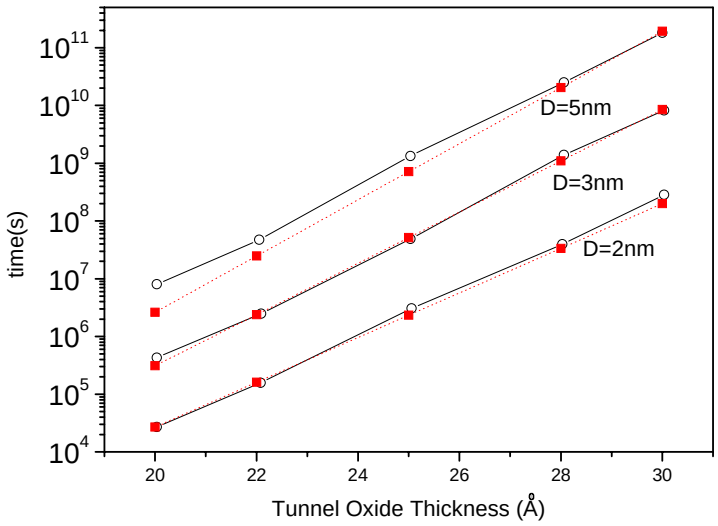


Fig.4.11 *Retention as a function of tunnel oxide thickness (red line means the published data and black line means our simulation result).*

4.6 Summary

This chapter presents the verification of some important simulation results. We perform the comparison between our results and the published results. The charging process, tunneling current (SiO2 and high-k dielectrics) and the programming/retention times are verified and their good agreements with published data are shown. The reasons for some slight differences between our simulation results and reported results are given. Through the verification work, the physical model that will be used in the subsequent chapters for various study and investigation is demonstrated to be sufficiently reliable and accurate.

Chapter 5
Simulation of Quantum Dot Floating Gate Flash Memory with SiO_2 Tunnel Oxide

5.1 Introduction

Nanocrystal flash memory was first introduced in the mid nineties [5], in which distributed charge storage elements are embedded between the tunnel oxide and the control oxide. The discrete, mutually isolated, crystalline nanocrystals or dots, typically made of semiconductor materials, can replace the conventional continuous floating gate. Compared to the conventional flash memory, nanocrystal charge storage offers several advantages，the main one being the potential to use thinner tunnel oxide without sacrificing nonvolatility as discussed in Chapter 2.2.2. It is a quite attractive property which provides a faster programming/erasing and longer retention time.

The nanocrystal flash memory has been widely investigated both experimentally and theoretically. However, in the theoretical aspect, many papers focus on only one or two characteristics, such as charging and discharging process [13], or programming and retention times [46]. It is necessary to give a comprehensive simulation to investigate a number of main characteristics of the nanocrystal flash memory. Therefore, in this chapter, using the physical model discussed in chapter 3, the main features of the Si nanocrystal flash memory with SiO_2 are explored and studied. The memory device

with alternative tunneling dielectrics, such as high-k materials, will be discussed in the next chapters.

In this chapter, section 5.2 gives a brief introduction to the simulator, named nanoFM-1.0. Section 5.3 focuses on the charge processing of the floating gate flash memory. Section 5.4 discusses the tunneling current through the tunnel oxide in the flash memory. Section 5.5 evaluates the programming and retention times. Section 5.6 gives a summary.

5.2 The Simulator nanoFM-1.0

The simulator nanoFM-1.0 is developed and its flowchart is shown in Fig.5.1. In this flowchart, firstly, a self-consistency method is used for solving potential profile and electrons distribution from the coupling Schrödinger and Poisson equations. Secondly, based on the potential and electrons density calculated, the tunneling current is evaluated by using a modified WKB approximation model. Thirdly, the programming and retention times are calculated. Finally, the simulator plots and outputs the results.

There are 17 routines in the nanoFM-1.0, which are implemented in Matlab 6.1. The average run time on PC for simulating a flash memory device ranges from 45mins~1.5hrs, which depends on the size of the flash memory device. In order to ensure the accuracy of the mode-space method, the channel thickness has to be less

than 5nm. It is found that an appropriate selection of the grid size according to different sizes of the devices is very important for achieving an efficient convergence.

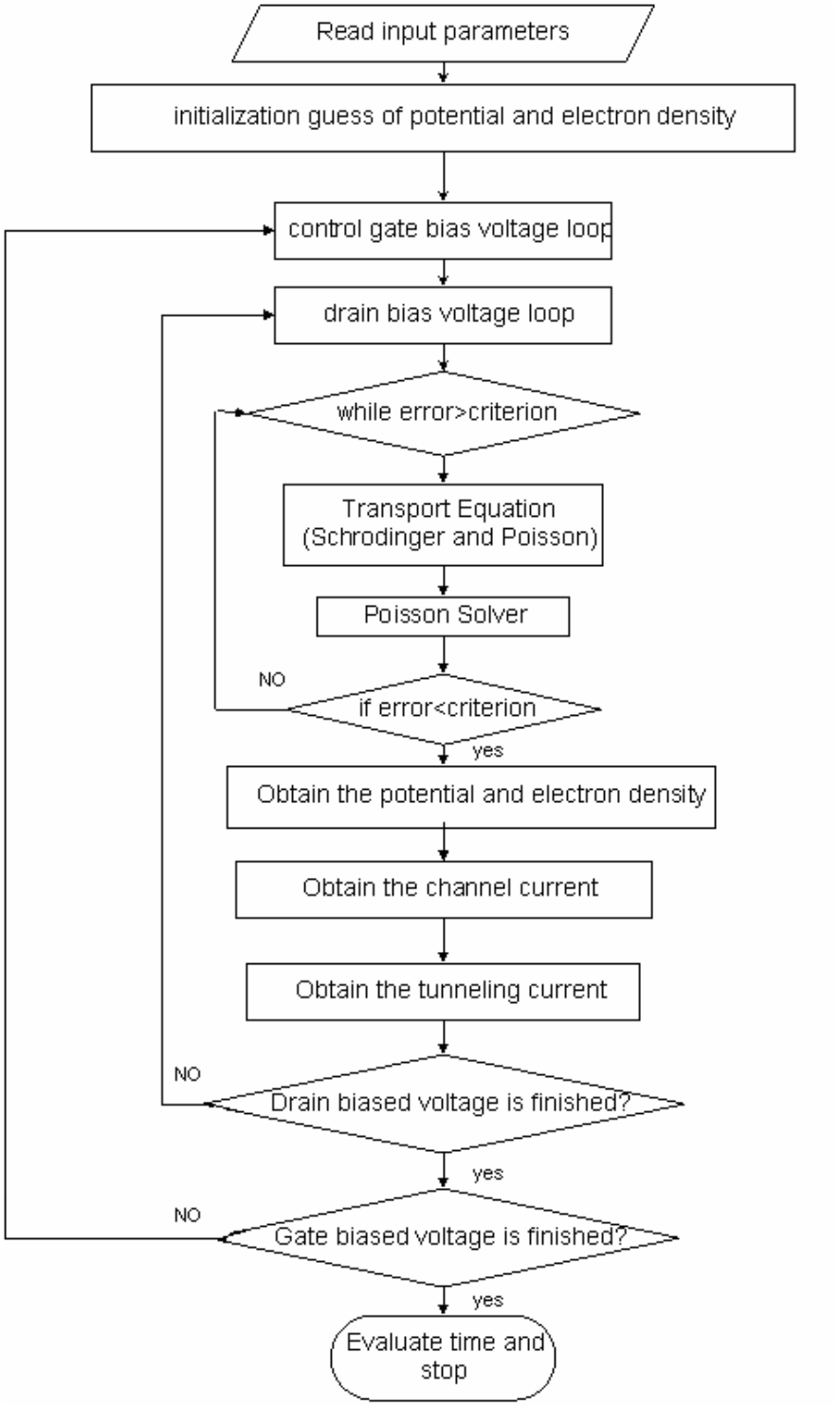


Fig.5.1 *The flowchart of nanoFM-1.0.*

5.3 The Charging Process of the Flash Memory Device

The device structure is illustrated in the Fig. 5.2. One silicon quantum dot as the floating gate is embedded between the control oxide and tunnel oxide. The control gate is polysilicon and the insulator is silicon dioxide. The dimensions of the dot are $6nm \times 10nm \times 10nm$. The length of the channel is 40 nm. The control oxide is fixed at 5 nm or 7 nm which will be indicated in results. The tunnel oxide thickness ranges from 1.5 nm to 4.5 nm in the simulation. The substrate doping density is assumed to be $5 \times 10^{17} cm^{-3}$.

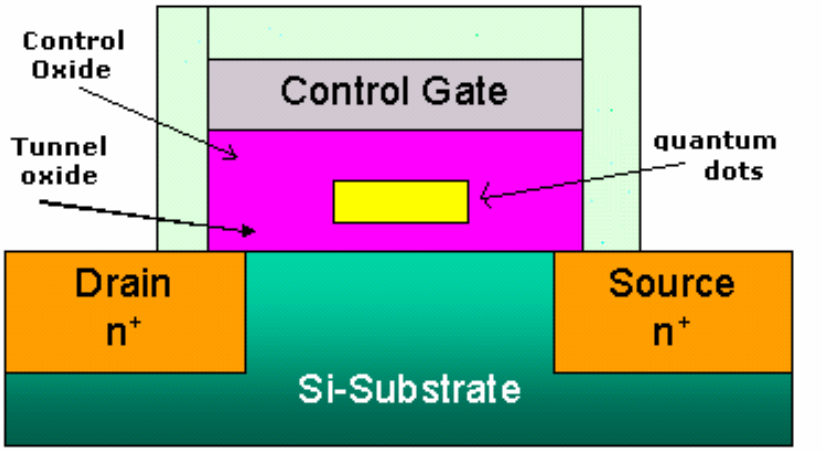


Fig.5.2 *The cross-section of the flash memory device.*

When a positive voltage is applied on the control gate and a very low positive voltage is applied on drain contract, the electrons will tunnel into the floating gate through the tunnel oxide from the silicon substrate. It is called charging processing or the programming. Fig.5.3 shows 2D electrons distribution of the memory device when control gate voltage is 0 V, 1.3 V, 1.9 V and 3 V. This model assumes a quantum dot

with dimension of $6nm \times 10nm \times 10nm$, control oxide thickness of 5nm, tunnel oxide thickness of 1.5nm and substrate doping of 10^{23} m^{-3}. It is obvious that with the increase of the control gate voltage, more and more electrons are trapped into the floating gate, and therefore more electrons appear in the region of the floating gate. During the charging process, the electrons distribution is symmetrical in the floating gate region and in the source/drain/silicon substrate regions.

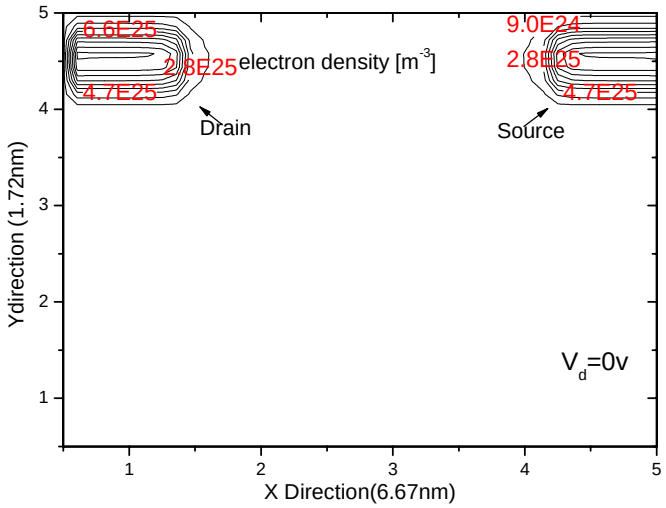


Fig. 5.3(a) *2D Electrons distribution of the flash memory at V_e=0V.*

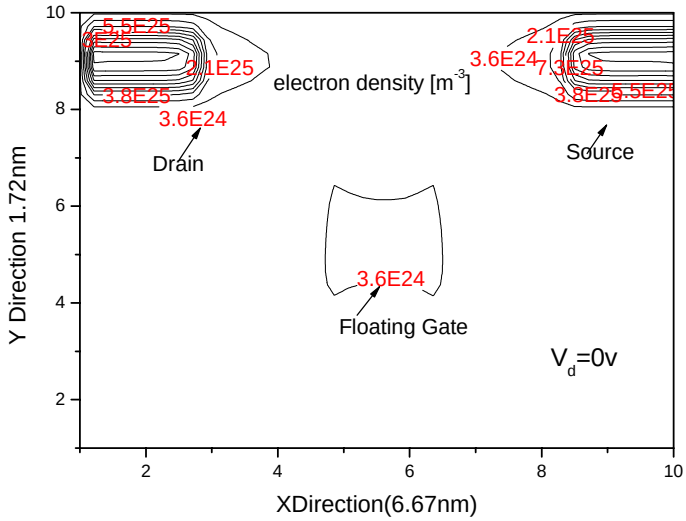


Fig. 5.3(b) *2D Electrons distribution of the flash memory at V_g=1.3V.*

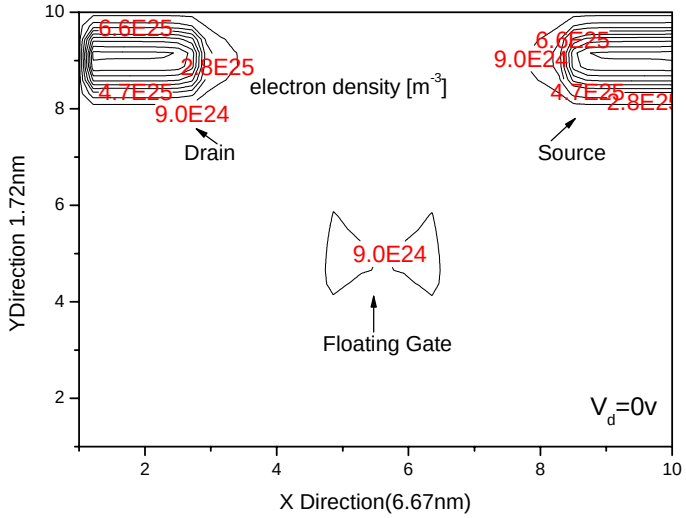


Fig.5.3(c) *2D Electrons distribution of the flash memory at V_e=1.9V.*

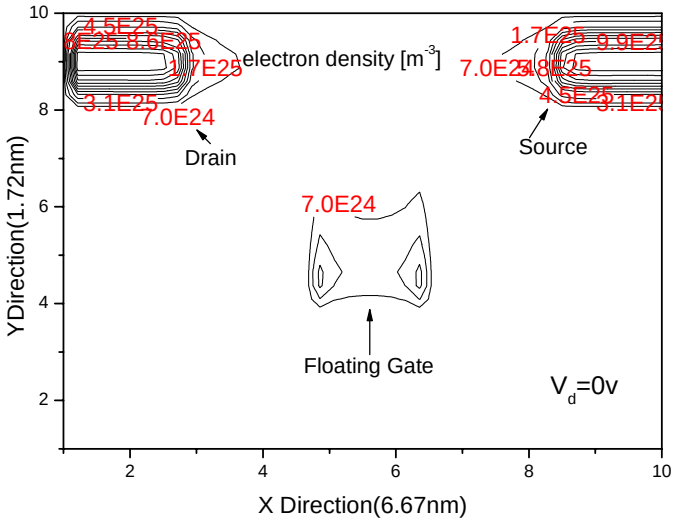
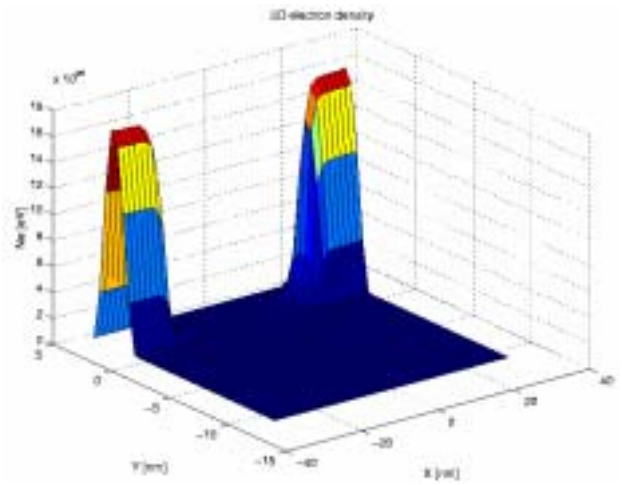


