Technology and reliability of sub-micron 1T-flash EEPROM.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

TECHNOLOGY AND RELIABILITY OF SUB-MICRON 1T-FLASH EEPROM

By

Franklin D. Nkansah

A Dissertation

Presented to the Graduate Committee

Of Lehigh University

In Candidacy for the Degree of

Doctor of Philosophy

In

Electrical Engineering

Lehigh University

January 2001

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

7/16/00
Date

Prof. Miltiadis K. Hatalis (Director)

7/20/00
Accepted Date

Committee Members:

Prof. Miltiadis K. Hatalis (Chairman)

Prof. Marvin H. White

Prof. Demetrios Christodoulides

Dr. Kolawole R. Olasupo

Acknowledgements

Many thanks to my thesis advisor Dr. Miltiadis K. Hatalis for his encouragement, advice and continued friendship. I will forever be grateful for his inspiration and continuous guidance during my entire research work. I also express my sincere gratitude to Professors M. White , D. Christodoulides and Dr. K. R. Olasupo for serving on my committee. Special thanks to Dr. K.R. Olasupo for the continuous encouragement and guidance. A special thanks to the Motorola Semiconductor Technology Laboratories (STL) pilot line for providing me with the silicon resources which was necessary to complete fabrication of devices needed for the this research work.

Lastly, I will like to extend my sincere thanks and love to my wife Mercy, children Ruth, Claribel and Franklin II, for their unweathered support and patience during the many years of my late night research and studies which always robbed them of precious family time.

To my wife, Mercy, children Ruth

Claribel and Franklin II for their love and support.

To my Father John B. Nkansah for his

continued encouragement.

Table of Contents

List of Tables

List of Figures

ABSTRACT

As device dimensions scale into the submicron regime, stability and reliability effects become increasingly important. As charges are cycled in and out of the isolated floating gate through the tunnel oxide, the effects of oxide wear-out, charge trapping and defects affects the reliability of Flash Electrically-Erasable Programmable Read-Only Memory (EEPROM) devices.

In this research study submicron One-Transistor Flash (1T-Flash) devices were fabricated using advanced LOCOS isolation and I-line lithography. Device parametrics were extracted to characterize the tunnel oxide and Oxide-Nitride-Oxide inter-poly dielectrics. The results indicate that the tunnel /ONO oxides for fast and normal bits are comparable. The 1T-Flash EEPROM device reliability was studied and quantified with discrete devices and a large array of bits. Memory threshold voltages have been characterized for all bits and the overerased or fast bits identified. Results indicates that fast bit threshold voltages can vary from –0.2V to 0.5V which is 1.7V to 1V lower than that of a normal bit. The characterization showed that a fast and normal bit had similar data retention characteristics, and hence the fast bit was not inherently defective. Experiments were performed to the assess the reliability of fast bits. It was demonstrated that floating gate polysilicon grain size plays a significant role in fast bit generation. Finally, analytical model for programming of fast bit threshold voltage including floating gate polysilicon area and field enhancement effects was developed for 1T-Flash EEPROM .

Chapter 1

Introduction

1.1 Nonvolatile Semiconductor Memory Market

Memory devices, being an integral part of a computer, are continuously being developed and enhanced in most of today's R&D labs. As software applications become more demanding for memory to store large data and lengthy codes, memory units with smaller feature sizes, higher density, higher performance and reliability are a must. With this increased demand for high-density sub-micron memories the reliability issues are becoming more important. A Flash memory is one in which the entire memory array or sector can be erased very quickly instead of single bytes at a time as in traditional EEPROM. I will discuss in brief the Flash memory market trends, the floating gate concept, the Flash cell technology features and requirements, the basic Flash memory architecture and the fundamentals of device operation.

A Flash memory is a class of non-volatile solid state memory devices consisting of an electrically isolated storage element called the floating gate and a means of placing and depleting charges from the floating gate. I will discuss the details of this operation later in other sections. In recent years, high density Flash memories have been expected to replace some parts of the external storage device market of computers, because of its ruggedness, fast random access and low power dissipation. The 1T-Flash Electrically-Erasable Programmable Read-Only Memory (Flash EEPROM) is a one-transistor storage

device as compared to active SRAM (6 or 4 transistors) and DRAM (1 transistor and a Capacitor). Another feature of Flash EPROM's is its non-volatile nature in comparison to SRAM and DRAM, both of which are volatile because the data stored is lost when the power supply is discontinued. However for Flash EEPROM the data stored is stored indefinitely after the power supply is discontinued. As a result the Flash EEPROM combines the high access speed and density of a DRAM with the non-volatility of a Hard Disk Drive (HDD).

The Flash market arena can be categorized into two main areas, namely the "Stand alone" and the "embedded" markets. The stand-alone market, which is driven by memory storage density, is the primary technology driver. This market is dominated by industry giant Intel with about 54% market share followed by AMD with about 24%. The embedded market is very broad-based and mostly driven by customized applications and embedded microcontrollers. This affords customers ardent flexibility with superior field programmability and speedy time-to-market[1]. Providing an understanding of fast bits and their behavior will be invaluable to the field of NVM research since the Fowler-Nordheim based cell physics is here to stay as we head into the next millennium, where low voltage applications are no longer a device option but a necessity for survival in the increasingly competitive portable market space.

1.2 Nonvolatile Semiconductor Memory Devices

Nonvolatile Semiconductor Memory (NVSM) devices are semiconductor memory devices which can be read from or programmed like static or dynamic Random Access

Memory (RAM) or Read-Only-Memory (ROM) devices but can retain data even without external power. Since the proposal of the first floating gate device [2] and Metal-Nitride-Oxide-Silicon (MNOS) memory [3] in 1967, tremendous progress has been made in realizing a reliable, highly user-friendly Electrically Erasable and Programmable ROM memory. The NVSM devices can be subdivided into two major categories i.e. (1) the floating gate and (2) the floating trap devices. Figure 1.1 shows the family of NVSM devices and their applications for Electrically Programmable ROM (EPROM), Electrically Erasable PROM (EEPROM).

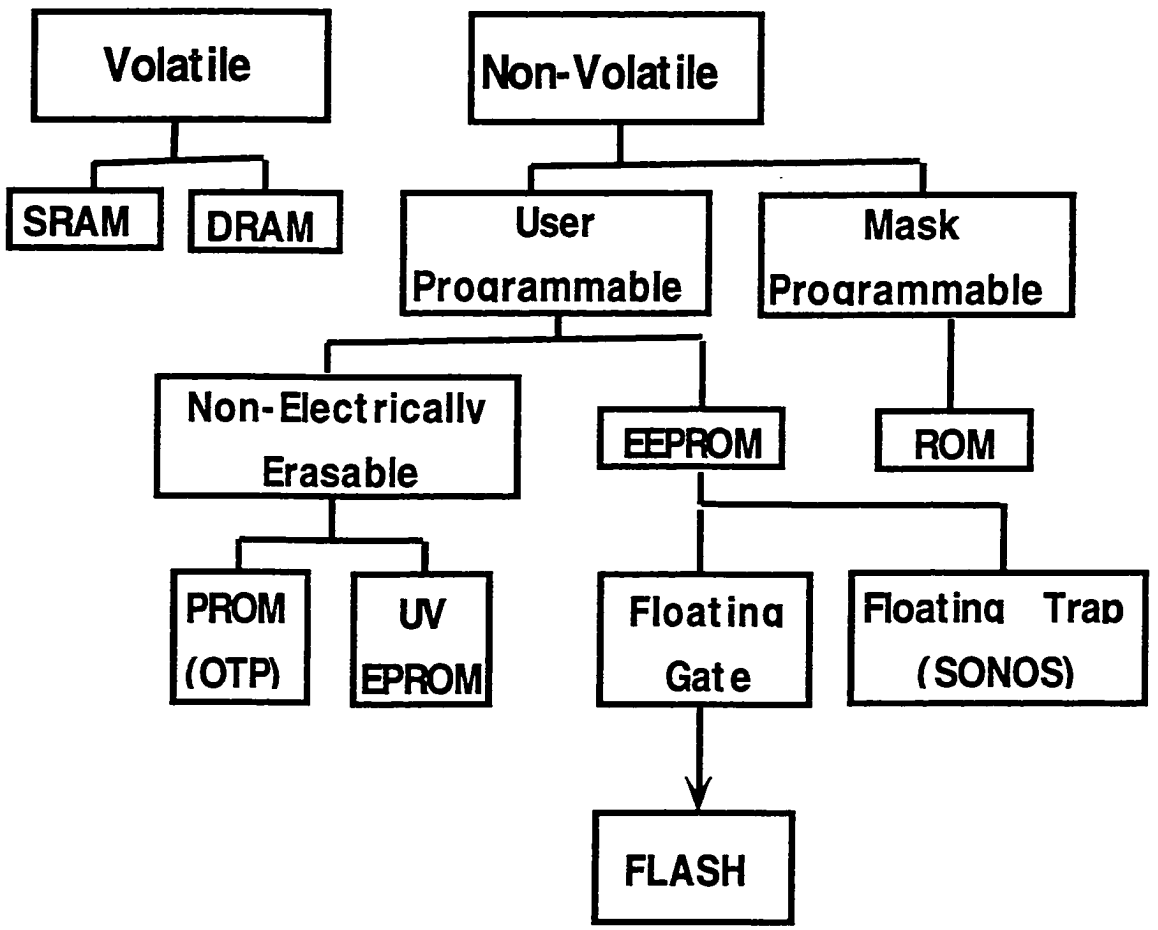


Figure 1.1: The Nonvolatile Semiconductor Memory Classification [64]

1.2.1 The Floating Gate Concept

To achieve the desired memory action in a nonvolatile semiconductor memory device, charge is introduced into the floating gate by electron tunneling or ejection from the drain edge to cause a threshold voltage shift of the MOS transistor [4]. The charge is stored inside the Floating Gate (FG), which is sandwiched between the two dielectrics as shown in Figure 1.2. The storage of negative charge gives rise to a positive shift in the transistor turn-on characteristics and puts the device in a high threshold state since more positive gate voltage will be required to turn it on. The charge retention requirements on the FG are usually ~10 years based on customer 1FIT requirements. The charge can be removed from the FG by applying a negative gate voltage to repel electrons into substrate or by using Ultra-Violet (UV) radiation whose photon energy can combine with electrons on FG to make them neutral [5], thus returning the device to its natural low threshold voltage state. Since the charge transport is from the silicon substrate through the tunnel oxide to the FG, the electric fields across the tunnel oxide have to be very high for F-N tunneling to occur. As a result the tunnel oxide is usually made very thin to enhance the field and thus the transport physics is primarily Fowler-Nordheim (F-N) tunneling. Charge transfer through the Insulator 2 (Oxide-Nitride-Oxide) stack is prevented by making the oxide thicker than Insulator 1 (tunnel oxide). For floating gate devices, there appears to be no intrinsic retention problem [6]: retention is limited by defect densities in the tunnel oxide, which are activated by temperature-bias stress and by large number of Erase/Write cycles [7]. Thus, many endurance failures are retention failures.

The Erase/Write endurance of floating gate devices is determined by four phenomena: (1) tunnel oxide breakdown,(2) gate oxide breakdown, (3) trap-up i.e. the increase of electron trapping in the oxide during programming operation and reduction in programming speed, (4) the degradation of the sense transistor characteristics. The failure rate due to tunnel oxide breakdown is determined by defects in the oxide. Memory cells, which contain defective tunnel oxides, will become increasingly leaky with cycling and will eventually breakdown. This phenomenon is responsible for almost all endurance failures in Floating-gate Tunnel Oxide (FLOTOX) memory devices [6,8,9]. Fowler-Nordheim tunneling (tunneling through a triangular barrier) is used to program these devices. The Channel Hot Electron (CHE) injection mechanism which is sometimes used for programming of Flash EEPROM's, results in the injection of high field electrons that causes gate oxide breakdown and the degradation in the memory transistors.

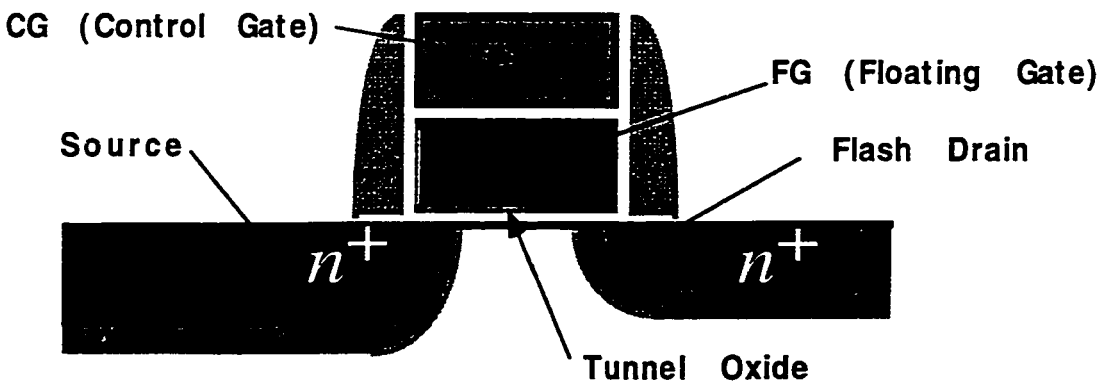


Figure 1.2: Cross-Sectional View of 1T-Flash Memory Cell

1.3 . Flash Cell Technology and Architectures

Summarized in the table 1.1 below is bitcell information compiled from numerous readings. As can be seen in table 1.1, the cell operational requirements and customer needs may necessitate a specific bitcell technology. The customer driven application are classified into three major categories namely: Embedded Flash, Low voltage Flash and Mass storage of "stand-alone" Flash. Not only must the characteristics be clearly defined, this must be combined with the desired customer application requirements to ensure the most appropriate bitcell choice.

BITCELL TECHNOLOGY					
CELL OPERATION	**ETOX**	**DINOR**	**AND**	**NAND**	**NOR**
Program / Erase Physics	CHI / FN	FN / FN	FN / FN	FN / FN	CHI / FN
Relative Cell Size	1	0.9	0.8	0.75	0.8
Read Operation	Random	Random/Page	Page	Page	Random/Page
Program Operation	Random	Page	Page	Random	Page
Average Block or Sector Size	64KB	64Kb	512KB	4Kb	64KB
Erase Time/Block or Sector	0.5s	1ms	1ms	10ms	0.5-1s
Programming Speed / Byte	10µs	17µs	15µs	16µs	9µs
Highest Internal Voltage	12V	10V	13V	20V	12V
Low Voltage Friendly	Difficult	Good	Good	Average	Difficult

Table 1.1: A Summary of Flash Memory Bitcell Requirements

1.3.1 Flash Memory Logic and Architecture

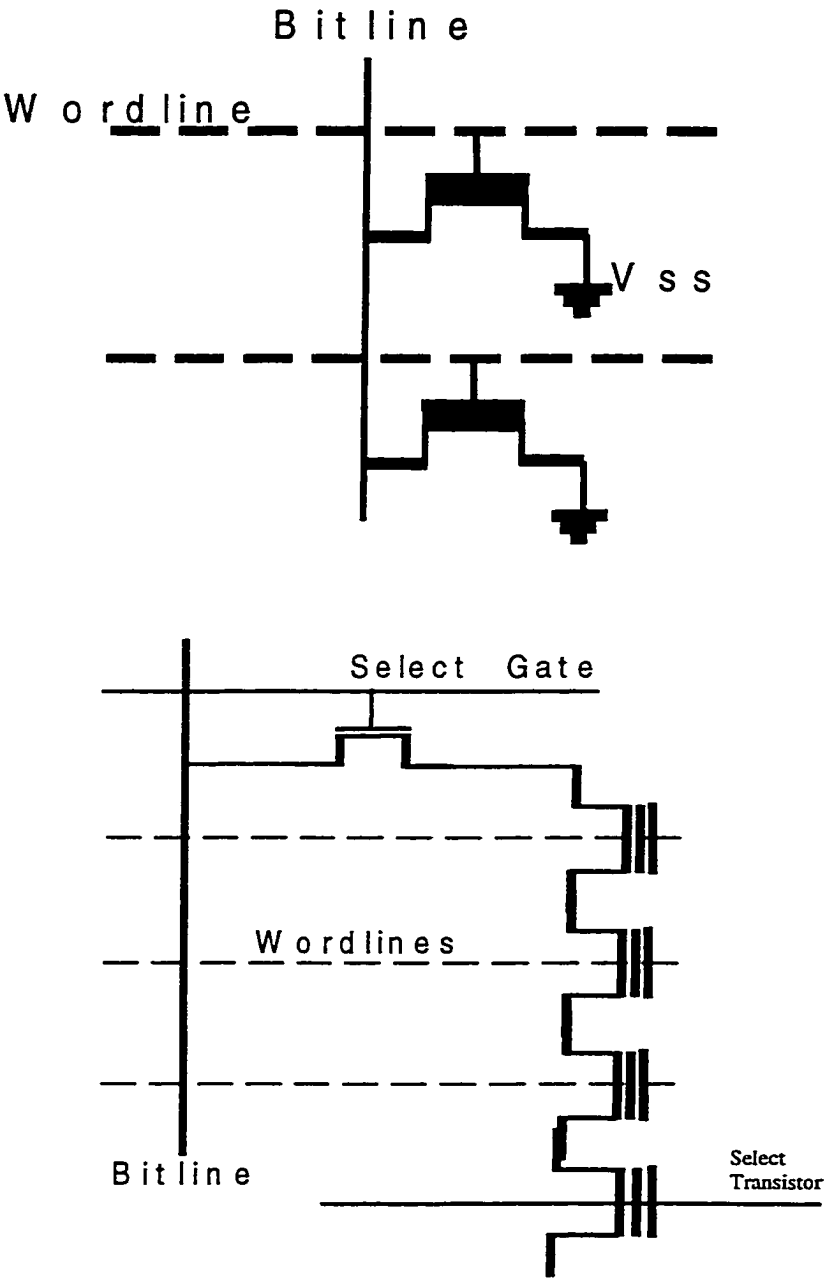
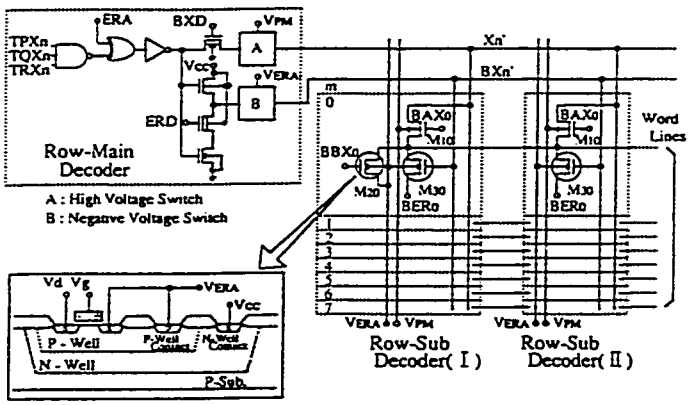


Figure 1.3: a) NOR Architecture; b) NAND Architecture for Flash Memory

The Flash memory chip has a multikilobit or megabit arrays that are surrounded by the decoders, current sensors, and other circuitry needed for the device operation. The wordlines and the bitlines are connected by X-Y decoders and are used to select a certain bit or groups of bits at a time. As shown in Figure 1.3a above for the NOR logic, which is used in the industry , the source nodes are shared by multiple bits. The column decoders access the bitlines (drain) and the row decoders access the wordlines or control gates. The programming of both types of logic can be done with byte (8 bits) or page, which is 256 bits wide. This flexibility which is very circuit dependent allows faster programming times to be realized by clever array sectoring. An example of the memory sectoring and decoding scheme is shown in Figure 1.4.

1.4 Fundamentals of Flash Device Operation

The Flash chip employs either a single supply voltage Vcc or a dual supply with a separate Vpp used for program (write) and erase. Other negative or positive voltages required are generated on chip using "charge pump' circuitry. The basic Flash bitcell can be operated by programming or writing logic level "1", erasing or writing logic level "0" and reading the stored information whiles maintaining the integrity of the stored charge. As shown in the Figure 1.5 below, the terminal biases influence the band structure of the non-volatile memory bitcell. In this thesis work, Fowler-Nordheim Channel Erase, and Drain-side programming will be used for characterization.



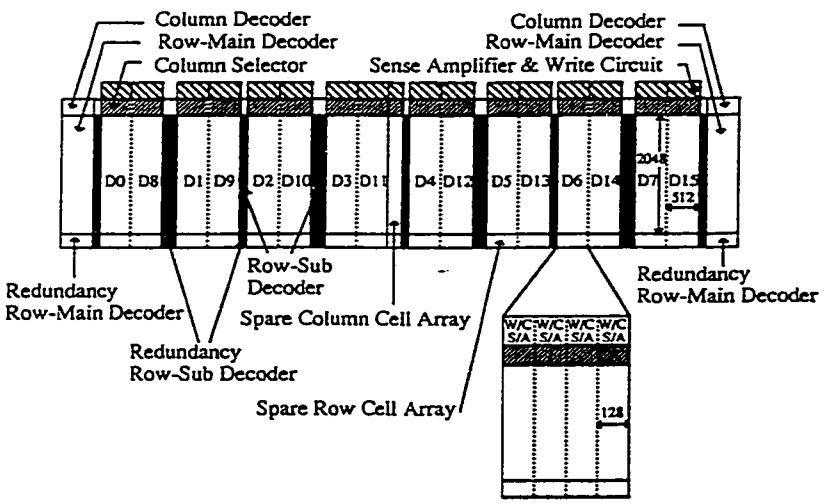


Figure 3: Array segmentation of 16Mb flash memory.

Figure 1.4: Flash Array Segmentation for a 2MB Memory

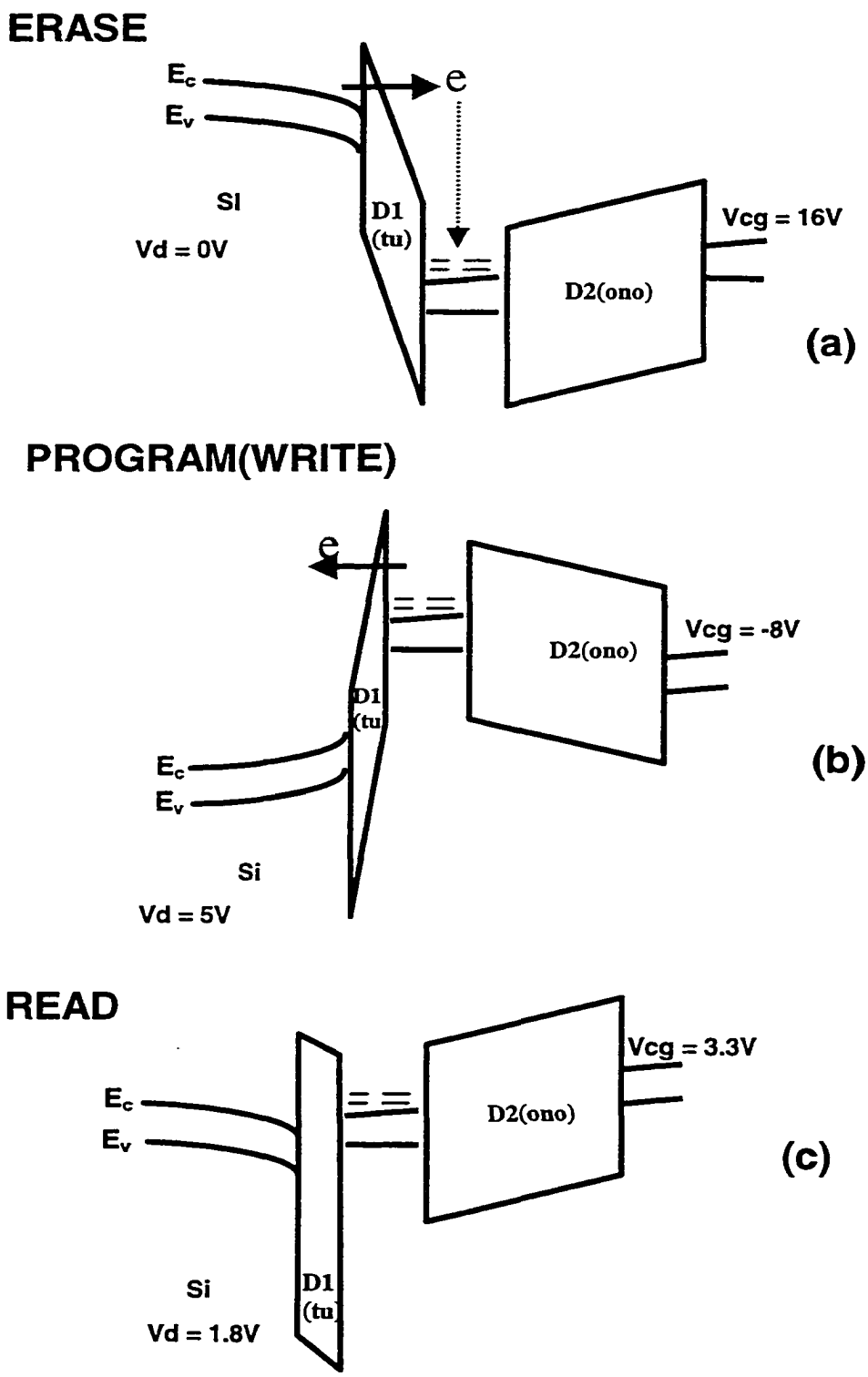


Figure 1.5: 1T-Flash Band Structure in Program, Erase and Read Operations

1.4.1 Programming (Write) Operation

There are two ways of writing to the memory cell: either by Fowler-Nordheim injection of electrons or by Channel-Hot-Electron injection. The physics of these mechanisms will be discussed later. As shown in Figure 1.6, the electron movements can occur over a variety of surfaces: whole channel area, drain side or via a special tunneling window. In this work my characterizations will be conducted using F-N tunneling operation (f) in Figure 1.6 for programming and F-N tunneling operation (a) in Figure 1.6 for erase.

1.4.2 Erase Operation

In all Flash and EEPROM-based non-volatile memories today, Fowler-Nordheim mechanism is always used to erase the bitcell. Again as shown in Figure 1.6 this can take place over various areas of the device channel regions. Generally in a Flash array this erase operation is performed on a block or sector which is a large subset of the array. Hence for 1T-Flash EEPROM byte erasability has been sacrificed for bulk erase which yields higher array bit density.

1.4.3 Read Operation

To read the memory content to determine whether a cell is at logic level "0" or "1", the drain(bitline) is biased at a low voltage ~1V and the control gate (wordline) of the row is raised to a positive potential ~3.3V. When a programmed cell is selected, the

control gate voltage will be larger than nominal memory threshold voltage, which is about 1-1.5V, thus causing the drain I_{DS} current to begin flowing. This source/drain current is then detected by the sense amplifiers which translates this signal into a logic "1". Similarly, if an already erased cell is selected, since its threshold voltage of 5-6V is much larger than the 3.3V on the control gate, the erased transistor (cell) does not turn on. As a result no I_{DS} current is detected by the sense amplifiers, which interprets this as logic "0". The read operation can be performed in a random, serial or page mode.

In this research work, F-N tunneling will be used for both Programming and Erase. This research presents impact of F-N operation on the reliability of the programmed bits. The research focuses on F-N based programming and erase because the trend of the industry towards low power and portable devices necessitates the use of programming physics, which requires lower currents for programming, and erase.

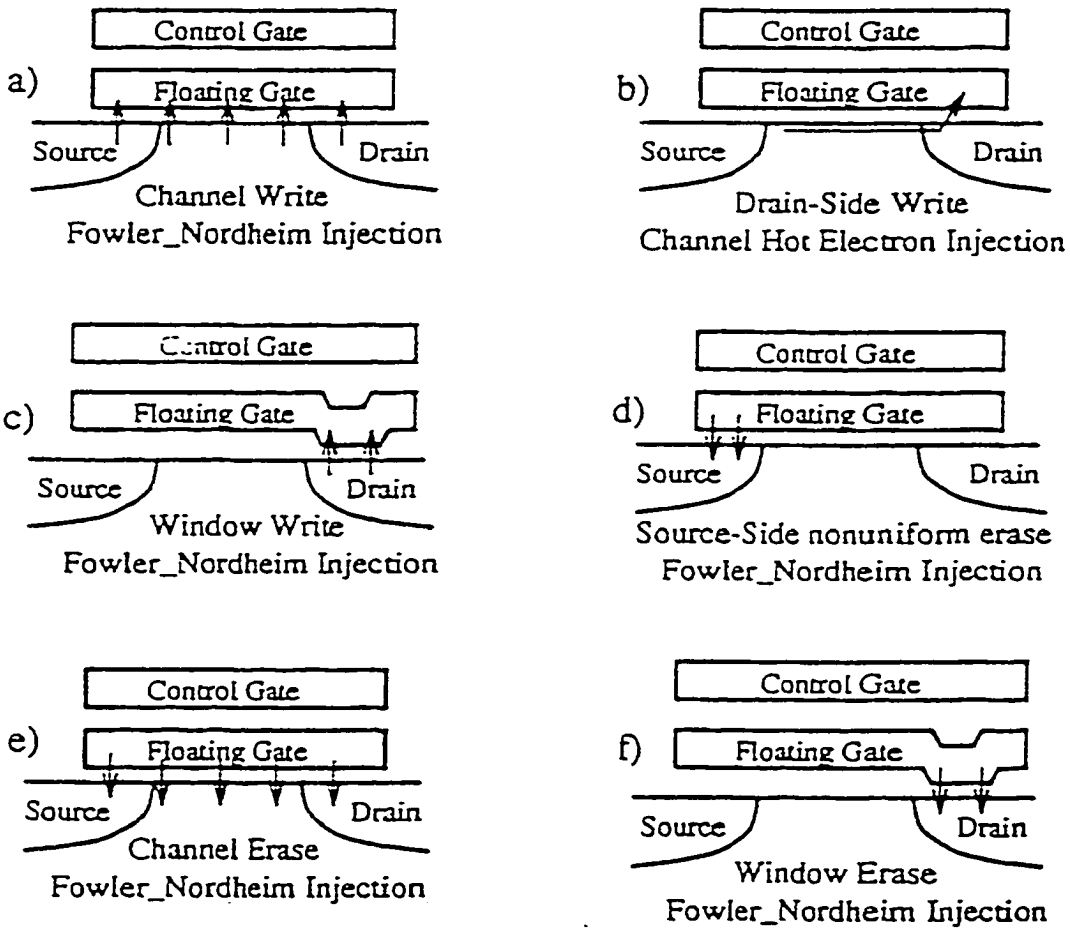


Figure 1.6: Various Types Program and Erase Operations for ETOX and Flash structures

1.5 Reliability Issues in Flash EEPROMs

The reliability issues associated with NVM is mostly determined by the properties of the non-volatile low threshold voltage operation i.e. (ETOX- Erase and F-N-Programming). The general criteria are the ability to move charges through the tunnel oxide dielectric with as little damage as possible, and also to keep these charges on the floating gate for as long as required. The Flash reliability is characterized by the cell endurance or window closure, disturb immunity, data retention and over-erase immunity.

1.5.1 Endurance

This is also commonly referred to as window closure. This is a cycling procedure used to determine the long-term programmability and erasability of the Flash cell. An ideal memory should be able to be written and erased over infinite cycles, however in reality this cycling is limited due to device degradation. As shown in the Figure 1.7 below the endurance curve have three distinct regions typically observed. The regions are characterized by the window opening which takes place at early stages of cycling, followed by the plateau and severe window closing.

1.5.2 Disturb Immunity

The most common Flash bitcell disturbs that makes device reliability requirements extremely stringent are: Bitline, Wordline and Read disturbs.

Bitline (Drain) Disturb: In an array environment, the unselected cell can experience stress and disturb due to the voltage at the drain when an adjacent cell is being written as shown in Figure 1.8 below. This results in an undesired erasing of the unselected cells during programming of the other cells on the same bitline.

Wordline (Gate) Disturb: As a result of an array environment, when the common wordline (control gate) of the array is biased during the write operation as shown in Figure 1.9, some charge on the floating gate of a written cell can be lost to the control gate or charges can be injected into the floating gate from the substrate. This could result in an undesired erasing of the unselected cell.

Read Disturb: This is simply an undesired programming of a neighboring bit during the read operation. Again this could result in the changing of the memory content inadvertently, thereby causing erroneous data output on the bitlines, as shown in Figure 1.9.

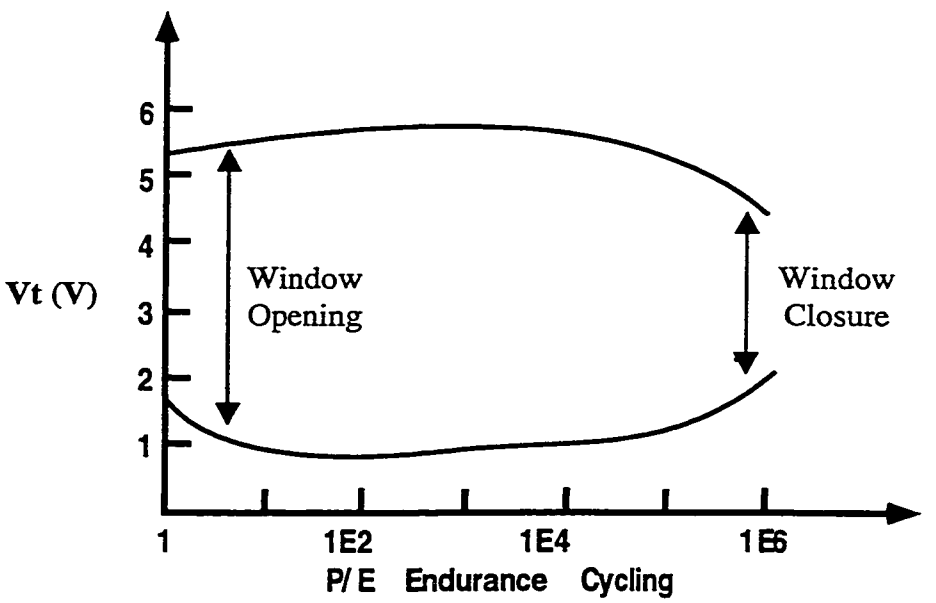


Figure 1.7: Flash Memory Threshold Voltage Write/Erase Endurance Cycling

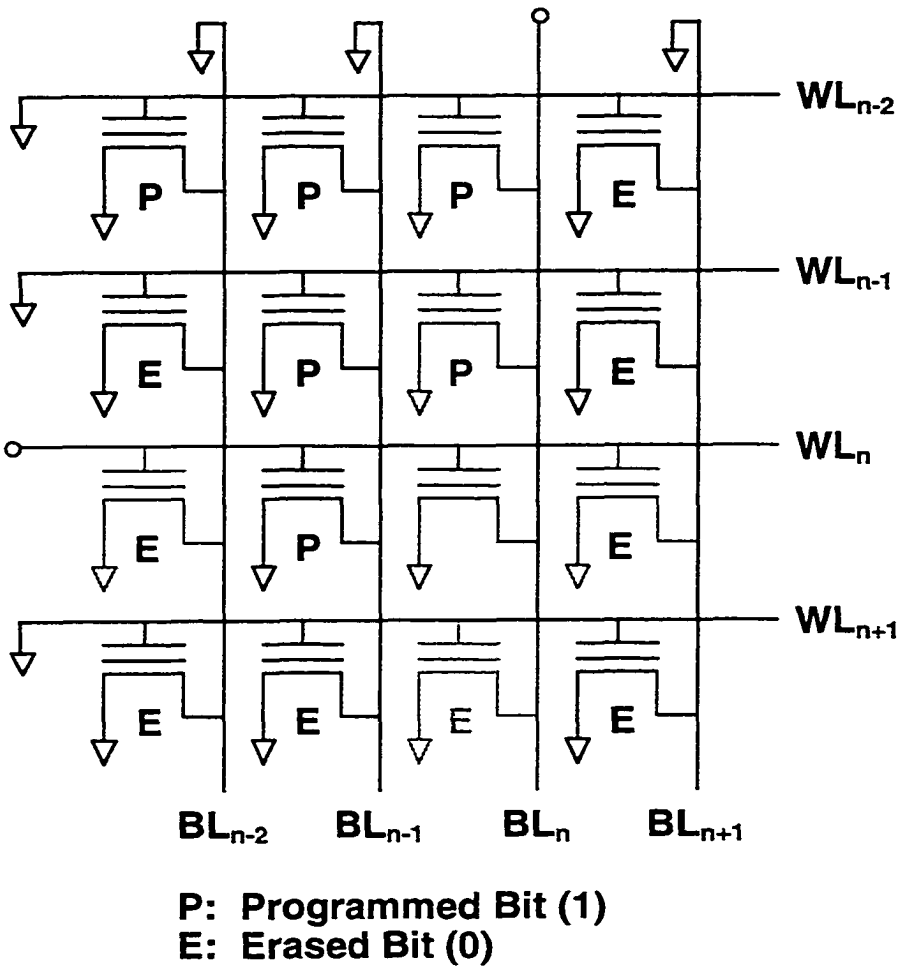


Figure 1.8 : Schematic Description of Flash Gate and Drain Disturb Mechanisms

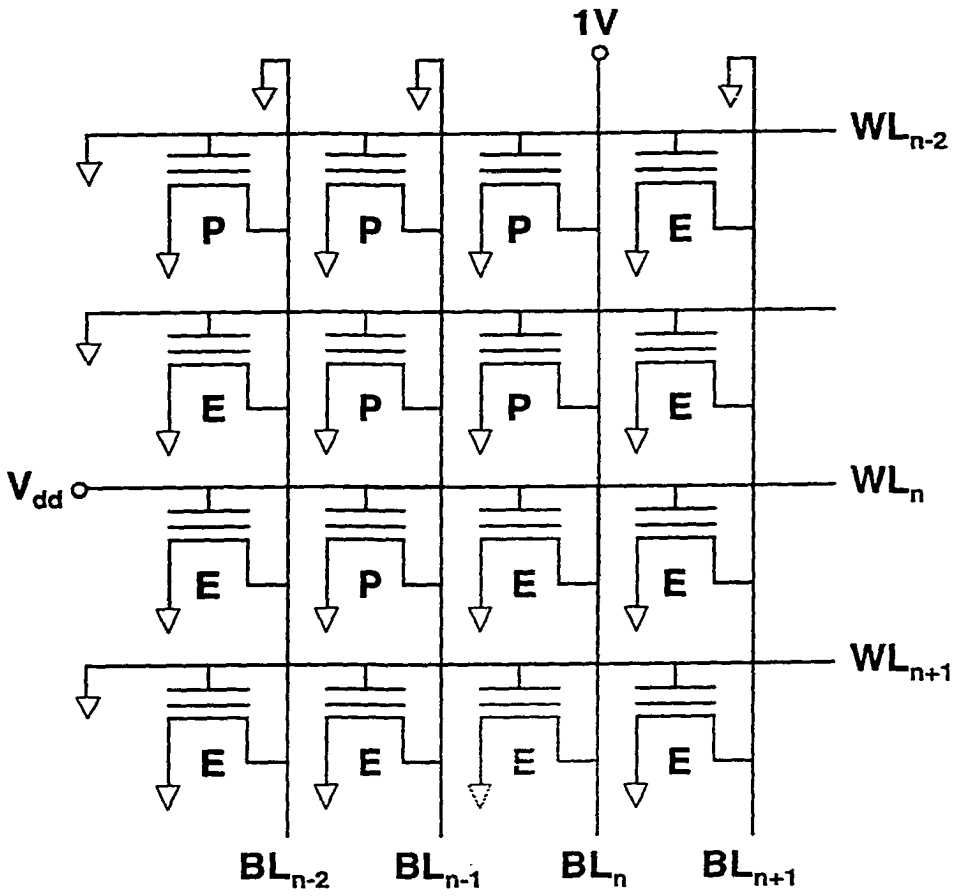


Figure 1.9 : Schematic Description of Flash Read Disturb Mechanism

1.5.3 Data Retention

Data retention is characterized by the loss of charge from the floating gate of a programmed device. As a result of the charge loss into the substrate or the control gate, the memory threshold voltage of the cell decreases. If the memory threshold voltage drops

below Vcc, the cell can no longer be read as a high and the stored data is lost. There are three distinct phases of charge loss namely: initial-phase < 10min, slow-phase usually 10-100hours and finally the non-saturating phase > 100hours. It has been reported that the initial-phase accounts for about 30% of the charge loss, slow-phase accounts for ~30-45% and non-saturating phase ~30-50% [10].

1.5.4 Over-erase or Over-Programming

ETOX researchers have reported in the past about the issues of "over-erase" [11]. This is the phenomenon by which after the threshold voltage (V_t) lowering operation, the memory array has a broad V_t distribution as shown in the Figure 1.10a below. This has been attributed to the way Flash V_t - lowering occurs i.e. sector or block, which involves multiple bits. It has been found also to be a non-uniform phenomena, thus making the absolute V_t control of all bits in a given array very difficult, since it is independent of the basic cell physics. Some bitcell threshold voltages could get so low that it could go negative thus resulting in a depletion mode device. This leaky device when in series with a good bit could send the wrong message to the sense amplifiers regarding the true value of the good bit as shown in Figure 1.10b.

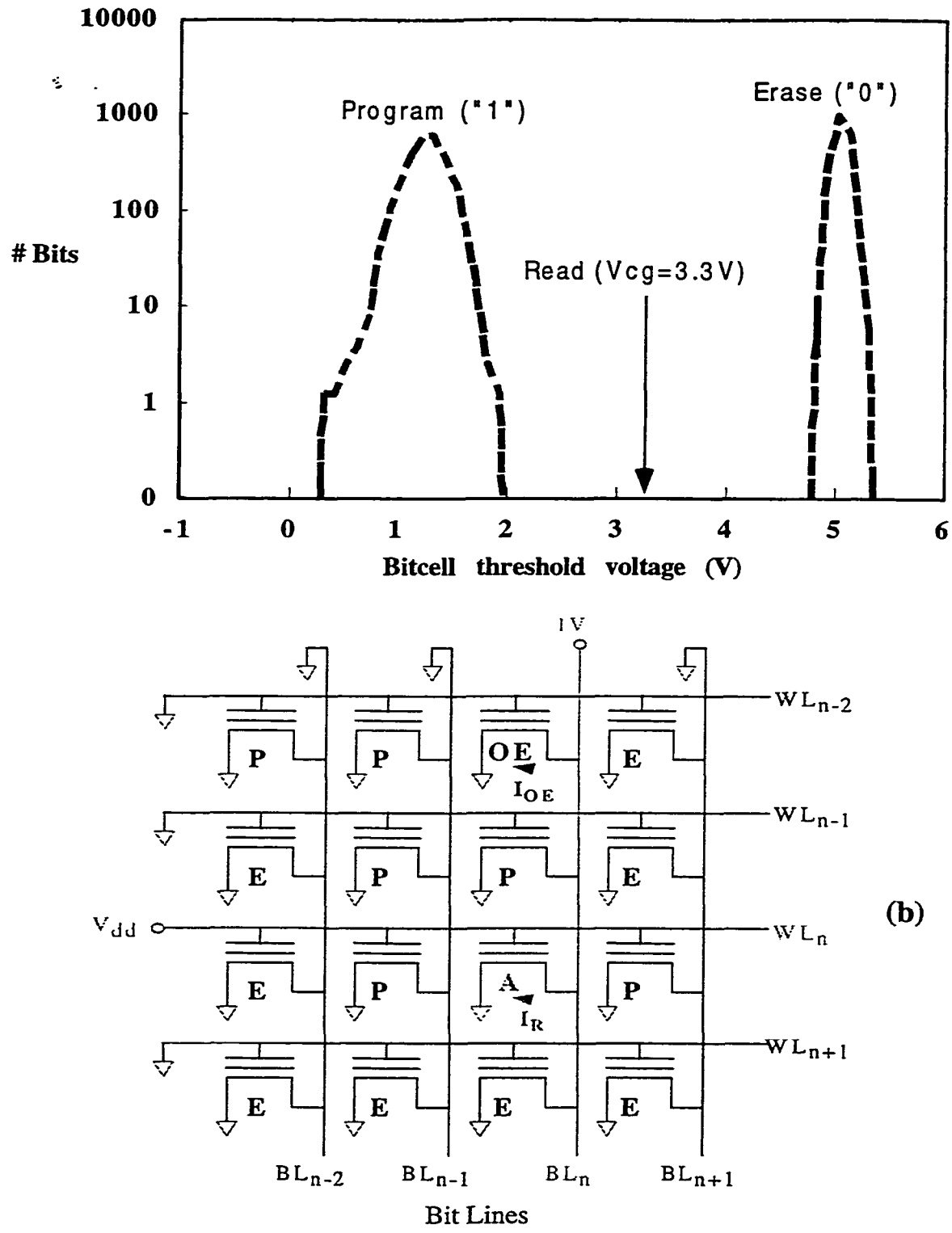


Figure 1.10: 1T-Flash Over-erase Vt Distribution and Bit Leakage

1.5.5 Non-Uniform Erase

As discussed earlier the electron injection during erase (ETOX) and program(F-N) takes place through a small area. Since this region of operation is so small, any statistical variations among cells can result in varied program/erase characteristics. It is also widely known that since the Fowler -Nordheim tunneling current is a strong function of tunnel oxide thickness, any non-uniformity and reliability could lead to memory failures. However these effects coupled with Flash drain profile types and Floating Gate structure has not been comprehensively quantified yet for the Fowler-Nordheim based devices.

1.6 Scope of Dissertation

An over-erased bit or over-programmed bit interferes with the read and programming of bits on the same column. Hence this leads to an erroneous data detection by memory sense amplifiers.

The scope of this dissertation is on the reliability of 1T-Flash EEPROM devices. I have explored two fields in this arena: (1) an analytical model of the fast programming bit characteristics (2) novel device structure with associated floating gate polysilicon engineering and technology. By analyzing the behavior of the fast programming bits, I have gained a better understanding of the effects of floating gate polysilicon micro-structural effects on the reliability of 1T-Flash NVSM devices.

This research encompassed a range of experimentation, device design, layout, characterization and modeling. Aggressive 1T-Flash bitcells were defined and laid-out

using Mentor Graphics IC-graph. The bitcell was then implemented in a parametric test structure for device characterization. The bitcell was also implemented in a 2MB 1T-Flash memory array which was employed in the extensive array memory reliability studies. The devices were then fabricated using the process flow detailed in Appendix-A. The 1T-Flash transistor was optimized for improved short-channel margin, programming and erase speeds. To achieve these the following criterion had to be met:

1) Short-channel margin greater than or equal to 100 nanometers
2) 1T-Flash programming speed of less than 1 mill-second (1msec)
3) 1T-Flash sector erase speed of 1 second (1sec)

After the devices were design to achieve the above specifications, they were then extensively characterized. Both normal and fast programming bits were quantified and intensively characterized. These included pre and post cycling bases, endurance cycling, data retention.

After establishing a knowledgeable baseline of fast programming bit behavior, the emphasis shifted to focus on understanding how the bits can be modulated. Experimentation centered on flash drain junction optimization, tunnel oxide types and floating gate polysilicon grain microstructure. It was subsequently established that the fast programming bits could be modulated by the floating gate polysilicon microstructure. This was further quantified by fabricating 1T-Flash devices with various floating gate grain sizes.Finally a compact model was developed to predict the memory threshold voltage of a 1T-Flash fast programming bit as a function of the floating gate polysilicon grain size.

Chapter 2

Theory and Model Development For 1T-Flash EEPROM Programming

The purpose of this chapter is to establish a theoretical background for the reliability study of 1T-Flash electrically erasable programmable read-only-memories (Flash EEPROM). Fowler-Nordheim injection of electrons and channel-hot-electron injection are operations that are frequently used for the programming of Flash memory arrays. In this research Fowler-Nordheim injection is used for both programming and erasing of the memory cell. Electrons are the preferred carriers for charge transport in these operations whereas holes are avoided because of their low mobility and high trapping rate in the oxide. Although the device is designed to operate under bias conditions that favor electron injection, it has been demonstrated that some reduced level of hole injection is difficult to avoid. For that reason, in addition to Fowler-Nordheim tunneling, hole generation due to band-to-band tunneling injection mechanisms will be discussed in this chapter.

2.1 Fowler-Nordheim Tunneling

As discussed in [12], there are several possible conduction processes in insulators, including Schottky emission, Frenkel-Poole emission [13], tunnel emission [14]; space-charge limited conduction, ohmic conduction and ionic conduction. It has been found that the primary conduction process in a good thermal SiO_2 film is due to Fowler-Nordheim (F-N) tunneling of electrons [14] through a triangular barrier into the oxide conduction band, as

shown in Figure 2.1 below. Therefore, this section is devoted to the theory of the Fowler-Nordheim tunneling process. Fowler-Nordheim tunneling of holes from the silicon into the valence band of the oxide is possible but much less likely, because of the higher energy barrier for holes compared to that for electrons.

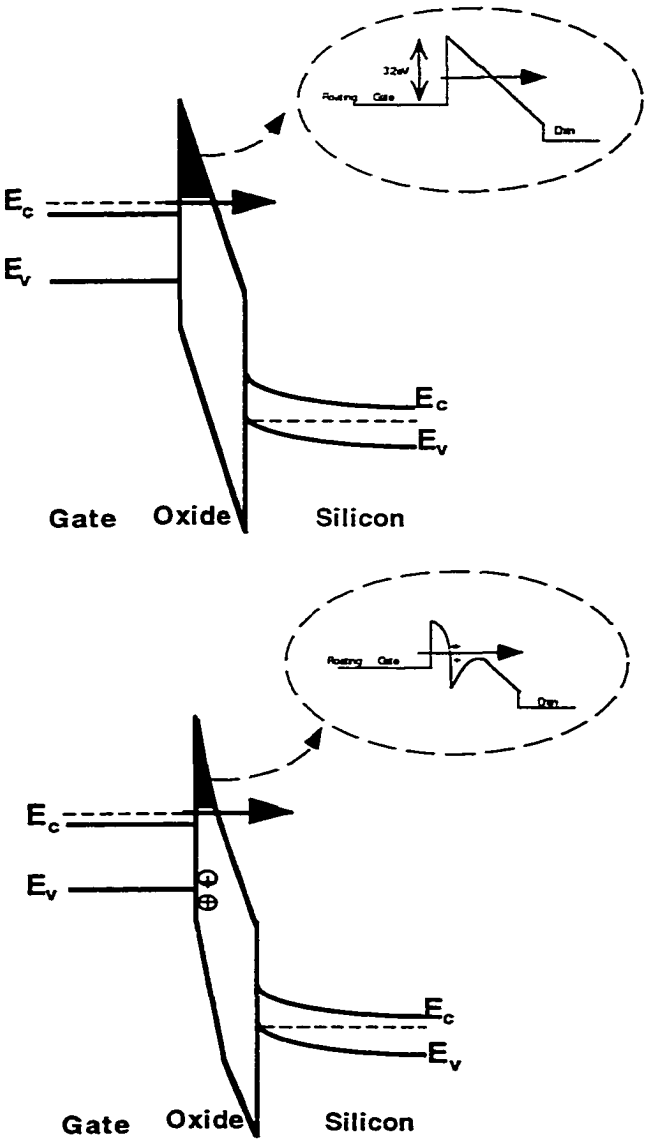


Figure 2.1: Effects of Barrier Height and Traps on Electron Tunneling

2.1.1 Quantum Mechanical Model

In this section we will derive the Fowler-Nordheim equation that relates the current density through the oxide to the electric field in the oxide. This will lead to a better understanding of some programming characteristics of the Flash EEPROM devices and the associated problems. The Fowler-Nordheim tunneling mechanism is a quantum mechanical process [14]. The current density due to Fowler-Nordheim tunneling in Figure 2.1 can be expressed analytically using the free-electron gas model for the metal and the Wentzel-Kramer-Brillouin (WKB) approximation[15] for the tunneling probability.

Starting from the Schrodinger equation:

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2\mu}\nabla^2\psi + V(r)\psi \quad \text{----------} \quad (2.1)$$

where ψ is the wave function solution, $h/2\pi$ Plank's constant, μ the reduced mass, r the coordinate, t the time, the solution is in the form of:

$$\psi(r,t) = A\, e^{iW(r,t)/\hbar} \quad \text{----------} \quad (2.2)$$

For the solution to be an eigenfunction:

$$\psi(r,t) = u(r)\, e^{iEt/\hbar} \quad \text{----------} \quad (2.3)$$

where E is the energy of the electron, W(r,t) should then consist of two terms:

$$W(r,t) = S(r) - Et \quad \text{----------} \quad (2.4)$$

For the one-dimensional case, when the expected solution is substituted into the Schrodinger equation 2.1, the following two equations are obtained:

$$\frac{d^2u}{dx^2} + \kappa^2(x)u = 0 \text{ ---------- } (2.5)$$

$$\frac{d^2u}{dx^2} - \kappa^2(x)u = 0 \text{ ---------- } (2.6)$$

where

$$\kappa(x) = +\frac{1}{\hbar}\{2\mu[E - V(x)]\}^{\frac{1}{2}} \quad V(x) < E, or \text{ -- } (2.7)$$

$$\kappa(x) = +\frac{1}{\hbar}\{2\mu[V(x) - E]\}^{\frac{1}{2}} \quad V(x) > E, \text{ ---- } (2.8)$$

where V(x) is the potential barrier. When $u(x)=Ae^{iS(x)/h}$ is inserted into the Schrodinger equation, the following equation is obtained:

$$i\hbar S''(x) - S'^2(x) + \hbar^2\kappa^2 = 0, \text{ ------- } (2.9)$$

where the prime denotes differentiation with respect to x. the variable S(x) can be expressed in terms of a power series of (h):

$$S = S_0(x) + \hbar S_1(x) + ..., \quad \text{------- } (2.10)$$

and when substituted into equation (2.9), the resulting equation is valid only if:

$$iS''_0(x) - 2S_0'(x)S_1'(x) = 0 \text{ ---- } (2.11)$$

$$-S'^{2}_{0}(x) + 2\mu(E - V) = 0 \text{ --------(2.12)}$$

Integrating equations 2.11 and 2.12 gives:

$$S_0(x) = \pm\hbar\int^{x}\kappa(x')dx', \text{ and} \quad \text{----------(2.13)}$$

$$S_1(x) = \frac{1}{2}i(\ln[\kappa(x)]) \quad \text{--------------(2.14)}$$

Thus the approximate equations for the electron wave function are obtained as:

$$u(x) = A\kappa^{-\frac{1}{2}}\exp(\pm i\int^{x}\kappa dx) \rightarrow V < E, \text{ and} \quad \text{----(2.15)}$$

$$u(x) = A\kappa^{-\frac{1}{2}}\exp(\pm\int^{x}\kappa dx) \rightarrow V > E \quad \text{------(2.16)}$$

The WKB approximation is useful for potential energies that change so slowly that the momentum of the particle is reasonably constant over many wavelengths; in other words, the ratio of hS_1/S_0 is small. This is assured if the local de Broglie wavelength λ, which relates the energy of the particle to the frequency of its wave, is such that:

$$\frac{\lambda}{4\pi}\left|\frac{d\kappa}{dx}\right| << \kappa \quad \text{------------------(2.17)}$$

This condition is violated near the turning points of the classical motion where V(x)-E, k and k are all zero, and the wavelength is infinite. These turning points are the two ends of the barrier where the tunneling particle/wave enters or leaves the potential barrier, as shown in Figure 2.2 below.

The tunneling probability of an electron through a potential barrier is the square of the wave-function ratio between the incoming and the exiting waves, and is expressed as:

$$P = \exp\left(-2 \int_{x_1}^{x_2} \kappa(x) dx \right) \text{----------(2.18)}$$

where x_1 and x_2 are the tunneling points. Using this derived WKB approximation and the free-electron gas model for metals, the following expression for the Fowler-Nordheim current density at 0K is obtained:

$$J(T = 0K) = (q^3 E_{ox}^2 / 8\pi\hbar\phi_B) x \exp[-4(2m)^{1/2} \phi_B^{3/2} / 3\hbar q E_{ox} \text{ --(2.19)}$$

where J is the current density, m the free electron mass, q the elementary charge, E_{ox} the electric field in the oxide and ϕ_B the potential barrier height.

For most applications , a simplified temperature-independent equation serves as a good approximation :

$$J = \alpha_{FN} E^2 \exp [-\beta_{FN} / E_{ox}] \text{ ----------(2.20)}$$

$$\alpha_{FN} = q^3 E^2 / 8\pi\hbar\phi_B) \text{ ----------(2.21)}$$

$$\beta_{FN} = -4(2m^*)^{1/2} \phi^{3/2} (v(y)) / 3\hbar q_B \text{ ----------(2.22)}$$

are the Fowler-Nordhiem constants , and are roughly 6.3E-7 A/v^2 and 2.21E+8 V/cm respectively.

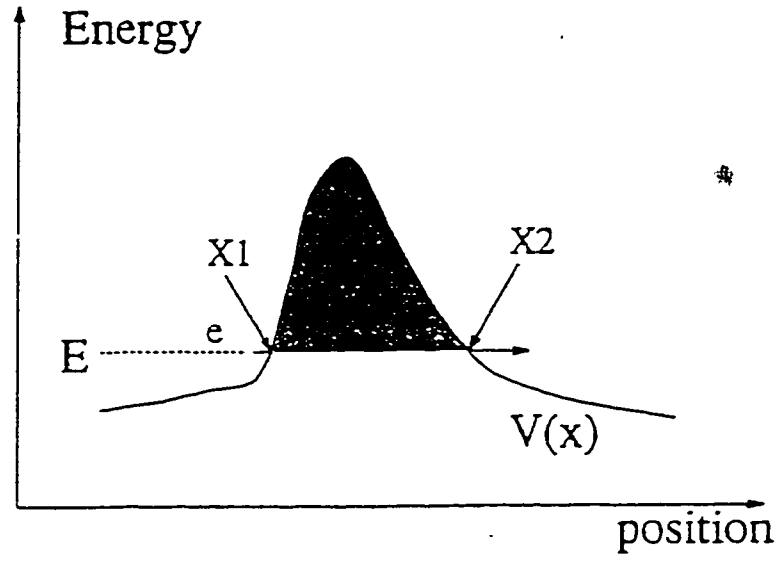


Figure 2.2: The Tunneling Probability Process between X_1 and X_2

2.2 Band-to-Band Tunneling

It is well known that electrons are capable of tunneling from the silicon valence band to the silicon conduction band if the band bending is more that the silicon bandgap, as shown in Figure 2.3. Band to Band tunneling is a quantum mechanical process, and can be modeled similarly to the Fowler-Nordheim tunneling. Using equation.(2.18) and the corresponding potential barrier for band-to-band tunneling, the tunneling probability for electrons can be calculated analytically[16]:

$$P(E_{si}) = \frac{q^2 m^{1/2} E_{si}^2}{18\pi\hbar^2 E_{gap}^{1/2}} \exp\left(-\frac{\pi m^{1/2} E_{gap}^{3/2}}{2\hbar q E_{si}}\right) \text{----------} (2.23)$$

Where m is the electron mass, E_{si} the electric field in silicon and E_{gap} the silicon bandgap. Since for each electron tunneling out of the silicon valence band a hole is left behind, the tunneling probability is thus the same as the hole generation rate. Therefore, the hole current corresponds to the volume integral of the generation rate in the entire silicon. As equation 2.23 suggests, this current is proportional to the electric field, which is a direct measure of the band bending. The empirical formula used for band-to-band tunneling is [17]:

$$J = \alpha_{bb} E_{si}^{\ 2} \exp[-\beta_{bb} / E_{si}] \text{ ---------- (2.24)}$$

As it can be seen from Figure 2.3 (b) when an electron leaves the valence band and tunnels into the conduction band, a hole is left behind. The energy of this hole depends on the degree of band-bending. The hole generation at point b, for instance, is more energetic compared to the hole generated at point a in Figure 2.3 (b). The excess energy of the hole will be transformed into kinetic energy and cause hot carriers. Another possible tunneling process for an electron is the trap-assisted band-to-band tunneling, which occurs when the electron at an interface state tunnels into the silicon conduction band, and the empty state is subsequently filled by an electron from the silicon valence band as shown in Figure 2.3. This process does not necessary require a band bending of 1.2eV, and becomes significant when the density of interface states are high.

2.2.2 Location of Band-To-Band Tunneling in MOSFETs

The band-to-band tunneling phenomenon is expected to occur in silicon where there is a sufficient band bending as seen in Figure 2.3 (a). In a MOSFET structure the location with the highest tunneling probability in silicon is near the Si/SiO_2 interface right around the junction transition region because of the combined effect of the gate and the junction biases. The band-to-band tunneling current can be experimentally observed in a MOS field-effect transistor as the off-state junction leakage current (see the region below valence in Figure 2.6) when the junction is reversed biased. The lateral field in the depletion region combined with the vertical field under the gate gives rise to the tunneling process.

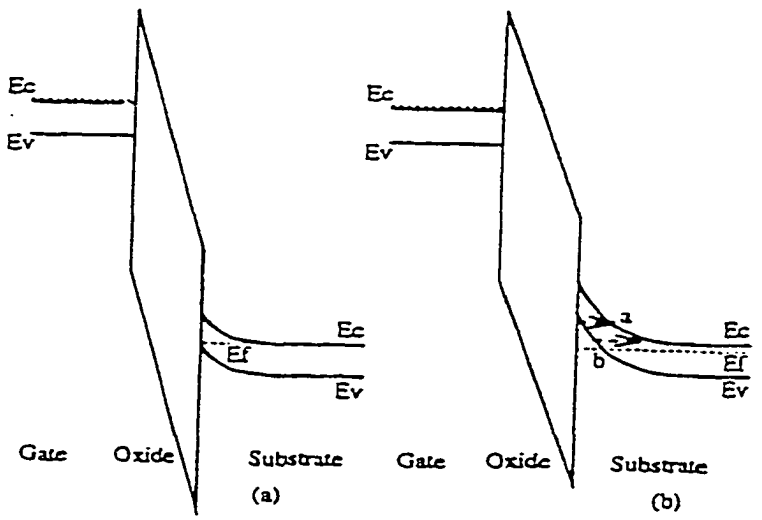


Figure 2.3: Energy Band Diagram of Band-To-Band Tunneling

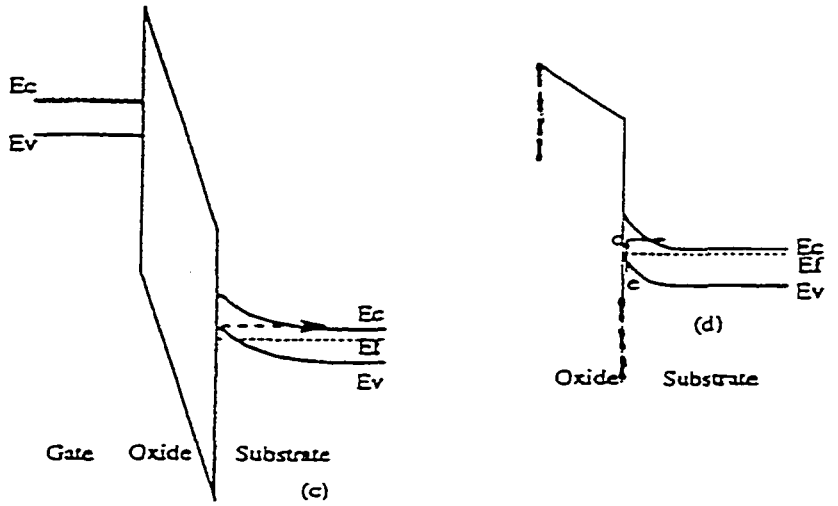


Figure 2.3: Energy Band Diagram of Band-To-Band Tunneling

The generated holes are swept away by the lateral field from the point of generation, leading to a constant supply of holes (see Figure. 2.5). While the electrons are collected by the junction, the holes are drifted to the substrate, thus contributing to the junction leakage current. Because of the critical role of the gate voltage in creating the necessary high field, this current is called the Gate-Induced Drain Leakage (GIDL). There have been several attempts [16,18-20] to model and simulate the band-to-band hole generation and the associated junction leakage current. The common approach has been to simulate the electric field distribution throughout the silicon using the Poisson equation, and determine the

tunneling probability at a given point. It's integration over the entire silicon will give the total leakage current that flows into the substrate. However, that approach is not adequate because it does not take into account the transport mechanism, which is quite complicated. For instance, the constant removal of carriers from the point of generation is likely to cause deep depletion, which further enhances the band-to-band tunneling process. This phenomenon also alters the electric field distribution, and thus the transport process. To correct this, a self consistent algorithm that incorporates the carrier generation by band-to-band tunneling into the Poisson and the continuity equations, should be employed. However this approach is computational intensive and a satisfactory convergence is not always attainable with today's computationally intensive computers. It should be noted that the locations of the peak vertical and lateral electric fields in silicon depend strongly on the doping concentration. Simulations show that peak lateral fields in silicon always occurs right under the gate where the doping concentration is around $1E19cm^{-3}$. The peak vertical electric field in silicon , on the other hand is approximately one third of the local oxide field, which is the ratio between the silicon permitivity and the oxide permittivity. For high doping concentrations the electric field does not penetrate much into the silicon and the band bending is very small. In contrast, for the same gate voltage, a wide depletion region is formed with a large band bending in the low doping region even though the electric field in silicon is lower. Figure 2.4 shows the electric field in silicon under the gate oxide at a distance of 50Å from the interface of a MOS capacitor as a function of the doping concentration.

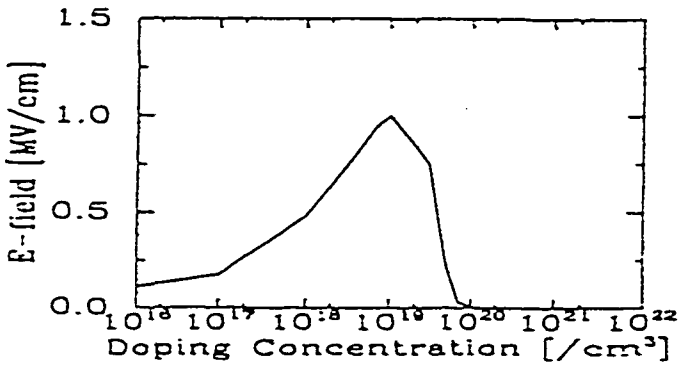


Figure 2.4: Flash Drain Junction E-field Dependence on Doping Concentration

This distance of 50 Å has been chosen for the simulation to avoid the mis-leading high electric field right at the interface that cannot penetrate into the silicon to create the required band bending. It is clearly seen that the peak vertical field is located at a doping concentration of around 1E19cm^{-3}. After the holes are generated at the junction corner, where the doping concentration is in the rage of (1E18-1E19) cm^{-3}, most of them are drifted into the substrate by the lateral field in Figure 2.5, but some of them may be injected into the oxide due to the vertical field. Only those, which have high enough energy to surmount the Si-SiO_2 barrier, are injected into the tunnel oxide. The rest are reflected by the potential barrier and eventually find their way into the substrate. At very high junction potentials, a large number of holes may be generated by the avalanche breakdown mechanism, which is

characterized by a sudden rise in the current as shown in Figure 2.6. Again some of these holes may be injected into the oxide, and the resulting hole trapping leads to a distortion of the oxide barrier and enhanced tunneling.

2.2.3 Band-to-Band Tunneling in Flash Cell

As mentioned above the main cause for the band-to-band tunneling and the subsequent hole generation is the high electric field in the junction transition region. Since high electric fields are needed to have substantial Fowler-Nordheim tunneling of electrons from the floating gate into the substrate or into the junction diffusion areas, and the vice versa, large voltage differences are applied to the terminals (source, drain and control gate). During the source-side Fowler-Nordheim erase of a floating gate Flash device, a high voltage is applied to the source terminal and the control gate is grounded with the substrate. The floating gate typically assumes a potential of roughly 0.3V if the device is programmed and the drain is kept floating, while the bias on the source junction must be high enough to produce an electric field of at least 10MV/cm in the tunnel oxide which is required for the Fowler-Nordheim injection of electrons from the floating gate into the source diffusion. This bias condition leads to the hole generation due to substantial band-to-band tunneling right under the Si/SiO_2 interface, especially near the junction corner inside the source diffusion where the doping concentration is around (1E18-0E19) cm^{-3}. The fact that both the vertical and lateral fields contribute to the silicon band bending needed for the band-to-band tunneling underlies the significance of the junction profile for minimizing the hole generation.

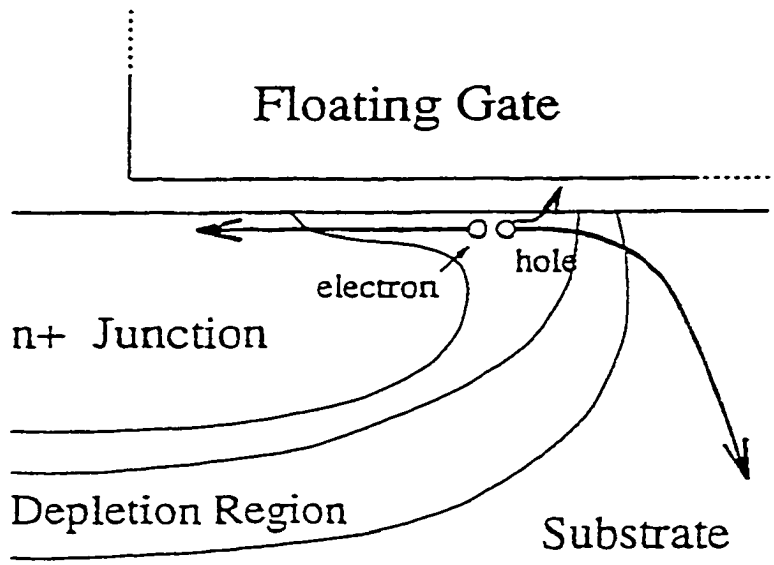


Figure 2.5: The generation of a hole (due to band-to-band tunneling of an electron) and its trajectory in silicon of a MOSFET with a gate biased in strong accumulation and the drain reverse biased.

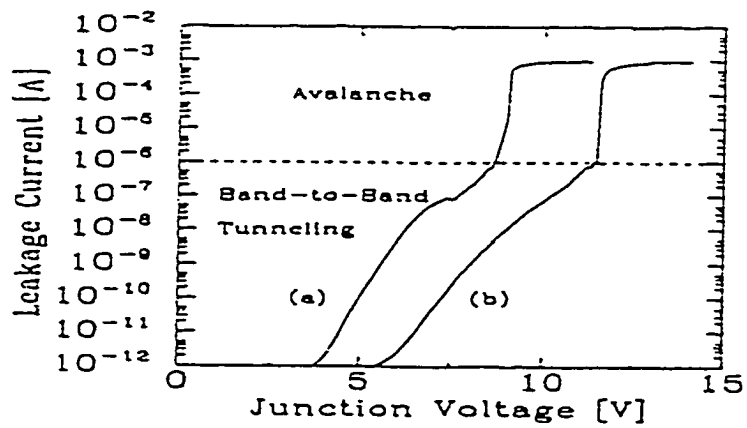


Figure 2.6: The reverse-biased junction current of a gated diode with the gate biased to strong accumulation for (a) an abrupt junction; (b) a graded junction.

The typical electric field distribution in silicon for a graded junction of a Flash device is shown in Figure 2.7. Though the vertical electric field cannot be significantly altered because of the voltage requirements in each programming operation, the lateral field can be reduced by grading the junction to increase the depletion width. This reduces the hole generation considerably compared to an abrupt junction. A comparison between the electric field distribution in an abrupt and a graded junction can be found in Fig. 2.7 and Fig. 2.8 and the corresponding junction leakage currents has been shown in Fig. 2.6. One can clearly see that the overall electric field (sum of vertical and lateral) components for an abrupt junction at a given bias condition is larger than that for the graded junction at the same bias. This is due to a difference in the peak lateral field, which also explains the larger junction leakage current in the abrupt junction. However, even with a graded junction the hole injection continues to be an important reliability issue in the Flash cells so long as the vertical field remains high. Holes may also be generated during bitline (read) disturb. As a typical example the drain voltage is roughly at 6V and the floating gate potential is roughly at 0.3V for a written cell. This bias configuration induces a large band bending and a hole-substrate current due to band-to-band tunneling of electrons. Since the drain junction is optimized for the write (program) operation , an abrupt junction profile is used to increase the carrier injection efficiency during the floating gate electron ejection process. Therefore the abrupt drain junction during the bitline disturb is likely to under-go hole injection. As seen from Fig. 2.8 the high peak lateral field located near the vertical peak enhances the tunneling probability and increases the hole generation. Therefore both junctions in the Flash cell could suffer from hole generation although there are several differences. In brief the lateral

the lateral field in the abrupt junction plays a significant role in hole generation. whereas the vertical field is the

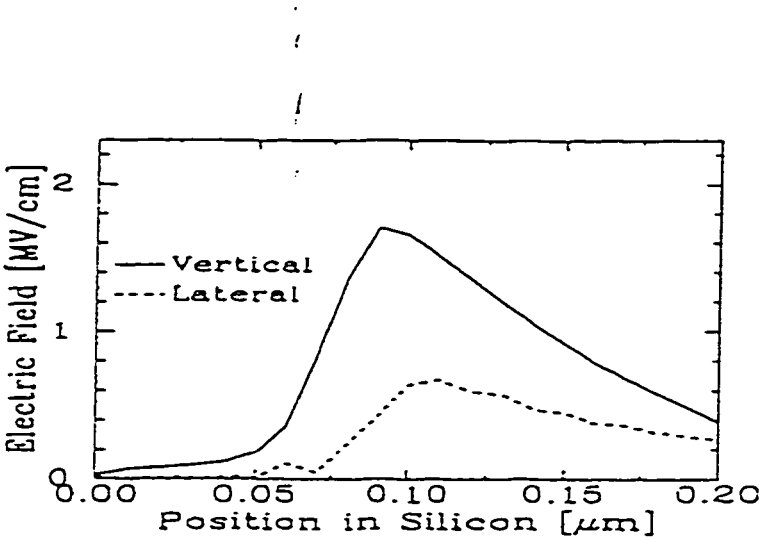


Figure 2.7: Simulated vertical (solid), and lateral (dashed) electric field in silicon at 50Å from the Si/SiO2 interface for a graded junction at 6V drain and 0V on the gate.