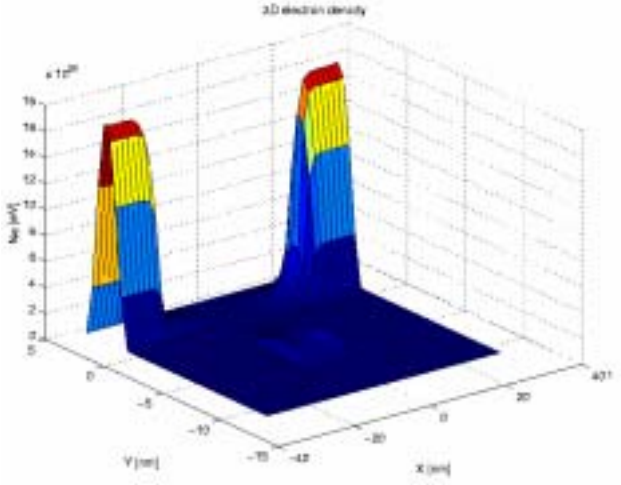
Fig.5.3 (d) *2D Electrons distribution of the flash memory at V_e=3V.*

For a better representation, Fig.5.4 (a)-(d) shows the 3D plots of 2D electrons distribution of the memory device system as a function of control gate voltage. Drain voltage is kept at 0 V. The doping density of silicon substrate is $10^{20}m^{-3}$ and the source/drain doping density is $10^{23}m^{-3}$. As same as shown in Fig.5.3 (a)-(d) with the increase of the control gate voltage, more and more electrons get trapped into the

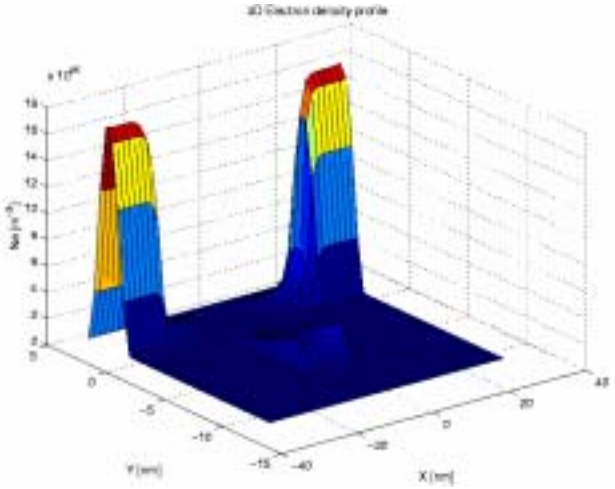
quantum dot floating gate.



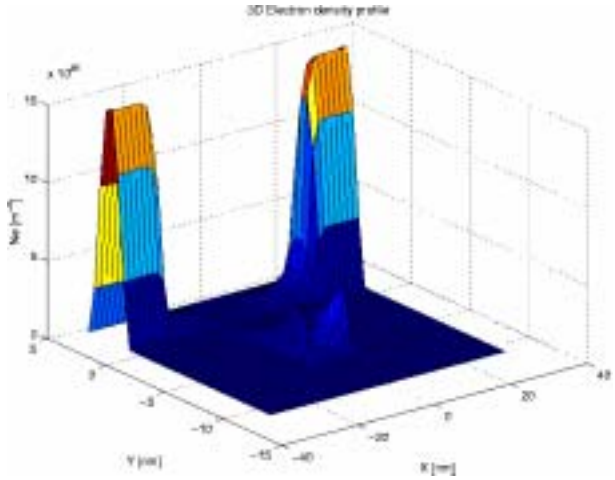
(a)V_g=0V



(b)V_g=1.3V



(c)V_g=1.9V



(d) V_g =3V

Fig.5.4 *3D Electron density distribution of the flash memory.*

In order to give a more detailed analysis for the charging behavior of the floating gate, the interaction between the charging behavior of the floating gate and the channel as a function of the control gate voltage is shown in Fig.5.5. The simulation assumes a

substrate doping density of $1.5\times10^{24}\,m^{-3}$. The number of electrons is calculated from the electron density in each grid and therefore it is not constrained to an integer. That is why Fig. 5.5 shows non-integral number of charges in the quantum dot. As number of electrons in the channel builds up, the rate of the increase of electrons in the quantum dot slows down. This is due to the increasing amount of electrostatic gate field energy required to sustain the inversion charge in the channel at the expense of the charging of the floating gate. The rate of the increase of number of electrons in the channel tends to saturate after 0.5 V. It is widely accepted that when more and more electrons in the channel tunnel into the floating gate, it will results in the saturation of the number of electrons in the channel.

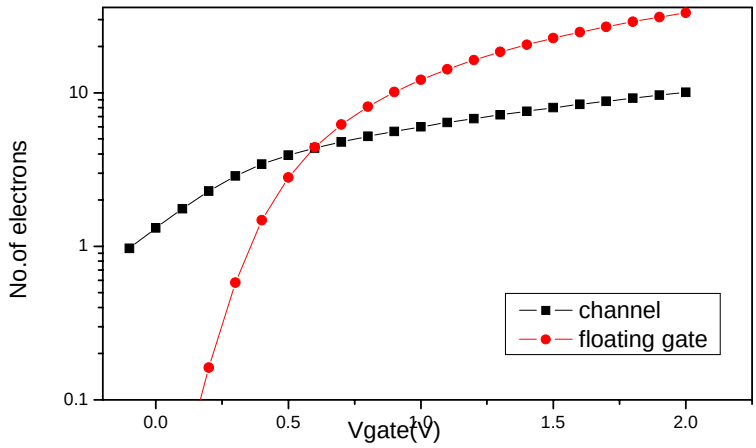


Fig.5.5 *Number of electrons in the channel and floating gate.*

Fig.5.6 plots the drain current as a function of control gate voltage for a number of electrons in the quantum dot ranging from 0 to 10. The number of electrons is constrained to an integer number during the simulation. The threshold voltage shift is obvious. When the quantum dot has 6 six electrons, the total threshold voltage shift is

0.3 V. When 10 electrons are in the quantum dot, the total threshold voltage shift is about 1 V. The result indicates the charging sensitivity of the device to a field modification by every additional electron in the quantum dot. It shows, in effect, the single electron charging phenomenon, and means Coulomb blockade effect is simulated approximately.

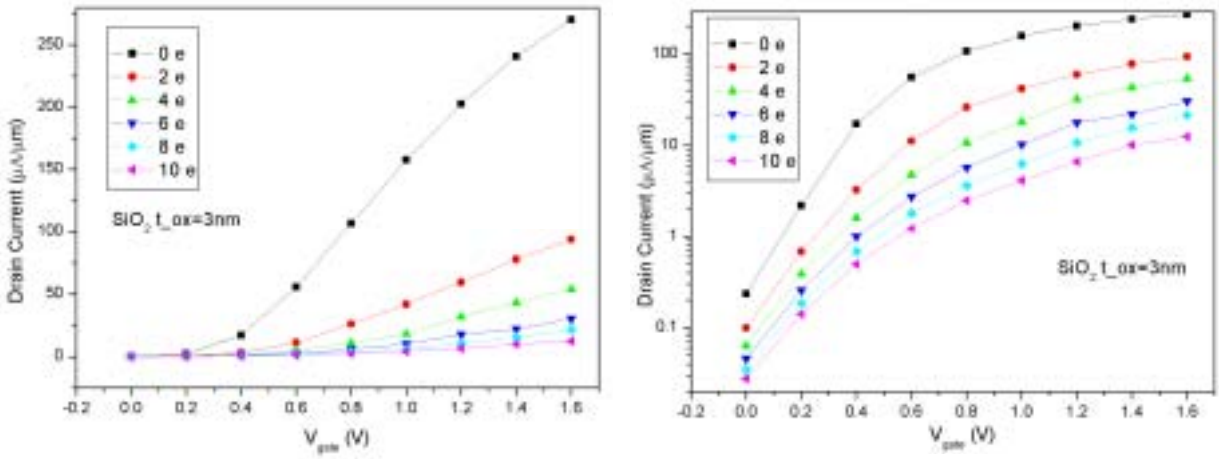


Fig. 5.6 *Drain current as a function of control gate voltage by keeping fixed number of electrons in the quantum dot (a) linear scale (b) log scale.*

5.4 The Tunneling Current through the Tunnel Oxide

In the flash memory, the tunneling current is dominated by direct tunneling and F-N tunneling mechanisms. At a low voltage, a large amount of current can pass through the tunnel oxide, as shown in Fig. 5.7. In the calculation, the conduction band offset between the silicon quantum dot and silicon substrate is fixed at 3.15 eV. When tunnel oxide thicknesses are 1.5 nm, 2.0 nm and 2.5nm, large tunneling currents at relatively low voltages (< 2.0V) are observed. The tunneling current is very sensitive to the tunnel oxide thickness. A 0.5 nm difference between tunnel oxide thicknesses results

in 2 orders of magnitudes difference in the tunneling current. With the increase of the control gate voltage, the rate of the increase of the tunneling current slows down. This is because when lower energy states in the quantum dot are occupied by electrons, it is more difficult for following electrons to occupy higher energy states in the quantum dot. Therefore, the probability of electrons tunneling into the quantum dot becomes smaller and thus the rate of increase of the tunneling current decreases despite of increase in electric field. In this result, the effective mass is assumed to be $0.51m_0$ for SiO_2.

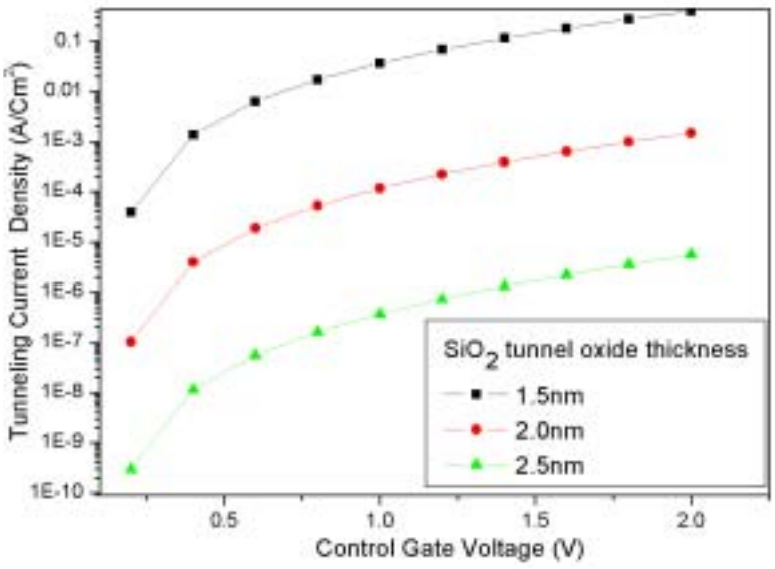


Fig.5.7 *Tunneling current as a function of control voltage.*

The tunneling currents with various tunnel oxide thicknesses at different control gate voltages are showed in Fig. 5.8. A small difference in tunnel oxide thicknesses results in a large variation of the tunneling current. This attribute makes thin tunnel oxide a very attractive candidate for achieving fast programming/erasing time at a low

operation voltage. In Fig. 5.8, compared to the tunnel oxide thickness, control gate voltage has less effect on the tunneling current at low voltage operation. However, when the tunnel oxide thickness increases from 1.5 nm to 4.5 nm, the impact of the control gate voltage on the tunneling current becomes significant, which is believed to be due to the effect of F-N tunneling under a higher control gate voltage 2.0 V.

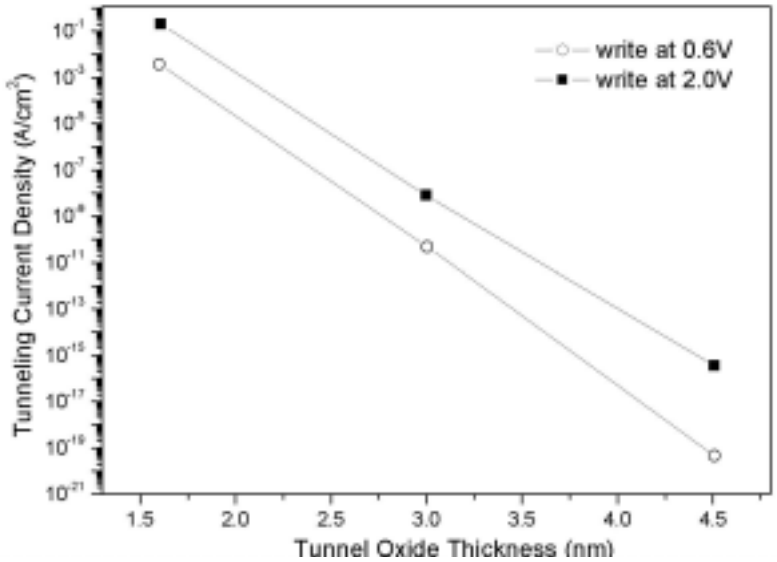


Fig.5.8 *Tunneling current as a function of tunnel oxide thickness.*

5.5 The Programming and Retention Times

Table 5.1 summaries the performance parameters defining state-of-art semiconductor memory devices. For the flash memory, an ideal device provides 10 years retention standard and 1 μs~1ms programming/erasing time. The operation voltage is less than 5V. A good memory device should have faster programming/erasing operation, longer retention and lower power operation ability. Hence, the programming/retention times

are important parameters for evaluating the device performance. In this section, the programming/retention times are quantified and their characteristics are evaluated.

Device	Operation Voltage	Write/Erase Time	Data Retention Time	Endurance
DRAM	3V	50~100ns	0.1~0.5 sec	No limit
EEPROM	-8V~5V	1 μs~1ms	10 years	10^5 cycles

Table 5.1 *Device parameters for different semiconductor memories. Each is optimized for either dynamic or non-volatile application.*

Fig.5.9 shows the number of electrons in the floating gate as a function of programming time when control gate voltage is 2V. The more the electrons in the floating gate, the longer it takes to add an extra electron into the floating gate. There are several reasons for explaining this general feature.

The first few electrons tunneling into the quantum dot will enter at lower energy states in the quantum dot with higher occupation probability, and therefore the following electrons have to occupy higher energy states with lower probability. Secondly, the presence of electrons in the quantum dot changes the threshold voltage of the device, and thus results in less electrons available in the channel which can be trapped into the quantum dot. Thirdly, because in the initial state there is no electrons in the quantum dot, therefore the control gate potential across the tunnel oxide is large, resulting in a large coupling constant. With the injecting of electrons in the floating gate, the potential drop across the tunnel oxide becomes smaller with a corresponding

reduction in the coupling constant. As a result, fewer electrons can tunnel into the quantum dot.

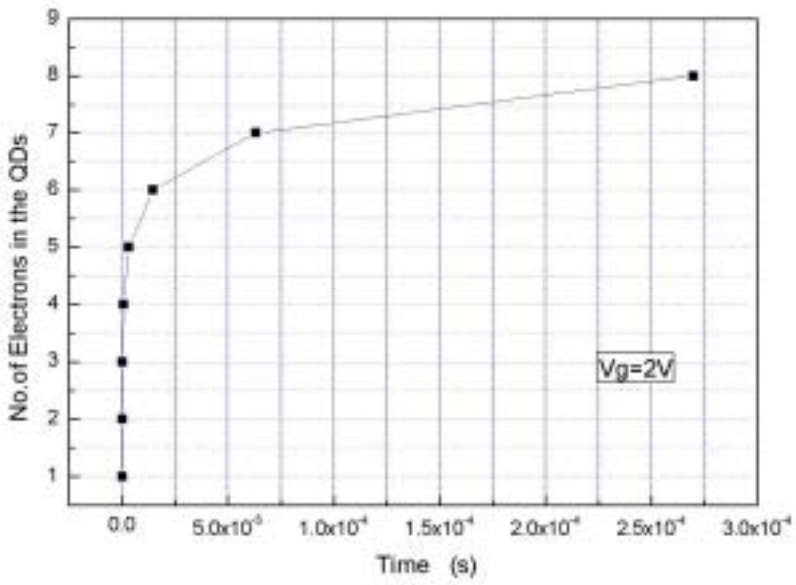


Fig.5.9 *The evolution of mean number of electrons in a Si quantum dot when control gate voltage is 2V.*

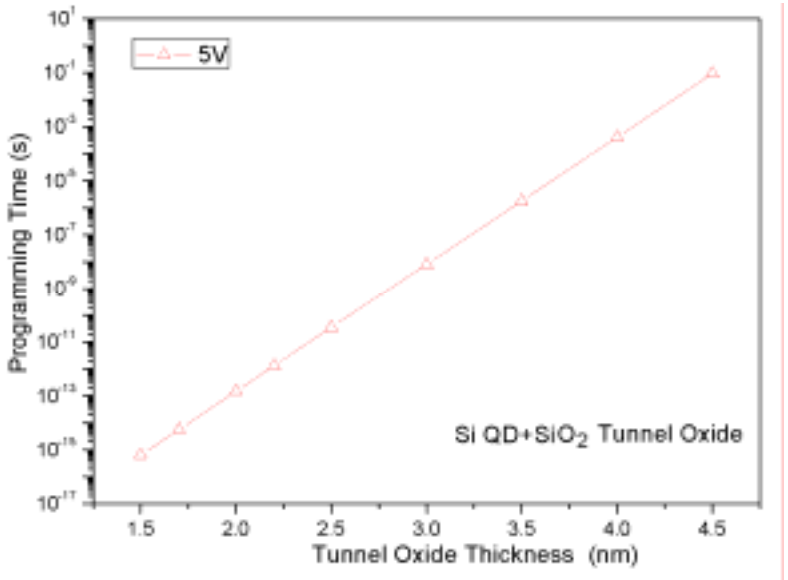


Fig.5.10 *Programming time as a function of the tunnel oxide thickness.*

The programming characteristic in terms of tunnel oxide thickness when low programming voltage of 5V is applied on the control gate is shown in Fig.5.10. With a

fixed control oxide thickness of 7nm, the programming time has reached nano-second range when tunnel oxide thickness is less than 2.75 nm. The programming time increases drastically with the increase of the tunnel oxide thickness. When tunnel oxide thickness is about 2.75 nm, the programming time can reach nanoseconds. It demonstrates again that the thin tunnel oxide is a very attractive candidate for achieving fast programming time at low operation voltage.

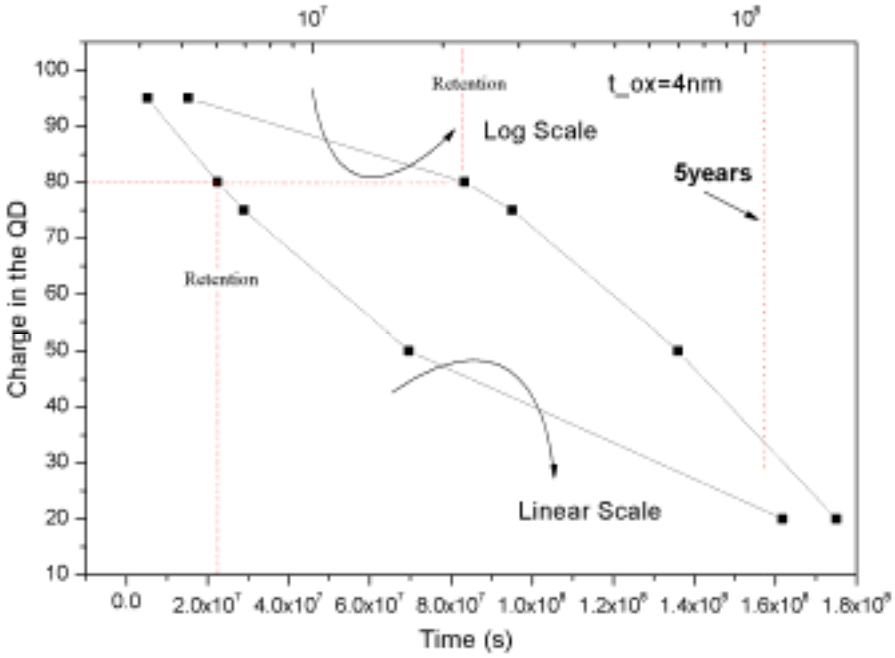


Fig.5.11 *The charge in the quantum dot as a function of time in the retention state.*

Fig. 5.11 represents the retention time vs the stored charge in the quantum dot with the tunnel oxide of thickness 4 nm. The dot line represents 5 years retention time, which is the half of the current requirement for nonvolatile flash memories. It shows that after 2.24×10^7 s, 20% charge in the quantum dot is lost which falls substantially short of 10 years standard. It indicates that the flash memory device with tunnel oxide thickness of 4 nm is not with thick enough SiO_2 for achieving 10 years retention

standard.

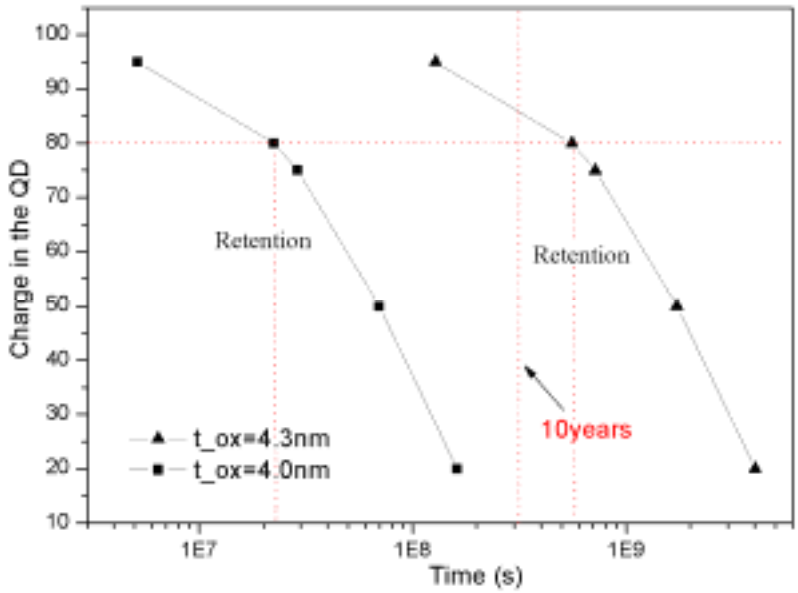


Fig.5.12 *The charge in the quantum dot as a function of time with different tunnel oxide thicknesses in the retention state.*

In Fig. 5.12, we show that the retention time of the memory device with different tunnel oxide thicknesses, 4.0 nm and 4.3 nm respectively. When 20% stored charge is lost, the device with tunnel oxide thickness 4.3 nm can reach 10 years retention time which is similar to the requirement mentioned in ITRS2003. The figure also indicates that the retention time is very sensitive to the tunnel oxide thickness. That is why high k materials with thicker physical thickness and thinner EOT will have a better potential to provide fast programming and longer retention time.

Figure 5.13 shows clearly the tradeoff between the retention time and programming time when control gate voltage is 5 V. When tunnel oxide thickness is 4.3 nm which can provide 10 years retention time, the programming time is 0.0109 *s* that can not reach nanoseconds range. When programming time reaches nanoseconds regime with

tunnel oxide thickness 2.82 nm, the retention time is 1×10^{2} *s* which is unacceptable and extraordinarily. Since both of fast programming time and good retention performance are desirable attributes for the flash memory, the alternative materials for tunnel oxide and the quantum dot are sought to improve the programming and retention characteristics simultaneously.

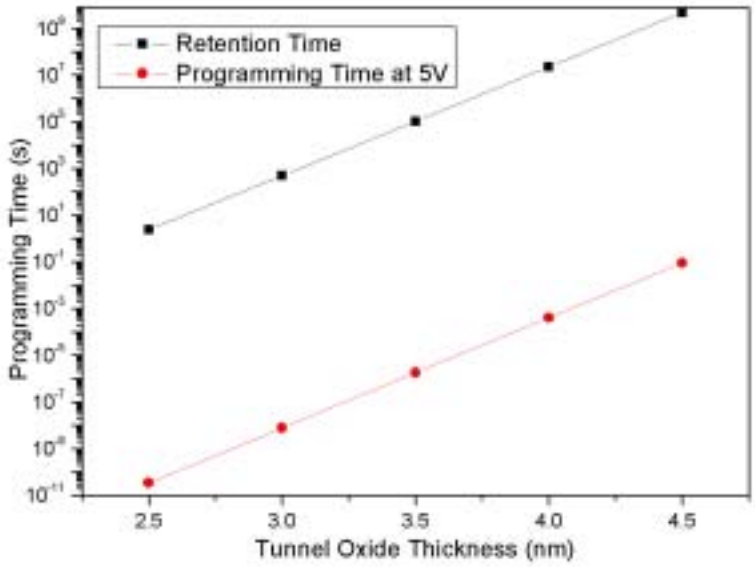


Fig.5.13 *Tradeoff between retention time and programming time as a function of tunnel oxide thickness.*

5.6 Summary

In this chapter, memory characteristics of a quantum dot floating gate structure are predicted by the theoretical model. The charging process and its impact on the memory device are studied. The interaction of the charging behavior between the quantum dot and the channel is discussed. The tunneling currents are calculated by a modified WKB approximation, including direct tunneling and F-N tunneling

mechanism. The impact of the tunnel oxide thickness on the tunneling current is studied and the result demonstrates the importance of tunnel oxide in improving programming efficiency. The programming characteristics which show quantum confinement phenomena are predicted by tunneling currents. The programming/retention times are investigated and used to examine the tradeoff between the programming and retention performance. By adjusting the tunnel oxide thickness, an ideal quasi-nonvolatile memory with high programming speed and longer retention can be achieved. The results and predictions in these chapters are essential for the design and further optimization of the flash memory device at low voltage operation. The Si quantum dot flash memory with silicon dioxide thickness 4.3 nm can reach 10 years retention standard while its programming speed is 0.0109 *s* at 5V, which is not good. It indicates less possibility of low voltage operation which is essential in future flash memory devices. Therefore, for further optimization of the flash memory device, new materials for both tunnel oxide and quantum dots are proposed and the simulation results will be discussed in the following chapters.

Chapter 6
Memory Device with High-k Dielectrics

6.1 Introduction

The ideal goal of the flash memory device is to make the programming speed as fast as possible and achieve a long retention time. However, because of the continuous scaling of the tunnel oxide thickness to achieve programming speed at lower voltage, the charge loss between the floating gate and silicon substrate prevents the further improvement of the retention time. As a result, requirements for fast programming/erasing time of the flash memory at low voltage are in direct conflict with the necessary long retention time, because both of them depend on the tunneling oxide thickness.

Recently, in order to overcome this problem, high-k dielectrics that can provide high dielectric constant and low electron barrier height are proposed for replacing conventional silicon dioxide in the flash memory [20, 48]. During the programming/erasing modes, the high k dielectric with low barrier height can provide fast and efficient programming/erasing operation and can have thicker tunneling dielectric with small EOT compared to SiO_2. During the retention mode, due to having thick tunnel oxide, the leakage current of high-k dielectric film is several orders of magnitude smaller than that of SiO_2 due to larger physical thickness

compared to SiO_2, thus resulting in superior data retention time [36]. Therefore, using the high-k dielectric, both longer retention time and efficient programming/erasing can be achieved.

Many experiments have demonstrated that both the low voltage operation and good retention time could be achieved by using high k dielectrics [33-35]. Although the performance of the nanocrystal flash memory with high-k dielectric has been extensively investigated in experiments, simulation and modeling are also very important and useful for guiding the design and fabrication of the nanocrystal flash memory with high-k dielectrics. Therefore, in this chapter, the flash memory with high-k dielectric is studied. We investigate characteristics of the nanocrystal flash memory with HfO_2 and HfAlO alternative dielectrics, respectively. Their performances are compared and contrasted with SiO_2 flash memory.

The properties of high-k dielectrics for memory application are presented in section 6.2. In section 6.3, the device characteristics of the flash memory with high-k dielectrics are explored and discussed. The tunneling current through the high-k dielectric is studied. The programming and retention times of the high-k dielectric flash memory are investigated. The advantages of faster programming/erasing time and longer retention of high-k dielectric flash memory are demonstrated theoretically. Finally, in section 6.4, a summary of this chapter is given.

6.2 High k dielectrics

The concept of the equivalent oxide thickness is introduced in this part. Considering a parallel plate capacitor $C = k\varepsilon_0 A/t$, where k is the dielectric constant, ε_0 is the permittivity of free space, A is the area of the capacitor, and t is the thickness of the dielectric. The expression of this parallel plate capacitor can also be shown in terms of t_{eq} (equivalent oxide thickness) and k_{ox} (dielectric constant) of the capacitor. t_{eq} represents the theoretical oxide thickness which can achieve the same capacitance density as the dielectric. The physical thickness of an alternative dielectric employed to achieve the equivalent capacitance density of t_{eq} can be obtained by $t_{eq}/k_{ox} = t_{high-k}/k_{high-k}$. Therefore, a high-k dielectric with a relative permittivity of 16 can afford a physical thickness of $\sim 40\overset{o}{A}$ to obtain $t_{eq} = 10\overset{o}{A}$. As a consequence, larger physical thickness of high-k dielectric provides longer retention, and its thin equivalent oxide thickness with low barrier height affords F-N tunneling at very low voltage, resulting in fast and efficient programming/erasing states.

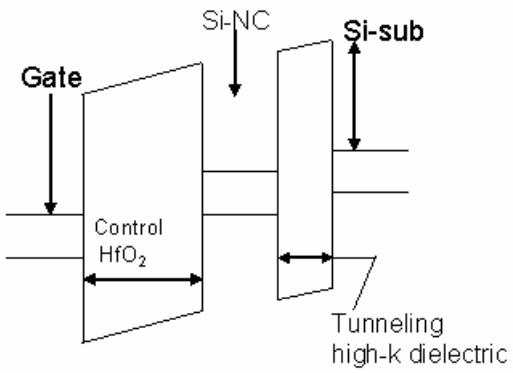
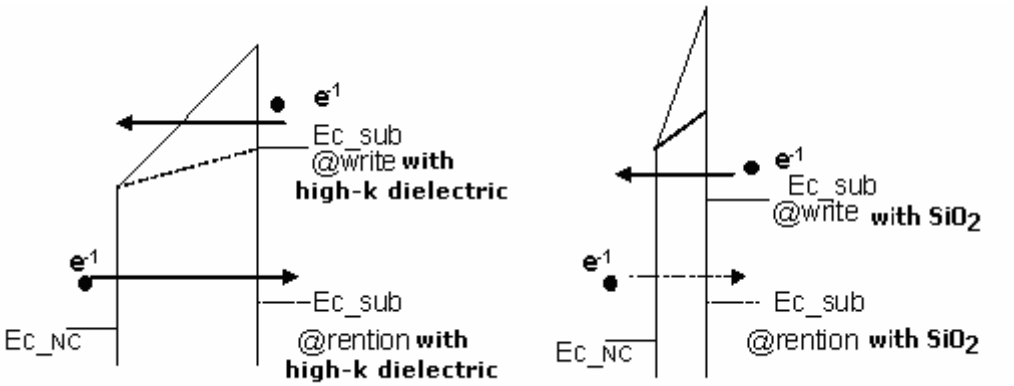


Fig.6.1 *Energy band diagram of silicon nanocrystal memory with high-k at equilibrium and enlarged conduction band edge profile at programming mode.*



(a) *(b)*

Fig.6.2 *(a) Enhanced electron injection by F-N tunneling in high-k dielectrics (b) Direct electron tunneling in SiO_2. Dashed line indicates conduction band edge profile at retention.*

In order to understand the roles of the electron barrier height and dielectric thickness in device operation, the conduction band edge profiles of silicon nanocrystal/tunneling dielectric/silicon substrate with 4.5nm HfO_2 (EOT 1.6nm) are illustrated in Fig. 6.1 and Fig. 6.2. The band bending of SiO_2 tunneling barrier is such that the whole dielectric acts as an insulator both in programming and retention states, as shown in Fig.6.2 (b). However, for high-k dielectric, the energy band is not symmetric in the programming and retention states, as shown in Fig.6.2 (a). Due to its lower electron barrier height, only part of the high-k dielectric offers a barrier to carrier flow in programming mode and the charge transport mechanism is changed from direct tunneling into F-N tunneling [36], as shown in Fig.6.2. (a). This gives rise to enhance electron injection from substrate to silicon nanocrystal, resulting in more efficient programming. During the retention, because of the use of larger physical thickness, the high-k dielectric substantially reduces the leakage current between the quantum dot and silicon substrate and thus enables the better retention in Fig.6.2 (a).