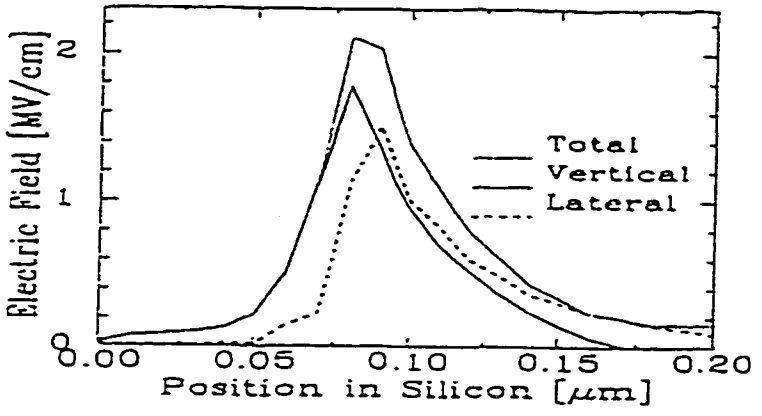


Figure 2.8: Simulated vertical (solid), lateral (dashed) and total (dotted) electric fields in silicon at 50Å from the Si/SiO2 interface for an abrupt junction at 6V drain and 0V on the gate.

key factor in the graded junction. Furthermore , the voltage employed during the erase operation and the bitline disturb are different. This may give rise to different hole injection efficiencies in the two junctions.

2.3 Charge Storage in Flash Memory

The purpose of this section is to introduce the concept of "Capacitive Coupling Coefficients" which is employed in Flash EEPROMs to quantitatively define the floating

gate potential and to discuss the experimental method of determining the coefficients. As discussed in the previous chapters, a non-volatile memory devices where the information is stored as extra charge in the floating gate electrode, the floating gate potential is set by the capacitive coupling from various terminals (control gate, source, drain and substrate), and the charge density in the floating gate. Since the floating gate potential is the determining factor in the carrier injection mechanisms, the proper design of Flash memory devices requires the prediction and verification of the associated coupling coefficients.

There are several published methods to extract the coupling coefficients; they usually rely on the comparison between the electrical characteristics of the actual memory cell and a (single-poly) test transistor which is used as the reference device, where there is a direct electrical contact to the gate [21,22,23]. To define the threshold voltage at the control gate, the capacitance model can be used to make the charge balance of the floating gate, using electrostatic potential ψ in the various regions of the cell.

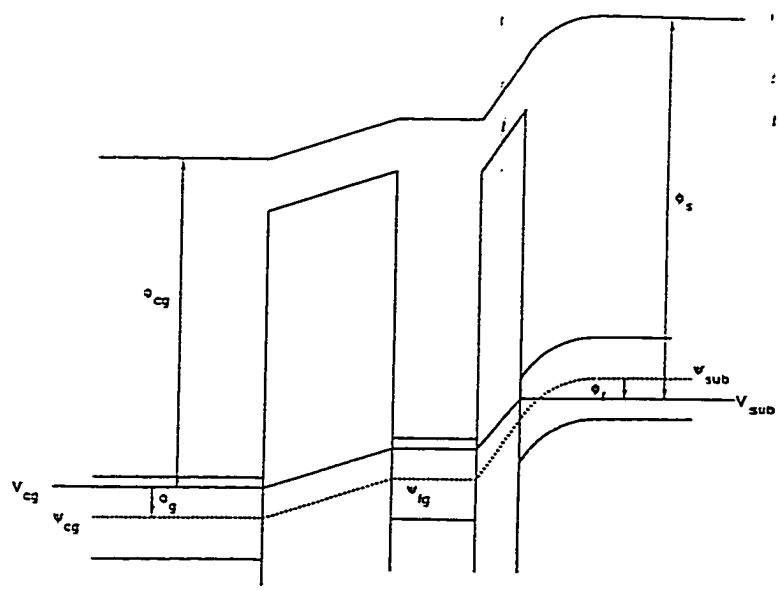


Figure 2.9: Energy band diagram for a floating gate transistor at the onset of inversion with no charge on the floating gate.

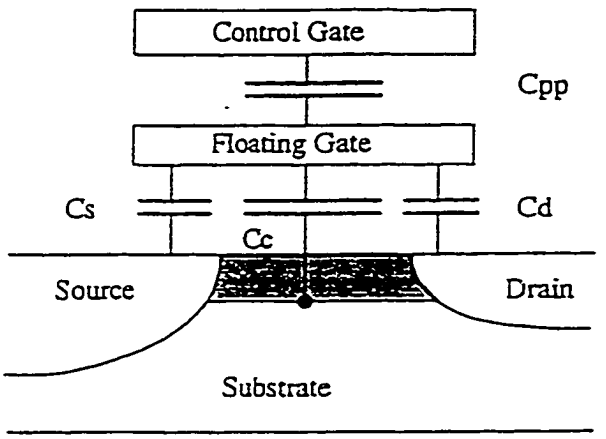


Figure 2.10: Capacitive network model for a floating gate structure.

2.3.1 Definition of Coupling Coefficients

The floating gate electrode is electrically isolated and assumes a potential when the other transistor terminals are biased appropriately. Its potential can be determined by the gaussian equation as discussed in the earlier section. The sum of the surface integral of the electric fields above and below the floating gate electrode equals to the charge on the floating gate, and an analytical expression for the floating gate charge density can be obtained as:

$$\varepsilon_{ox} A_{pp} \left(\frac{V_{fg} - V_{cg}}{d_{pp}} \right) + W_{chan} \varepsilon_{ox} \int_0^L E(V_{fg}, V_s, V_d, x) dx = Q_{fg} \text{ --------(2.7)}$$

Where the first term comes from the field above the floating gate and the second term comes from that below the floating gate. Since $C_{pp}= \varepsilon_{ox} A_{pp}/d_{pp}$, equation 2.27 can be simplified to:

$$Q_{fg} = C_{pp}\left(V_{fg} - V_{cg}\right) + W_{chan}\,\varepsilon_{ox}\int_0^{L_g} E(V_{fg}, V_s, V_d, x)\,dx \text{ - - - - - - - - (2.28)}$$

where $E(V_{fg},V_s,V_d,x)$ is the vertical electric field distribution in the tunnel oxide pointing from the floating gate to the substrate. This is a function of the applied voltages as well as the position in the oxide, ε_{ox} the oxide permitivity, A_{pp} the area formed between the floating gate poly and control gate poly. C_{pp} is the interpoly capacitance, d_{pp} is the thickness of the inter-poly dielectric, W_{chan} is the width of the channel , L_g is the gate length, V_d the drain bias, V_{cg} the control gate bias, Vs the source bias, V_{fg} the floating gate potential, Q_{fg} the floating gate charge, and x the position in the tunnel oxide. In the rest of the chapter E_{ox} will be used instead of $E(V_{fg},V_s,V_d,x)$ for simplicity. This leads to a recursive equation that describes the relationship between the electric field distribution in the tunnel oxide and the floating gate potential in terms of the external biases and the initial floating gate charge, which may be called the "electric field model":

$$V_{fg} = V_{cg} - \frac{W_{chan}\,\varepsilon_{ox}}{C_{pp}} + \int_0^{L_g} E_{ox}\,dx + \frac{Q_{fg}}{C_{pp}} \text{ - - - - - - - - - - - (2.29)}$$

In fact , this equation has been used to simulate the erase characteristics shown in Figures 5.24a & b in chapter 5. Besides the electric field model, one can use a simplified capacitive

network as shown in Figure 2.10 to model the floating gate structure and the associated electric field distribution[21]. This model, however, is not as accurate as the exact electric field model, but provides a great deal of simplification in its application. The following equation for the floating gate potential can be easily obtained from the capacitive network model:

$$V_{fg} = \frac{C_s}{C_T}(V_s - \Phi_{sf}) + \frac{C_d}{C_T}(V_d - \Phi_{df}) + \frac{C_{ch}}{C_T}(\psi_{chs} - \Phi_{cf}) + \frac{C_{pp}}{C_T}(V_{cg} - \Phi_{pp}) - \frac{1}{C_T}Q_{fg} \quad \text{---(2.31)}$$

where $C_T=C_{pp}+C_s+C_{ch}+C_d$, C_s is the source overlap capacitance, C_d the drain overlap capacitance, C_{ch} the channel area capacitance, Φ_{sf} the work function difference between the source and the floating gate, Φ_{df} the work function differnence between the drain and the floating gate, Φ_{pp} the work function difference between the control gate and the floating gate (poly-1 to poly-2), Φ_{chf} the work function difference between the substrate and the floating gate, and ψ_{ch} the surface potential of the channel. Each pre-factor in equation (2.31) is called a capacitive coupling coefficient. Thus the floating gate potential can be expressed in terms of these coefficients as follows:

$$V_{fg} = \alpha_s(V_s - \Phi_{sf}) + \alpha_d(V_d - \Phi_{df}) + \alpha_{ch}(\psi_{chs} - \Phi_{cf}) + \alpha_g(V_{cg} - \Phi_{pp} - \Delta V_t) \text{----(2.32}$$

where $\alpha_s+\alpha_d+\alpha_{ch}+\alpha_g=1$, and α_s is the source coupling coefficient, α_d is the drain coupling coefficient, α_g is the gate coupling coefficient, α_{ch} is the channel coupling coefficient, and ΔV_t is the change in the threshold voltage of the memory cell from its unprogrammed value.

Since the capacitance values in equation 2.31 are bias dependent, the corresponding coupling coefficients are also expected to be bias dependent:

$$V_{fg} = \alpha_s^{\circ}(V_s - \Phi_{sf}) + \alpha_d^{\circ}(V_d - \Phi_{df}) + \alpha_{ch}^{\circ}(\psi_{chs} - \Phi_{cf}) + \alpha_g^{\circ}(V_{cg} - \Phi_{pp} - \Delta V_t) \text{----(2.32)}$$

where the sign (°) denotes that the coupling coefficients are a function of V_{fg}, V_s, and V_d. Since the change in coupling coefficients with respect to the floating gate potential is greater than their change with respect to the junction biases, the 1T-Flash device coupling coefficients are more dependent on the floating gate potential. Further, the work function difference between the junctions and the N+ doped poly can be neglected. Thus the equation for the floating gate potential can be simplified to:

$$V_{fg} = \alpha_s^{*}V_s + \alpha_d^{*}V_d + \alpha_{ch}^{*}(\psi_{chs} - \Phi_{cf}) + \alpha_g^{*}(V_{cg} - \Phi_{pp} - \Delta V_t) \text{--------(2.34)}$$

Where the sign (*) indicates that the coupling coefficient is a function of V_{fg}. The change in the floating gate potential with respect to the changing external biases and floating gate charge density is:

$$dV_{fg} = \alpha_s^{*}dV_s + \alpha_d^{*}dV_d + \alpha_{ch}^{*}(\psi_{chs} - \Phi_{cf}) + \alpha_g^{*}dV_{cg} - \alpha_g^{*}\frac{Q_{fg}}{C_{pp}} \text{-------(2.35)}$$

This differential equation holds if the coupling coefficients are relatively constant around the operating point, which is usually the case when the device is in accumulation or inversion. On the other hand, the floating gate potential around an operating point can be expressed in terms of the following partial derivatives:

$$d\dot{V}_{fg} = \frac{\partial V_{fg}}{\partial V_s} dV_s + \frac{\partial V_{fg}}{\partial V_d} dV_d + \frac{\partial V_{fg}}{\partial V_{ch}} d\psi_{ch} + \frac{\partial V_{fg}}{\partial V_{cg}} dV_{cg} + \frac{\partial V_{fg}}{\partial Q_{fg}} dQ_{fg} \text{ ----(2.36)}$$

By incorporating the electric field model into equation (2.36) above, the change in the floating gate potential with respect to the terminal voltages around the operating point can be expressed in terms of the following equation:

$$dV_{fg} = \frac{1}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}} dx} dV_{cg} + \frac{\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_s} dx}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}} dx} dV_s +$$

$$\frac{\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_d} dx}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}} dx} dV_d + \frac{\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial \psi_{ch}} dx}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}} dx} d\psi_{ch} + \frac{1}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}} dx} \frac{dQ_{cg}}{C_{pp}} \text{ -------(2.37)}$$

Using the previously defined capacitive coupling terminology, the coupling coefficients can thus be expressed as:

$$\alpha_g = \frac{1}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}} dx} = \frac{C_{pp}}{C_{pp}+C_s+C_{ch}+C_d} \text{ ----------(2.38)}$$

These derived results are similar to published methods of extracting coupling coefficients. In most literature, the coupling coefficients are defined as partial derivatives in the electric field model, and are used to find the actual floating gate potential within a certain error that is acceptable by most NVSM applications.

$$\alpha_s = \frac{\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_s}dx}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}}dx} = \frac{Cs}{C_{pp}+C_s+C_{ch}+C_d} \text{---------(2.39)}$$

$$\alpha_d = \frac{\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_d}dx}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}}dx} = \frac{C_d}{C_{pp}+C_s+C_{ch}+C_d} \text{---------(2.40)}$$

$$\alpha_{ch} = \frac{\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial \psi_{ch}}dx}{1+\frac{W\varepsilon_{ox}}{C_{pp}}\int_0^L \frac{\partial E_{ox}}{\partial V_{fg}}dx} = \frac{C_{ch}}{C_{pp}+C_s+C_{ch}+C_d} \text{--------(2.41)}$$

2.4 Analytical Model for Fast Programming Bits in 1T-Flash

This analytical model is developed and reveals that the behavior of fast programming bits in the 1T-Flash array can be predicted. This new model is based on a combination of the conventional Fowler-Nordheim conduction mechanism, a first-order kinetic trapping model, and a well-chosen non-uniform distribution of injection current; it is able to explain the experimental observations of floating gate conduction including trapping phenomena which is crucial to understanding and quantifying the 1T-Flash reliability. This model will be used to explain the Flash EEPROM degradation and the Fast programming bit behavior observed in the figures of chapter 5.

Electron conduction in thin oxides is a quantum mechanical process that can be described by Shroedinger equation with a 1D potential barrier U(x) to an electron with an energy E:

$$\left[\frac{-\hbar^2}{2m^*}\left(\frac{d^2}{dx^2}\right)+U(x)\right]A(x)=EA(x)\text{----------(2.42)}$$

Using a solution of the form $A(x)=e^{\alpha(x)}$, and tunneling probability P(ε) and the W.K.B. approximation, the Fowler-Nordheim tunneling current as shown in equation 2.43:

$$J=\alpha_{bb}E_{si}^{\ 2}\exp[-\beta_{bb}/E_{si}]\text{----------(2.43)}$$

The amount of floating gate(FG) charge available for programming is a function of the F-N tunneling current, is given by:

$$\frac{dQ_{FG}}{dt}=C*J_{FN}\left[E_{inj}(r_C)\right]=CE^2{}_{inj}(G_S)e^{\frac{-E_C}{E_{inj}(G_S)}}\text{----------(2.44)}$$

From the measured experimental data . the number of grains (N) and polysilicon grain size (G_S) can be modeled as a Gaussian distribution with a normal probability density function. Hence the number of grains(y) as a function of the grain size:

$$y(pdf)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-(G_s-\overline{G_s})^2/2\sigma^2\right]\text{----------(2.45)}$$

Ignoring minute trapped electrons in the tunnel oxide during a programming operation, expressions can be developed for the injection field and threshold voltages as a function of time. Too explain the observed experimental results in chapter 5, these expressions must comprehend the "area-effects" due to floating gate polysilicon grain size (Geff) and some field enhancement (μ) resulting from sidewall asperities. From TEM micrographs shown in chapter 5 one can make the case that the probability of asperities present at terminating polysilicon edges is high. As a result the derivations are developed with Geff and μ factors incorporated.

As seen from Figure 2.13, the effective tunnel area of a polysilicon grain is reduced as a result of grain-boundary charge and resultant depletion region in the grain.

From Polysilicon Physics:

$$Q_{Trap} = N_t * T_{grain} * L_{grain} = 2X_d * N_D * T_{grain} * L_{grain} \quad \text{-- (2.46)}$$

$$\Rightarrow X_D(\textit{depletion} \text{ width in grain}) = \frac{N_t}{2 * N_D} \quad \text{---------- (2.47)}$$

assuming $L_{grain} = W_{grain} = G_S$ then

$$A_{inj} / (cm^2) = (G_S - 2X_D)^2 = (Lg - 2X_D)^2 \quad \text{---------- (2.48)}$$

$$A_{inj}(cm^2) = ((G_S - 2X_D)^2 * N_g) = \frac{(G_S - 2X_D)^2}{G_s^2} * W * L_{overlap} \quad \text{---- (2.49)}$$

$$G_{eff} = \frac{A_{inj}}{A_{overlap}} = \frac{A_{inj}}{W * L_{overlap}} = \frac{(G_S - 2X_D)^2}{G_s^2} = \left(1 - \frac{2X_D}{G_s}\right)^2 \quad \text{------ (2.50)}$$

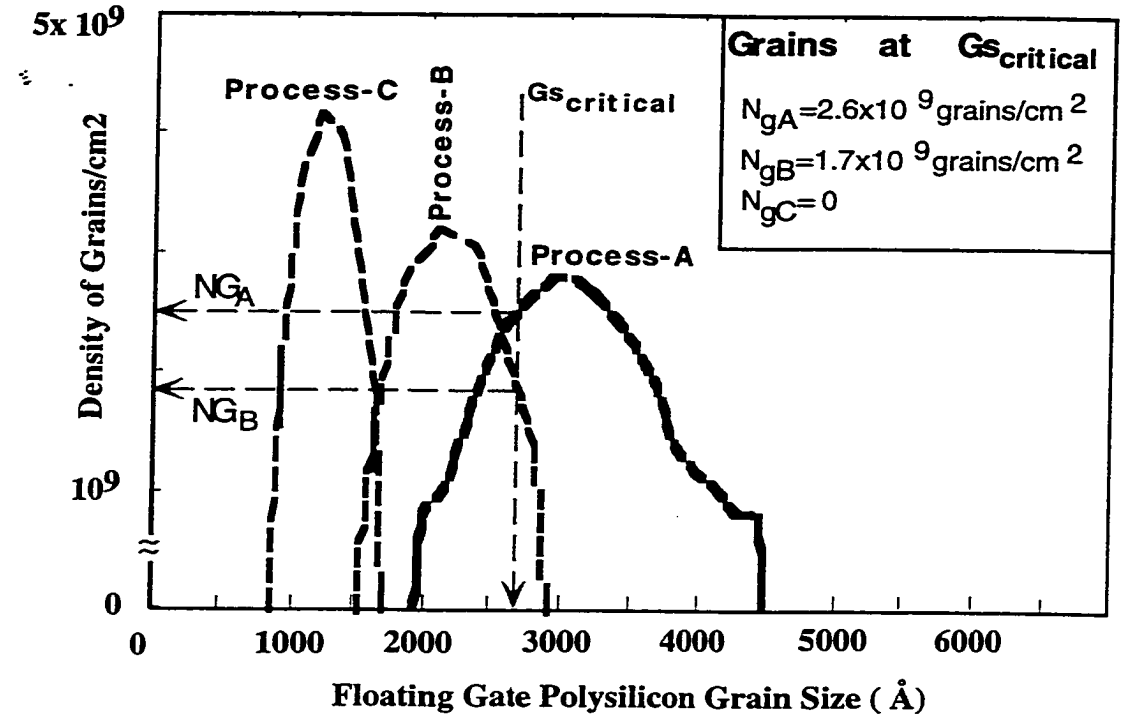


Figure 2.11: Floating gate polysilicon grain density as a function of grain size for process-A(Phos. Implanted poly), B (Phos. In-situ doped poly) and C (udoped+ Phos. Insitu doped stack)

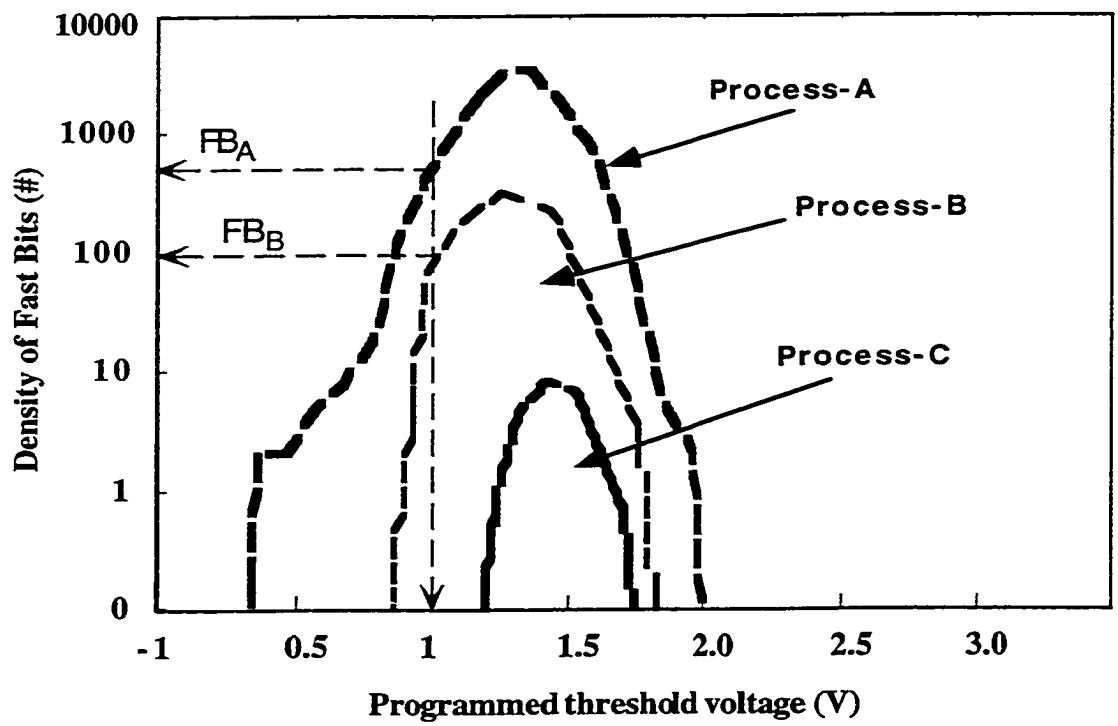


Figure 2.12: Fast Programming bit density vs. threshold voltage for process-A(Phos. Implanted poly), B (Phos. In-situ doped poly) and C (udoped+ Phos. Insitu doped stack).

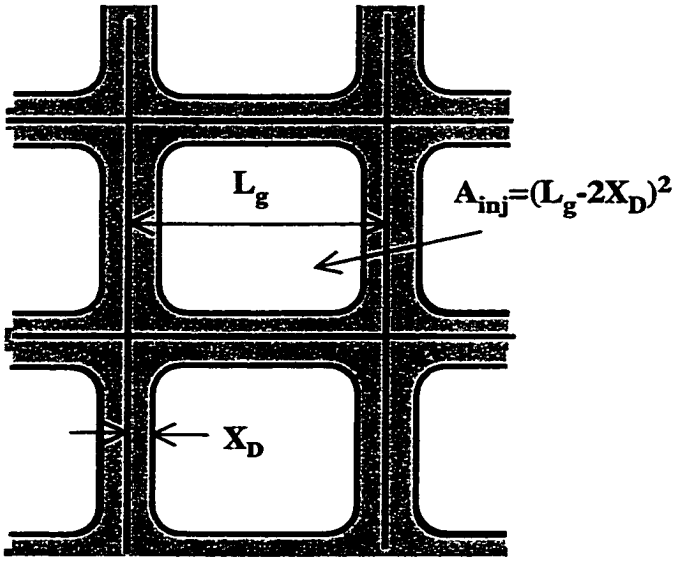


Figure 2.13: A Planar view of polysilicon grain showing area of grain that injects carriers

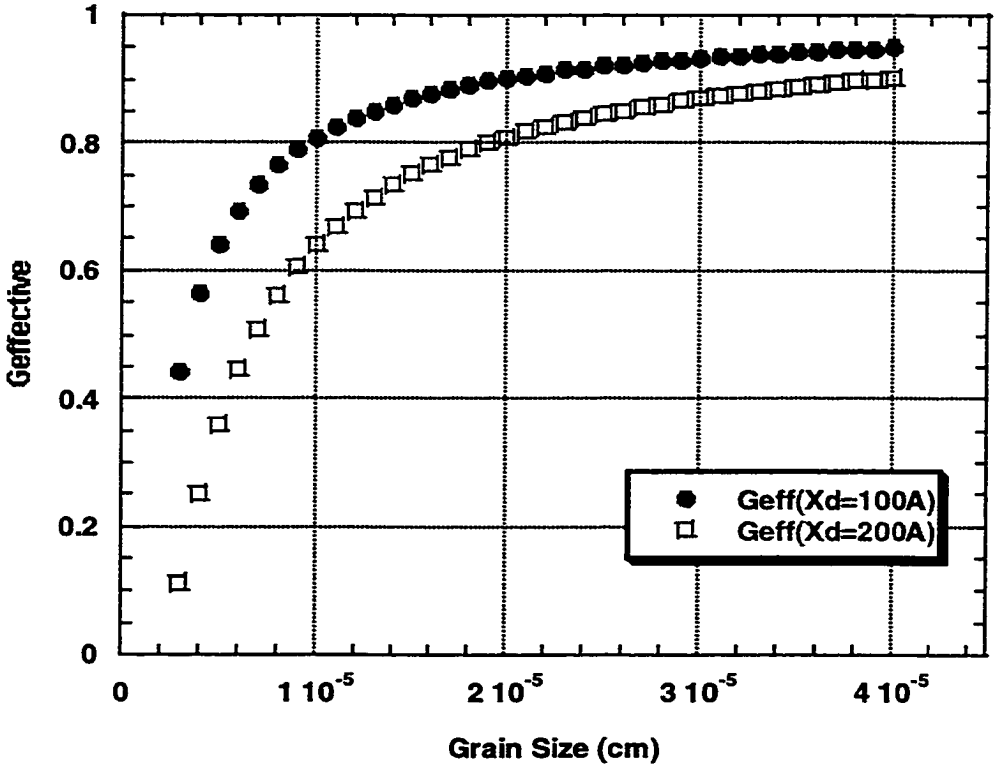


Figure 2.14: Grain Injection Area Factor (Geff) vs. Grain Size and Depletion Width

We can now proceed to establish the expressions for the injecting field necessary for the 1T-Flash programming as a function of time (t). Since we have established the effective grain area that influences the tunneling current, the injecting current from a polysilicon grain will be modified by the grain factor , whiles the field will be modified by an enhancement factor μ(Rc). Hence the F-N tunneling equation can be re-written as:

$$J_{FN} = CE_{inj}^2 e^{-\frac{E_C}{E_{inj}}} = \alpha_{fn} G_{eff} E_{inj}^2 e^{-\frac{E_C}{E_{inj}}} \quad \text{--------- (2.51)}$$

$$E_{inj} = \frac{V_{fg}}{X_{tun}}\left(1 + \frac{X_{tun}}{Rc}\right) \quad \text{---------- (2.52)}$$

$$\text{where } \mu(Rc) = \left(1 + \frac{X_{tun}}{Rc}\right) \quad \text{---------- (2.53) [66]}$$

Re-calling some basic NVM threshold voltage and floating gate discharge expressions:

$$\frac{dV_T}{dt} = \left(\frac{-1}{C_{fg}}\right) * \frac{dQ_{fg}}{dt} \quad \text{---------- (2.54)}$$

$$\frac{dQ_{fg}}{dt} = -J_{FN} * A_{tun} \quad \text{---------- (2.55)}$$

$$Q_{fg} = V_{fg} C_t - V_{cg} C_{fg} \quad \text{---------- (2.56)}$$

Combining equations 2.51,2.52,2.55 and 2.56 yields:

$$\frac{dQ_{fg}}{dt} = \frac{d}{dt}(V_{fg}C_t - V_{cg}C_{fg}) = C_t\frac{dV_{fg}}{dt} - C_{fg}\frac{dV_{cg}}{dt} \text{ -------- (2.57)}$$

substituting the other equations and re - arranging:

$$\frac{dQ_{fg}}{dt} = \frac{C_t X_{tun}}{\mu} * \frac{dE_{inj}}{dt} = -A_{tun}\alpha_{fn}G_{eff}E_{inj}^2 e^{\frac{-Ec}{Einj}} \text{ ------ (2.58)}$$

Let variable $K = -A_{tun}\alpha_{fn}G_{eff}$ and re - arrange equation for Einj:

$$\frac{dE_{inj}}{dt} = \frac{K\mu}{C_t X_{tun}} * E_{inj}^2 e^{\frac{-Ec}{Einj}} \text{ ---------- (2.59)}$$

Solving the equation 2.59 integral for Einj(t) yields:

$$E_{inj}(t) = \frac{E_{inj}(0)}{1 + \frac{E_{inj}(0)}{E_c}\ln\left(1 + \frac{t}{t0}\right)} \text{ ---------- (2.60)}$$

where the characteristic turn - on time t0 is defines as:

$$t0 = \frac{X_{tun}C_{fg}e^{Ec/Einj(0)}}{\alpha cgA_{tun}\mu E_c Geff} \text{ ---------- (2.61)}$$

$$E_{inj} = \mu\left(\frac{Vcg - \alpha cgVcg + \alpha cgV_T - \alpha cgVt0}{X_{tun}}\right) \quad \text{----------(2.62)}$$

if :

$$E1 = \frac{1}{X_{tun}}(Vcg(1-\alpha cg) - \alpha cgVt0 \quad \text{---------- (2.63)}$$

substituting E1 into equation 2.62 and differentiating results in:

$$\frac{dE_{inj}}{dt} = \mu\frac{\alpha cg}{X_{tun}}\frac{dV_T}{dt} \quad \text{---------- (2.64)}$$

the resulting integral solution is:

$$V_T(t) = \frac{\alpha cg}{\mu X_{tun}} * E_{inj}(t) + C \quad \text{----------(2.65)}$$

at t = 0 , Einj(t) = Einj(0) hence:

$Vt0 = \frac{\alpha cg}{\mu X_{tun}} * E_{inj}(0) + C$, which when substituted back into equation 2.65 yields:

$$V_T(t) = \frac{\alpha cg}{\mu X_{tun}}\left[E_{inj}(t) - E_{inj}(0)\right] + Vt0 \quad \text{---------- (2.66)}$$

Substituting equations 2.60, 2.63, and 2.62 at t=0 into equation 2.66 yields the final equation that relates the memory threshold voltage $V_T(t)$ to the injecting field, the enhancement factor $\mu(Rc)$ ant t0 which is a function of the polysilicon grain size:

$$V_T(t) = \left[\frac{\left(E1 + \frac{\alpha cg Vt0}{X_{tun}} \right)}{1 + \frac{\mu(Rc)\left(E1 + \frac{\alpha cg Vt0}{X_{tun}} \right)}{E_C} \ln\left(1 + \frac{t}{t0} \right)} - E1 \right] * \frac{X_{tun}}{\alpha cg} \quad \text{----------} (2.67)$$

All the critical equations used in the development of equation 2.67 were used in model validation, the results of which will be discussed in chapter 5.

Chapter 3

Measurements Techniques

In this chapter, I will discuss measurement techniques used in this dissertation. We will describe the Capacitive coefficient extractions, Program/Erase measurements, Data Retention and Endurance measurements that are employed to characterize and quantify the reliability of the 1T-Flash Memory devices.

3.1 Capacitive Coefficients Extractions

The values of the different capacitive coupling coefficients discussed in chapter 2 can be calculated from measured data. This is achieved by comparing the threshold voltages of the floating gate device under test (Vtcg measured at control gate), to that of the equivalent contacted floating gate (Vt0 measured at the contacted floating gate). This is achieved by comparing the I-V characteristics of the floating gate device to the contacted transistor. The coupling coefficients is then calculated from the ratio of the currents and from the dependence of the threshold voltage Vtcg on the drain voltage.

3.2 Program/Erase Measurements

During this mode of operation, I varied the control gate voltage V_{cg} from -10V, -9V, -8V and the drain voltage V_d from 5,6,7V. These were performed with instrumentation in device lab with the schematic diagram as shown in Figure 3.1. The source and substrate nodes will

be left floating so as to maximize the F-N tunneling at the drain junction. This biasing condition will induce F-N tunneling of electrons from the floating gate to the drain via the drain-floating gate overlap region, thus resulting in a net positive charge on the floating gate. We will then quantify the program V_t versus time characteristics as shown below in Figure 3.2.. As discussed earlier the electron injection during erase (ETOX) and program(F-N) takes place through a small area. Since this region of operation is so small, any statistical variations among cells can result in varied program/erase characteristics. It is also widely known that since the Fowler -Nordheim tunneling current is a strong function of tunnel oxide thickness, any non-uniformity and reliability could lead to memory failures.

The erase characterization will be carried out by applying 15,16,17V on the control gate with the source and drain nodes floating, whiles the substrate node is grounded or negatively biased. This will cause electrons to be attracted to the floating gate from the substrate since it is at a positive potential. A net negative charge is therefore stored on the floating gate which defines logic level "0". These characterization shown in Figure 3.3 will also be used to evaluate the experiments previously described.