Until now, many advanced high-k dielectrics have been employed and their most relevant properties are summarized in Table.6.1. The most commonly studied high-k gate dielectric candidates have been materials systems such as HfO_2 , Al_2O_3 and Ta_2O_5 , which have dielectric constant ranging from 10-80.

material	dielectric constant(κ)	band gap *Eg* (eV)	barrier height(eV)
SiO_2	3.9	9	3.5
Si_3N_4	7	5.1	2
HfO_2	25	5.7	1.5
Al_2O_3	9	8.7	2.8
Y_2O_3	15	5.6	2.3
La_2O3	30	4.3	2.3
TiO_2	80	3.5	1.2
ZrO_2	25	7.8	1.4

Table 6.1 *The main parameters of various high-k dielectrics.*

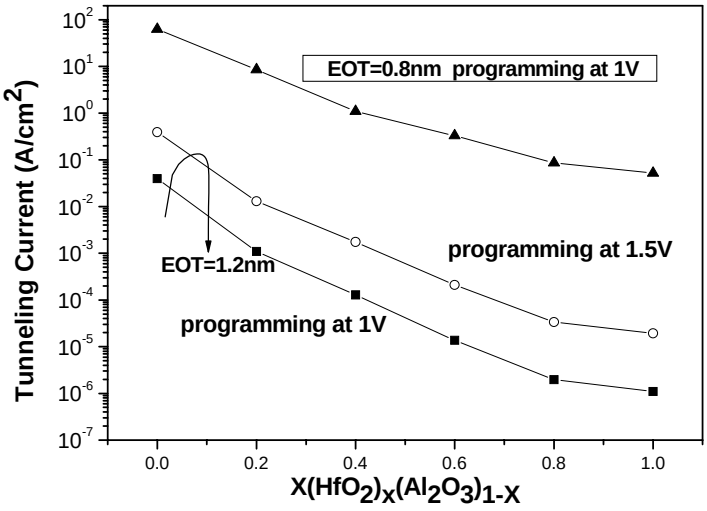


Fig.6.3 *Tunneling current of $(HfO_2)_x(Al_2O_3)_{1-x}$ for various Hf compositions.*

In experiments, because HfO_2 is not compatible with the high temperature processing, typically used for source-drain post implant anneal in any standard CMOS process, a potential solution that adds Al into HfO_2 to form HfAlO is proposed [45]. For HfAlO, the band offset data are determined from the XPS (X-ray Photoelectron Spectroscopy)

experiments. Their dependences on the Hf composition are demonstrated to be in a linear relationship [40]. The electron effective mass and the dielectric constant of HfAlO are interpolated linearly between those of HfO_2 and Al_2O_3. The tunneling current through $(HfO_2)_x (Al_2O_3)_{1-x}$ is calculated by nanoFM-1.0 and showed in Fig.6.3. It is seen from Fig.6.3 that the higher the Al composition the higher tunneling current. In the following simulation results of $(HfO_2)_x (Al_2O_3)_{1-x}$, x=0.3 is selected and used to determine the relevant parameters.

6.3 Characteristics of the Flash Memory Device with High-k Dielectrics

6.3.1 Basic characteristics of flash memory with high-k dielectrics

In this simulation work, the size and structure of the memory device is similar to the device discussed in chapter 5(Fig. 5.1). The control gate thickness is fixed at 7 nm. The tunnel oxide thickness ranges from 1.5 nm to 4.5 nm according to different dielectrics. The length of the channel is 40 nm.

In Fig. 6.4, the gate current density as a function of control gate voltage with different dielectrics is shown. It is obvious that the gate currents of high-k materials (HfO_2 and HfAlO) are higher than that of SiO_2 with the same physical oxide thickness. Because high-k dielectrics have low barrier height, therefore, due to the F-N tunneling, they

can provide higher tunneling current. Since 2 eV barrier height difference between SiO_2 and HfO_2 results in near 12 orders of magnitude difference in tunneling current, it confirms that the barrier height has great impact on tunneling current.

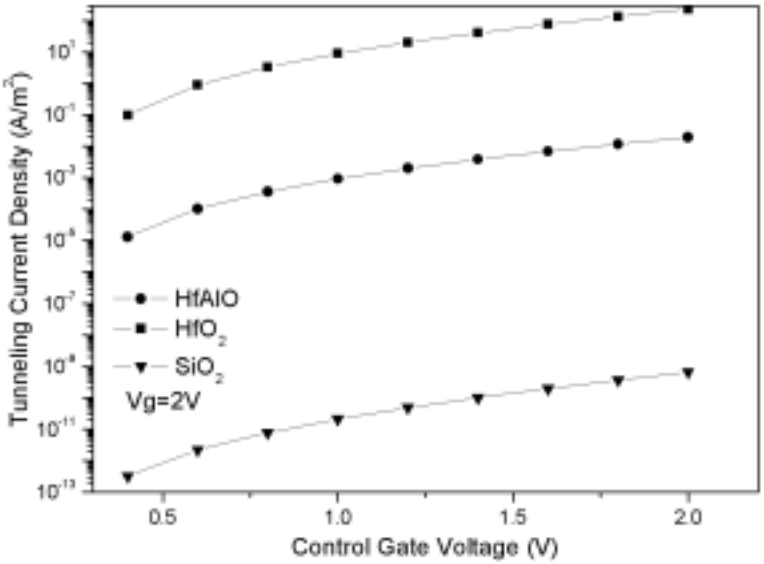


Fig.6.4 *Simulated tunneling current as a function of gate voltage with SiO_2, HfO2 and HfAlO dielectrics with t_ox=4.5nm.*

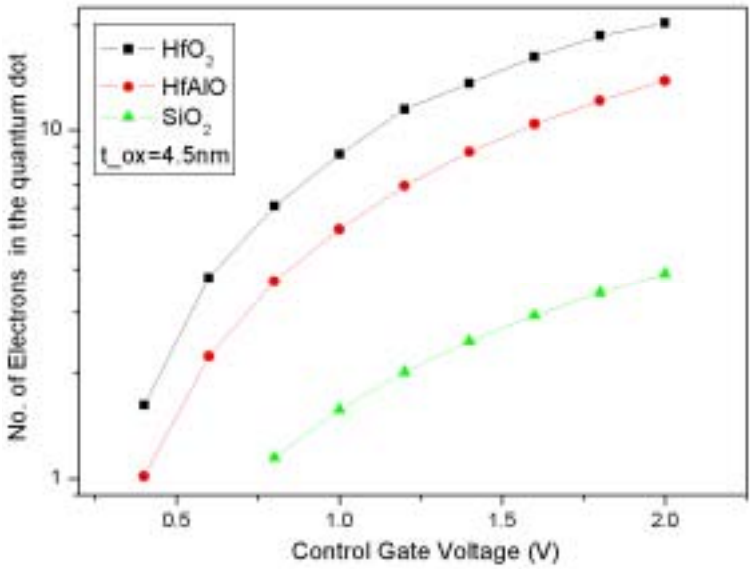


Fig.6.5 *Number of electrons in the quantum dot as a function of gate voltage with SiO_2, HfO2 and HfAlO dielectrics and t_ox=4.5nm.*

Fig. 6.5 shows the number of electrons in the quantum dots as a function of control

gate voltage with various dielectrics. Notice that the number of electrons in the quantum dots of high-k dielectric flash memory is more than that of with SiO_2 flash memory. It is explained that at the same programming voltage, due to low barrier height, more electrons are allowed to tunnel into the quantum dot easily. Therefore, the results demonstrate that low voltage and efficient programming mode can be provided by the high-k dielectric. As discussed in Chapter 5, because the number of electrons is calculated from electron density, so Fig 6.5 shows non-integral number of electrons in the quantum dot.

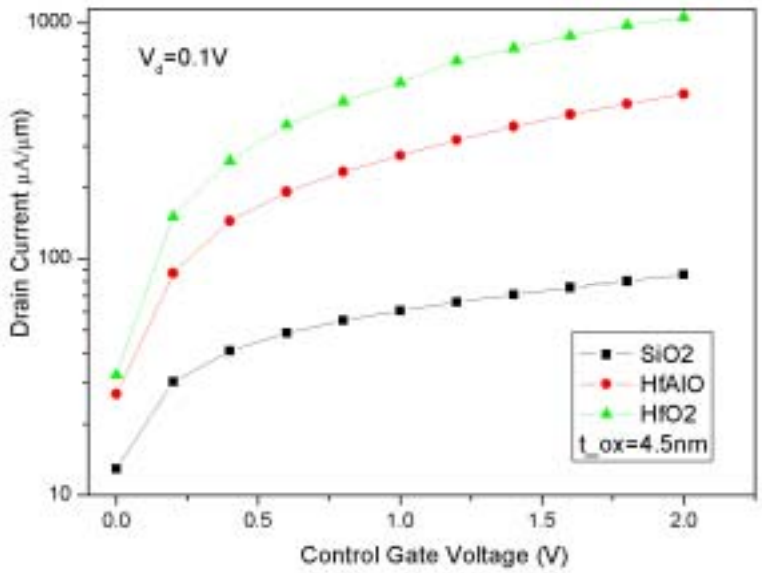


Fig.6.6 *Simulated drain current as a function of gate voltage with SiO_2, HfO2 and HfAlO dielectrics and t_ox=4.5nm.*

The channel currents with the different high-k dielectrics are given in Fig. 6.6. High-k dielectrics proved much larger oxide capacitance between the control gate and the floating gate compared to SiO_2. As a result, the channel currents of high-k dielectrics become higher.

Fig.6.7 plots the drain current as a function of control gate voltage for a number of additional electrons in the quantum dot ranging from 0 to 10 when HfO_2 is used. When the quantum dot contains 10 electrons, the threshold voltage shifts is about 0.2 V. It indicates that the charging sensitivity of the device channel to the charging of the quantum dot. As discussed in Chapter 5, this result also emulates single electron charging and the Coulomb Blockade effect.

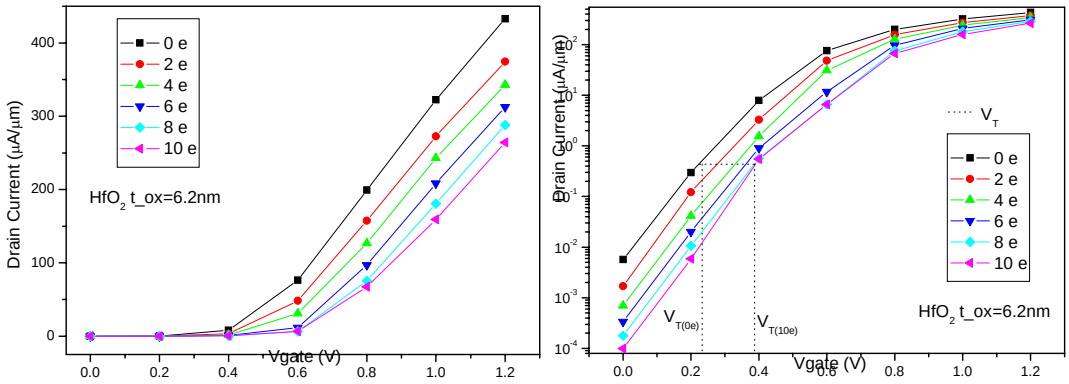


Fig. 6.7 *Drain current as a function of control gate voltage by keeping fixed number of electrons in the quantum dot (a) linear scale (b) log scale.*

6.3.2 Tunneling current of flash memory with high-k dielectrics

In Fig.6.8 and Fig.6.9, the gate currents as a function of dielectric thicknesses with different dielectrics at control voltage 0.6V and 2V are simulated. The gate current increases with the decrease of the dielectric thickness at a relative low operation voltage. With the same physical thickness, the tunneling current of the high-k dielectric is much higher than that of SiO_2, which shows agreement with Fig.6.4. The results show significantly that the impact of the programming voltage on the tunneling current is less than the impact of the tunnel oxide thickness, since the increase of the

control gate voltage doesn't result in as much difference of the tunneling current. Therefore, it is possible to implement efficient programming operation at the low voltage.

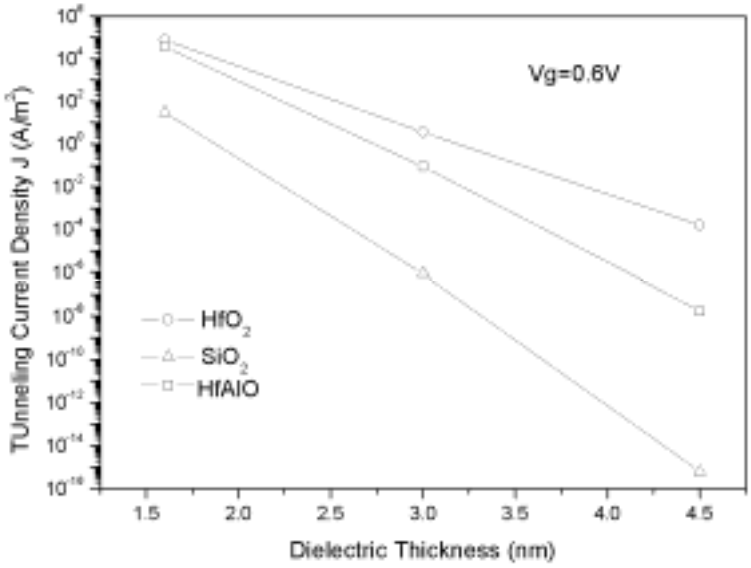


Fig.6.8 *Simulated tunneling current as a function of dielectric thickness with different high-k dielectrics at programming mode when control gate voltage is 0.6V.*

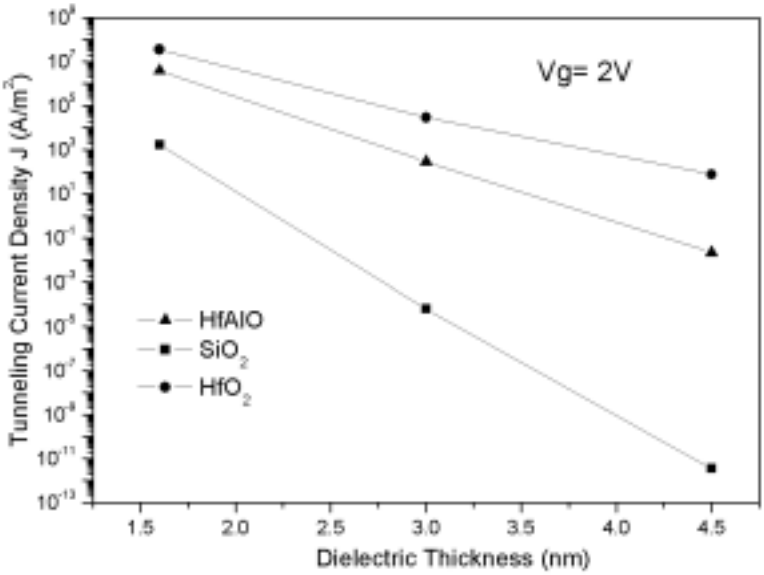


Fig.6.9 *Simulated tunneling current as a function of dielectric thickness with different high-k dielectrics at programming mode when control gate voltage is 2V.*