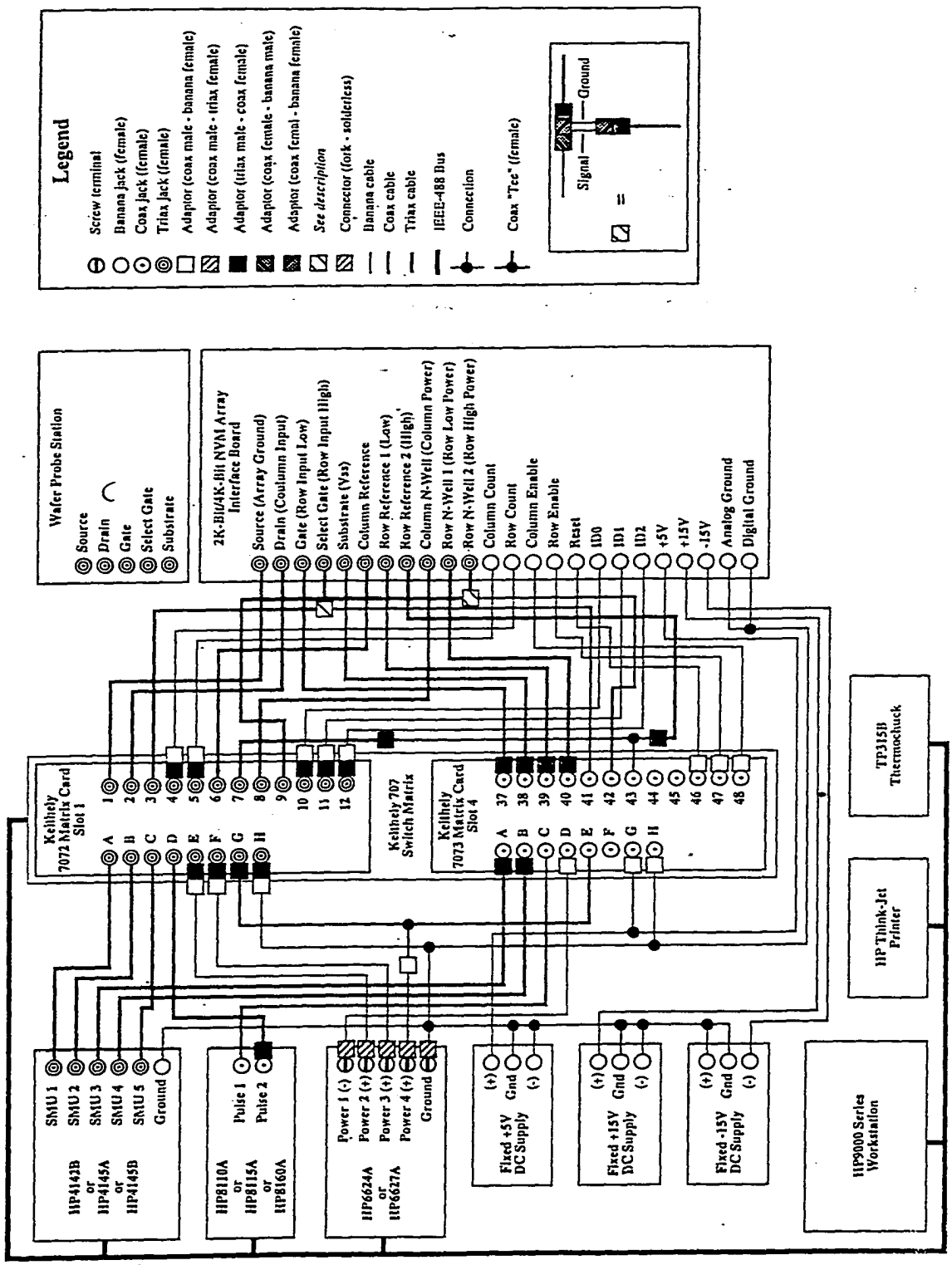


Figure 3.1: 1T-Flash Program and Erase Automated Measurement Instrumentation

3.5 Array Disturb Characterization

We will use the disturbs previously defined i.e. bitline, wordline and read disturbs to quantify the operating windows for the 1T-Flash The gate or wordline disturb will be performed with Vcg at -8,-9,10V and Vd at 0V over the entire time it takes to perform one read operation to investigate effects on unselected bits. The bitline or drain disturb will be performed with Vcg at 0V and Vd at 5,6,7V and evaluated in the context of the defined experiments. Below in Figure 3.7 are examples of the various disturb measurements.

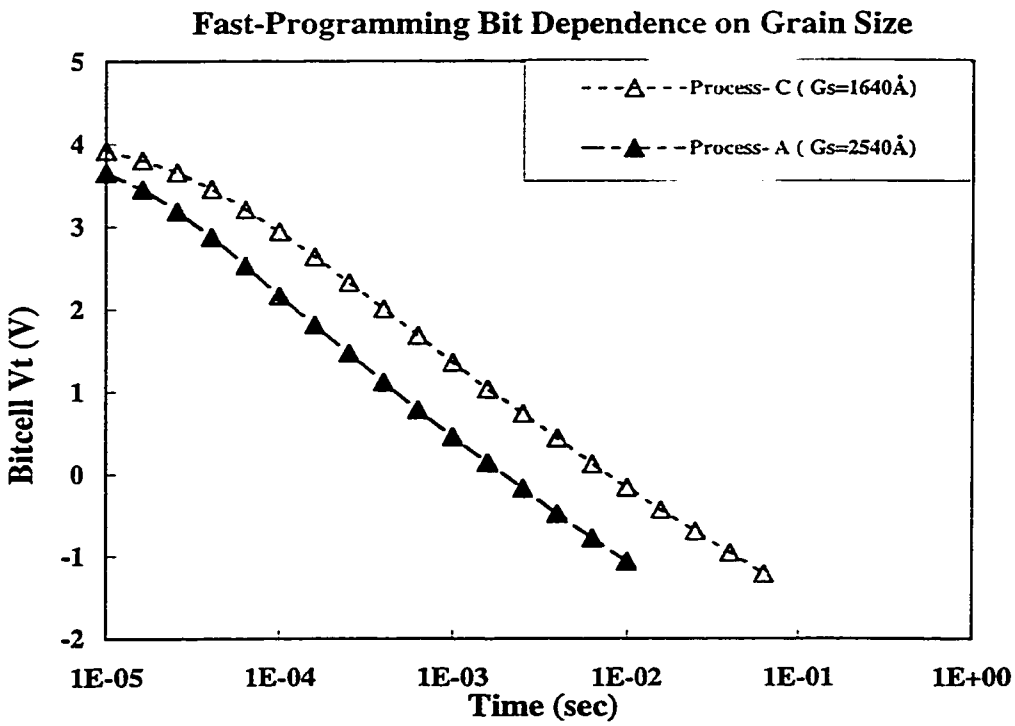


Figure 3.5: Program vs Log T (pulse-width) plot for flash characterization.

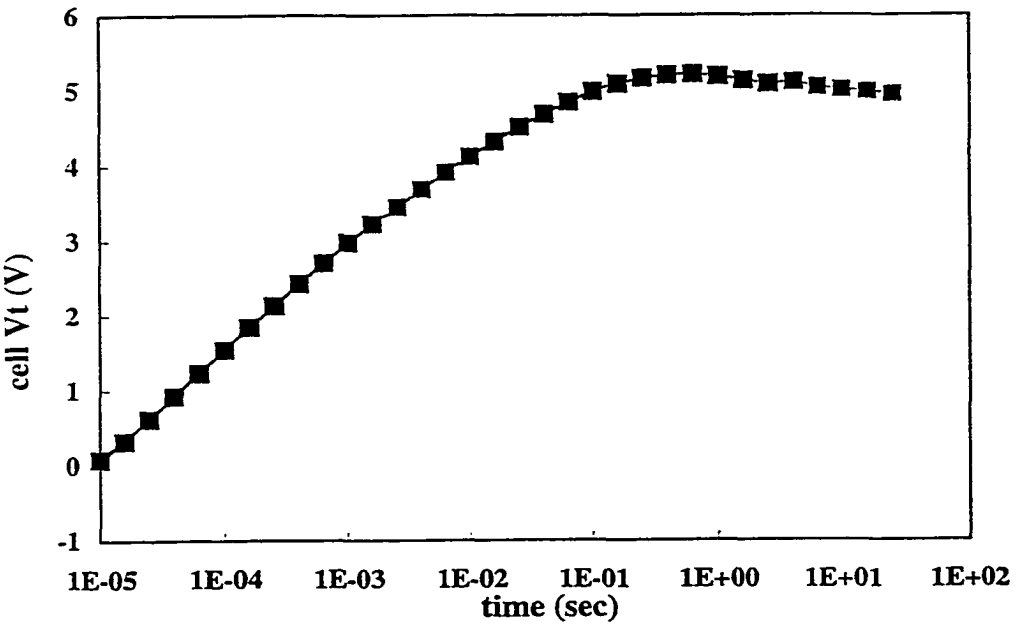


Figure 3.5: Erase vs Log T (pulse-width) plot for flash characterization.

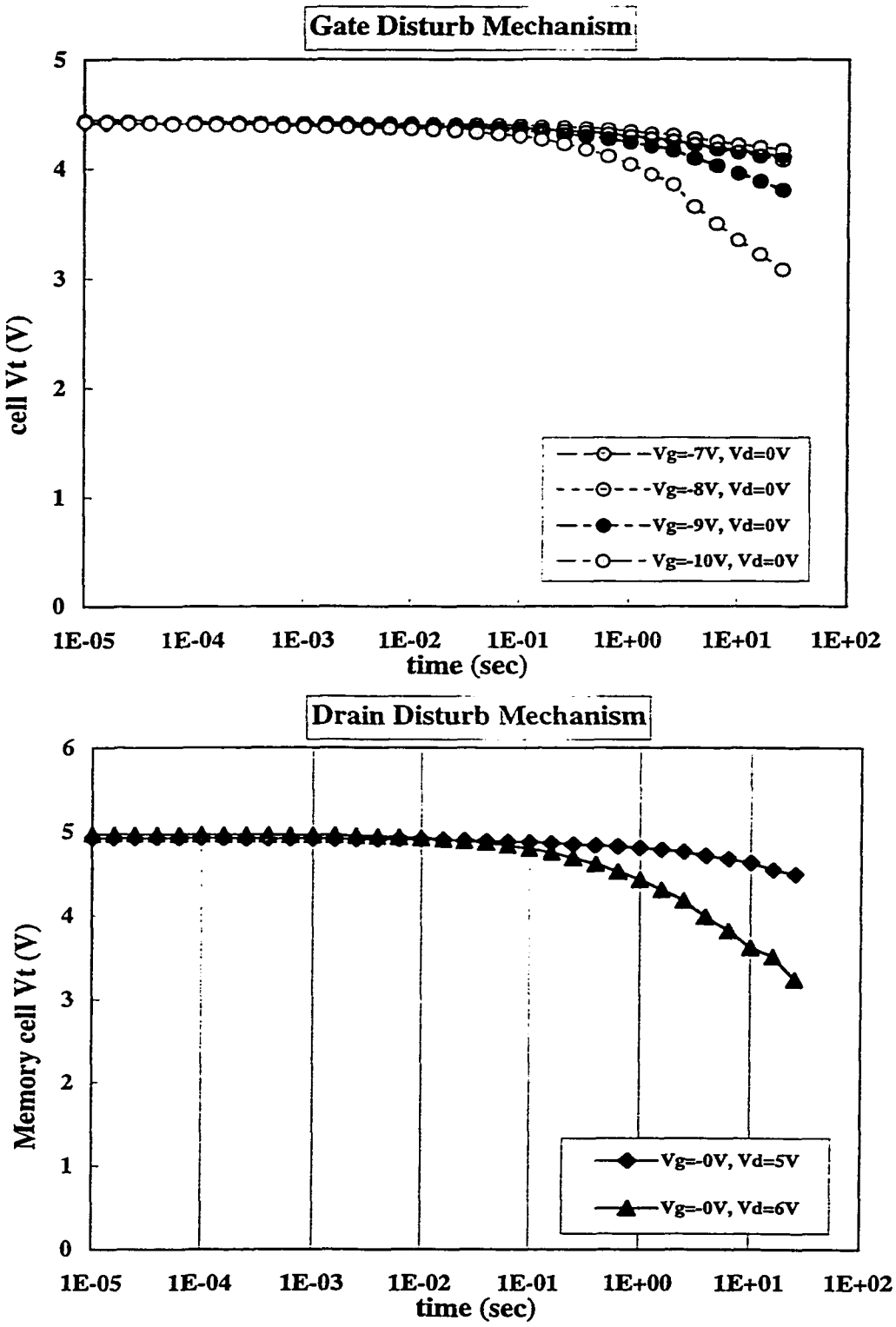


Fig. 3.7: Disturb Mechanisms in Flash EEPROM

3.6 Endurance Measurement

As discussed earlier this is a Flash reliability test in which numerous programs and erase operations are performed alternatively at some fixed biases. I will employ this technique to evaluate the discrete device where appropriate and also the 4096 bit array as needed. The appropriate conditions for cycling will be established after the program and erase characterizations. During my research I will generate the typical bathtub cycling curve as shown below in Figure 3.8 and will use this as an aid in explaining the results of the above-specified experiments.

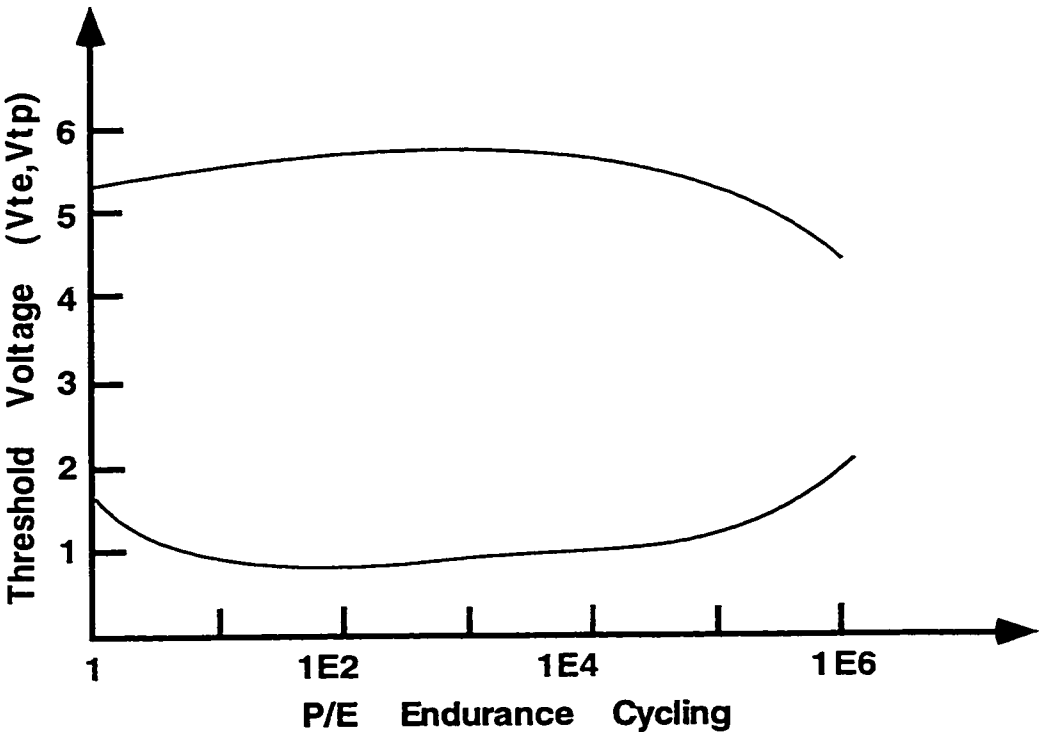


Figure 3.8 Schematic of Endurance for 1T-Flash EEPROM

3.7 Data Retention

Data retention is characterized by the loss of charge from the floating gate of a programmed device. As a result of the charge loss into the substrate or the control gate, the memory threshold voltage of the cell decreases. If the memory threshold voltage drops below Vcc, the cell can no longer be read as a high and the stored data is subsequently lost. There are three distinct phases of charge loss namely: initial-phase < 10min, slow-phase usually 10-100hours and finally the non-saturating phase > 100hours. It has been reported that the initial-phase accounts for about 30% of the charge loss, slow-phase accounts for ~30-45% and non-saturating phase ~30-50% [13], during the bake test used to evaluate charge loss from the floating gate. I will use this as needed to understand charge loss mechanisms during this study. The 4096 bitcell array were packaged and stressed at various temperatures 150,200,250 and 300C. An example of data retention characterization is shown in Figure 3.9.

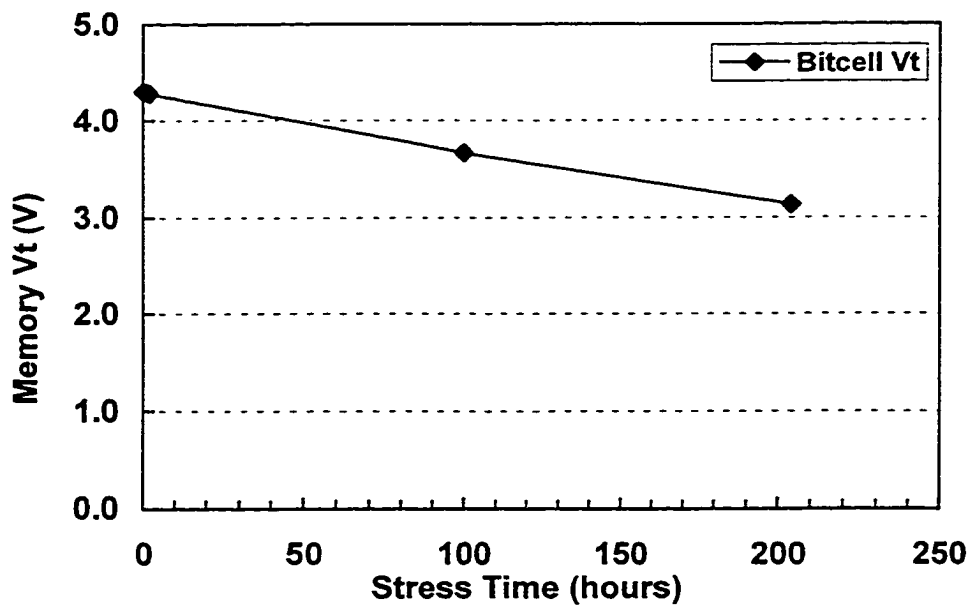


Figure 3.9: Typical data retention plot for 1T-Flash EEPROM.

3.6 Tunnel Oxide Characterization

Growth of defect –free ultra-thin tunnel oxides reproducibly, typically SiO_2 or Oxy-Nitrides, is necessary for the development of floating gate memories which can operate at low programming voltages and provide reliable long-term charge/data storage. The key properties of interest include high breakdown fields as shown in Figure 3.7 (J vs. E) data, low electron trap density, and low leakage. The tunnel oxide quality is typically measured as the total injected charge per unit area (C/cm^2) required to cause breakdown, referred to as Qbd as shown in Figure 3.8 for this work. The characterization of the tunnel oxides revealed that the intrinsic quality was adequate for full 1T-Flash device design and integration.

During tunnel oxide formation the silicon surface preparation becomes extremely critical for high-quality thermal oxides in the thickness range of 50-100A. The surface needs to be free of native oxides, microscopic reactive impurities and particles. The native oxides contain more pinholes and defects than thermal oxides and will degrade the final quality of a thermally grown oxide. Surface micro-roughness at the $Si-SiO_2$ interface correlates with lowers breakdown fields, lower Qbd and increased oxide leakage currents. Studies have been conducted by previous researchers have shown that the presence of a thin oxide layer prior to carrying out an oxidation is critical for obtaining good quality oxide. This fact appears to be due to roughening of the silicon surface. The surface micro-roughness has been attributed to the etching of the silicon surface by trace O_2 present during the temperature ramps in inert ambient. [67].

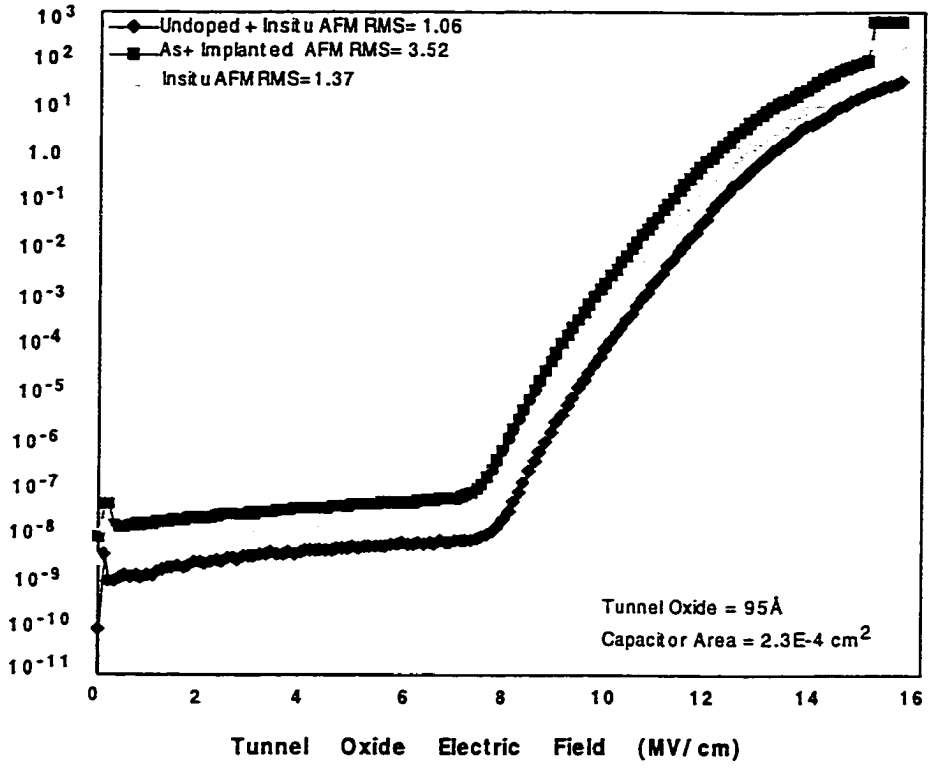


Figure 3.7: Tunnel Oxide Current Density as a Function of Applied Fields

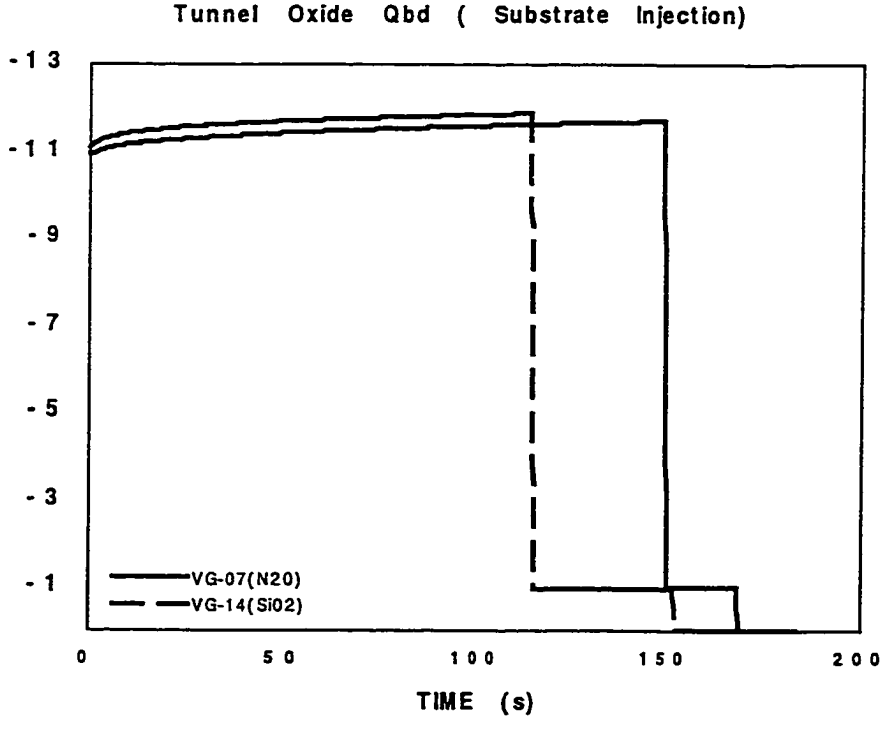


Figure 3.8: Typical Tunnel Oxide Charge Breakdown

Growth of defect –free ultra-thin ONO oxides reproducibly, typically SiO_2 or Oxy-Nitrides, is necessary for the development of floating gate memories which can operate at low programming voltages and provide reliable long-term charge/data storage. The key properties of interest include high breakdown fields, low electron trap density, and low leakage. The ONO quality is typically measured as the total injected charge per unit area (C/cm^2) required to cause breakdown, referred to as Qbd. The characterization of the ONO oxides revealed that the intrinsic quality was adequate for full 1T-Flash device design and integration.

During ONO dielectric formation a thin layer of oxide is grown on top of the floating gate polysilicon. This high quality oxide is followed by the deposition of furnace nitride of about 180Å – 250Å. The nitride is then densified with a high temperature oxidation which grows a thin protective oxide layer of 30Å – 50Å which passivates the nitride. This ONO dielectric acts as the key insulator between the control gate and the charge storage floating gate in the 1T-Flash EEPROM memory.

Chapter 4

1T-Flash Device Fabrication

The fabrication of the 1T-Flash devices begins with an isolated Pwell which is followed by the Polysilicon Encapsulated LOCOS (PELOX) [29] isolation. The detailed processing sequence of this N-channel 1T-Flash devices is listed in Appendix A.
In this chapter, we will go through briefly the fabrication sequence for the N-channel 1T-Flash transistor.

4.1 Starting Material

The starting material is a 6-inch p-type <100> Si wafers with a 3.4μm epitaxial silicon on a P+ substrate. The thick epi is required for NVM devices to ensure adequate immunity as a result of the integration of high voltage wells in conjunction with the low voltage wells for the periphery devices.

4.2 Advanced PELOX Isolation

The wafers are first cleaned using a standard RCA clean and native oxide etched in a dilute hydrofluoric (HF) acid. The initial 300Å pad oxide is thermally grown followed by a low-pressure chemical vapor deposition of 1400Å silicon nitride. The nitride layer then serves as the oxidation mask. After the nitride photo and etch process, HF is used to under-cut the nitride by about 45Å per side. This is followed by the deposition of ~300Å of polysilicon which serves as an encapsulation to minimize the lateral oxidation of the active

area under the nitride. This is followed by the thermal oxidation to grow a field oxide of 6700Å. This is then followed by the removal of the nitride from the device active regions using hot phosphoric acid, followed by a buffered hydroflouric (BHF) acid to remove the 300Å pad oxide as shown in Fig. 4.1.

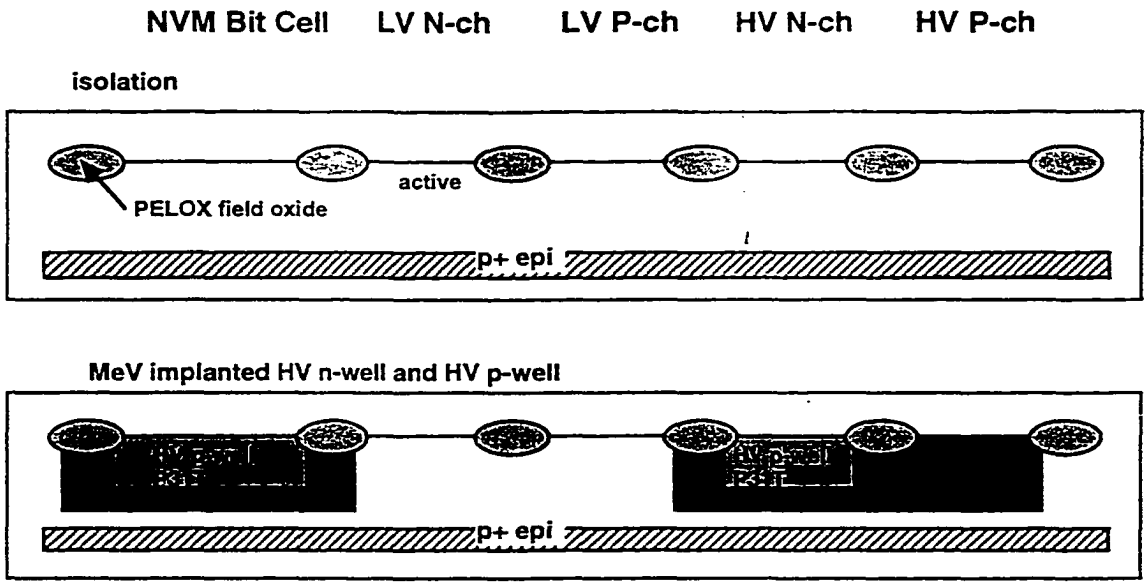


Figure 4.1: Formation of active 1T-Flash transistor regions

4.3 Sacrificial Oxidation

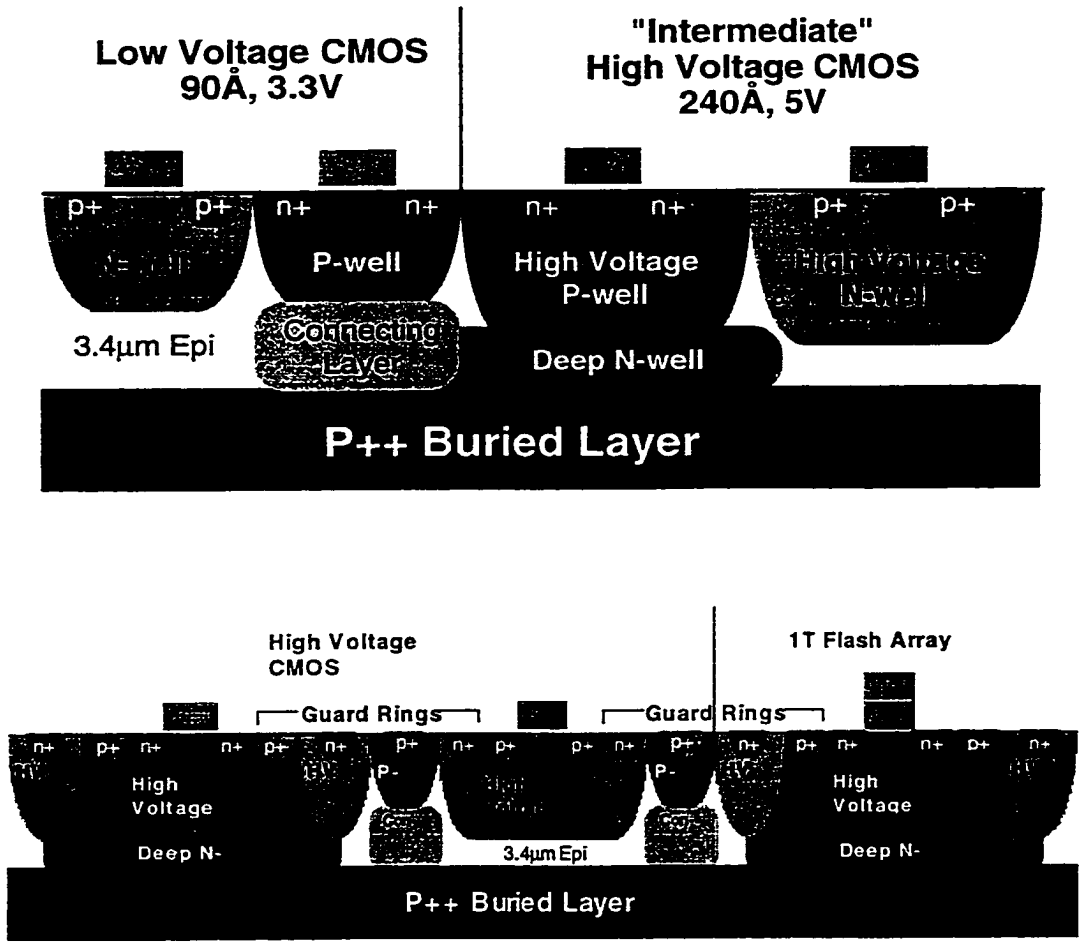


Fig 4.2: Quadruple Well formation for 1T-Flash and Periphery logic

It is well known that LOCOS processes produce a compound of oxynitride of unknown proportion (Si_xN_y)[30] at the edges of the active areas (bird's beak) which will

result in a region of thinner oxide in a subsequent oxidation, such as gate oxide growth. A sacrificial oxide is therefore grown in a wet atmosphere, after the pad oxide removal, to remove the SixNy composite. The thickness of the sacrificial oxide used in this work was ~500Å. This oxide is then etched away using HF, and is followed by a second sacrificial oxidation used as a surface preparation before the transistor gate oxide growth.
The sacrificial oxidation is followed by the isolated high-voltage Pwell Boron implant using MeV implanters to achieve the desired depth at about 400KeV. The implant chain is terminated by the N-channel threshold voltage adjust implant using BF2. This bitcell Vt-adjust implant on the order of low 3E12 is tailored to achieve a natural threshold voltage of the 1T-Flash bitcell post UV-erase of ~0.6V. After the Pwell resist strip, the high voltage Nwell is also defined followed by a high energy Phosphorus implantation at 600KeV. The Nwell chain is also terminated with a phosphorus threshold voltage adjust implant on the order of 2E12 at 60KeV as shown in Figure 4.2.

4.4 Tunnel Oxide Dielectric Formation

After the High Voltage Well implants and threshold voltage adjust implants, the 250Å sacrificial oxide is removed using an HF pre-clean. This is followed by the growth of a 60Å tunnel oxide at 900C in a dilute O_2 ambient, followed by a 50A N_20 anneal. The purpose of incorporating the N_20 anneal is to improve the tunnel oxide resistance to the extensive high electric field stress necessary for the Flash operations.

4.5 Floating Gate, ONO and Control Gate Formation

Immediately following the tunnel oxide formation, a 1500Å phosphorus in-situ doped film of polycrystalline silicon is LPCVD deposited at 650C at 0.85 Torr. This is followed by the photolithography step to define the floating gate poly in the memory array regions. The polysilicon in the periphery regions outside the memory array is then plasma etched using an HBr and Chlorine chemistry. This is followed by the ONO bottom oxide growth of ~100Å followed by a 120Å LPCVD nitride deposition. The ON stack is then patterned and etched to ensure removal of the (ON) bottom oxide and nitride stack from the periphery logic areas. A sacrificial oxide growth and etching, a final 240Å HV gate oxide is grown using an 830C steam oxidation cycle. This is followed by the deposition of 2000Å of Control gate and HV gate polysilicon using LPCVD deposition at 650C. The gate polysilicon is then patterned and a single stacked gate etch is employed to define the stacked gate bitcell transistor (control gate / ONO/floating gate) and the periphery high voltage transistor gates. After the gate etch and resist removal the 1T-Flash bitcell sidewalls are protected with a light 100Å 950C sidewall oxidation which also serves as the sacrificial oxide for the subsequent flash bitcell implants. The final post gate etch and poly re-oxidation is shown in Figure 4.3 below.

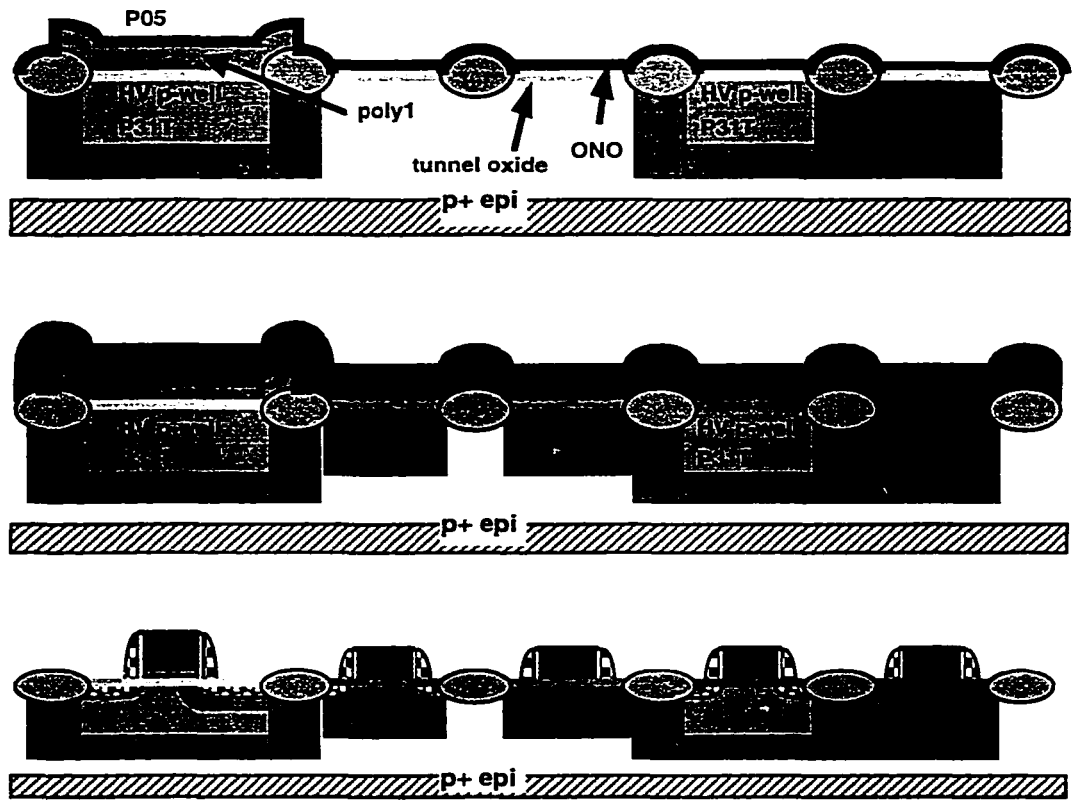


Fig 4.3: 1T-Flash Gate Stack and Tunnel Oxide Formation

4.6 Flash Drain and Source Engineering

The flash drain and source implants are designed to achieve an abrupt junction at the drain side with a sufficient high surface concentration of dopants to enhance Fowler-Nordheim tunneling which is crucial for effective and fast programming time. The memory bitcell region is photo lithographically defined for the source drain implants. The drain implant comprises of Arsenic on the order of 5E15 dose at 50KeV, followed by the co-implant of Phosphorus at 1E14 30KeV which serves to grade the Arsenic junction and thus reduce Band-to-Band leakage as shown in Figure 4.4. The Phosphorus co-implant is followed by a Boron Halo of 1E13 50KeV which serves an effective transistor punch-through suppression implant thus ensuring an adequate transistor leakage. Following the Flash implants is the P-channel high voltage transistor LDD implants. This is a 20KeV 3E13 Boron implant. This is then followed by an LPCVD deposition of 200Å of TEOS liner and 600Å of Silicon Nitride which is subsequently etched to form the spacer both to protect the Lightly-Doped regions (LDD) and simultaneously enable the formation of self-aligned silicidation. The post spacer processing is followed by the N-channel source/drain implants of 6E15 Arsenic at 60KeV with a Phosphorus co-implant of 5E13 at 50KeV. Following the N-channel source/drain implant is a short RTA of 50 sec to activate the dopants. This is followed by the masking and formation of the P-channel source/drain implants of BF2 at 3E15 30KeV which is also followed by a short RTA of 20 sec for dopant activation.

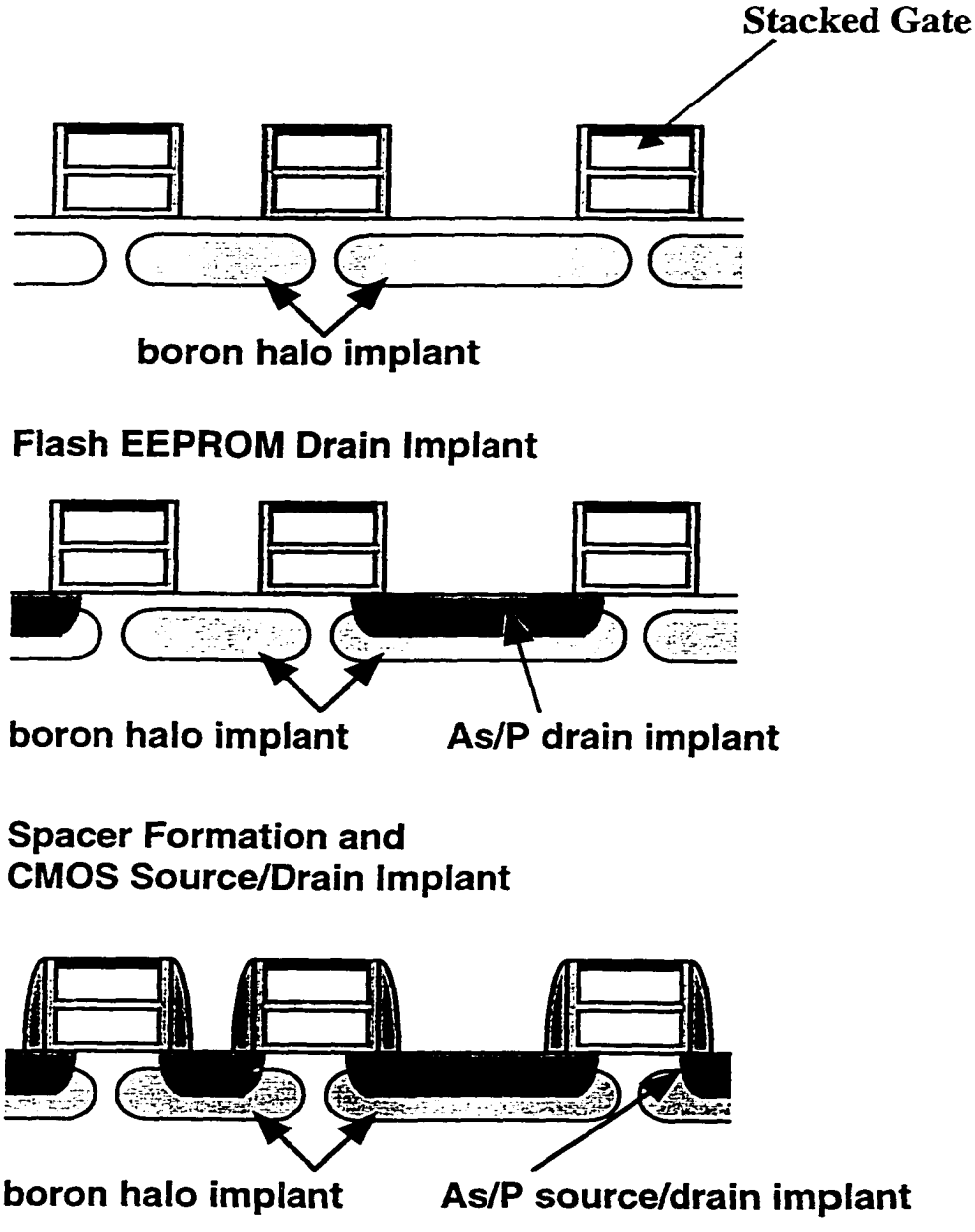


Fig. 4.4: 1T-Flash Drain Formation

Following the Flash transistor formation a 200Å of screen oxide is removed using a dilute HF, which is followed with the deposition of 400Å of Titanium. This is annealed in an RTA tool at 650C for 60 sec in Nitrogen ambient. After the selective removal of un-reacted titanium from spacers, the $TiSi_x$ salicide of C-49 phase is then annealed again in the RTA tool at 850C for 20sec in Nitrogen ambient to stabilize and transform the SALICIDE to the C-54 phase thus achieving a lower sheet resistance.

4.7 Contact Formation

Following the SALICIDE formation, a 1000Å layer of Plasma-Enhanced Tetra-Ethyl-Orthosilicate (PETEOS) which is deposited followed by 18000Å of BPSG to serve as the isolation between metal-1 and the polysilicon. This deposited oxide composite is then annealed at 700C for 30sec to improve planarization over topography and also to densify and stabilize the glass. This is followed by Chemical Mechanical Polishing (CMP) of the oxide which is a necessity for advanced sub-micron process technology, to improve lithography depth-of-focus and minimize tungsten stringers from the subsequent metallization integration. After the contact lithography process the contacts to all devices and array terminals are etched using an advanced AMAT5000 etcher. This is followed by the deposition of Ti/TiN glue layer of 500Å/700Å thickness. The barrier layer is then RTA annealed at 650C in NH3 , after which is the PVD deposition of Tungsten to fill the contacts and then polished to produce plugs in the contacts. This is shown in Figure 4.5 below.

Figure 4.5: Contact Formation

4.8 Metalization and Passivation

The metallization begins with the sputtering deposition of Ti/TiN/Aluminum/TiN stack of 400Å/800Å/8000Å/400Å using an advanced Applied Materials Endura Chamber. This is followed by the lithography of the metal and an RIE etch using BCl3 chlorine chemistry which defines the metal lines. This is followed by the deposition of 1000Å of PETEOS, 4000Å of O3-TEOS and capped with 17000Å of PETEOS. This Inter-Level Dielectric (ILD) between metal-1 and metal-2 is then planarized using CMP. The planarized ILD is then patterned using the via mask to open the via connections. The via is then etched and subsequently filled with 400Å of TiN glue followed by 6000Å of Tungsten which is also

planarized using CMP. This is followed by the deposition of the metal-2 stack of 200Å Ti/300A TiN/ 6000Å of AlCu/ 100Å Ti/ 700Å TiN. The stack is lithographically patterned and again etched using the exact same chemistry as metal-1. Since three levels of metal is required to complete the addressable Flash memory arrays the metal-2 ILD dielectrics and metal thickness were repeated for metal-3. This is followed by a forming gas anneal of 390C 30min in a hydrogen ambient to passify interface states and dangling bonds. This is followed the deposition of 7000Å of Silicon-Oxynitride (SiON) as passivation layer using AMAT5000 PVD deposition chamber. The passivation is etched out from the bondpad areas to expose the metal-3 aluminum to facilitate probing of the devices. The addressable 2MB arrays are then diced and packaged for array characterization. The final device is shown in Figure 4.6.

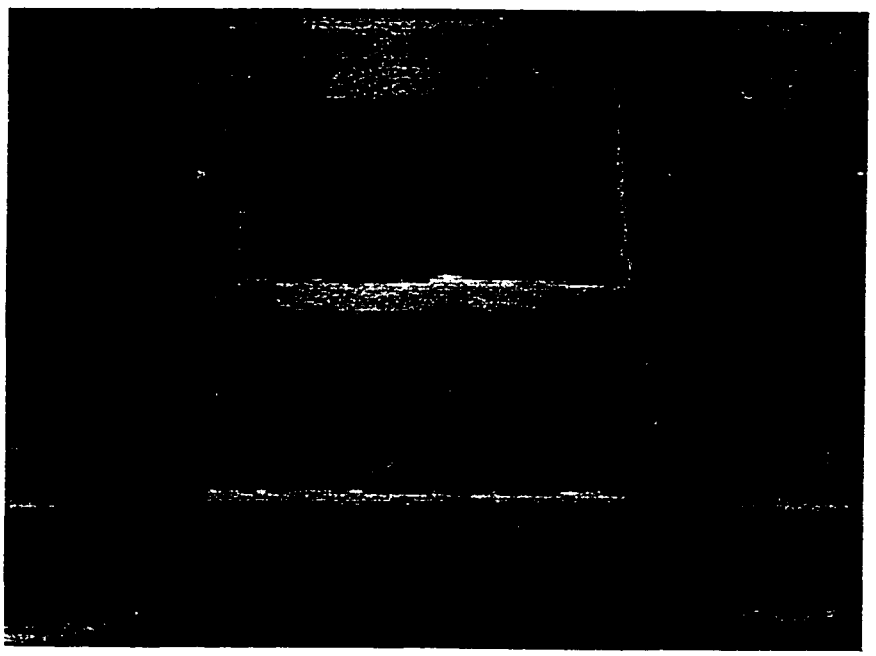


Figure 4.6: Fully fabricated Transistor

Chapter 5

Theoretical Analysis of 1T-Flash Memory Behavior

5.1 Introduction

The erasing method of ejecting electrons from the Floating Gate (FG) to a graded source junction by Fowler-Nordheim tunneling current is widely used in NOR-logic based Flash EEPROM. In a Flash cell, the electron ejection can also be performed on the drain side and thus used for cell programming [31]. The tunneling operation takes place over a small FG to Drain junction overlap ~ 0.6 x 10^{-9} cm^2, which is less prone to defects. As a result, identical program/erase behavior is expected for all bits. The use of F-N tunneling as a programming mechanism facilitates low voltage read operation without the added complications of word-line boost voltage circuitry. The F-N tunneling current is a sensitive function of the drain to FG overlap, and the associated junction process parameters [32]. The problems of over-erase and fast programming bits have plagued non-volatile memory manufacturers for sometime, but have been contained with intelligent programming/erase techniques at the expense of program/erase times. In the classical ETOX™ Flash arrays, F-N tunneling is used for the bulk erase on an entire array or sectors via a tunnel window in the source - FG overlap region. The resulting sub-threshold characteristics may not be the same for each bitcell, which results in non-uniform threshold voltage(V_t) distribution [33]. This class of memory bitcells with abnormally low V_t is referred to as over-erased bits,

however when F-N tunneling is employed during programming such bits are considered fast bits.

A fast bit is defined as a bit whose programmed memory V_t leads the main population by ~0.5V or less, while a normal bit is one whose programmed V_t is close to the median of the population. Fast and Normal bits are quantified and characterized by examining bitcell programming and erase stability, behavior after UV-erase and Data Retention. To further understand the role of FG poly grain morphology effects on tunneling currents, various FG -poly deposition and doping techniques were examined, which were further correlated to the memory programmed V_t distribution of bits in a 2MBit Flash EEPROM array.

In this research, we report for the first time another class of over-erased bits associated with 1T- Flash EEPROM memory arrays. A number of these bits were found to be insensitive to annealing and do not exhibit the type of erratic behavior previously reported [11]. We report the effects of flash "channel" programming, or severe gate disturb, on the threshold voltage of fast or over-erased bits. Experiments were performed to establish that this class of fast-programming bits are non-erratic and remain fast after 250°C bake. These fast bits exhibit identical sub-threshold characteristics similar to that of a normal bit after UV-erase, thus establishing that the initial charge stored on the floating gate is the same for both normal and fast bits. Polysilicon grain boundary enhanced electric fields which results in impact ionization by tunneling electrons, thus generating trapped positive charges in the grain boundary oxide ridges is believed to play an important role in the generation of fast bits.

In this chapter we will discuss the fundamentals of 1T-Flash device design approach used during this research, the 1T-Flash reliability and conclude with a comparison of compact model results with experimental data. We will begin with a quantitative description of the transistor design. The tunnel oxide and ONO effects on the Flash programming and erase characteristics will also be discussed. In addition the floating gate polysilicon engineering and its effect on the F-N tunneling and subsequently its influence on programming characteristics of the 1T-Flash bitcell will also be discussed. Furthermore the device leakage optimization and the careful design and trade-off of short-channel margin for abrupt junction to enhanced programming speeds will be analyzed. The device design will center on the use of advanced halo drain engineering structure to control the transistor off state leakage. In the sub-section for 1T-devices, the effects of phosphorous and Arsenic flash drain junction optimizations will also be discussed. This will encompass the trade-off between fast programming time as a results of a very abrupt junction and band-to-band leakage.

In the 1T-Flash reliability section, the flash programming and erase endurance characteristics will be discussed. The unique behaviors of the observed fast programming bits will be discussed including the associated characterization results. This will include the effects of the floating gate polysilicon microstructure on the program and erase characteristics of the fast programming memory bitcells.

In the modeling section we will compare the results from the compact models developed in chapter 2 to the experimental results. We will then attempt to explain the observed differences in the characteristics and how they compare to the compact model.

5.2 1T-Flash Device Design

5.2.1 Charge Storage in the Floating Gate

The floating gate in a Flash device forms an equi-potential plane between the control gate and the substrate, and is parallel to both of them. This kind of equi-potential plane does not exist in ordinary MOSFETs. As previously discussed in chapter 2, the floating gate voltage is determined by capacitive-coupling between the floating gate and the externally applied voltages more specifically the control gate and the drain voltages as shown below in Figure 5.1.

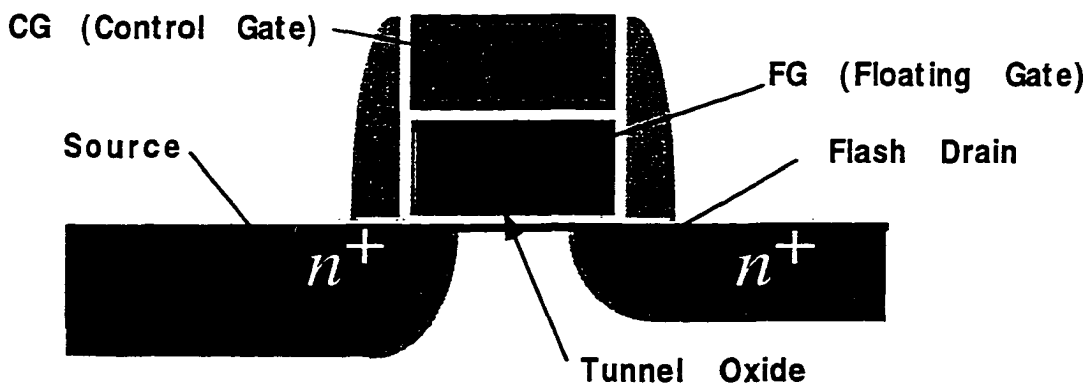


Figure 5.1a: Perpendicular cross section across a 1T-Flash EEPROM cell wordline.

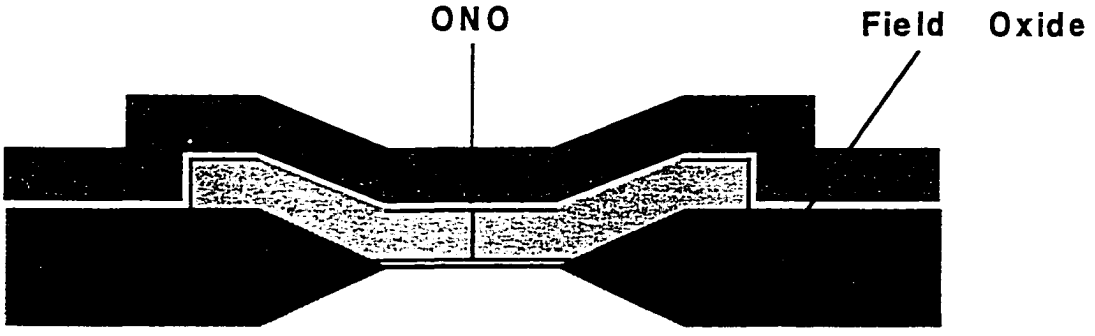


Figure 5.1b: Perpendicular cross section along a 1T-Flash EEPROM cell wordline.

The tunnel oxide, ONO and floating gate form a potential well which "traps" the floating gate charge as shown in Figure 5.2 below using a pictorial representation of the resulting electron probability density function. The presence of the floating gate charge leads to a modified transistor characteristics, since the stored charge induces an electric field across the dielectrics to obey Gauss' law:

$$\oint EdA = \varepsilon_{ox} Q_{FG} \text{ ----------} \quad (5.1)$$

The induces field changes the threshold voltage of the NMOS 1T-Flash bitcell transistor:

$$V_T = \Phi_{GS} + 2\phi_F + \frac{\sqrt{4\varepsilon_{si} q N_B \phi_F}}{C_{eff}} - \frac{Q_{FG}}{C_{total}} \quad \text{-------} \quad (5.2)$$

The delta shift in threshold voltage (ΔV_t) is linearly proportional to the amount of charge (Q_{fg}) stored on the floating gate. Hence for an NMOS-based 1T-Flash bitcell, if Q_{fg} is positive then it implies that the cell is in the programmed "1" state since the effective Vt from equation (5.2) will be lower. Conversely when Qfg is negative, then the threshold voltage is high and the bitcell is in the erased "0" state. This is very evident in Figure 5.3 below which compares the transistor drain currents as a function of the applied control gate voltage.

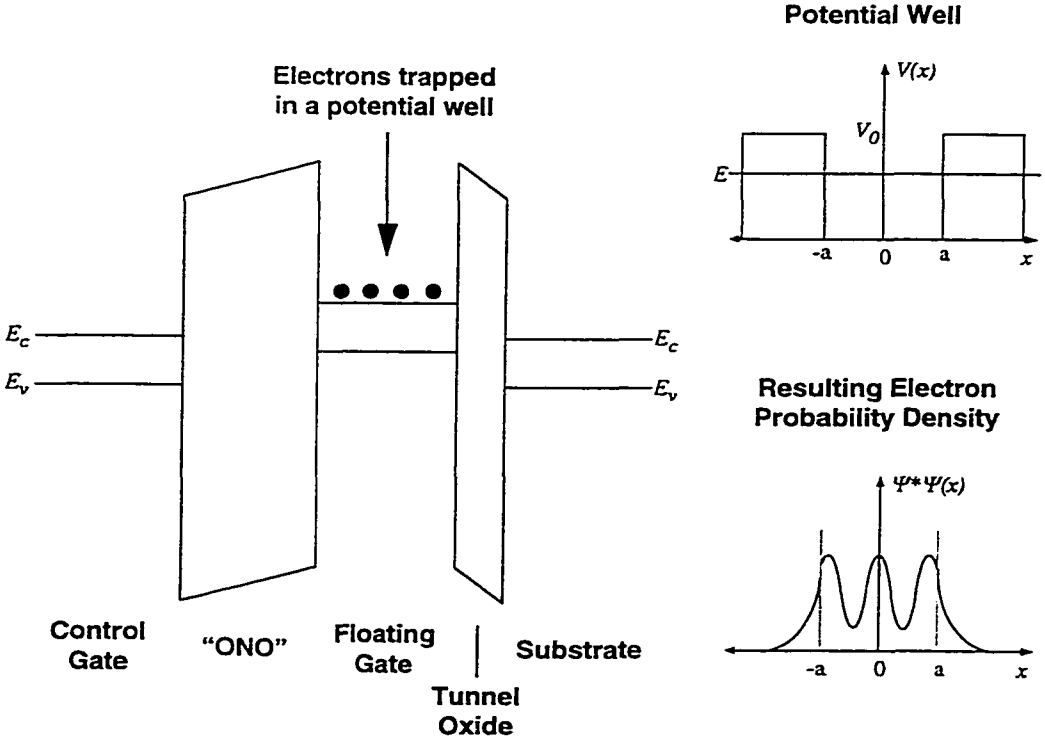


Figure 5.2: Potential Well Diagrams of the Floating gate Structure in Flash EEPROM

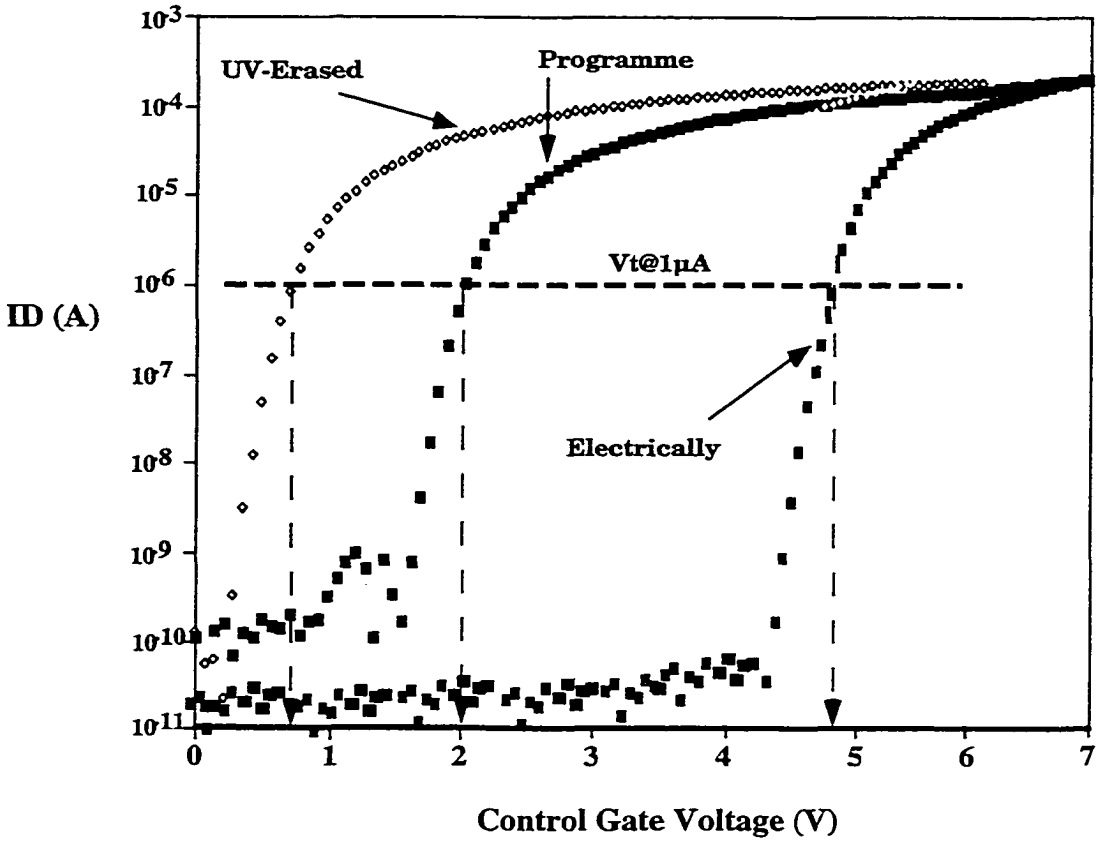


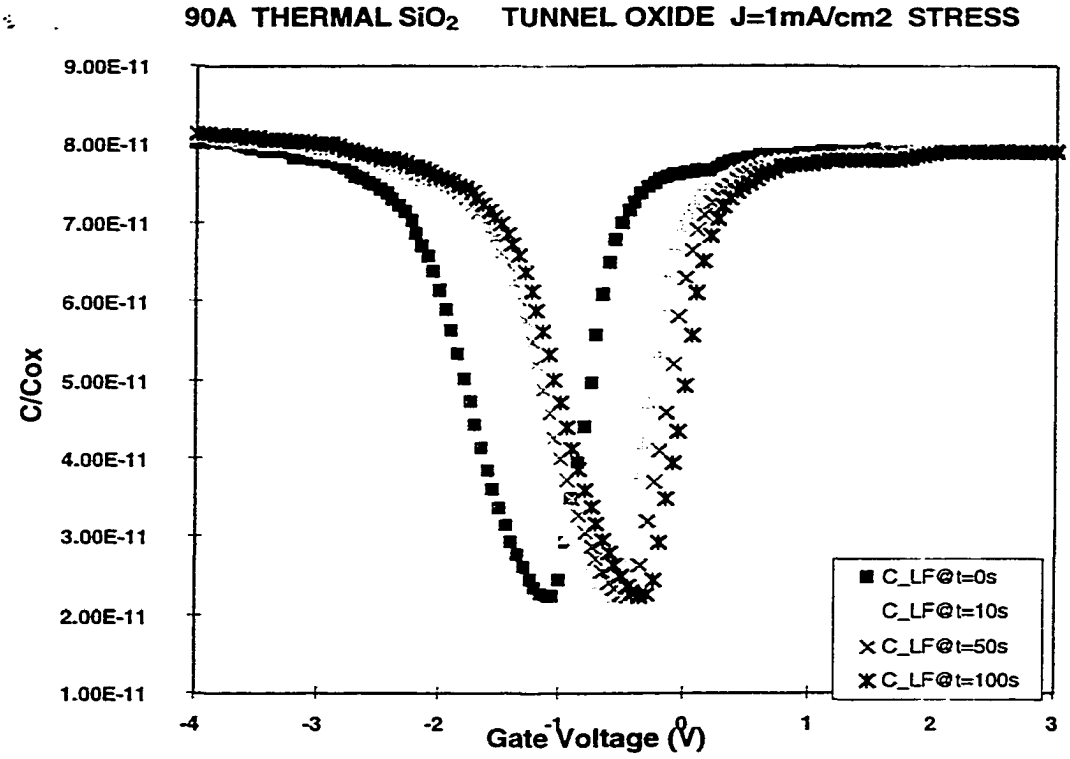
Figure 5.3: Subthreshold characteristics of 1T-Flash Memory NMOS Transistor

5.2.2 Tunnel Oxide

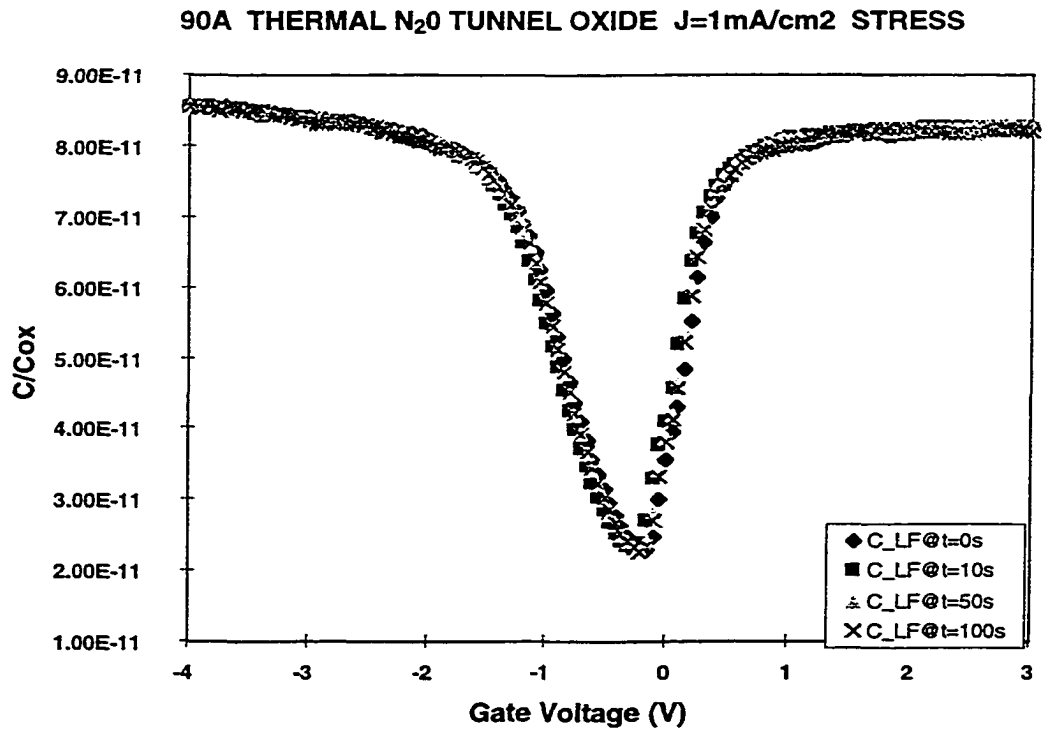
The growth of defect free ultra thin (< 100Å) tunnel oxide dielectrics reproducibly, using SiO_2 or oxy-nitrides, is a necessary requirement for the development of floating gate non-volatile semiconductor memories. These memories are required to operate at low programming voltages whiles simultaneously providing access times and reliability for long-term storage. The key physical property requirements for tunnel oxides includes a high

breakdown fields, a low electron trap density for high cell endurance and low gate leakage currents. For this research work, both SiO_2 (pure) and Nitrided Oxides were investigated. The stringent programming and erase fields of F-N tunneling requires a tunnel oxide that can withstand the high fields . As can be seen in Figure 5.4a and 5.4b, the N_2O based tunnel oxide provides the superior resistance to oxide damage, which is required for a reliable 1T-Flash operation. The Low-Frequency CV plots show that after some constant current stressing the thermal oxide traps more electrons compared to the N_20- based tunnel oxide. The N_20 tunnel oxide also appears to have a better capacitance recovery. The threshold voltage shift with increasing stress time leads to a lower charge fluence for the pure SiO_2-based tunnel oxides, thus rendering it susceptible to poor endurance cycling performance.

In the nitrided tunnel oxide formulation, N_20 is used as the nitrogen source. The stronger Si-N covalent bonds improves the interface by replacing the terminating Si-O bonds. This bonding replacement is effective and subsequently results in a significant reduction of electron traps. The lower electron trapping of the N_20-based tunnel oxide in comparison to thermal SiO_2 is shown in Figure 5.5a. The gate voltage shift as a function of stress time which is a measure of electron trapping clearly reveals that the over a long stress time the electron trapping in the N_20-based oxide is significantly less than thermal oxide. This inherently enhances the program/erase endurance of the 1T-Flash memory as shown in Figure 5.5b.

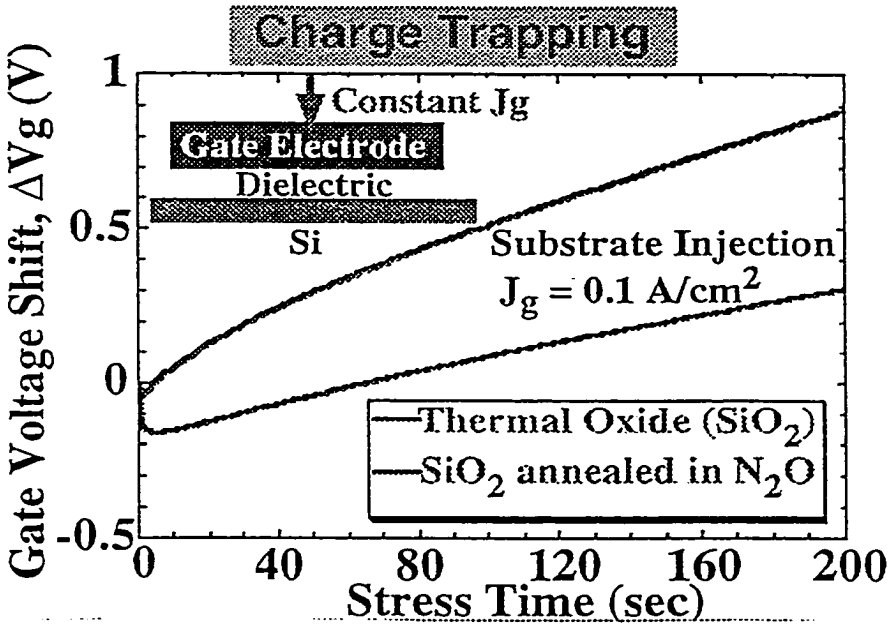


(a)



(b)

Figure 5.4: a) Low Frequency C-V plots for thermal SiO_2 b) Low Frequency C-V plots for N_2O tunnel oxide for 1T-Flash EEPROM.



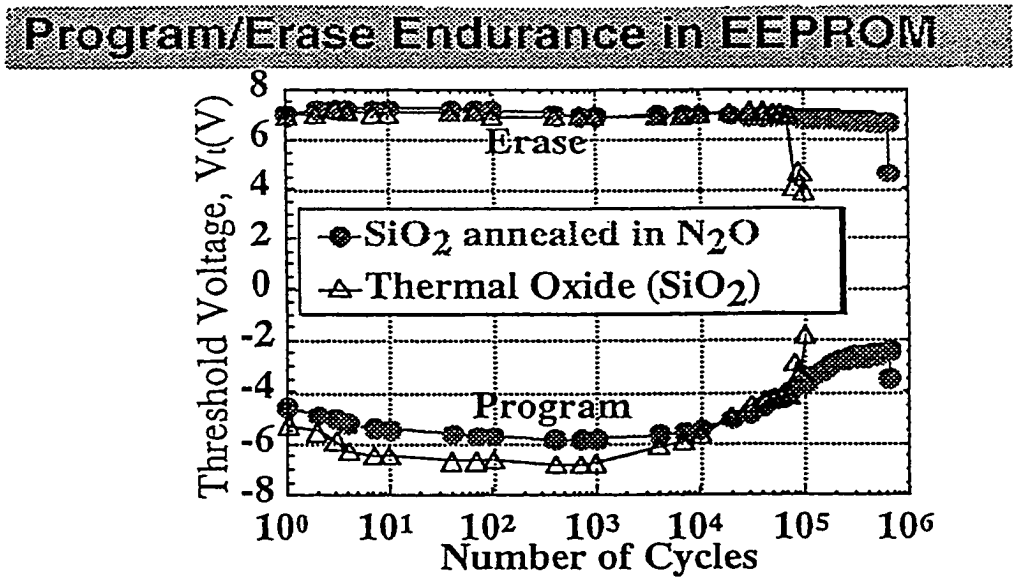


Figure 5.5: a) Gate voltage shift vs. Stress time b) Threshold voltage shift vs. P/E Cycles

Arhenius plots were also established to characterize the activation energies of the different tunnel oxides. As shown in Figure 5.6a, the N_20-based tunnel oxide shows about .05eV higher activation energy, which validates the longer program erase endurance cycling shown in Figure 5.5b. In addition the normalized erase currents remains fairly constant over longer P/E endurance cycles as compared to the thermally grown tunnel oxide as shown in Figure 5.6b, again validating the superiority of the N20-based tunnel oxides.