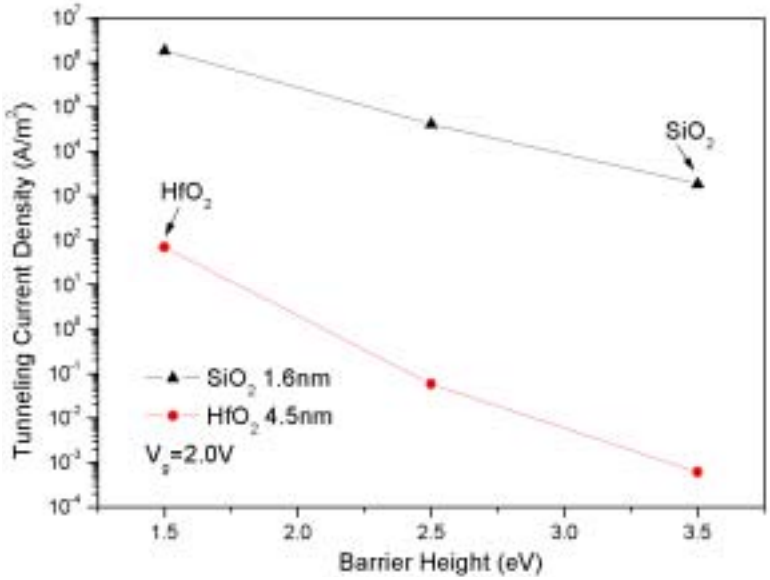


Fig.6.10 *Simulated tunneling current as a function of barrier height with different materials at programming mode.*

When we compare the Fig. 6.8 and Fig.6.9, notice that the impact of the tunnel oxide thickness of SiO_2 on tunneling current is more obvious than that of HfO_2 and HfAlO on the tunneling current. It is believed that the high tunneling current of high-k dielectrics during the programming mode is mainly provided by lower electron barrier height.

During the programming mode, due to the low barrier height, high-k dielectric can provide higher tunneling current as shown in Fig.6.4. The sensitivity of tunneling current to electron barrier height is observed in Fig. 6.10, which shows the simulated tunneling current as a function of barrier height when control gate voltage is 2 V. With the decrease of the barrier height of the high-k dielectrics, the tunneling current increases exponentially as expected. In this Fig. 6.10, the dielectric with lower barrier height and larger thickness (HfO_2) can achieve a comparable performance as the

dielectric with higher electron barrier height and smaller thickness (SiO_2) under programming mode. For example, in Fig. 6.10, the 4.5 nm HfO_2 (1.6nm EOT) with 1.5 eV electron barrier height has comparable programming time to the SiO_2 dielectric with barrier height 3.5 eV with thickness 1.6nm. Comparing the Fig. 6.4 with Fig. 6.10, that the 0.5 eV difference of barrier height results in more than 3 orders of magnitudes change of tunneling current demonstrates barrier height is a more important factor in providing higher tunneling current than tunnel oxide thickness under programming regime.

The permittivity of the high-k dielectric is measured on bulk samples and in some cases even on thin films, while, for the more complex dielectrics, the dielectric constant may not be as well known. Therefore, it is necessary to explore the relationship between the tunneling current and the high-k dielectric constant. The tunneling currents of HfO_2 and HfAlO as a function of high-k dielectric constant are plotted in Fig.6.11. The tunneling current decreases with the increases of dielectric constant, which can be demonstrated theoretically from the tunneling current calculation mode (WKB) (Eq. (3.10)) discussed in chapter 3. The result indicates the dielectric constant has less effect on the tunneling current, especially for high-k dielectrics, while its effect on programming performance is less compared to that of the effect of tunnel oxide thickness and barrier height. Because of the high dielectric constant, the physical thickness of the high-k dielectric enables thin EOT, which leads to efficient programming. Under this case, the dielectric constant is a very important factor in designing good memory device.

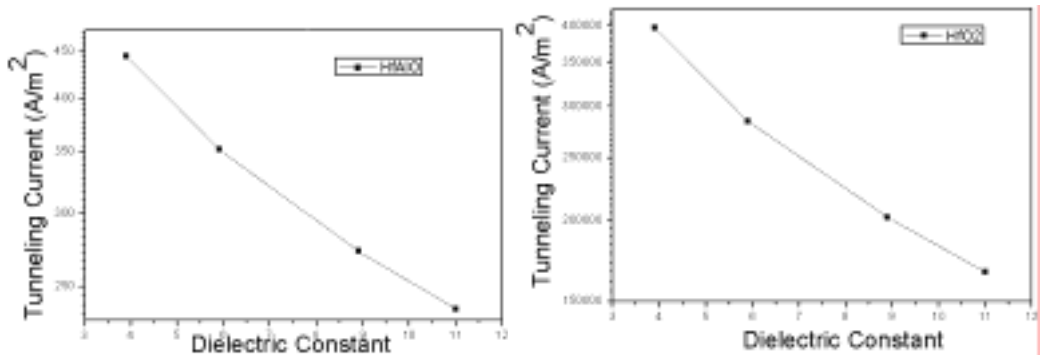
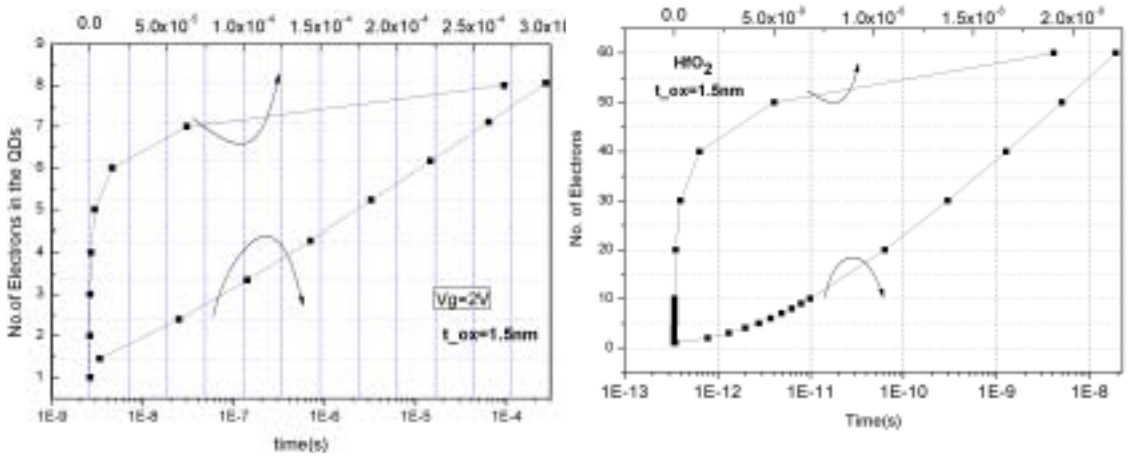


Fig.6.11 *Simulated tunneling current as a function of dielectric constant with different dielectrics at programming mode and t_ox=4.5nm.*

6.3.3 Programming and retention times

The advantages of the flash memory with high-k dielectrics are low voltage operation, fast programming time and longer retention time. The efficient programming of high-k dielectric flash memory at a relative low voltage has been demonstrated by the results discussed in section 6.3.2. During the programming mode, the high-k dielectric with low barrier height enables fast programming/erasing speed, and during the retention mode, the high-k dielectric can prevent leakage current to provide good retention performance due to larger physical thickness. As a result, considering the ratio of programming current and leakage current during retention, the high-k dielectrics are expected to provide higher ratio than SiO_2. The programming and retention times of high-k dielectric flash memory will be discussed in this part.



(a) *(b)*

Fig.6.12 *The programming time as a function of stored charge in the quantum dot when V_g=2V (a) SiO_2 (b) HfO_2.*

The programming time as a function of stored charge in the quantum dot for SiO_2 and HfO_2 is presented in Fig.6.12. The tunnel thickness is taken to be 1.5 nm for SiO_2 and HfO_2. Compared with SiO_2, HfO_2 provides significantly faster programming time and more electrons in the quantum dot with the same physical thickness 1.5 nm. It demonstrates that high-k dielectric has potential to provide faster programming time and more efficient programming operation. Both in SiO_2 and HfO_2 dielectrics, the general feature that the more electrons tunnel into the quantum dot the longer time it needs to add extra electrons into the quantum dot is similar.

The relationship of the programming time and dielectric thickness for HfO_2, HfAlO and SiO_2 is plotted in Fig. 6.13. With the same dielectric thickness, the programming time of the flash memory with high-k dielectrics is significantly faster than that of SiO_2. When the thickness increases, the difference between the high-k dielectrics and SiO_2 becomes more significant. This result is similar to the result shown in Fig.6.8

and Fig.6.9. This shows that the low barrier height results in F-N tunneling and hence speeds up the programming time. The 0.4 nm difference of tunnel oxide thickness leads to near 100 times difference of the programming time. Hence the thickness is a key factor in optimizing the programming and retention times.

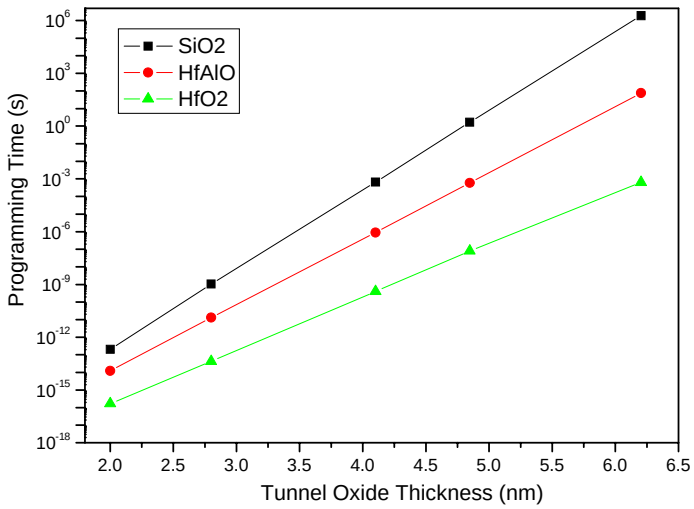


Fig. 6.13 *The programming time as a function of tunnel oxide thickness with different dielectrics.*

The retention time of the flash memory with SiO_2 and HfO_2 determined from the charge lost is shown in Fig. 6.14. It is significant that the retention characteristics of high-k dielectric flash memories are better than SiO_2 flash memory. The dot line presents 10 years retention time standard. The retention times of the flash memory with SiO_2 4.3 nm, SiO_2 4.5 nm and HfO_2 6.2 nm are 9.009×10^7 *s*, 1.603×10^9 *s* and 4.009×10^9 *s*. The SiO_2 with tunnel thickness 4.5 nm, HfO_2 with thickness 6.205 nm can reach the 10 years retention standard. In this case, the EOT for SiO_2 and HfO_2 are 4.5 nm and 2.2 nm, in which the thin EOT of 2.2nm of HfO_2 can enable faster

programming time at the same time.

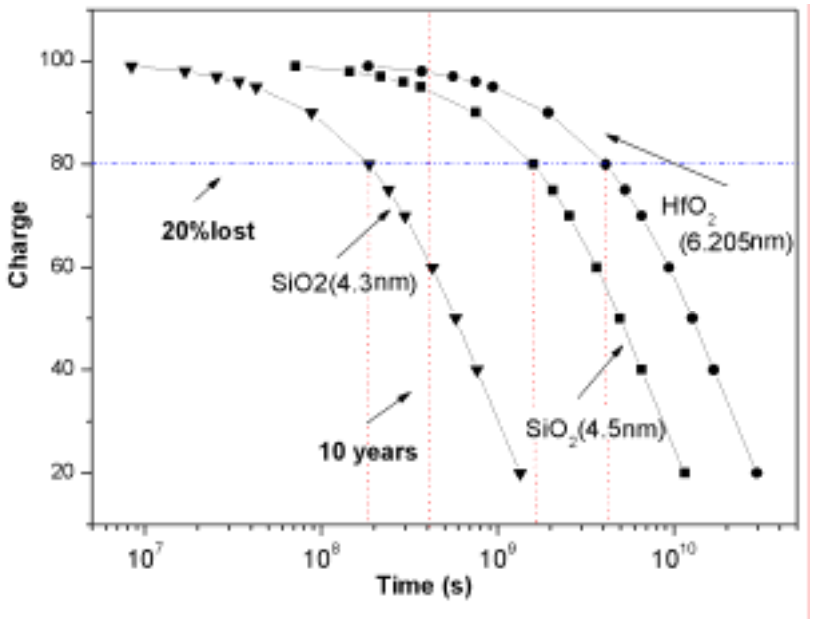


Fig.6.14 *The charge in the quantum dot as a function of time with different dielectrics in the retention state.*

The charge loss shows exponential attenuation as a function of time and a suddenly drops at around 80% mark, as shown in Fig.6.14. However, a smooth charge loss as a function of retention time is expected and showed in the experimental result [11]. This is because that the transmission probability across the tunnel oxide from the quantum dot to silicon substrate is assumed to be a constant value according to the residual charge in the quantum dot in this model. Thus, by recalling the Eq. 3.31 and Eq.3.32, the exponential attenuation of charge loss as a function of time can be explained and the suddenly drop observed in Fig. 6.14 is possible.

However, in fact, with more electrons escaping from the quantum dot flash memory, the potential in the quantum dot changes and results in the change of the effective barrier for tunneling. As a consequence, the potential in the quantum dot becomes

close to the silicon substrate and the transmission probability becomes smaller. Therefore, in general, the transmission probability will become smaller and smaller during the retention mode due to the change of the potential in the quantum dot. As a result, a smooth charge loss as a function of time in retention state is expected and shown in the experimental results [11].

If we assume that the potential in the quantum dot decreases linearly with more electrons tunneling out of the quantum dot during the retention, the charge loss as a function of time is shown in Fig. 6.15. In this result, we assume that the potential has a linear increase with the charge loss in the quantum dot which is a reasonable assumption. For example, 5% charge loss results in 5% decrease of the potential in the quantum dot. The loss is now more gradual as compared to Fig.6.14 as expected.