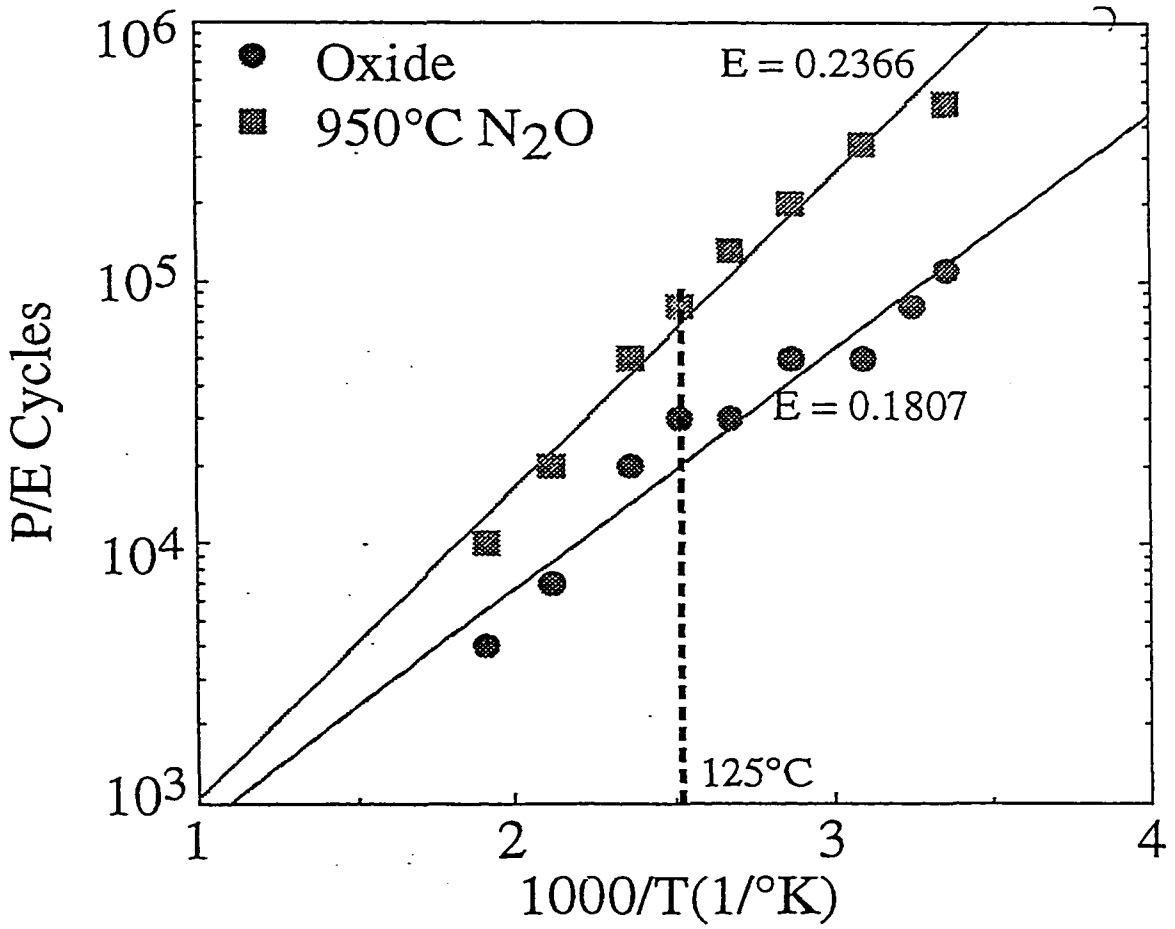


Figure 5.6a: Arhenius of P/E cycles as a function of temperature for N_2O and thermal SiO_2 tunnel oxides.

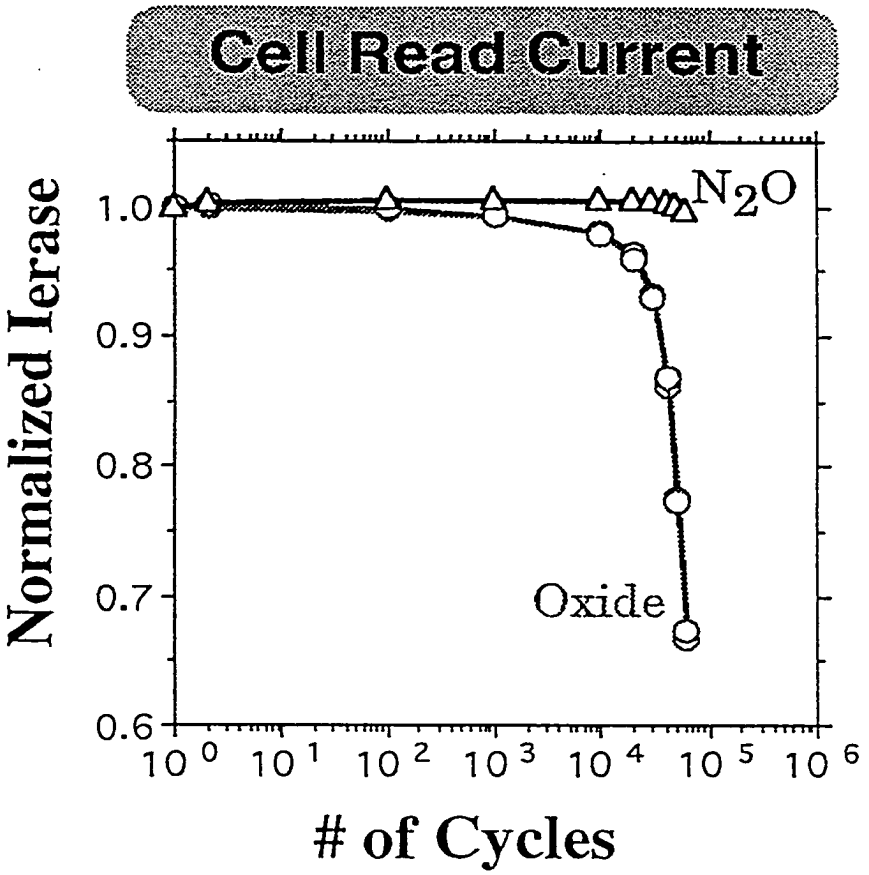


Figure 5.6b: 1T-Flash cell Read Current as a function of P/E Cycles.

To complete the tunnel oxide characterization, the oxide high field breakdowns for the N_20-based and thermal $Si0_2$ tunnel oxides were evaluated. As shown in Figure 5.7a both oxides exhibited similar breakdown characteristics. However keen differences were observed in the charge –to-breakdown (Q_{bd}) measurements. The N_20 oxide clearly can withstand more stress with electron injection from the floating gate polysilicon to the substrate as compared to the thermal oxide. This also supports the superior cycling endurance of the N_2O oxide as discussed earlier. The program and erase vs. time characteristics of both types of tunnel oxide are comparable over a short duration of stressing, as evident in Figure 5.8a and 5.8b.

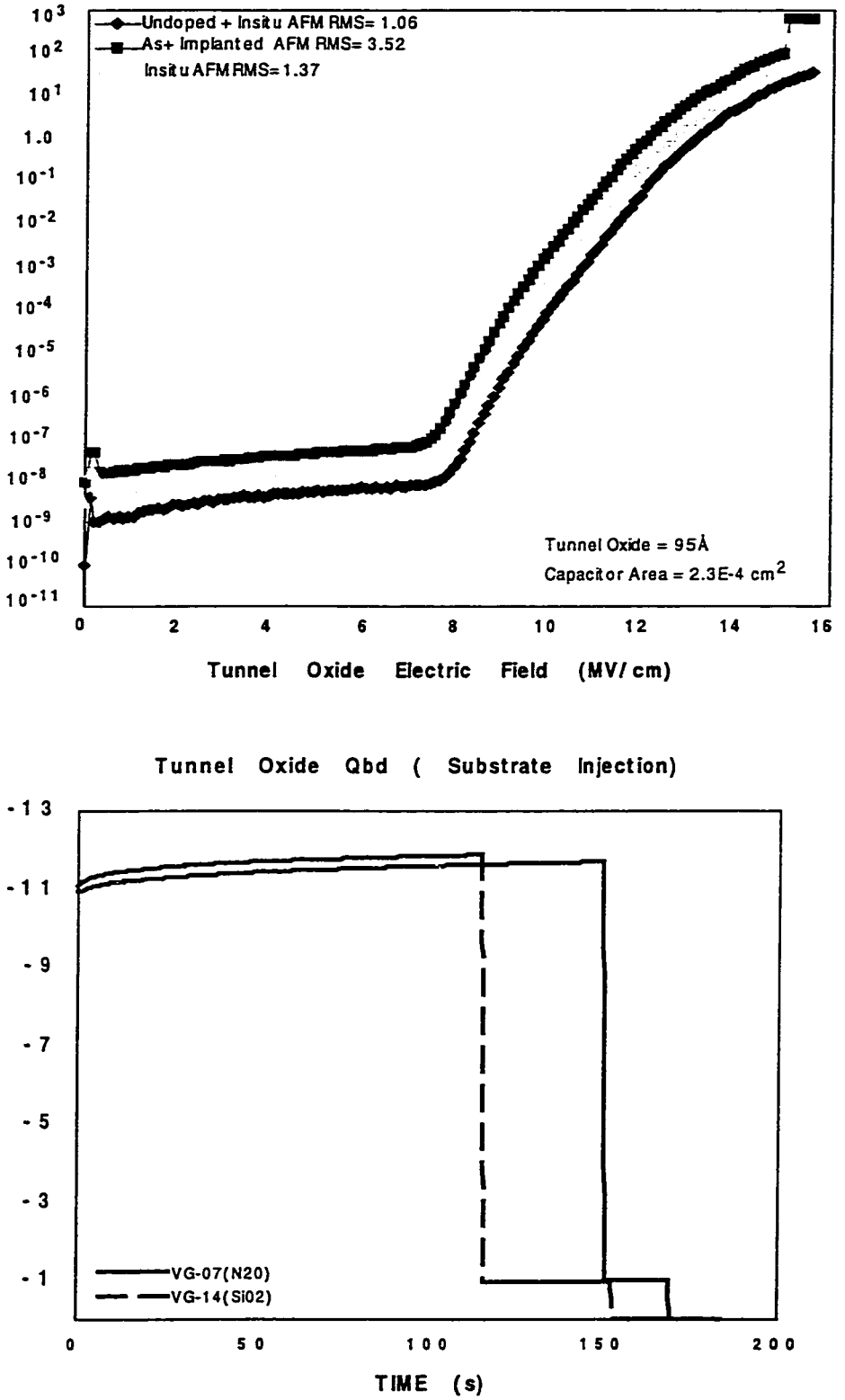
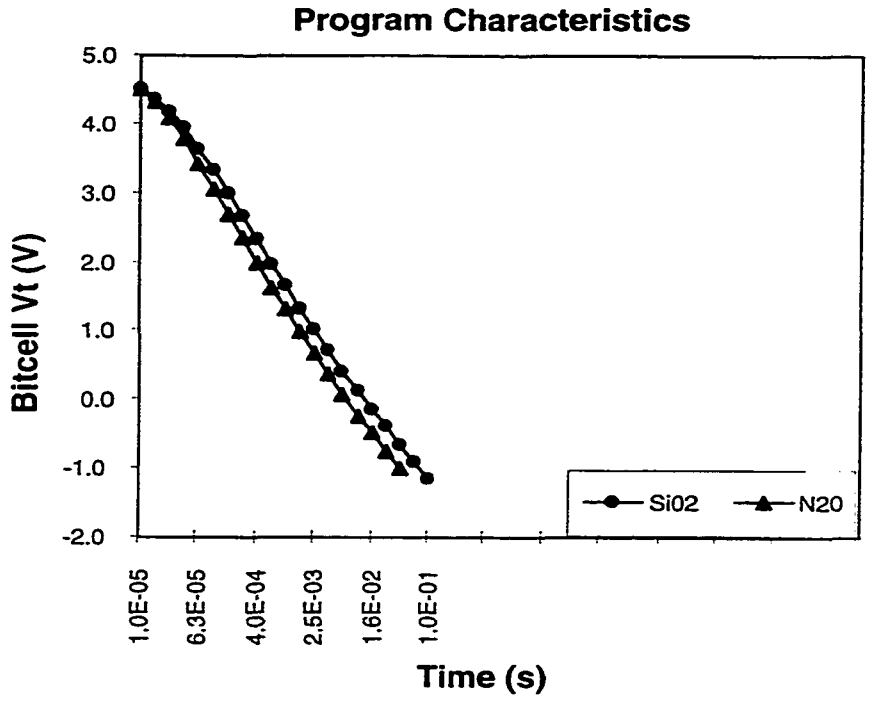


Figure 5.7: a)Tunnel oxide breakdown voltages for SiO_2 and N_2O tunnel oxides b) Qbd for N_20 and SiO_2 tunnel oxides for 1T-Flash EEPROM



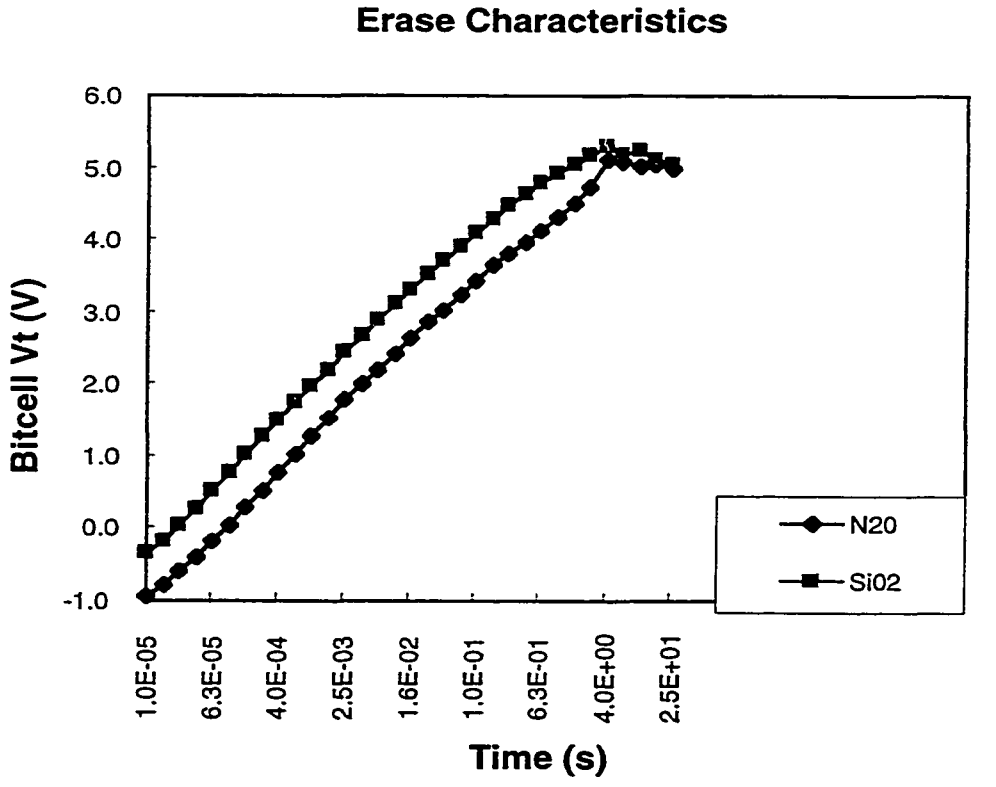


Figure 5.8: a) 1T-Flash memory program threshold voltage vs. time for thermal SiO_2 and N_2O tunnel oxides. b) memory erase threshold voltage vs. time for thermal SiO_2 and N_2O tunnel oxides.

5.2.3 Inter-Poly Dielectric (ONO)

The inter-poly dielectric (ONO) as shown in Figure 5.1a provides the necessary isolation between the control gate polysilicon and the floating gate polysilicon. This ONO (Oxide-Nitride-Oxide) formation is previously discussed in chapter 4. The ONO thickness was optimized during this research to balance adequate control gate to floating gate isolation, which tends to drive thicker ONO stack. This was balanced with the need to effectively couple a higher potential from the control gate to floating gate which means a higher ONO capacitance which drives thinner ONO stack. As shown in Figure 5.9a and 5.9b the ONO stack was optimized to achieve adequate programming and erase speeds with appropriate gate and drain disturbs. Any compromises of these variable could lead to poor data retention and subsequently memory failure. As shown thinner ONO stack increases the programming and erase speeds considerably, however as will be shown later in the reliability section, this tends to drive poor data retention characteristics.

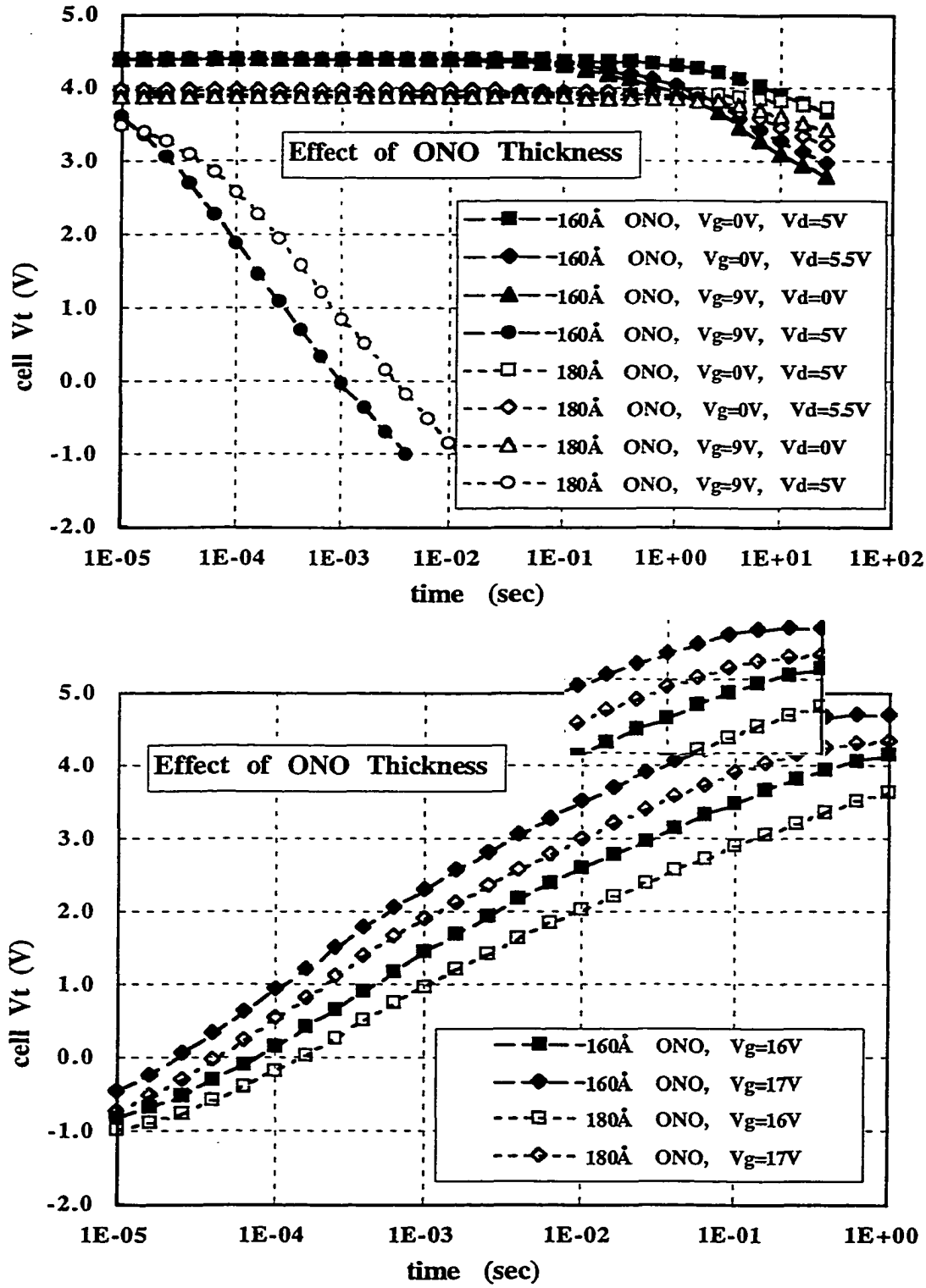


Figure 5.9: Program and Erase characteristics of 1T-Flash cell showing effect of ONO thickness on program and erase speed.

5.2.4 Floating Gate Polysilicon Engineering

Experiments were conducted to investigate the possible causes of the fast programming bits. "Channel" programming or severe gate disturb with V_{cg} = -16V, V_d, V_s floating and V_{sub}=0V, was used to modulate the V_t state of the fast bits. As previously quantified, a fast bit appears normal if the programming takes place over the entire channel region instead of the FG-drain junction overlap region. Further measurements of low B-B tunneling leakage at the flash drain junction indicates that the enhanced F-N tunneling is influenced at the FG-tunnel oxide interface as evidenced in Figure 5.10.

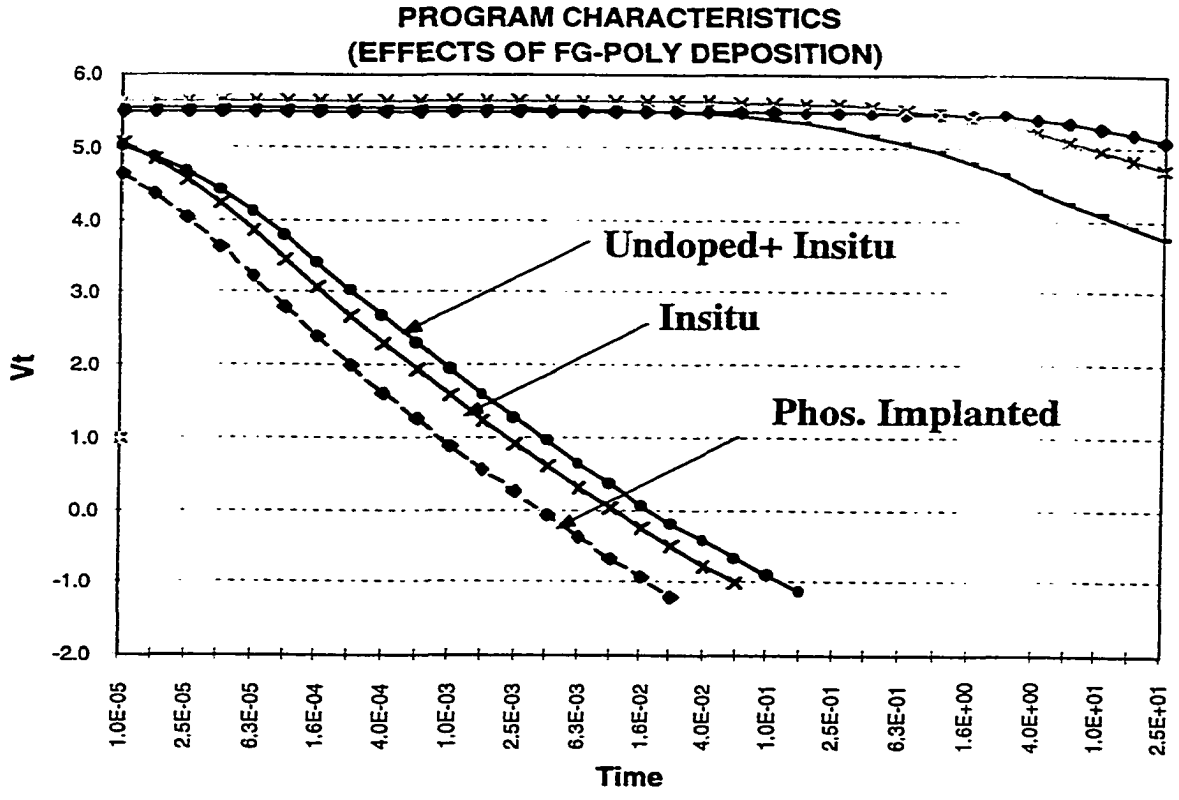


Figure 5.10: Effects of Floating gate polysilicon deposition process on program threshold voltage of 1T-Flash EEPROM.

To further this understanding, MOS capacitors were fabricated using the flash floating gate as the cathode on tunnel oxide. The 1500Å FG poly was deposited and doped in three ways: 1) sandwich of undoped plus phosphorus in-situ doped, 2) undoped followed by an implant and 3) in-situ doped. Figure 5.7 shows data for the three capacitors of the F-N tunneling currents J vs. E-field across the tunnel oxide. From AFM measurements, the undoped plus in-situ doped stack for the FG poly exhibited the smoothest morphology with the smallest RMS value of 1.06 which is a measure of the FG poly surface roughness. This was also confirmed with TEM grain size characterization as shown in table 2 below. The average grain size is largest for process - A (implanted) poly and smallest for process-C (undoped plus in-situ doped) stack, which correlates well with the AFM surface roughness characterization and the corresponding TEM micrographs in Fig. 5.21. The J vs. E-field characteristics in Fig. 5.7 shows that the undoped plus in-situ sandwich has reduced tunneling currents for a given E-field compared to the other stacks.

GRAIN SIZE INFORMATION	PROCESS-A (Å)	PROCESS-B (Å)	PROCESS-C (Å)
Median Grain Size	2540.3	2219.1	1643.4
Grain Size Range	1900 - 4564	1530 - 3360	962 - 1731
# Measured Points	53	53	53
Std Deviation	805.6	560.1	352.7
Variance	78.1	37.6	17.3

5.2.5 1T-Flash Leakage Optimization

The high terminal voltages required for an effective flash operation , imposes a very stringent limitation on the inherent ability to design an effective leakage suppression in the 1T-Flash design. The competing effects of fast programming times necessitates high drain doping concentrations with enhanced abrupt junctions. These requirements however conflict directly with the need for acceptable transistor leakage required for a low standby current. In this research Halo drain engineering approach is implemented to help optimize the flash performance for both a fast programming time and low device leakage. The process integration of the halo is shown schematically in Figure 5.11 below. After the gate etch process, the Boron halo is implanted using an optimized dose derived from simulations. This is followed by the implantation Arsenic and Phosphorus co-implants to provide a graded abrupt junction. This junction is optimized for fast programming and also minimal ban-to-band tunneling. As shown in Figure 5.12 Supreme simulations were used to optimize the halo dose and energy . The device was designed to place the peak laterally across the Arsenic peak thus keeping dopants in place and minimizing sub-surface leakage. The effects of the halo optimization on transistor leakage is shown in Figure 5.13. The halo dose of 1E13 at 50KeV provides the most optimum control of leakage. This was effective in controlling short-channel margins down to an Lgate of 0.35um and thus providing an Lmin margin of 80nm.

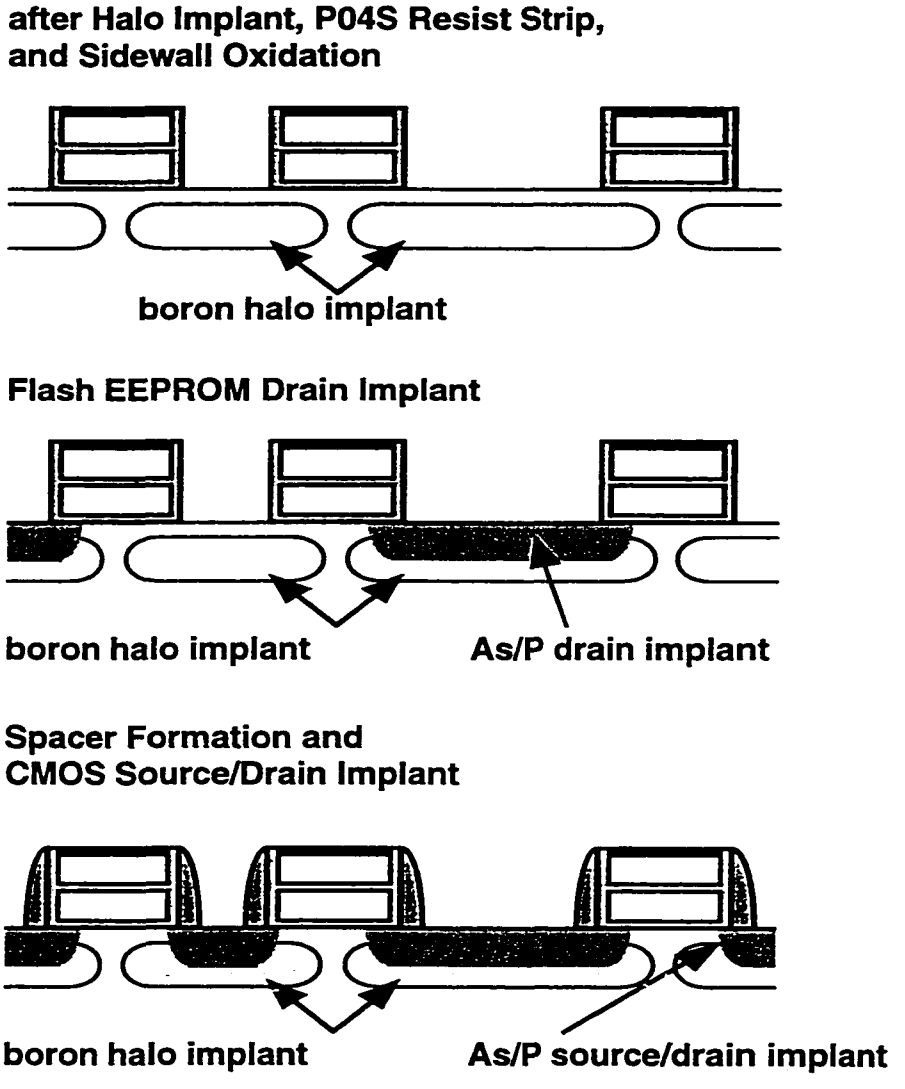


Figure 5.11: Schematic representation of Flash Drain junction Halo Optimization

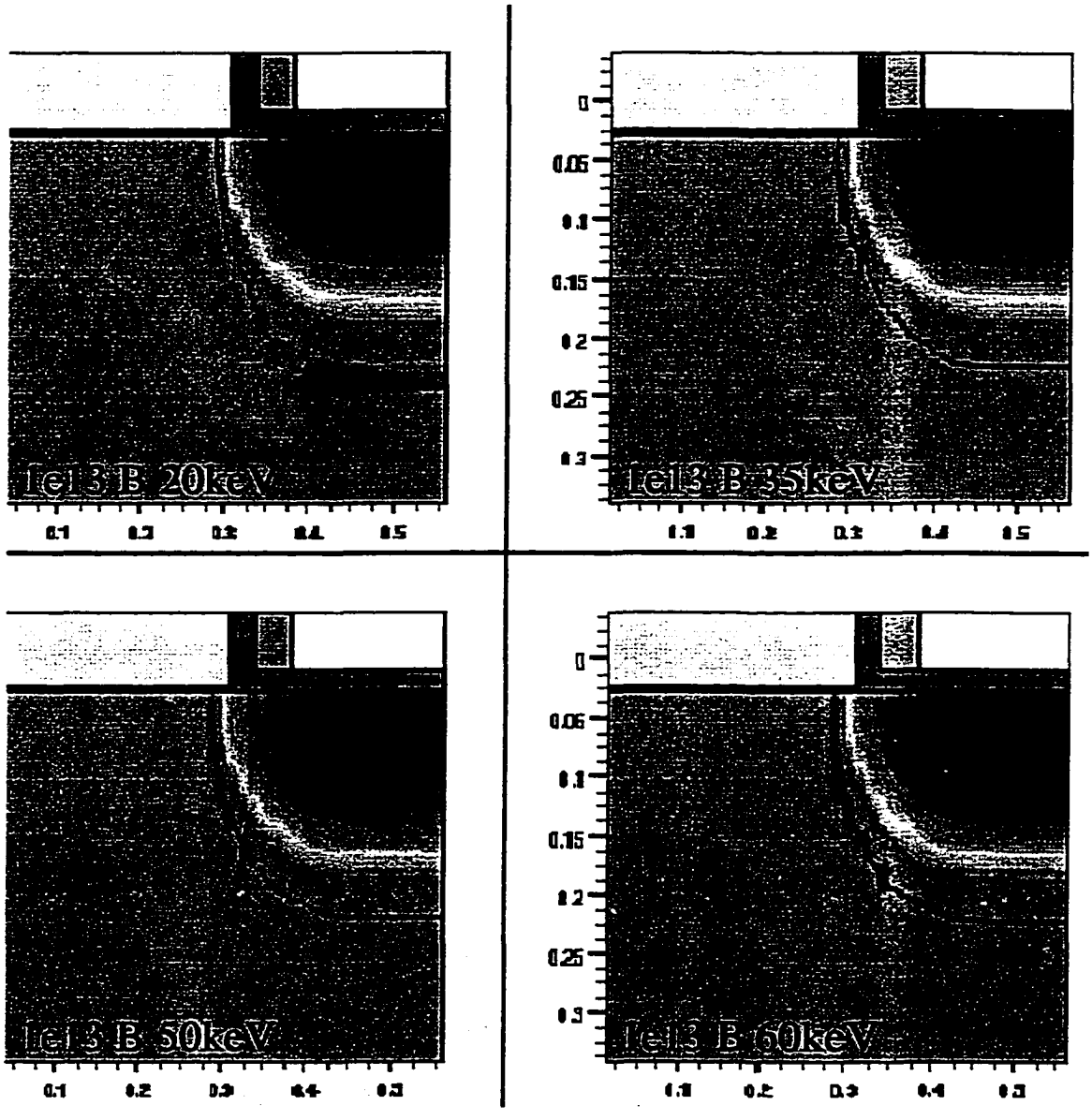


Figure 5.12: 2-D Simulation results for doping profiles of Boron Halo Flash Drain Optimization.

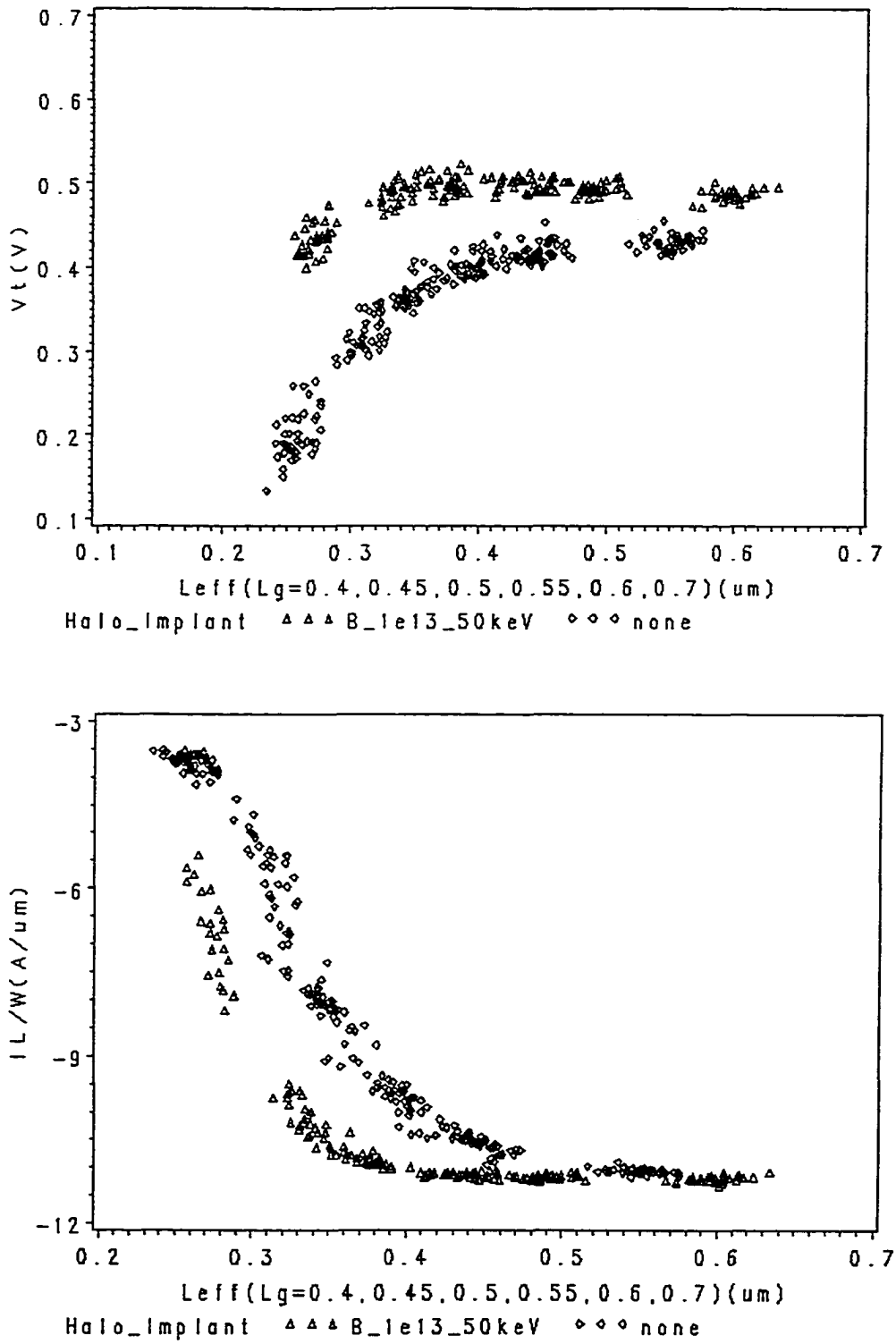


Figure 5.13: Electrical results 1T-Flash with Boron Halo effects on threshold voltage and leakage.

5.2.5 1T-Flash Source - Drain Junction Optimization

During the drain-side Fowler-Nordheim programming of a floating gate Flash device, a high voltage is applied to the source terminal and the control gate is grounded with the substrate. The floating gate typically assumes a potential of roughly 0.3V if the device is programmed and the source is kept floating, while the bias on the drain junction must be high enough to produce an electric field of at least 10 MV/cm in the tunnel oxide which is required for the Fowler-Nordheim injection of electrons from the floating gate into the drain diffused junction. This bias condition leads to the hole generation due to substantial band-to-band tunneling right under the Si/SiO_2 interface, especially near the junction corner inside the drain diffusion where the doping concentration is around (1E18-0E19) cm^{-3}. The fact that both the vertical and lateral fields contribute to the silicon band bending needed for the band-to-band tunneling underlies the significance of the junction profile for minimizing the hole generation.

In this research, the 1T-Flash drain was optimized using Arsenic and phosphorus co-implants. This junction engineering was optimized to achieve a fast programming and erase speed without increasing junction leakage and breakdown. Experiments were carried out using Arsenic only junctions in comparison to Phosphorus only and the Arsenic with Phosphorus co-implanted junction. Figure 5.14, compares the Arsenic with Phosphorus co-implant junction to the Phosporus only junction. The Phosphorus only junction exhibits a slower programming speed as shown in Figure 5.14b and a higher bitcell transistor leakage. As shown in Figure 5.14a the Arsenic with co-implanted Phosphorus junction can achieve optimum performance with a program threshold voltage of 1.5V at 1ms and a gate /drain disturb immunity of ~900ms which meets the requirements for page programming. Figure

5.15 compare the erase characteristics for the same junctions described above. Again the Arsenic with phosphorus co-implanted junction meets the requirements for the 1T-Flash sector erase implementation.

Band-to-band measurements were made following the drain junction optimization to ensure that the optimized flash device drain leakage's were still acceptable. As shown in Figure 5.16 at a drain potential of ~5V which is required for programming the B-B leakage is less than 100pA for the nominal transistor Lpoly of 0.5um.

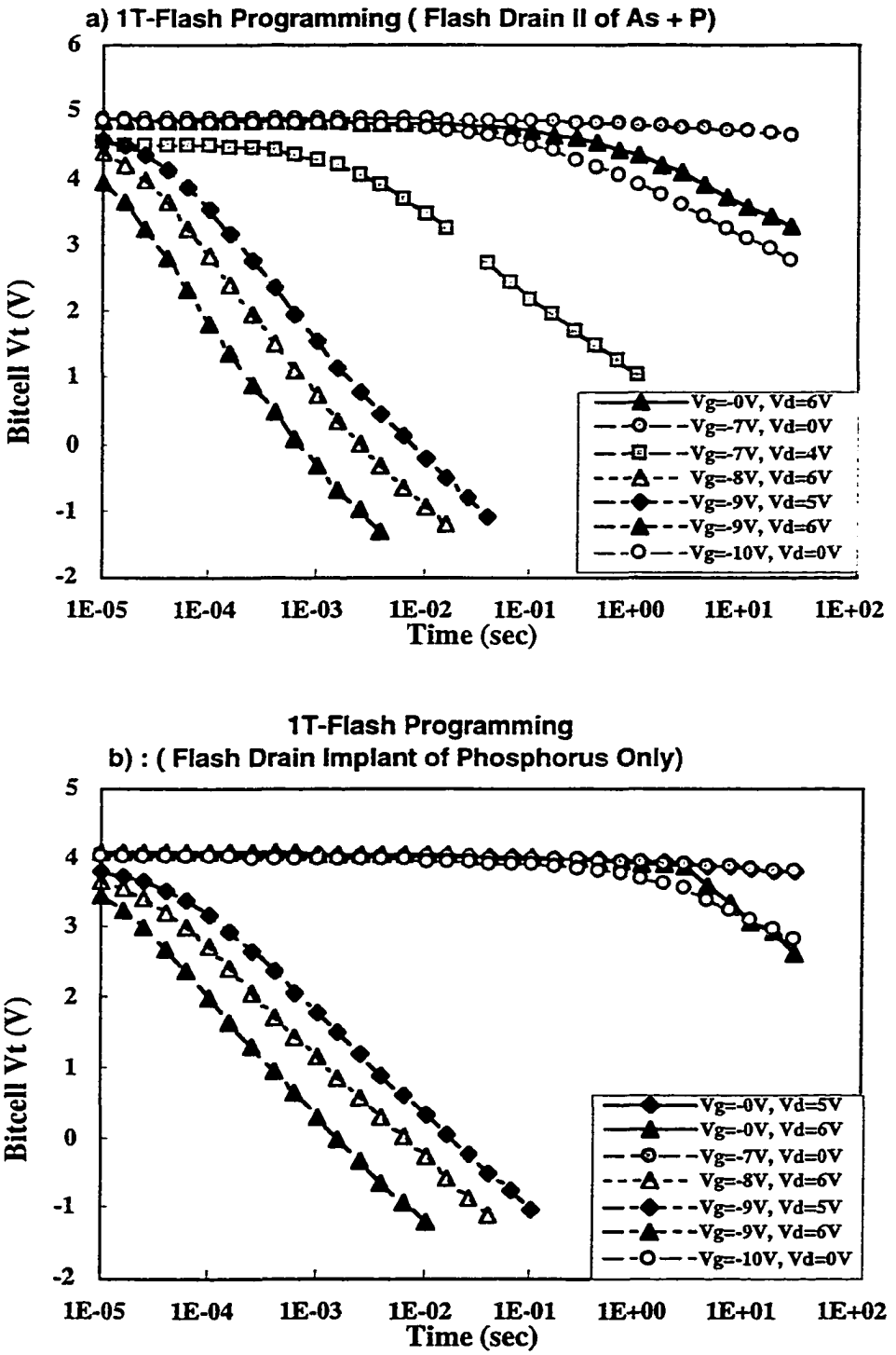
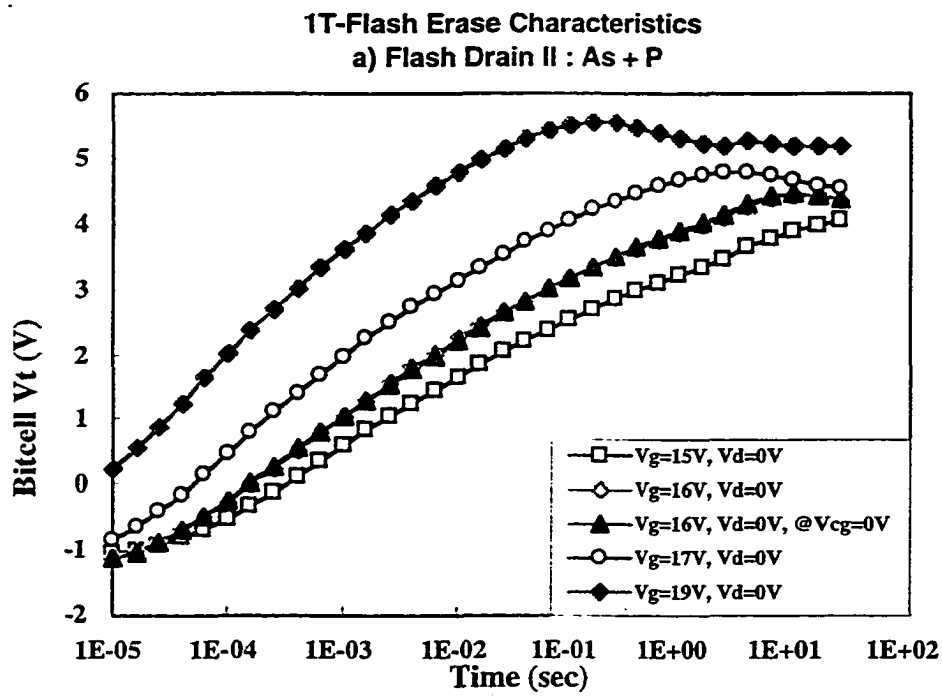


Figure 5.14: 1T-Flash Program characteristics for Phosphorus only vs. Arsenic/Phosphorus co-implanted junction.



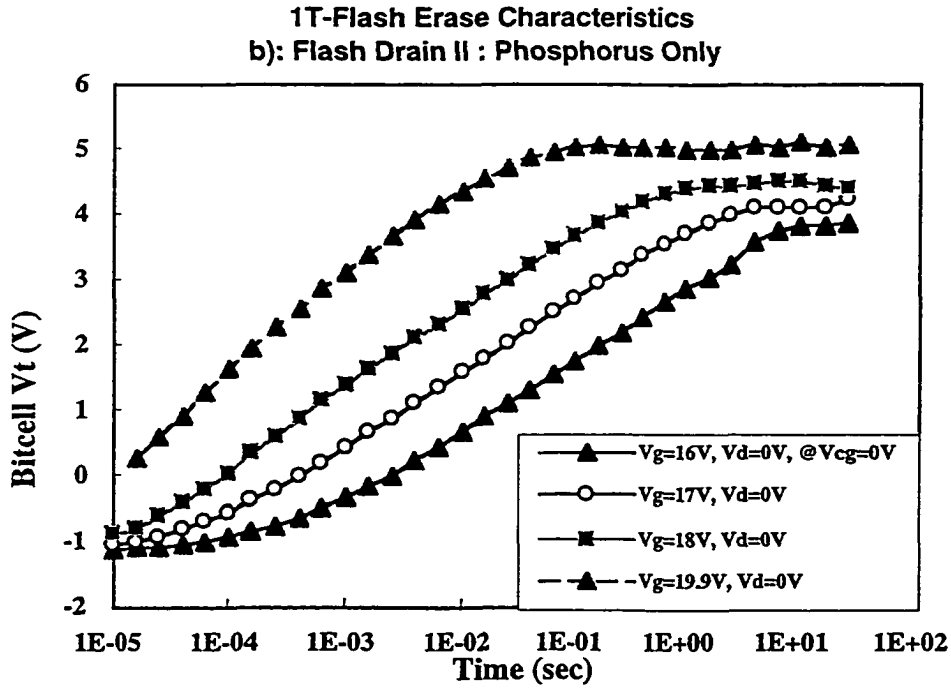


Figure 5.15: 1T-Flash Erase characteristics for Phosphorus only vs. Arsenic/Phosphorus co-implanted junction.

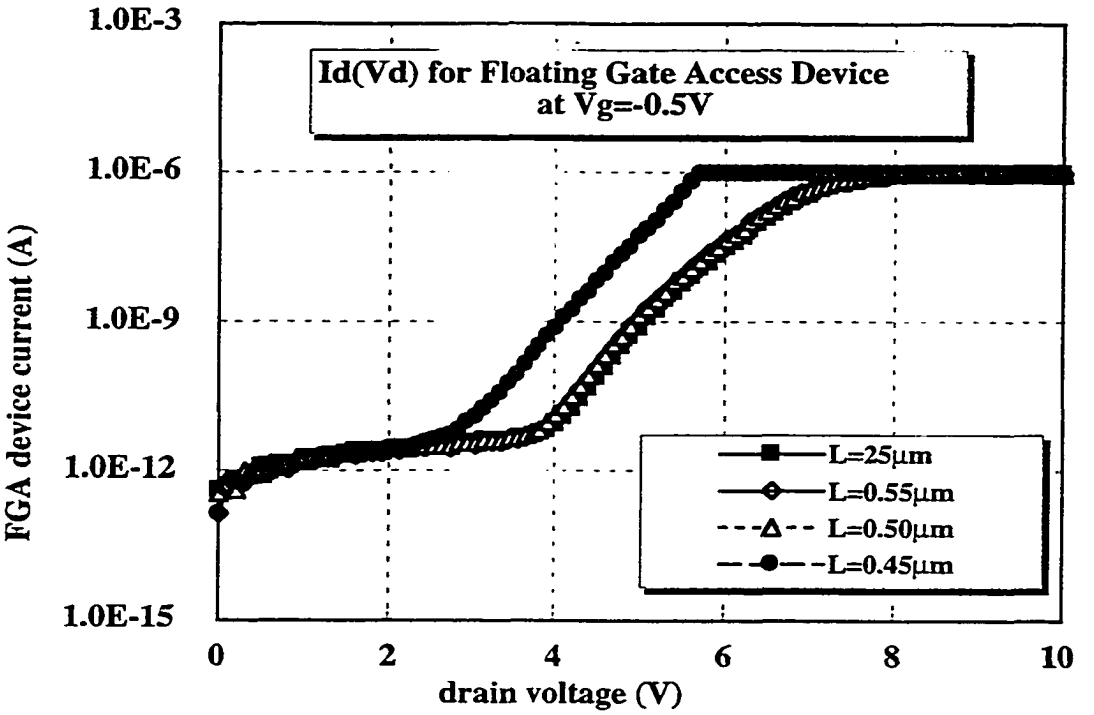


Figure 5.16: 1T-Flash Band-to-Band tunneling leakage current characterization.

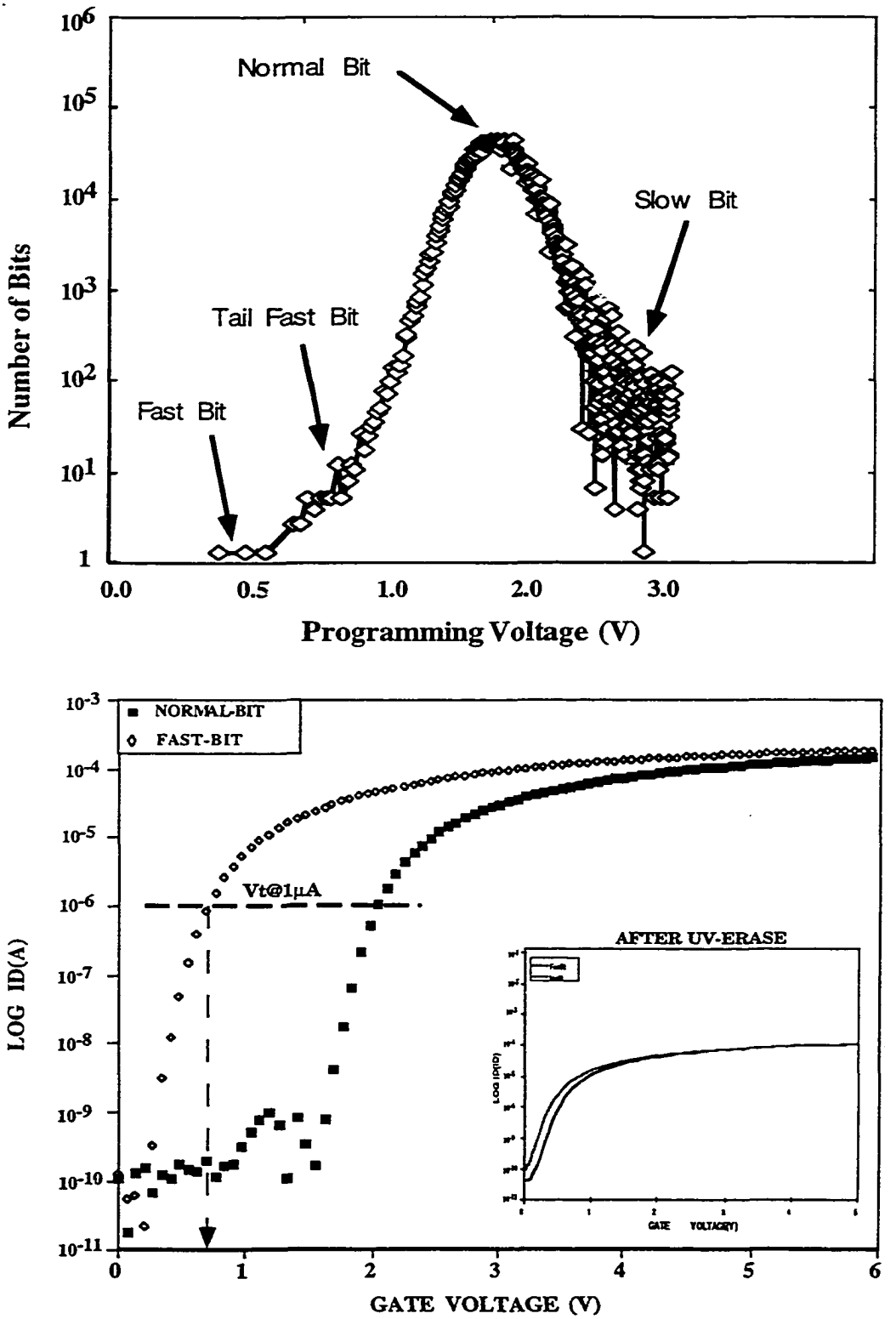


Figure 5.17: a) F-N tunneling programmed Vt distribution for 2-Mbit Flash array. b) subthreshold characteristics comparing programmed fast and normal bits pre/post UV-erase.

5.3 1T- Flash Reliability

As can be seen in Figure 5.17a, the V_t distribution of the fast or over-erased bits leads the normal population by ~1V after an initial programming with V_{cg}=-8V and V_d=5V. As shown in the insert of Figure 5.17b, the post UV-erase V_t for both over-erased and normal bits are essentially the same, thus indicating that the initial amount of charge on the floating gate of both fast and normal bits after UV-erase is the same. The program vs. time curves in Figure 5.18 compares the fast and normal bits both pre-cycling and post 24hr 250°C bake. It is evident that the bake had no effect on fast bit characteristics thus indicating that a fast programming bit has the same inherent stability as a normal bit. We report here a behavior of over-erased bits where after a few cycles (N > 50), the programmed V_t of the fast bit asymptotically increases to a higher steady state value where it remains, even after 6000 P/E cycles, as shown in Figure 5.19. Although the observed V_t spread of the fast bit is larger than the normal bit, there is no evidence of an erratic V_t behavior. We have consistently observed that the fast bits are always fast. Although the V_t increased with cycling, there was no evidence of the programmed V_t toggling between fast and normal V_t states.

Experiments were conducted to investigate the possible causes of the new class of fast bits. We observed that "channel" programming or severe gate disturb with V_{cg} = - 16V, V_{sub}=0V, V_d=0V, with V_s and floating, modulated the V_t state of the fast programming bits. As shown in Figure 5.20, a previously identified fast bit appears normal if the programming (i.e. the ejection of electrons from floating gate) takes place over the entire channel region instead of the floating gate-drain junction overlap region. This enhanced tunneling is observed to be

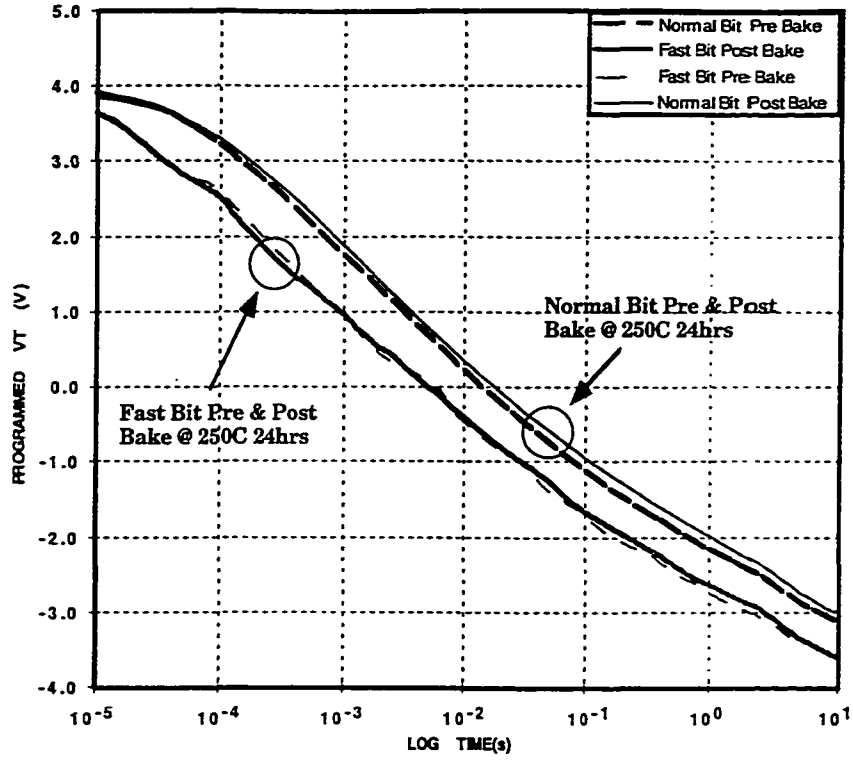


Figure 5.18: Program versus time characteristics of fast programming and normal 1T-Flash bits pre and post 250°C 24hrs bake.

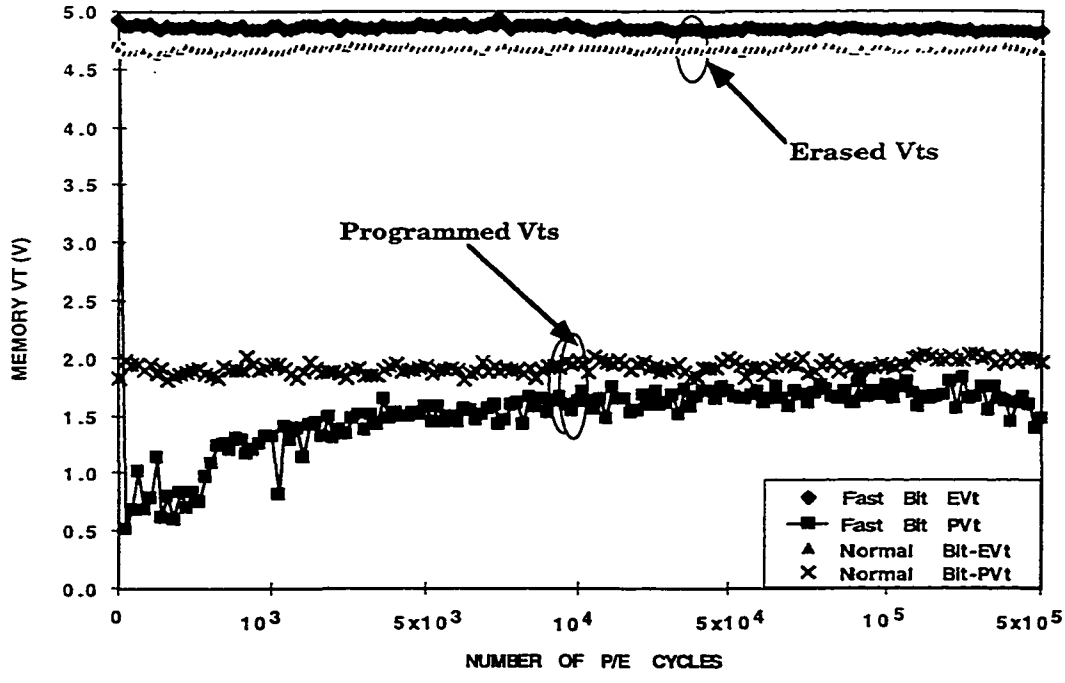


Figure 5.19: Behavior of Fast Programming bit memory threshold voltage as a function of P/E cycles.

more evident in the drain-side programming because, the drain tunneling area, of about 0.6 x 10^{-9} cm^2 is more sensitive to the floating gate polysilicon grain boundary electric field in comparison to the larger channel region.

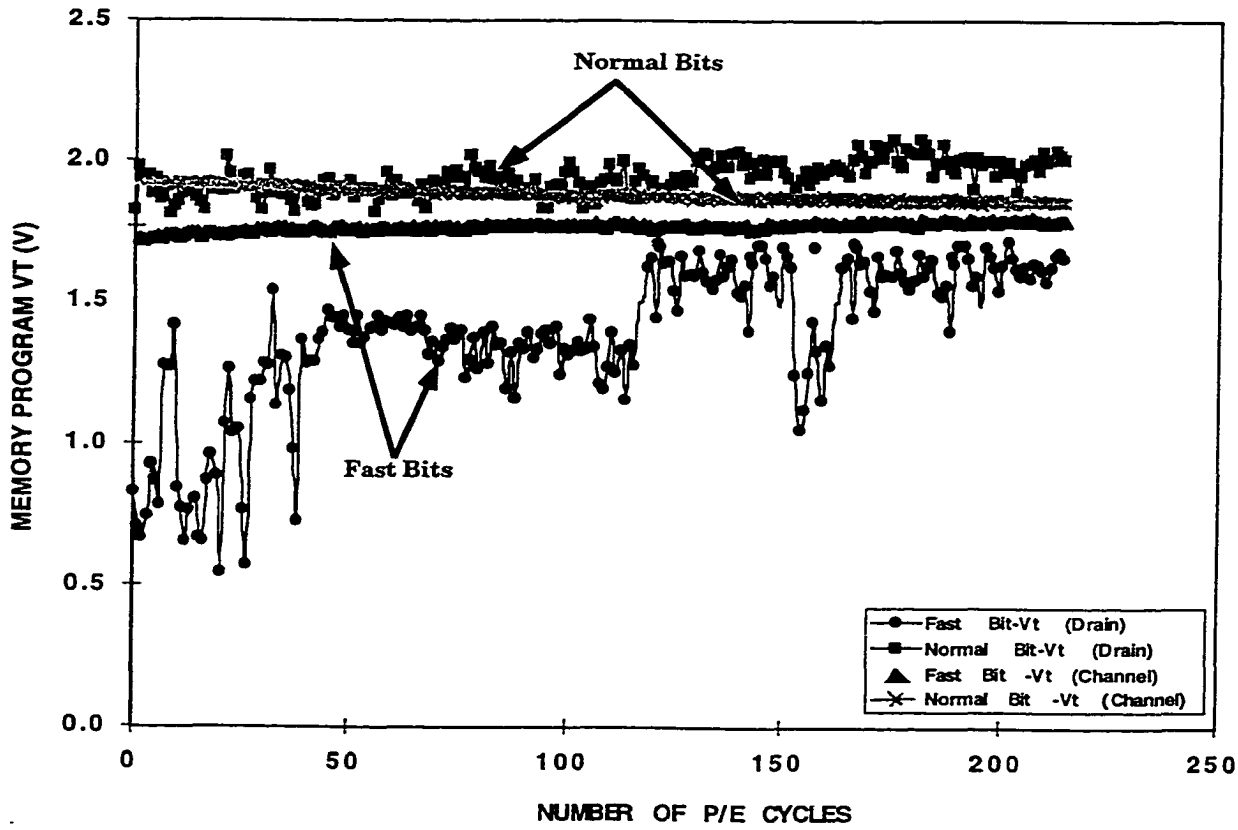


Figure 5.20: Effects of Channel and Drain programming (low Vt) on fast and normal bits.

5.3.1 Bitcell Characterization

Programming (writing) is performed by Fowler-Nordheim ejection of electrons from FG to drain overlap region, while erasure is achieved by F-N tunneling of electrons from substrate to FG via the entire channel region of the device. Fig. 5.1 illustrates how the terminal biasing is performed, with the voltage conditions of the various operation modes summarized in table 3. A detailed description of the operation and underlying device physics of the F-N-based Flash cell is given in [34 -36]. Throughout this study, the threshold voltage is defined as the control gate voltage(V_{cg}) required to allow a drain current of 1μA with a V_d=1V as shown in Fig. 5.17b. The classical means for carrying out P/E degradation studies is an endurance measurement in which the bitcell threshold voltages are monitored as a function of P/E cycling. In any floating gate memory cell, a shift in the programmed (low-V_t) or erased (high-V_t) state during cycling can only be induced by two possible mechanisms: 1) a change in the amount of charge transferred to/from the FG under constant program/erase conditions; 2) a shift in the intrinsic threshold voltage V_{ti} of the cell due to uniform and or local charge trapping within the tunnel oxide.

A technique that enables the separation of trapped and FG charges is by illumination of the cell with ultra-violet (UV) light pre and post P/E cycling. A UV erase operation is a self- limiting process and the device always ends up in the same equilibrium state with the same reference charge on the FG. Thus any change in the cell characteristics after UV erase will reveal only the damage that must have occurred within the transistor, and will not be affected by the changes in the amount of charge stored on the floating gate. A second method that is used to identify and separate the different components contributing to P/E degradation is the charge pumping technique [37], which has been previously shown to be a

powerful tool for monitoring stress induced degradation in MOS transistors. Furthermore data retention bakes of 250°C for 24hrs which is traditionally employed to explore the charge retention capabilities, were used to examine the inherent charge stability of the fast programming bits.

Operation	Source	Drain	CG	Pwell
PROGRAM	0	5	- 8	0
ERASE	-	-	16	0
READ	0	1	3.3	0

Table 3 : Typical operating voltages for a 1T-Flash EEPROM cell

5.3.2 Characteristics and UV Erasure

Figure 5.17a shows the V_t distribution of a 2MBit Flash EEPROM using F-N programming in which the fast or over-programmed bits leads the normal population by ~0.5V after an initial programming with V_{cg}=-8V and V_d=5V. The I-V characteristics of normal and fast bits in Fig. 5.17b reveal different V_t for both bits with same programming

time and biases. However in the insert of Fig. 5.17b , the post UV-erase V_t for both fast and normal bits are the same, thus indicating that the initial amount of charge on the FG at post UV-erase is the same. Hence V_t can be expressed as:

$$V_t = V_{tuv} + \Delta V_{tq}$$

where V_{tuv} is the threshold voltage after UV erase and ΔV_{tq} is the V_t shift induced by additional charge on the FG with respect to the UV-erased state. Hence a change in ΔV_{tq} in a programmed or erased state is thus indicative of a change in the amount of charge transferred to or from the floating gate.

5.3.3. Program/Erase Endurance Cycling

Electrical experiments shown in Fig. 5.19 reveals that the programmed V_t of the fast bit asymptotically increases to a higher steady state value where it remains and exhibit characteristics comparable to the normal bit. This could be explained by the fact that the tunneling electrons cause impact ionization in the oxide thus creating positive charges at the FG –SiO_2 interface and subsequently enhancing the F-N tunneling. However after a long period of P/E cycling these positive trap sites are filled with electrons hence reducing the F-N tunneling process which leads to an asymptotic increase of Vt, which is confirmation of a similar behavior as reported in [10]. We consistently observed that the fast bits are always fast and there is no evidence of the programmed V_t toggling between fast and normal V_t states, however the V_t asymptotically increases as the positive trap sites are filled as a result of increased net negative charge in the tunnel oxide.

5.3.4. Data Retention Analysis

Programming characteristics have been measured before cycling and after a 24hr bake at 250°C bake for comparing fast and normal bits. As evident in Fig. 5.18, the bake had no observable effect on fast bit characteristics. To further enhance the understanding of the fast bit behavior, 2000 fast and normal bits were cycled and baked for 24hrs at 250°C to characterize the effects of cycling on fast programming bits. The data shift of ~0.2V after bake for a fast programming bit was comparable to the shift of the normal bit, hence once a fast bit is programmed to a desired V_t it has the propensity to retain its charge. The resultant reduction in F-N tunneling currents is manifested as an increase in programmed threshold voltage.

5.3.5. Effects of Floating Gate Poly Morphology

Experiments were conducted to investigate the possible causes of the fast programming bits. "Channel" programming or severe gate disturb with V_{cg} = -16V, V_d, V_s floating and V_{sub}=0V, was used to modulate the V_t state of the fast bits. As shown in Fig. 5.20, a previously identified fast bit appears normal if the programming takes place over the entire channel region instead of the FG-drain junction overlap region. Further measurements of low B-B tunneling leakage at the flash drain junction indicates that the enhanced F-N tunneling is influenced at the FG-tunnel oxide interface. This drain vs. channel programming behavior indicates that a locally enhanced tunneling field as a result of the FG poly grain structure, plays a crucial role in the fast bit generation. To further this understanding, MOS capacitors were fabricated using the flash floating gate as the cathode on tunnel oxide. The 1500Å FG poly was deposited and doped in three ways: 1) sandwich of

undoped plus phosphorus in-situ doped, 2) undoped followed by an implant and 3) in-situ doped. Fig. 5.21 shows data for the three capacitors of the F-N tunneling currents J vs. E-field across the tunnel oxide. From AFM measurements, the undoped plus in-situ doped stack for the FG poly exhibited the smoothest morphology with the smallest RMS value of 1.06 which is a measure of the FG poly surface roughness. This was also confirmed with TEM grain size characterization as shown in table 2. The average grain size is largest for process – A (implanted) poly and smallest for process-C (undoped plus in-situ doped) stack, which correlates well with the AFM surface roughness characterization and the corresponding TEM micrograpghs in Fig. 5.21. The J vs. E-field characteristics in Fig. 5.7a shows that the undoped plus in-situ sandwich has reduced tunneling currents for a given E-field compared to the other stacks. The programming distributions of 2Mbit arrays as a function of FG poly deposition shown in Fig. 5.23 were also evaluated. It is observed that the spread of the V_t distribution is reduced for the FG stack that showed a reduced F-N tunneling current in Fig. 5.7a, a further confirmation of the smaller polysilicon grains at the FG – tunnel oxide interface. A physical explanation for the observed results centers on the FG polysilicon morphology. As the FG polysilicon grain size decreases more dopants segregate to the grain-boundaries. Since smaller grain FG poly has more grain boundaries the effective electrons available for conduction is reduced and thus resulting in reduced F-N tunneling currents as evident in Fig. 5.7a.