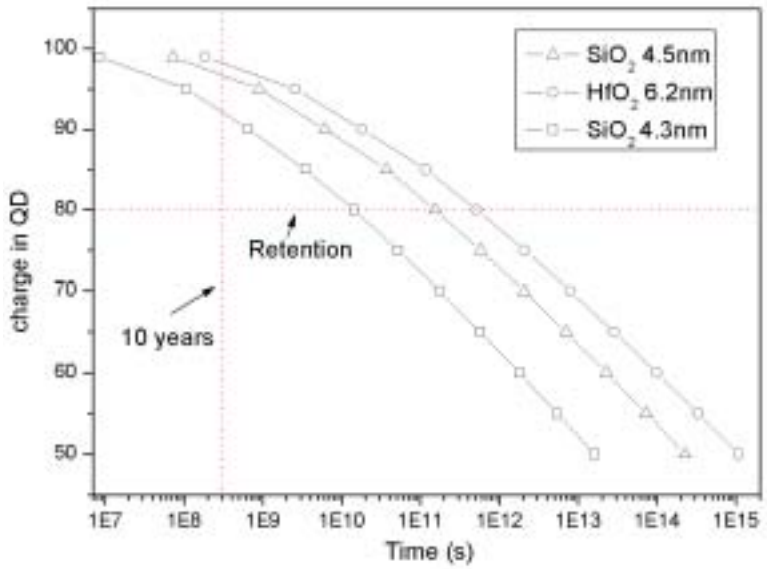


Fig.6.15 *The retention time as a function of charge lost in the quantum dot with different dielectrics simulated by barrier height approximation.*

The retention characteristics of HfO_2 and SiO_2 as a function of time are plotted in

Fig.6.16 and Fig.6.17. In this simulation result, the change of the barrier height of the dielectric with the loss of charge in the quantum dot is considered. In the result, the V_T of the flash memory with SiO_2 dielectric is obtained from the Fig.5.6. The V_T of the flash memory with HfO_2 is obtained from the Fig. 6.7. Here, the ΔV_T is defined as

$$\Delta V_T = V_{T(10e)} - V_{T(0e)} \tag{6.1}$$

where $V_{T(10e)}$ is the threshold voltage when 10 electrons are in the quantum dot and $V_{T(0e)}$ when no electron in the quantum dot. The retention time is defined as the loss of 20% of ΔV_T. We simulate the number of electrons as a function of control gate voltage as shown in Fig. 5.6 and Fig. 6.7, from which the threshold voltages, considering different number of electrons in the quantum dot, is found. We assume that 10 electrons in the quantum dot represents 100% charge in the quantum dot, therefore 8 electrons in the quantum dot represents 80% charge in the quantum dot and so on, as shown in Fig.6.15. Since the charge loss as a function of retention time can be found, as shown in Fig. 6.15, therefore the relationship between threshold voltage and retention time is found in Fig.6.16 and Fig. 6.17. The threshold voltage as a function of retention time for HfO_2 with thickness of 6.2 nm is shown in Fig.6.17 and the same result for SiO_2 with thickness of 4nm is shown in Fig.6.16. For HfO_2 dielectric flash memory device, when tunnel oxide thickness is 6.2nm, the 10 years retention time can be achieved and the programming time is 6.36e-4 s. For SiO_2 in order to reach 6.36e-4 *s*, the thickness should be 4 nm, as plotted in Fig.6.16.

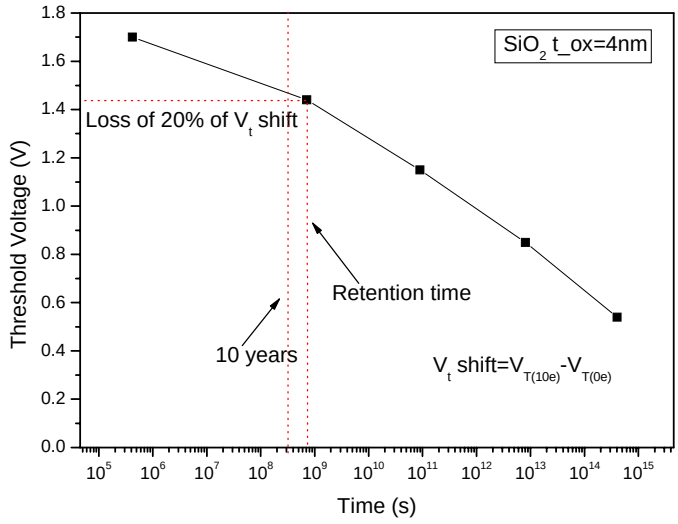


Fig.6.16 *Retention time for SiO_2 flash memory with tunnel oxide thickness 4 nm.*

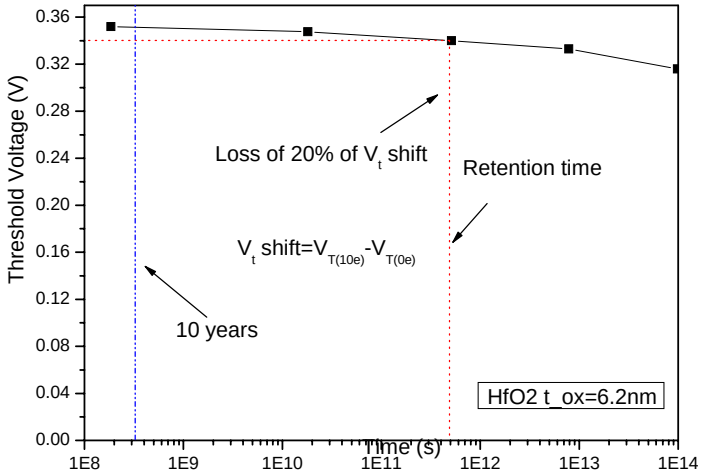


Fig.6.17 *Retention time forHfO_2 flash memory with tunnel oxide thickness 6.2 nm.*

When we compare these two figures, the high-k dielectric shows better retention performance because the charge loss rate of HfO_2 with thickness of 6.2nm is slower than that of SiO_2. For HfO_2 dielectric, no significant shrinkage of threshold voltage is observed at up to 8×10^{13} *s*, due to its sufficient physical thickness. However, for SiO_2, the rate of the decrease of the threshold voltage is fast from the beginning. It

means that the rate of charge loss of the flash memory with HfO_2 dielectric is slower than that of the flash memory with SiO2 dielectric. This memory characteristic shows a good agreement with the experimental result [11].

Fig. 6.18 shows the retention time as a function of EOT for SiO_2, HfO_2 and HfAlO. With the increase of EOT, the difference between the SiO_2 and high-k dielectrics becomes larger. In order to ensure 10 years retention standard, the tunnel oxide EOT of HfO_2 and HfAlO should be 2.0 nm and 2.2 nm, respectively. As we discussed previously, the EOT of SiO_2 has to be around 4.5 nm. The difference of the retention time between different dielectrics becomes larger with the increase of the EOT. It indicates that the thickness has more prominent effect on the retention time during the retention mode, which is different compared to the programming mode where the barrier height dominates the programming time.

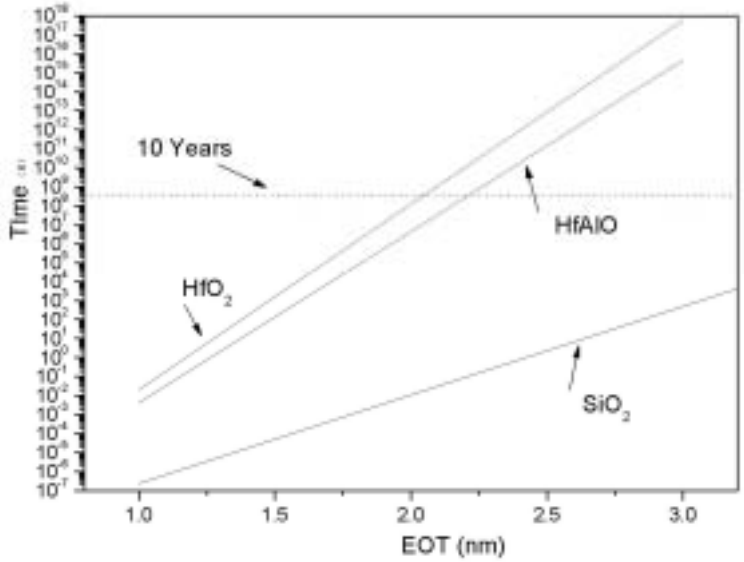


Fig.6.18 *The retention time as a function of EOT with different high-k dielectrics.*

6.4 Summary

This chapter focuses on the investigation of the flash memory incorporating high-k dielectrics. The basic properties of the high-k dielectrics are discussed and explored. The performance of the flash memory with high-k dielectrics is simulated and compared with the SiO_2 flash memory. The efficient programming mode at low voltage operation is demonstrated. Through the calculation of tunneling current and programming/retention time, the advantages of low voltage operation, fast programming speed and long retention time are predicted by our simulation results. The appropriate thickness for different dielectrics in order to ensure 10 years retention standard is proposed. In this chapter, the high-k dielectrics are demonstrated theoretically that they have great potential to replace conventional SiO_2 in the future application of the flash memory device.

Chapter 7
Flash Memory Device Using Ge/SiGe/Si Quantum Dot

7.1 Introduction

In the quantum dot flash memory device, electrons can be stored either in the traps or in the conduction band of the quantum dot. Evidences show that if electrons are stored in interface states or bulk traps, rather than the conduction band, good retention property can be provided. Because narrower band gaps can provide lower conduction band edge, resulting in better confinement of electrons in quantum dot and therefore better retention performance [47], quantum dot with narrow band gap materials can be advantages for flash memory devices. Compared to Si quantum dot, Ge quantum dot has a narrower band gap and a similar electron affinity. Therefore, Ge quantum dot is expected to provide better retention memory characteristics. Ge quantum dot flash memory devices using thin SiO_2 tunneling oxide have been demonstrated a few years ago and good retention performance was observed [51].

However, because of the low evaporation temperature of Ge quantum dot and the difference in surface energy with respect to the oxide, it is difficult to assemble Ge dots on insulators compared to Si dots. An alternative technique to take advantage of Ge smaller band gap is to grow the $Si_{1-x}Ge_x$ dots directly on the tunneling oxide using rapid thermal chemical vapor deposition ([48], [20]). The availability of this method has

been demonstrated experimentally [11]. The characteristics of the flash memory device using SiGe quantum dot need to be explored and studied theoretically. Therefore, SiGe flash memory is also investigated. In this chapter, both pure Ge quantum dot and SiGe quantum dot devices are considered. Their retention performance are explored and compared with Si quantum dot flash memory device.

In this chapter, section 7.2 explores basic properties of SiGe dots flash memory and try to give the relative important parameters of SiGe dots. Section 7.3 investigates programming and retention times of pure Ge quantum dots flash memory, including the impact of dot size on the programming and retention. The impact of the trap energy on retention performance is studied. Section 7.4 discusses and predicts an ideal flash memory device using theoretical simulation results. Section 7.5 gives a summary of this chapter.

7.2 Investigation of Si_xGe_{1-x} Dots

Dot	Effective Mass	Trap Energy	Dielectric permittivity
Ge	0.22 m_e	0.51 eV	16
Si	0.32 m_e	0 eV	11

Table 7.1 *Important parameters of Si and Ge dots.*

The important parameters for Ge and Si quantum dots are shown in Table 7.1. The conduction band shift for Si, Ge and SiGe quantum dots is assumed to the same. The conduction band shifts for 5nm, 3nm and 2 nm quantum dots are taken as 0.15 eV, 0.5

eV and 1 eV, respectively. For larger size of the quantum dot, the conduction band shift is smaller and assumed to be neglected.

In this work, the electron effective mass and the dielectric constant values for SiGe dots are assumed to be interpolated between those of Si and Ge. The trap energy level of pure Ge dots is extracted from experimental data [49]. For Si_xGe_{1-x}, the trap energy is kept at 0.51eV. Because the electron affinity of Si quantum dot is a little higher than Ge quantum dot, hence there is a slight difference between barrier heights of SiO_2 and HfO_2 tunnel dielectrics. In this simulation, the barrier height of the tunnel oxide for Ge quantum dot is assumed as 1.45 eV which is 0.05 eV less than that of tunnel oxide for Si quantum dot. 0.2 eV difference of the barrier height is also used in the simulation for SiO_2. Hence, the value of the barrier height of the tunnel oxide for SiGe quantum dot is interpolated between those of Si and Ge. The parameters of SiGe quantum dot as a function of Ge composition are assumed in Table 7.2.

	Ge	$Si_{0.25}Ge_{0.75}$	$Si_{0.5}Ge_{0.5}$	$Si_{0.75}Ge_{0.25}$	Si
effective mass (m_e)	0.22	0.245	0.27	0.295	0.32
barrier height(HfO_2)(eV)	1.3	1.350	1.4	1.45	1.5
dielectric permittivity	16	14.75	13.5	12.25	11

Table 7.2 *Parameters of SiGe.*

The retention time as a function of the composition of Ge quantum dot in the retention state is illustrated in Fig. 7.1, in which HfO_2 dielectric of thickness 4.6 nm is considered. The dot size is assumed to be fixed at 5 nm in diameter. The difference of

the barrier height of the tunnel oxide is assumed as 0.05 eV and 0.2 eV, respectively. In Fig. 7.1, with the increase of Ge composition, the retention time is reduced. This is because that the electron affinity of Si is 0.05 eV higher than that of Ge quantum dot and therefore the barrier height of the tunnel oxide for Ge is less than that of the tunnel oxide for Si quantum dot. As a result, a bit smaller barrier height of the tunnel oxide results in the reduction of the retention time. However, because the trap energy of SiGe quantum dot is 0.51eV and the trap energy of Si is 0 eV, so SiGe quantum dot is still expected to have better retention than Si quantum dot. If we assume a SiGe quantum dot flash memory with 50% Ge composition, the retention time is *5.25e8 s*. When a pure Si quantum dot flash memory is used, the retention time is *68.548 s* which is significantly poorer than that of Ge quantum dot. Since the Fermi level of the SiGe quantum dot is above the mid-band gap of the Si channel in the retention, SiGe quantum dot is also expected to have better retention performance. In our theoretical model, the Ge concentration is assumed to be 75% in SiGe quantum dots.

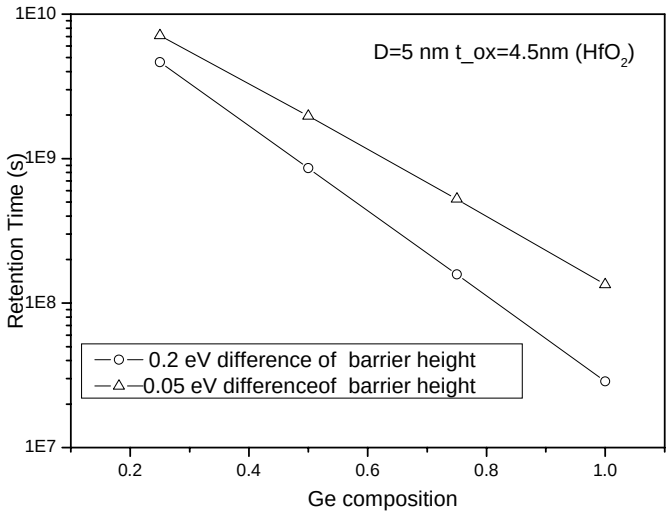


Fig.7.1*Retention time of SiGe quantum dot flash memory.*

When Ge is taken to be 75%, the retention time of SiGe quantum dot flash memory is simulated in Fig.7.2. Among SiGe, Si and Ge quantum dots flash memory, SiGe quantum dot flash memory shows much better retention characteristic than Si quantum dot flash memory, and pure Ge quantum dot flash memory provides the best retention than SiGe flash memory. In our simulation model, because the main difference between Si, SiGe and Ge quantum dot flash memory is the trap energy, it demonstrates that the change of trap energy results in great difference in the retention time. The trap energy is an important factor which results in good retention performance. When tunnel oxide thickness increases from 3.5 nm to 4.5 nm, the retention time increases about 3 orders of magnitude. With the same tunnel oxide thickness, the difference between Si and Ge quantum dot is about 7 orders of magnitude. As a result, for Ge quantum dot flash memory, the impact of the trap energy on the retention time is more important than the tunnel oxide thickness.

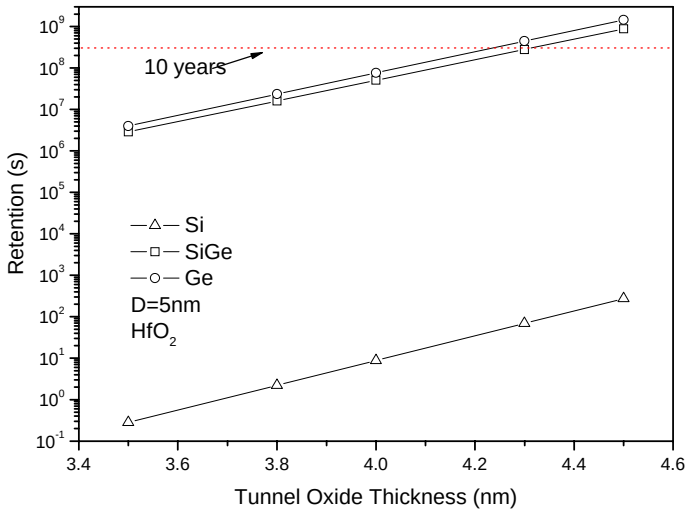


Fig. 7.2 *Retention time as a function of tunnel oxide thickness for Si, SiGe and Ge quantum dot.*

7.3 Ge Quantum Dot Flash Memory

Ge nanocrystal flash memory devices using SiO_2 tunneling oxide were demonstrated a few years ago and good retention performance was shown [51, 52]. The fabrication of Ge quantum dot on silicon dioxide became possible in recent years.

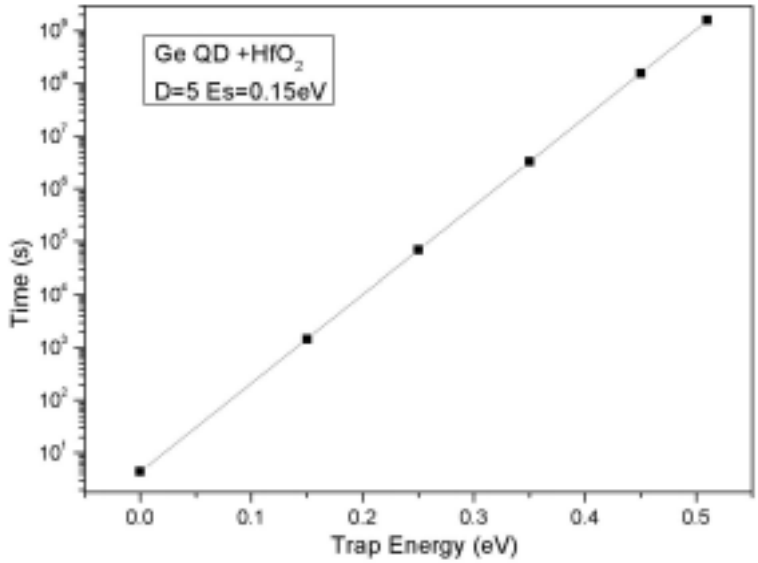


Fig.7.3 *The impact of trap energy on the retention time of Ge flash memory using HfO_2 dielectric.*

For pure Ge quantum dot flash memory, the impact of trap energy on the retention time is illustrated in Fig.7.3, in which the diameter of the quantum dot is 5 nm and the conduction band shift is 0.15 eV. The result demonstrates again that the trap energy has a very important effect on the retention time. A 0.2 eV difference between trap energies results in near 4 orders of magnitudes difference of the retention time. The explanation is that the electron is localized in the traps of the quantum dot, and hence it is difficult to be injected from the quantum dot. As a result, retention time becomes longer and the information can be stored longer.

When the trap energy is fixed, we study the impact of the barrier height on the retention time in Fig. 7.4 with tunnel oxide thickness 4.5 nm. Compared with Fig. 7.3, it is obvious that the barrier height has less effect on the retention time compared to the trap energy. It is believed that, with the same conduction band shift, the contribution of Ge quantum dot may be larger than the contribution of high-k dielectric to the good retention. Therefore, it is possible that Ge quantum dot flash memory with HfAlO will provide a better retention than Si quantum dot flash memory with HfO_2 as shown in Fig.7.7.

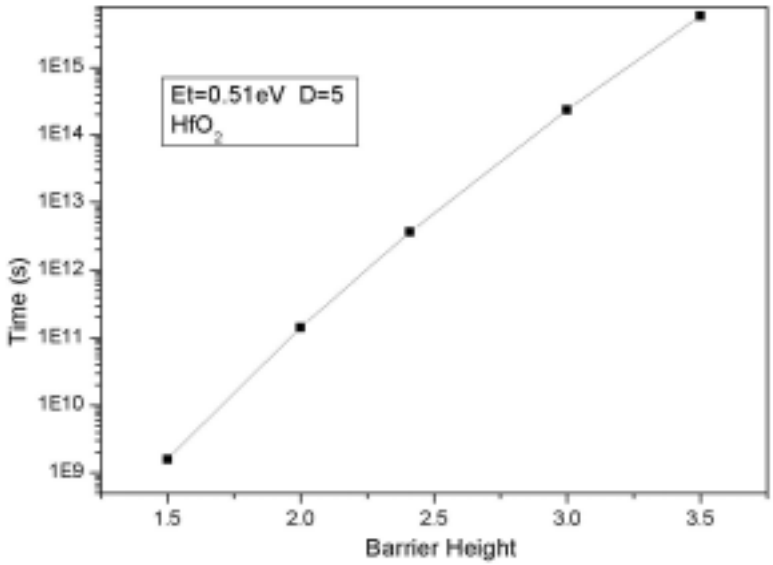


Fig.7.4 *The impact of barrier height on the retention time.*

The programming and retention times of Ge quantum dot flash memory with HfO_2 are illustrated in Fig.7.5, considering a large square quantum dot with conduction band shift 0 eV. For Ge quantum dot flash memory with HfO_2 dielectric, the tunnel oxide thickness of near 4.5 nm can provide 10 years retention time. As we discussed previously in chapter 6, for Si quantum dot flash memory with HfO_2 dielectric, the tunnel oxide thickness is required to be 6.1 nm. Therefore, with the use of Ge

quantum dot, the HfO_2 tunnel oxide thickness can be scaled down from 6 to 4.5 nm, which is a great improvement for the scaling of tunnel oxide thickness.

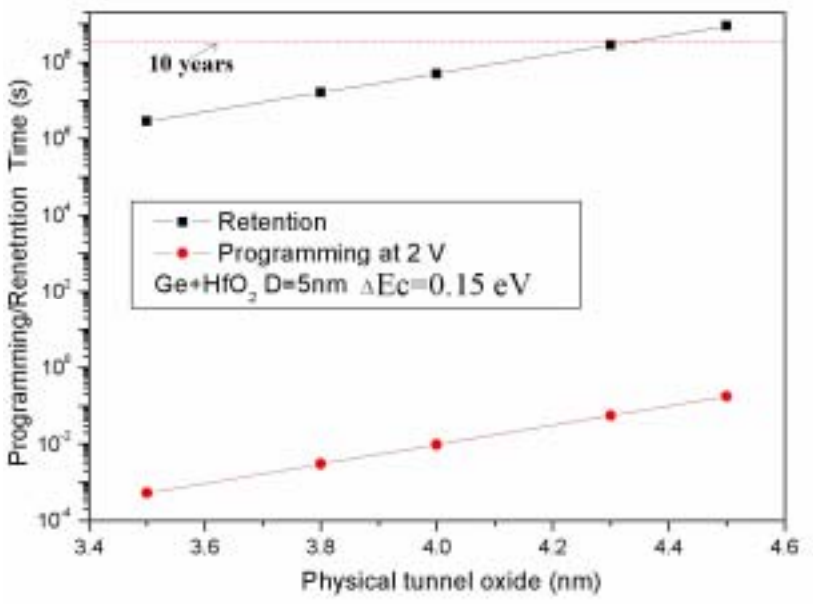


Fig.7.5 *Programming and retention times of Ge quantum dot flash memory (dot line is ten years retention standard).*

The impact of the dot size on the programming and retention times are discussed in Fig. 7.6. The programming and retention performance of HfO_2 flash memory with the quantum dot of 2 nm, 3 nm and 5 nm diameter is simulated in Fig.7.6. The result shows that 5nm quantum dot provides faster programming time and better retention time at all tunnel oxide thicknesses. The reason is that the larger the size of the quantum dot, the smaller is the quantum confinement effect, therefore it results in larger tunneling current probability. Hence, for faster programming time, larger size quantum dot is suggested. Therefore, 5 nm quantum dot flash memories with 4.3 nm HfO_2 tunnel oxide are selected to provide 10 years retention times and at the same time the programming time 2×10^{-2} *s* at 2 V control gate voltage. However, the larger

size of the quantum dot will decrease the reliability of the flash memory. Hence, for the flash memory with larger size of the quantum dot, it will not be good for providing better reliability compared with the flash memory with a smaller quantum dot. And in Fig.7.6, when the size of the quantum dot increases to 7 nm which the conduction band shift is assumed to be 0 eV, the gain in programming/retention is insignificant. For the quantum dot with larger size (larger than 7 nm), the quantum effect is reduced and the memory performance will be worse than that of the dot with smaller size.

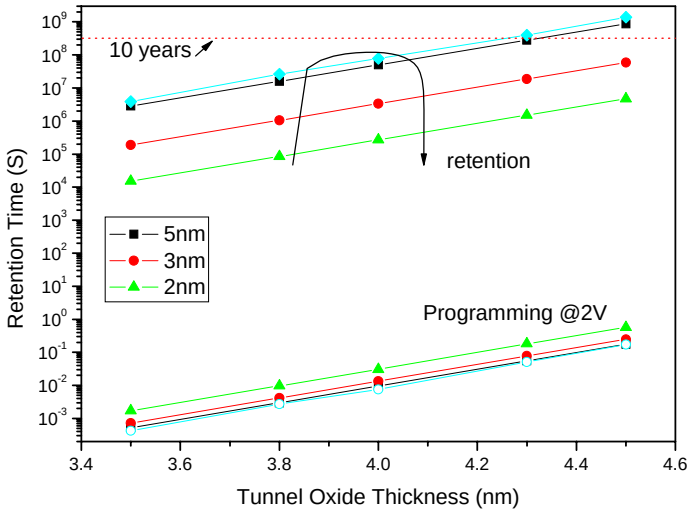


Fig.7.6 *The impact of dot size on programming and retention times.*

The comparison of the retention time of flash memories with different dielectrics, considering Si and Ge quantum dots are illustrated in Fig. 7.7. A larger quantum dot ($6nm \times 10nm \times 10nm$) is embedded between control oxide and tunnel oxide and the conduction band shift is taken as 0 eV. It is significant that the use of Ge quantum dot improves the retention time greatly. Using Ge quantum dot, the flash memory with HfAlO dielectric can even have better retention than the device with HfO_2 as we have

predicated in Fig. 7.4. That means Ge quantum dot has more contribution to optimizing the retention time than the use of high-k dielectrics under the same condition. With the increase of EOT of tunnel oxide, it seems that the contribution of high-k dielectrics becomes larger. When EOT is less than 1.6 nm, the retention time of Ge quantum dot flash memory with SiO2 is better than Si quantum dot flash memory with HfO_2 and HfAlO. However, when EOT is larger than 1.6 nm, the high-k dielectric shows obvious predominance and gives more contribution to optimizing retention time. With the continuous increase of tunnel oxide thickness, the difference between high-k dielectrics is enlarged. Hence, Ge quantum dot will play important role in providing good retention time in the flash memory with smaller dimension.

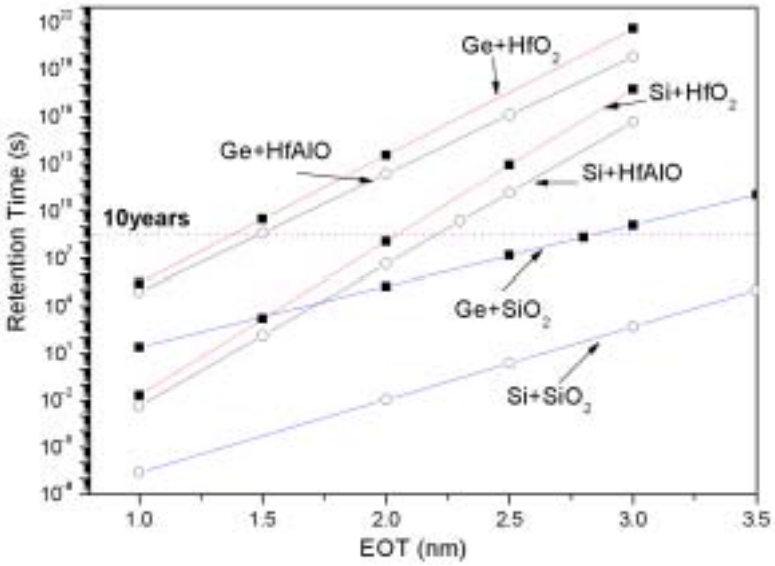


Fig.7.7 *The comparison of retention time of flash memories with various dielectrics and quantum dots.*

7.4 The Ideal Flash Memory Devices

In this section, using our simulation model, we try to predict an ideal flash memory

device. In our research work, the Si and Ge quantum dot with SiO2 and high-k dielectrics are studied. With the comparison of HfO_2, HfAlO and SiO_2 dielectrics, HfO_2 is believe to provide good memory characteristics, including faster programming time and longer retention time at a relative low voltage. Therefore, HfO_2 is proposed in the predication of a good flash memory device. Because it is not easy to compare the programming efficiencies between Si (without trap energy) and Ge quantum dot flash memory as we discussed in chapter 6, we try to predict the good Si quantum dot flash memory with a hypothetical dielectric and good Ge quantum dot flash memories with a hypothetical dielectric, respectively.

The simulation model assumes a diameter of 5 nm quantum dot embedded between the control oxide and tunnel oxide, channel length 40 nm and the substrate doping density of $5\times10^{20}m^{-3}$. Based on the standard in which the programming time should be 1 *ms* and retention time should be 10 years, we try to predict a good flash memory device.

Considering a Si quantum dot flash memory device with HfO_2 dielectric, in order to achieve 10 years retention time, the tunnel oxide thickness should be 6.743 nm. With same thickness, the programming time at 2 V write voltage is 0.0064 s which is acceptable. The dielectric constant and barrier height for HfO_2 are taken as 11 and 1.5 eV, respectively. For 6.743 nm HfO_2 tunnel oxide, the equivalent thickness is 2.39 nm.

When the Si quantum dot is changed by Ge nanocrystal, HfO_2 tunnel oxide thickness of 4.563 nm can reach 10 years retention standard and at the same time, the programming time at *0.0048 s* at the write voltage 2 V is achieved. The EOT of HfO_2 is only 1.62 nm. Hence, with the use of Ge quantum dot, the tunnel oxide can be scaled from 6.743 nm to 4.563 nm.

7.5 Summary

In this chapter, characteristics of SiGe and pure Ge quantum dot flash memory are investigated. For SiGe quantum dot flash memory, we try to study its relative important parameters and properties. The effect of Hf composition on the retention time is explored and the research work concentrates on the retention study of SiGe quantum dot. For pure Ge quantum dot flash memory, the impact of trap energy on the retention is examined and is demonstrated to be a most important factor for good retention time, especially for a relatively smaller device. The results show that within a range of tunnel oxide thickness, the impact of Ge on the retention time is larger than that of high-k dielectric. Therefore, Ge is seen as a possible material for replacing Si quantum dot in flash memory device.

Chapter 8
Conclusions and Recommendations

8.1 Summary

This thesis addressed device physical, modeling and design issues of the quantum dot floating gate flash memory with nanoscale dimension. The characteristics of the quantum dot flash memory are studied, considering high-k dielectrics, Si and Ge quantum dot. A simulation tool is developed to accomplish these objectives. The fundamental device physics, including charging process of the floating gate and the channel, are solved by self-consistent solution of Schrödinger-Poisson equations, in which the electron distribution is solved by Schrödinger equation and the potential profile is solved by Poisson equation. The tunneling current mechanism that dominates the programming/erase characteristics is examined by a modified semi-classical WKB approximation. Using the trap model, considering quantum confinement in the quantum dot, the programming and retention times are evaluated. These theories constitute the main methodology used in the research work.

Si quantum dot flash memory with conventional silicon dioxide tunnel oxide is simulated and studied. Its basic physical characteristics are examined. The obvious charging phenomenon is observed through simulated 2D and 3D electrons distribution. The effect of the charging process of the quantum dot on the inversion layer is

examined. The results show that the increasing amount of electrostatic gate field energy required to sustain the inversion charge in the channel is at the expense of the charging of the floating gate. The simulated tunneling current demonstrates that the scaling of the tunnel oxide thickness is very important for optimizing the programming performance of the flash memory. From the calculation of programming time, a time that can be of the order of tens of nanoseconds shows good agreement with experimental result on the quantum dot flash memory. A trap energy model based on quantum confinement is used to predict the retention time of the flash memory device. The tunnel oxide thickness is demonstrated as a key factor for providing good retention. A 4.3 nm silicon dioxide thickness is suggested in order to enable 10 years retention standard.

Flash memory with high-k dielectrics, including HfO_2 and HfAlO, are investigated and compared with the silicon dioxide flash memory. The basic device characteristics of high-k dielectric flash memory are studied. The simulation concentrates on the programming and retention performance of the high-k dielectric flash memory. The investigation shows the significant advantages of high-k dielectrics. It provides more efficient programming operation at a relatively low control gate voltage compared to SiO_2 flash memory. With the same tunnel oxide thickness, the programming time of high-k dielectric (HfO_2) flash memory is 4 times faster than SiO_2 flash memory. For good retention mode, tunnel oxide EOT 2.2 nm of HfO_2 flash memory device is proposed and with the same condition, the SiO2 flash memory needs 4.3 nm tunnel

oxide thickness. Therefore, our simulation results show that high-k dielectrics are the promising candidates for replacing the conventional SiO_2 in the flash memory device.

Ge quantum dot is proposed recently due to its large trap energy. The main advantage of Ge quantum dot flash memory is its good retention characteristic. This research work examines the impact of the trap energy of Ge quantum dot on the retention time and shows that the trap energy plays critical role for Ge quantum dot to provide longer retention time. The contribution of Ge and high-k dielectrics to the retention time is explored. For a device with a smaller tunnel oxide thickness, Ge quantum dot has more pronounced effect on the retention time, while, with a larger tunnel oxide thickness, high-k dielectrics make more contributions to the longer retention time.

Finally, we predict ideal quantum dot flash memory based on our theoretical studies, considering the high-k dielectrics and Ge/SiGe quantum dot. A good device which provides efficient and faster programming, longer retention and low voltage operation is proposed. Based on the simulation model of this thesis, some parameters of the device are suggested.

8.2 Recommendations for Future Work

Though the main characteristics have been studied and investigated in this thesis, there are some immediate extensions to this research work, as follows.

The self-consistent solution of Schrödinger-Poisson method is an appropriate method for simulating the electrons distribution and potential profile of the device systems. However, the convergence problem results in inefficient computation and is time consuming. Especially the use of mode-space method in solving the Schrödinger equation restricts the thickness of substrate to be less than 6 nm; otherwise, the mode-space method will be broken down because. Therefore, a more efficient numerical implementation of this approach is required by an extremely large computational capability.

As presented in Chapter 7, Ge and SiGe have been demonstrated as promising materials for quantum dot in the flash memory. However, there are still fewer studies on the programming time. Especially for SiGe quantum dot, the accurate determination of the alloy composition and its parameters is still difficult at present stage. A further study in such alloy quantum dot will be meaningful for the flash memory.

The coulomb blockade is very prominent in quantum dot flash memory, especially in a very small quantum dot flash memory. We emulate it in this thesis approximately, while we suggest that it needs to be simulated accurately and the simulation model needs to be enhanced. There are some quantum simulation models which consider coulomb blockade effect in quantum dot flash memory [23, 54, 55], but there is further scope in improving their accuracy.

References:

[1] http://www.bccresearch.com/editors/RG-277.html

[2] J.Blauwe, "Nanocrystal nonvolatile memory devices," *IEEE Trans. Nanotechnol.*, vol.1, pp.72-77, Mar.2002

[3] B.De Salvo *et al*, "How far will silicon nanocrystals push the scaling limits of NVMs technologies?" in *IEDM Tech. Dig.*, 2003, pp.597-600

[4] Piero Olivo, Enrico Zanoni, "Flash Memory," Boston, Mass Kluwer, Academic Publihsers, 1999

[5] S.Tiwari, F.Rana, K.Chan, H.Hanafi, C.Wei, and D.Buchanan, "Volatile and nonvolatile memories in silicon with nano-crystal storage," in *IEEE Int.Electron Devices Meeting Tech.Dig.*, 1995, pp.521-424

[6] Y.Shi, K.Saito, H.Ishikuro, and T.Hiramoto, "Effects of interface traps on charges retention characteristics in silicon-quantum-dot-based metal oxide semiconductor diodes," *Jpn, J.Appl.Phys.*, vol.38, pp.425-428, Jan.1999

[7] J.A.Wahl, H.Silva, A.Gokirmak, A.Kumar, J.J Welser and Sandip Tiwari, "Write, erase and storage times in nanocrystals memories and the role of interface states," in *IEDM Tech. Dig.*, 1999, pp.375-378.

[8] R.Muralidhar *et al*, "A 6V embedded 90nm silicon nanocrystal nonvolatile memory," in *IEDM Tech. Dig.*, 2003, pp.26.2.1-26.2.4

[9] M.Saitoh, E.Nagata, and T.hiramoto, "Effects of ultra-narrow channel on characteristics of MOSFET memory with silicon nanocrystal floating gates," in *IEDM Tech.Dig.*, 2002, pp.181-184

[10] Angus I.Kingon, Jon-Paul Maria, S.K.Streiffer, “Alternative dielectrics to silicon dioxide for memory and logic devices,” *Nature*, vol.406, pp.1032-1038, 2000

[11] D.-W Kim, T.Kim and S.K.Banerjee, “Memory characterization of SiGe quantum dot flash memories with HfO_2 and SiO_2 tunneling dielectrics”, *IEEE trans. Electron Devices*, vol.50.pp.1823-1829, 2003

[12] J.J.Lee, X.Wang, W.Bai, N.Lu, J.liu, and D.L.Kwang, “Theoretical and experimental investigation of Si nanocrystal memory device with HfO_2 high-k tunneling dielectric,” *in Proc.VLSI, Technol.Symp,* 2003, pp.33-34

[13] C.Monzio Compagnoni, D.Ielmini, A.S.Spinelli,,QA.L.Lacaita, C.gerardi, L.Perniola, B.De, Salvo and S.Lombardo, “Program/erase dynamics and channel conduction in nanocrystal memories,” in *IEDM Tech.Dig.*, 2003, pp.549-552.

[14] G.D.Willk, R.M.Wallace, J.M.Anthony, “High-k gate dielectrics: Current status and materials properties considerations,” *J.Appl.Phys.* vol.89, pp.5243-5275, May.2001

[15]A.Chatterjee,R.A.Chapman,K,Joyner,M.otobe,S.Hattangady,M.Bevan,G.A.Brown, H.Yang, O.He, D.Rogers *et al.*, *Tech.Dig.Int.Electron Devices Meet*.1998,pp.777

[16] P.K.Roy and I.C.Kizilyalli, *Appl.Phys.Lett.72*, pp.2835 (1998)

[17]I.C.Kizilyalli, R.Y.S.Huang, and P.K.Roy, *IEEE Electron Device Lett*.19, pp.423(1998)

[18] Wilk, G.D.,Wallace, R.M.&Anthony, J.M. “Hafnium and zirconium silicates

for advanced gate dielectrics," *J.Appl.Phys*.87, pp.484 2000

[19] Y.Shi, K.Saito, H.Zshikuro, and T.Hiramoto, "Effects of interface traps on charges retention characteristics in silicon-quantum-dot-based mental-oxide-semiconductor diodes," *Jpn.J.Appl.Phys.*, vol.38, pp425-428, Jan,1999

[20] D.-W.Kim, Hwang, T.T.Edgar, and S.Banerjeea, "Characterization of SiGe quantum dots on SiO2 and HfO2 grown by rapid thermal chemical deposition for nanoelectronic device," *J.Appl.Phys*.vol.81, pp.2384,1997

[21] A.thean and J.P.leburton, "Three-Dimentional self-consistent simulation of silicon quantum-dot floating gate flash memory device," *IEEE Electron Device letter*, vol.20, No.6, pp.286, 1999

[22] G.Iannaccore, A.trellakis and U.Ravaioli, "Simulation of a quantum-dot flash memory," *J. of Appl.Phys*, vol.84, pp.5032, 1998

[23] Farhan Rana, Sandip Tiwari, J.J.Welser, "Kinetic modeling of electron tunneling processes in quantum dots coupled to field-effect transistors", *Superlattices and Microstructures*, vol. 23, No. 3/4, 1998

[24] J.S.de Sousa, A.V.Thean, J.P.leburton, V.N.Freire, "Three-dimensional self-consistent simulation of the charging time response in silicon nanocrystal flash memories", *J.Appl.Phys*, vol.92, pp.6182-6187, 2002

[25] H.G. Yang, Y.Shi, H.M.Bu, J.Wu, B.Zhao, X.L.Yuan, B.Shen, P.Han, R.Zhang, Y.D.Zheng, "Simulation of electron storage in Ge/Si hetero-nanocrystal memory," *Solid state electronics*, vol.48, pp.767-771, 2001.

[26] G. Ianacone and P.loli, “Three-dimentional simulation of nanocrystal flash memories,” *Appl. Phys.Letter*, vol.78, pp.2046-2048, 2001

[27] Richard D.Pashley, Stefan K.Lai, ”Flash Memories: the best of two words”,*IEEE Spectrum*, Dec.1989,pp. 30.

[28] E. Burstein and S. Lunqvist, *Tunneling Phenomena in Solids*, Plenum press: New York, 1969.

[29] Khairurrijal, W.Mizubayashi, S.Miyazaki, and M.Hirose, “Analytic model of direct tunnel current through ultrathin gate oxide,” *J.Appl.Phys.*, vol.87,pp.3000,2000

[30] J.Cai and C.T.Sah, ”Gate Currents in ultrathin oxide metal-oxide-silicon transistors,” *J.App.Phys.*, vol.89,pp.2272,2001

[31] S.-H.Lo, D.A.Buchanan, Y.Taur, “Modeling and characterization of quantization, polysilicon depletion, and direct tunneling effects in MOSFETs with ultrathin oxides”, *IBM J.Res.Develop.* vol.43 , pp.209-211, May 1999

[32] Leland Chang, Kevin J.Yang, Yee-Chia Yeo, Igor Polishchuk, Tsu-Jae King, Chenming Hu, *IEEE Trans. Electron Device*, vol.49, pp.2288, 2002

[33] A.N.Khondker, M.Rezwan khan, A.F.M.Anwar, “Transmission line analogy of resonance tunneling phenomena: the generalized impedance concept”, *J.Appl.Phys.*, vol.63, No.10,May 1988

[34] E.Merzbacher, *Quantum Mechanics*, New Yok:Wiley, 1970,ch.7

[35] F.Rana, S.Tiwari, and D.A.buchanan, “self-consistent modeling of accumulation layers and tunneling currents through very thin oxides,”

Appl.Phys.Lett., vol.69, pp.1104-1106,1996.

[36] L.F.Register, E.Rosenbaum, and K.Yang, “Analytic model for direct tunneling current in polyscrystalline-silicon-gate metal-oxide-semiconductor devices”, *Appl.Phys.Lett*, vol 74, pp. 457, Jan 1999.

[37] 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 meansurements in MOS devices,” *IEE Trans.Electron Device*, vol.46, pp.1464, 1999

[38] W.K.Shin, E.X.Wang, S.Jallepalli, F.Leon, C.M.Maziar, and A.F.Tasch, “Modeling gate leakage current in nMOS structures due to tunneling through an ultra thin oxide,” *Solid-State Electron.*, vol.42, pp.997, 1998

[39] A.Dalla Serra, A.Abramo, P.Palestri, L.Selmi, F.Widdershoven, “Closed-and open-boundary models for gate-current calculation in n-MOSFETs,” *IEEE Trans.Electron Device*, vol.48, pp.1811, 2001

[40] Ren Zhi Bin, “Nanoscale MOSFETs: Physics, Simulation and Design”, *Ph.d thesis*, 2001, Purdue University.

[41] Ya-Chin King, *Phd thesis*, University of California, Berkeley, 1999

[42] L.F.Register, E.Rosenbaum, and K.Yang, “Analytic model for direct tunneling current in polyscrystalline-silicon-gate metal-oxide-semiconductor devices”, *Appl.Phys.Lett*, vol 74, pp. 457, Jan 1999.

[43] 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

meansurements in MOS devices," *IEEE Trans.Electron Device,* vol.46, pp,1464, 1999

[44] Min She, Tsu-Jae King, "Impact of crystal size and tunnel dielectric on semiconductor nanocrystal memory performance", *IEEE Trans.on Electron Device*, vol.50, pp.1934, 2003

[45] Hou Yong Tian, *Phd thesis,* National University of Singapore, 2003

[46] C.Monzio Compagnoni, D.Ielmini,A.S.Spinelli, A.L.lacaita, C.Previtali and C.Gerardi, "Study of data retention for nanocrystal flash memories," *IEEE 41st AIRPHS,* pp.506, 2003

[47] Y.Shi, K.Saito,H.Ishikuro,and t.Hiramoto,"Effect of interface traps on charge retention characteristics in silicon quantum dot based mental oxide semiconductor diodes"*Jpn.J.Appl.Phys*., vol.38,pp.425-428,1999

[48] Dong-Won Kim, Taehoon Kim, Sanjay K.Banerjee, "Memory characterization of SiGe quantum dot flash memories with HfO_2 and SiO_2 tunneling dielectrics," *IEEE Trans. on Electron Devices,* vol.50, pp.1823-1829, 2003

[49] Y. King, "Thin dielectric technology and memory devices," *Ph.D. dissertation*, Univ.California, Berkeley, CA, 1999.

[50] S.M.Sze, *Physics of Semiconductor Devices.* New York: Wiley, 1981.

[51] H.I.Hanafi, S.Tiwari, and I.Khan, "Fast and long retention-time nanocrystal memory," *IEEE Trans. Electron Devices*, vol.43, pp. 1553-1558, Sept. 1996.

[52] Y.-C. King, T.-J. King, and C.Hu, "MOS memory using germanium

nanocrystals formed by thermal oxidation of $Si_{1-x}Ge_x$," in *IEDM Tech.Dig*., 1998, pp.115-118.

[53] H.Y.Yu *et al*, "Thermal stability of $(HfO_2)_x(Al_2O_3)_{1-x}$ on Si," *Appl.Phys.Lett.*, vol.81,pp.3618-3620, 2002

[54] A.Thean and J.P.leburton, "3-D Computer Simulation of Single-Electron Charging in Silicon Nanocrystal Floating Gate Flash Memory Device", *IEEE Electron Device Letters*, Vol.22, No.3, March 2001.

[55] Min She, Ya-Chin King, Tsu-Jae King; Chenming Hu, "Modeling and Design Study of Nanocrystal memory device", *Device Research Conference*, 2001,25-27, pp.139-140, June 2001

[56] H.Grabert and M.H.Devoret, Eds., "Single Charge Tunneling-Coulomb Blockade Phenomena in Nanostrutures", .Ser. NATO ASI Series B. New York, Plenurn,1991

List of Publications

(1) Zhou Kai Hong, Bai Ping, Samudra Genash S, "Self-consistent Schrödinger-Poisson simulation of quantum dot flash memory," *International Conference on Scientific & Engineering Computation*, accepted (2004)

(2) Zhou Kai Hong, Bai Ping, Samudra Ganesh S, Chong Chee Ching, "Modeling characterization of silicon quantum dot flash memory with HfO_2 tunneling dielectric," *International Symposium on Integrated Circuits, Devices & Systems*, accepted(2004)

(3) Chong Chee Ching, Zhou Kai Hong, Bai Ping, Samudra Genash S, "Self-consistent simulation of quantum dot flash memory device with SiO_2 and HfO_2 dielectrics," *International Journal of nanoscience*, accepted(2004)