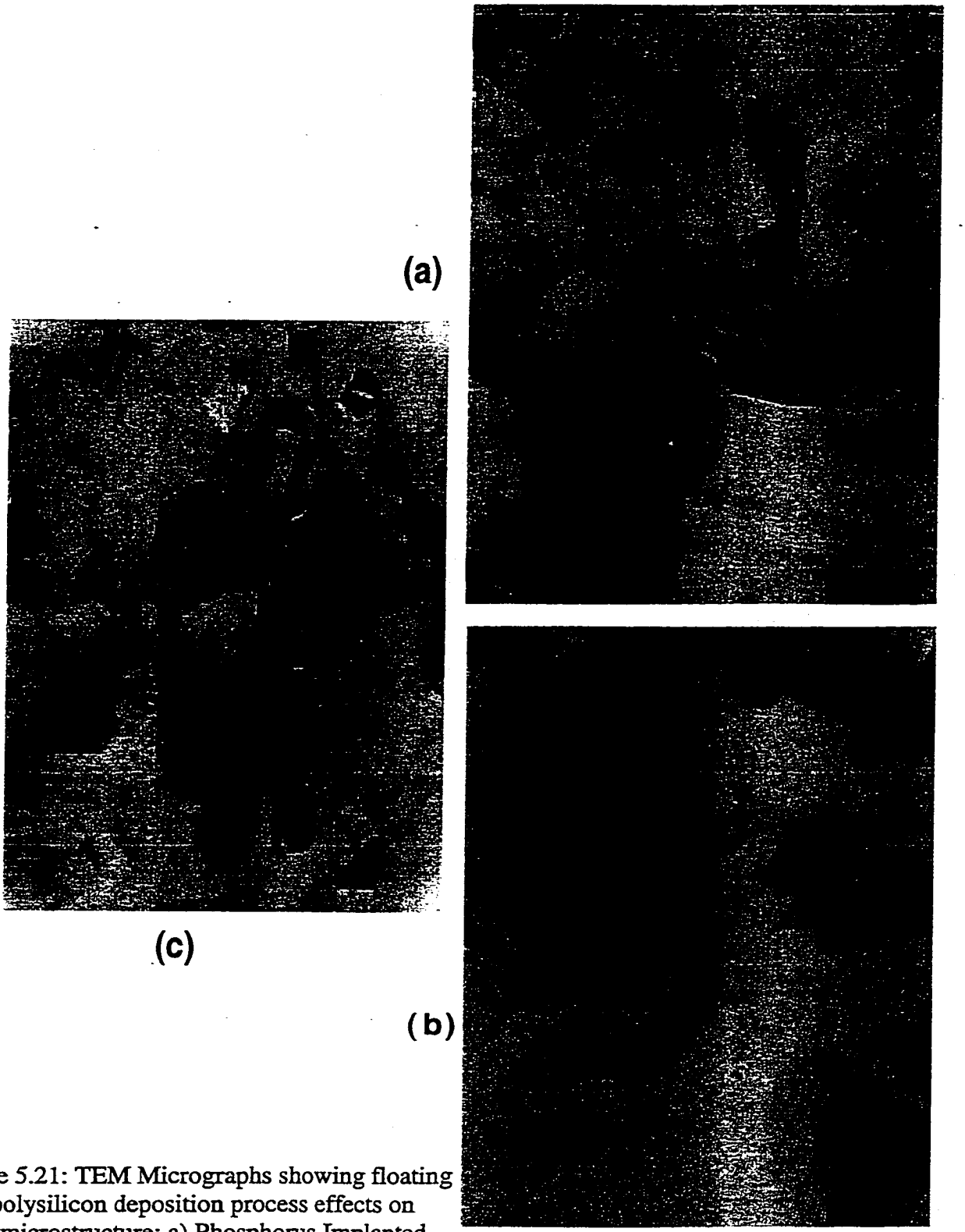


Figure 5.21: TEM Micrographs showing floating gate polysilicon deposition process effects on grain microstructure: a) Phosphorus Implanted
b) Phophorus In-Situ doped
c) Undoped+ Phosphorus Insitu stack

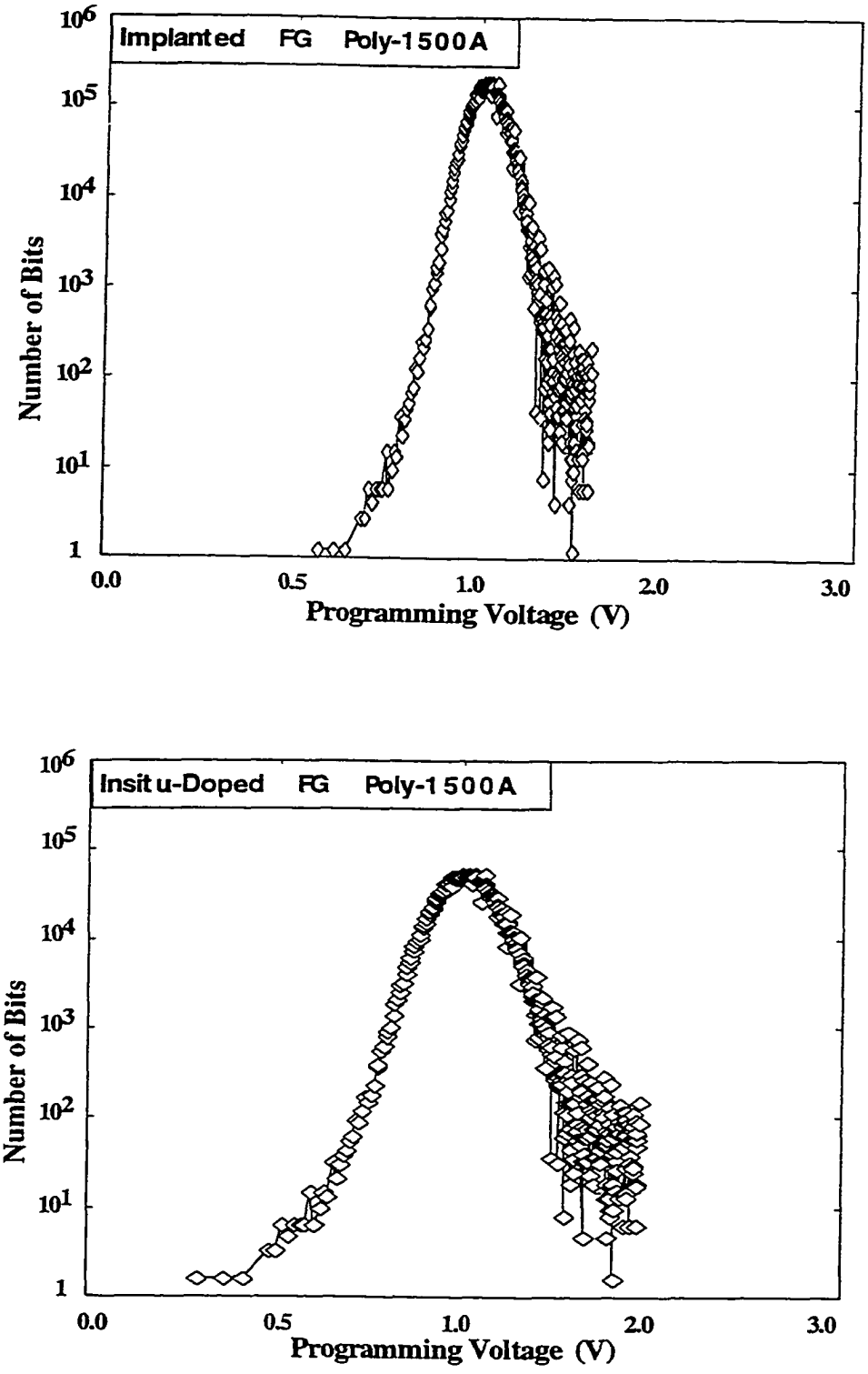


Figure 5.22: Effects of Floating gate polysilicon on the programmed memory threshold voltage distribution for a 2-Mbit 1T-Flash array.

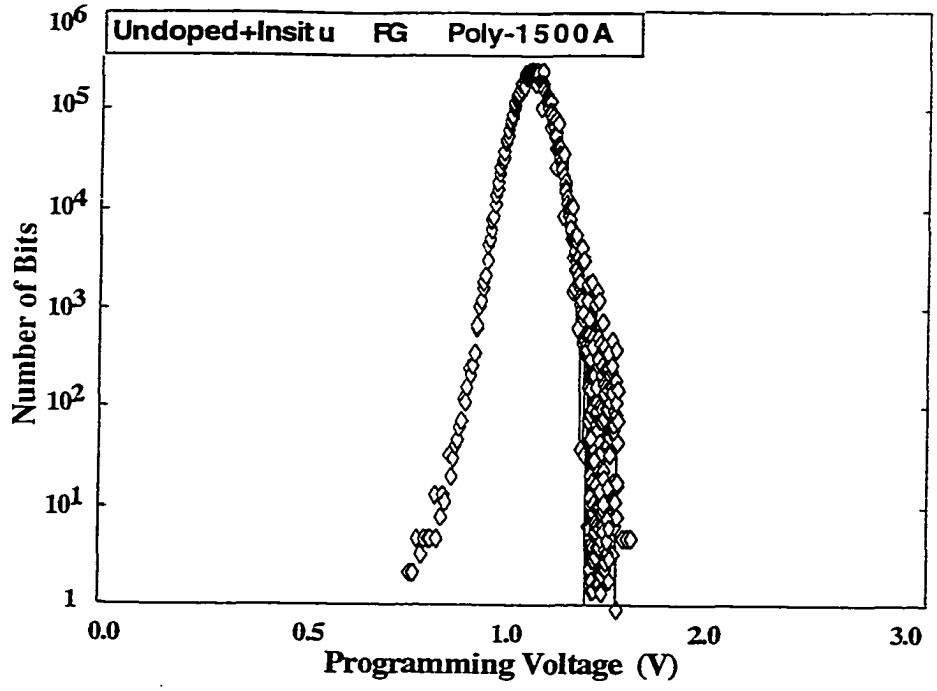


Figure 5.22: Effects of Floating gate polysilicon on the programmed memory threshold voltage distribution for a 2-Mbit 1T-Flash array.

5.4 1T-Flash Fast Programming Model

In this research the experimental data showed convincingly that the Fast programming bit behavior could be modulated with the Floating gate polysilicon microstructure. As a results of the findings, the modeling of the fast bit programming characteristics was focus on the incorporation of polysilicon physics and the associated physiological effects into the 1T-Flash EEPROM memory programming characteristics, the results of which will be compared to the experimental results. Besides the use of the capacitor model and I-V characteristics of the floating gate device for extraction and quantification of the coupling coefficients, they can also aid in the modeling of the memory bitcell behavior. The memory model allows the use of the programming or erases transient

characteristics. A comparison of measures and calculated characteristics reveals the validity of the assumptions made concerning the physical mechanisms governing the programming and erase behavior of both the fast-programming bits and the normal bits. In this section the basic equations developed in chapter 2 will be discussed. The model comparison to measured data will also be discussed and explanations offered for the experimental observations.

When a charge is injected or emitted from the floating gate, the floating gate potential will change as given by equation 2.29:

$$V_{fg} = V_{cg} - \frac{W_{chan}\,\varepsilon_{ox}}{C_{pp}} + \int_0^{L_g} E_{ox}\,dx + \frac{Q_{fg}}{C_{pp}} \text{ ---------- (2.29)}$$

Due to the change of the floating gate potential, the electric fields in the surrounding oxides change as well, by which the injected Fowler-Nordheim current increases or decreases. This process will continue until a steady state is reached. Thus it is important to remember that the injection currents during a programming operation is not constant, but changes rather rapidly as a function of time. The modeling of the memory behavior of floating gate cells is based on the following elements:

1. The basic memory starts from the expression for the charge on the floating gate:

$$\frac{dQ_{FG}}{dt} = \int_{AFG} J_{FN}(t)\,dA \text{ ---------- (5.3)}$$

This expression can further be arranged where A_i is the injection area to the

$$Q_{FG}(t) = \int_0^t \sum_{A_i} J_i\left[E_i(t)\right]dt \quad \text{---------} \quad (5.4)$$

floating gate. The currents J_i, through the different dielectrics are a function of time because the electric fields, E_i in the dielectrics are changing with time. By integrating equation 5.3 with respect to time, an expression for the charge that is accumulated on the floating gate is obtained:

2. In order to truly understand the effect of floating gate morphology, a relationship had to be established between the measured grain density and the average grain size. Since the experimental results concluded that process-C (Floating gate sandwich of undoped + in-situ doped) polysilicon film exhibited no fast-programming bits, the compact model was focused on results of process-C, and how it compares with that of process-A which had the most fast bits.

The above equation 5.4 cannot be solved as such because the electric fields, E_i, are dependent on the floating gate potential, and thus also on the floating gate charge $Q_{fg}(t)$. Therefore the model needs to allow for the calculation of the fields occurring inside the device during programming as a function of time. The following relationships between the floating gate polysilicon grain size and programming were developed in chapter 2:

$$J_{FN} = CE_{inj}^2 e^{-\frac{E_C}{E_{inj}}} = \alpha_{fn} G_{eff} E_{inj}^2 e^{-\frac{E_C}{E_{inj}}} \quad \text{---------} \quad (2.51)$$

These equations are the fundamental equations that govern the relationships between the floating gate polysilicon morphology and the enhancement of the F-N tunneling injecting

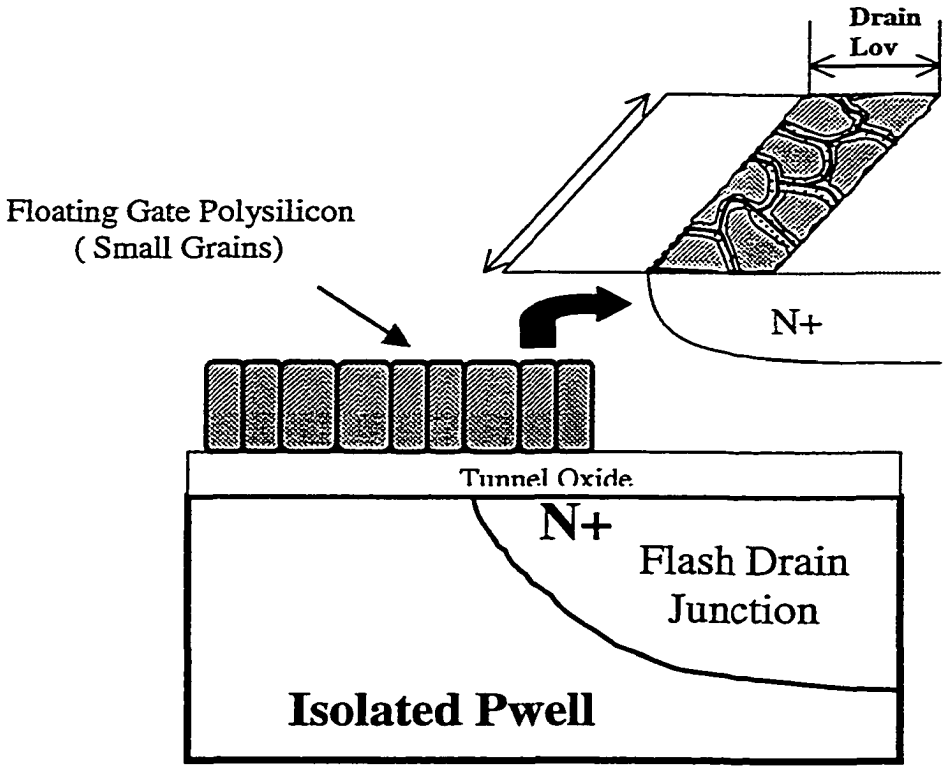
$$E_{inj} = \frac{V_{fg}}{X_{tun}}\left(1 + \frac{X_{tun}}{Rc}\right) \quad \text{----------} \quad (2.52)$$

field.

5.4.1 Physical Model

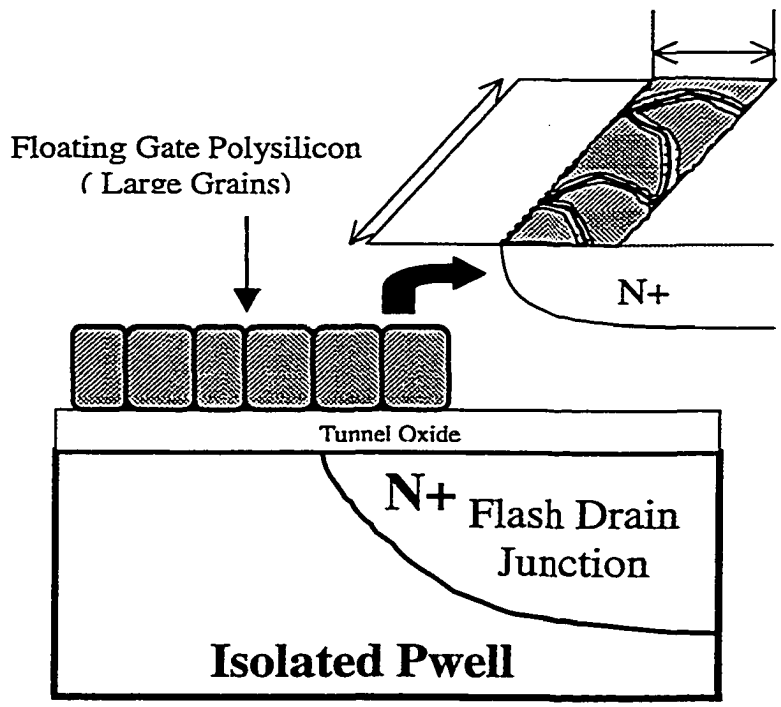
To enhance the understanding of the floating gate polysilicon effects in 1T-Flash devices, a physical model has been developed based on experimental observations. This model comprises of a grain area effect (Geff) and tunneling field enhancement ($\mu(Rc)$), that is a result of sidewall asperities a result of large grain.

Figure 5.23a &b shows schematically the area model of how floating gate polysilicon grain size influences the effective grain tunneling area. It is evident from these figures that process-A has fewer grains than process-C in the same drain overlap tunneling region. To further quantify these effects, programming threshold voltage measurements were made on a wide 1T-Flash transistor, comparing the variance of program voltage as a function of the floating gate grain size as shown in figure 5.24a. Again it is evident that as the grain size decreases, the Vt variance also reduces, a further indication of the lower probability of fast bit generation in process-C. Also in figure 5.24b the effects of the programming tunneling area is quantified. This is in effect a simulation of how a large grained floating gate polysilicon would show an increased incidence of fast bits if the tunneling area is increased. This is also confirmed with increased grain injection area (Ainj) as shown in Figure 5.26.



Process - C (Undoped(500Å) + Insitu doped(1000Å)

Figure 5.23a: Physical Model of Floating Gate Grain Size Effects in Process-C (Normal Bit)



Process - A Phosphorus Implanted - (1500Å)

Figure 5.23b: Physical Model of Floating Gate Grain Size Effects in Process-A (Fast Bit)

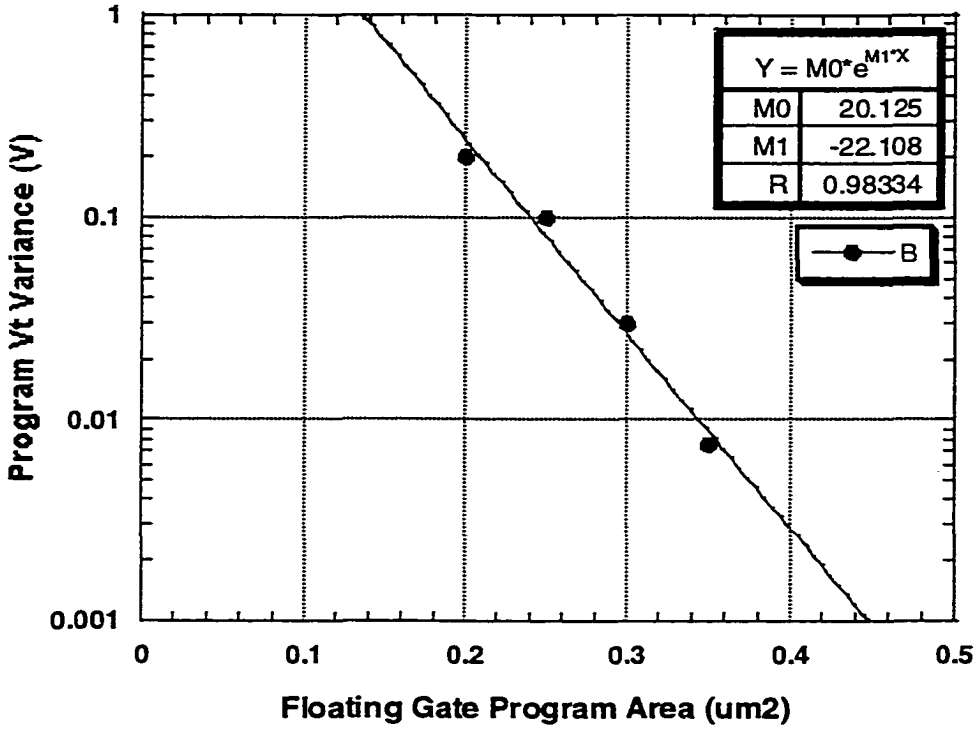


Figure 5.24a: Effect of Floating Gate Grain Size on Program Threshold Voltage

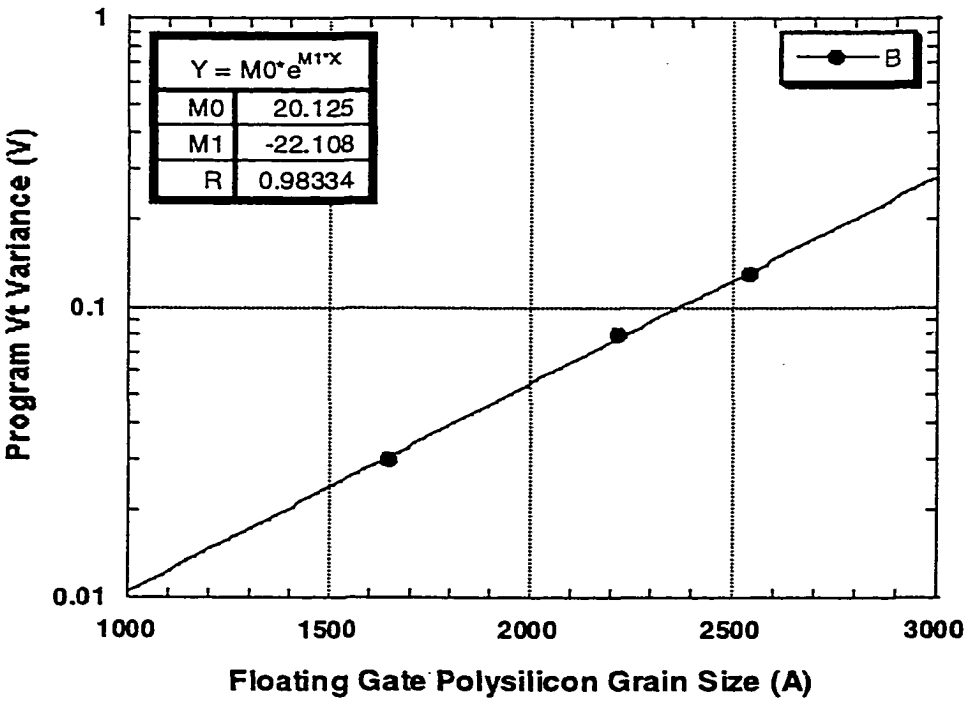


Figure 5.24b: Effect of Floating Gate Tunneling Area on Threshold Voltage

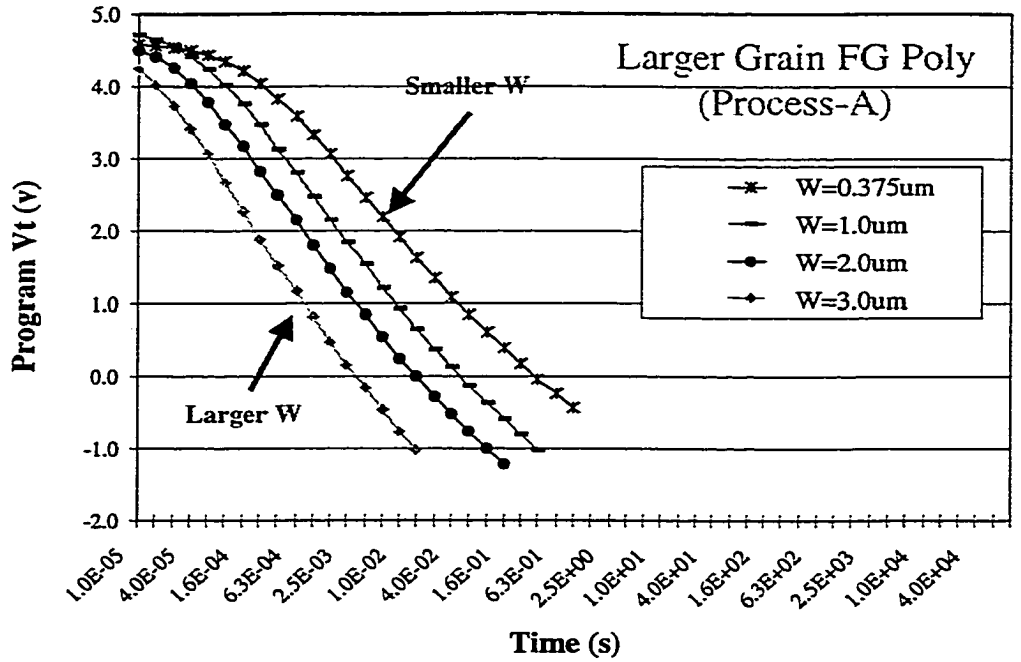


Figure 5.25: Effects of 1T-Transistor Width on Programming Threshold Voltage

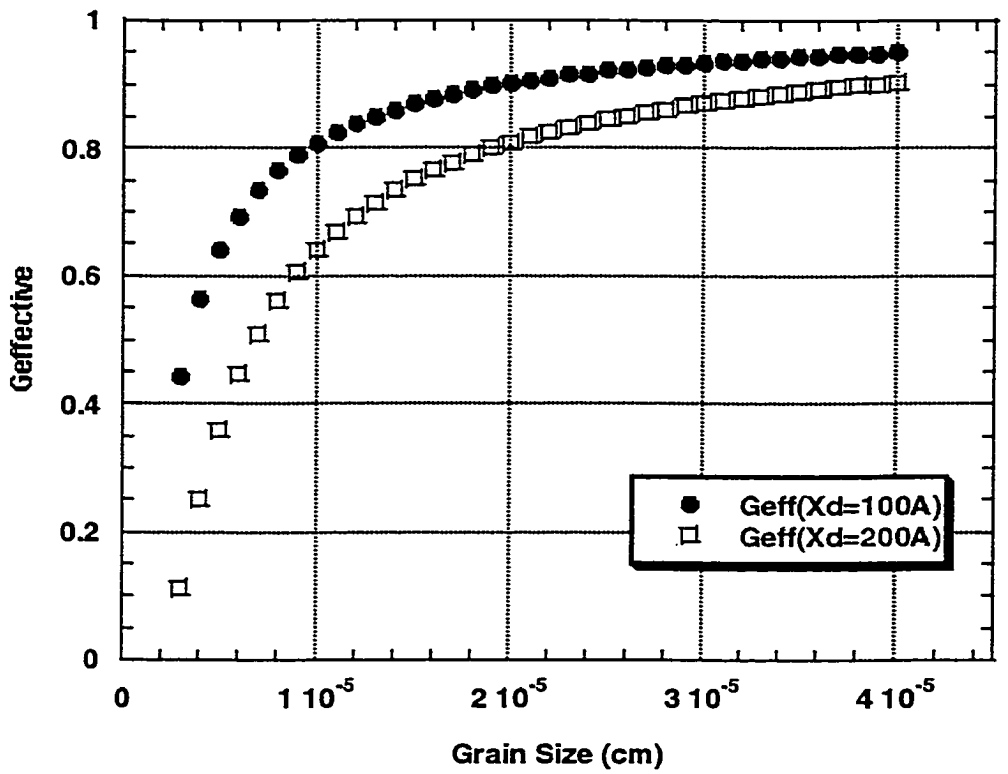
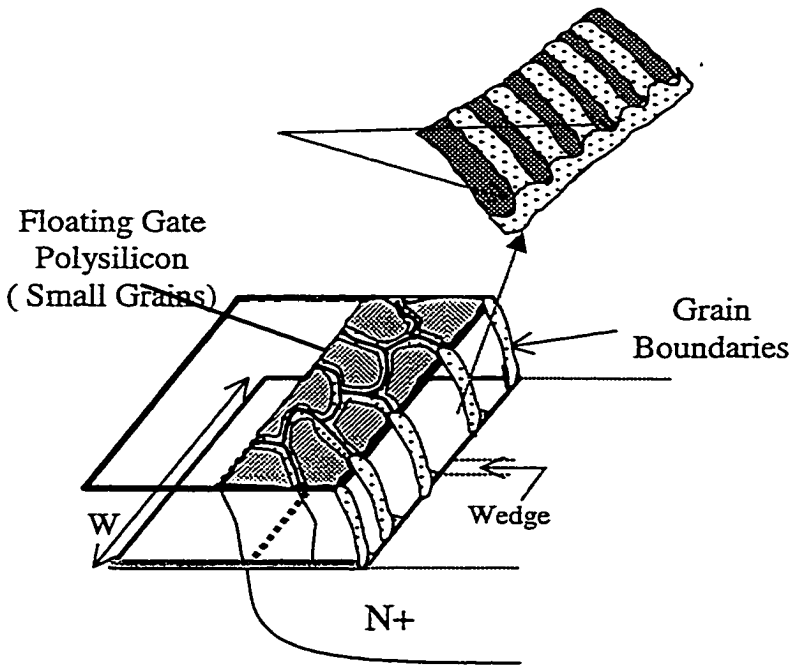


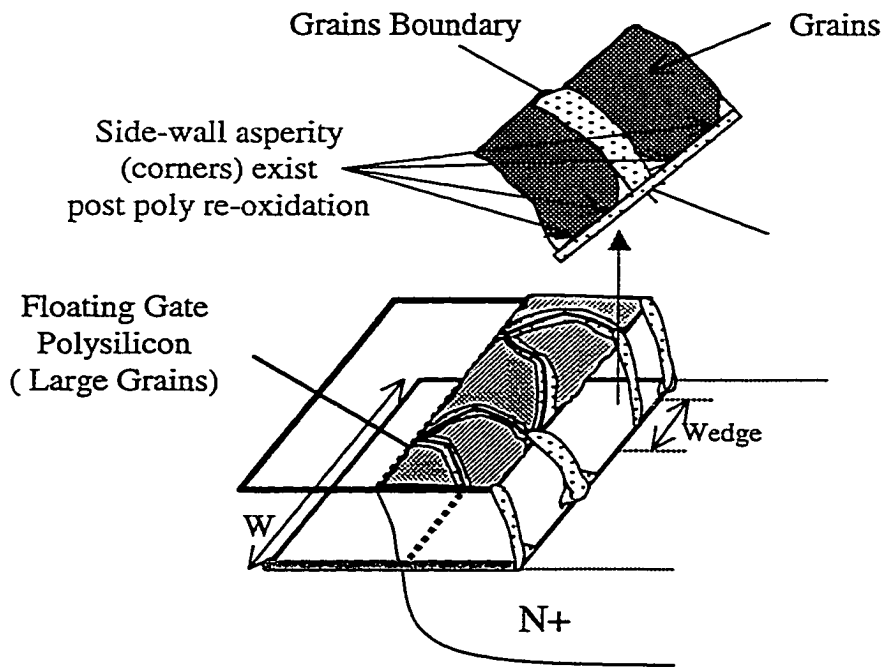
Figure 5.26: Grain Injection Area Factor (Geff) as a function of Grain Size

The area contribution model is complemented by the further indication that the edge effects of larger grain polysilicon are also playing a role. This we believe can be attributed to sidewall asperities which results in enhanced tunneling fields and hence a reduced program threshold voltage. The asperity contribution model is further confirmed by measurements of program versus time characteristics of 1T-Flash EEPROM transistors as shown in figure 5.25. The physical model of the asperity field enhancement is shown in Figure 5.27 and 5.28. Asperities have been reported in previous publication [66], as resulting from sharp edges in polysilicon. In this work it is believed that the floating gate etch can leave sharp polysilicon edges along the width direction of device. After polysilicon re-oxidation which results in enhanced sidewall grain-boundary oxidation, the sharp edges of the smaller grain floating gate material are oxidized away, whiles the large grain still has asperities a result of the greater "Wedge" for the large grain remaining after oxidation. As a result the critical radius (distance between charge & tunneling point) Rc is much larger for small grains than that for the larger grain FG polysilicon because of the larger "w-edge" of the large grain. This thus increases the probability that larger grain material will have the asperity since fewer GB exist along the edge which means more "Wedge" with asperity forming potential.



Process - C (Undoped(500Å) + Insitu doped(1000Å)

Figure 5.27: Physical Model of FG Polysilicon Asperity in Process-C (Normal Bit)



Process - A Phosphorus Implanted - (1500Å)

Figure 5.28: Physical Model of FG Polysilicon Asperity in Process-A (Fast Bit)

5.4.2 Grain Size Distribution Analysis

The fundamental assumption made was that the floating gate polysilicon grain size distribution could be approximated using a gaussian distribution function:

$$y(pdf) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(G_s - \overline{G_s})^2 \Big/ 2\sigma^2 \right]$$

From the data plotted in figure 2.11 and 2.12, there exists a relationship between the fast bits generated and the density of grains for process-A,B and C. Below is a summary of the relationship observed. Note that for a given process to produce fast bits as a results floating gate grain size, there has to be a critical grain size (Gscritical), beyond which the probability of fast bit occurrence increases. However the (Gscritical) must be independent of the polysilicon deposition process. From fast bit vs. program vt plots one can estimate the probability of producing fast bits for each process By integrating for Gcritical of each process which must be the same. Below is a summary of findings for each of the processes:

Process-A

1. Has ~700 Fast Bits at Vtp < 1V (FB_A)
2. y_A=700/4096 bits ~0.175 is probability of fast bits in process-A
3. $N_{gA}(G_{scrit}) = 2.6 \times 10^9$ grain/cm^2

Process-B

1. Has ~100 Fast Bits at Vtp < 1V (FB_B)
2. y_A=100/4096 bits ~0.02 is probability of fast bits in process-B
3. $N_{gB}(G_{scrit}) = 1.7 \times 10^9$ grain/cm^2

Solving all three probability equations for Gscritical yields:

$$y_A = 0.175 = \frac{1}{\sigma\sqrt{2\pi}} \int_{Gcrit}^{\infty} \exp\left[-(G_s - \overline{G_{sa}})^2 \Big/ 2\sigma_a^2 \right]$$

$$y_B = 0.02 = \frac{1}{\sigma\sqrt{2\pi}} \int_{Gcrit}^{\infty} \exp\left[-(G_s - \overline{G_{sb}})^2 \Big/ 2\sigma_b^2 \right]$$

$$y_C = 0 = \frac{1}{\sigma\sqrt{2\pi}} \int_{Gcrit}^{\infty} \exp\left[-(G_s - \overline{G_{sc}})^2 \Big/ 2\sigma_c^2 \right]$$

The solution for $G_{scritical}$ is an error-function of form:

$$G_{crit} = \frac{y\sqrt{2}}{\left\{ \mathrm{Erf}\left[\frac{(G_{su} - \overline{G_{sa}}}{2\sigma} \right] - \mathrm{Erf}\left[\frac{(G_{sl} - \overline{G_{sa}}}{\sigma} \right] \right\}}$$

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n X^{(2n+1)} / (n!(2n+1))$$

After numerical analysis the critical grain size of Gscritical = 2683.2 Å can be established. As evident from figure 5.28 below from experimental data, process-A has more $G_{scritical}$ grains that process-B and C, further evidence of why process-A produces more fast programming bits.

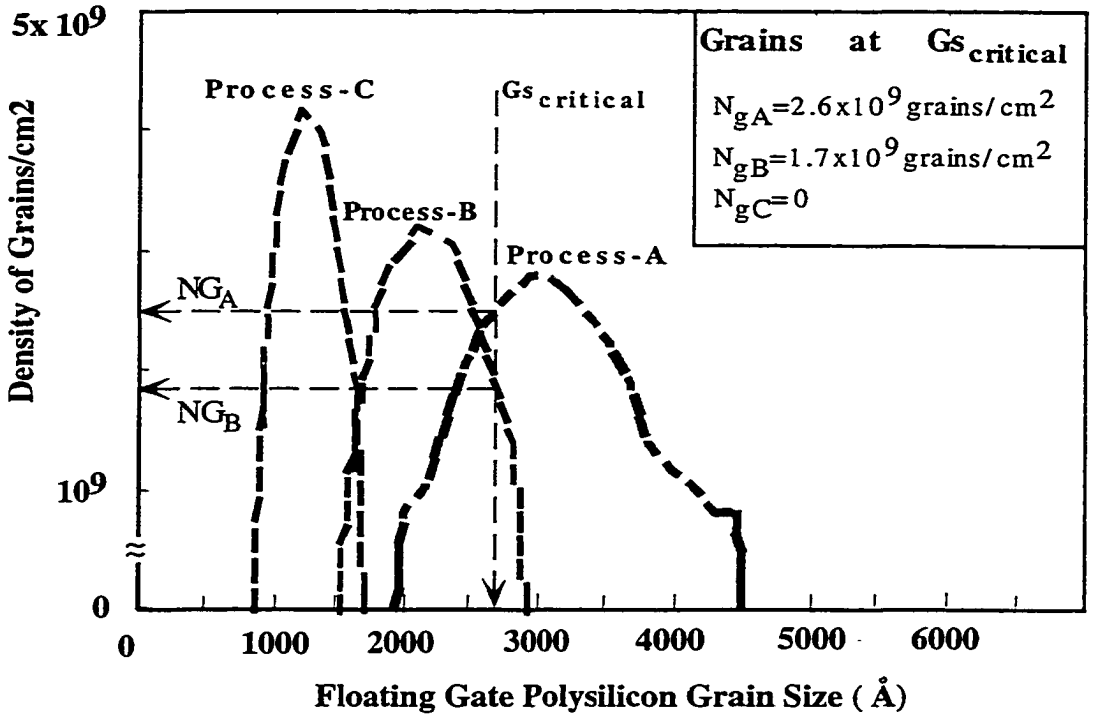


Figure 5.28: Floating gate Polysilicon Grain Size Distribution

5.4.3 Grain Depletion Width Estimation

In an attempt to develop the relationship between the floating gate grain size effects on the tunneling fields, the effective area of the polysilicon grain available for effective injection must be quantified. Since the negative charges from the grain are trapped by grain boundary, the depletion width in the grain must be known. The following equations were developed in the theory section of chapter 2 and are repeated here for discussion.

From Polysilicon Physics:

$$Q_{Trap} = N_t * T_{grain} * L_{grain} = 2X_d * N_D * T_{grain} * L_{grain} \quad \text{--(2.46)}$$

$$\Rightarrow X_D(\textit{depletion} \text{ width in grain}) = \frac{N_t}{2 * N_D} \quad \text{----------(2.47)}$$

assuming $L_{grain} = W_{grain} = G_S$ then

$$A_{inj} / (cm^2) = (G_S - 2X_D)^2 = (Lg - 2X_D)^2 \quad \text{---------- (2.48)}$$

$$A_{inj}(cm^2) = ((G_S - 2X_D)^2 * N_g) = \frac{(G_S - 2X_D)^2}{G_s^2} * W * L_{overlap} \quad \text{---- (2.49)}$$

$$G_{eff} = \frac{A_{inj}}{A_{overlap}} = \frac{A_{inj}}{W * L_{overlap}} = \frac{(G_S - 2X_D)^2}{G_s^2} = \left(1 - \frac{2X_D}{G_s}\right)^2 \quad \text{------ (2.50)}$$

As shown in Figure 5.26, as the grain size increases, the effective floating gate polysilicon area available for F-N tunneling current injection increases exponentially. Hence the Geff term which is proportional to t0 also exhibit similar exponential dependence as shown in Figure 5.29.

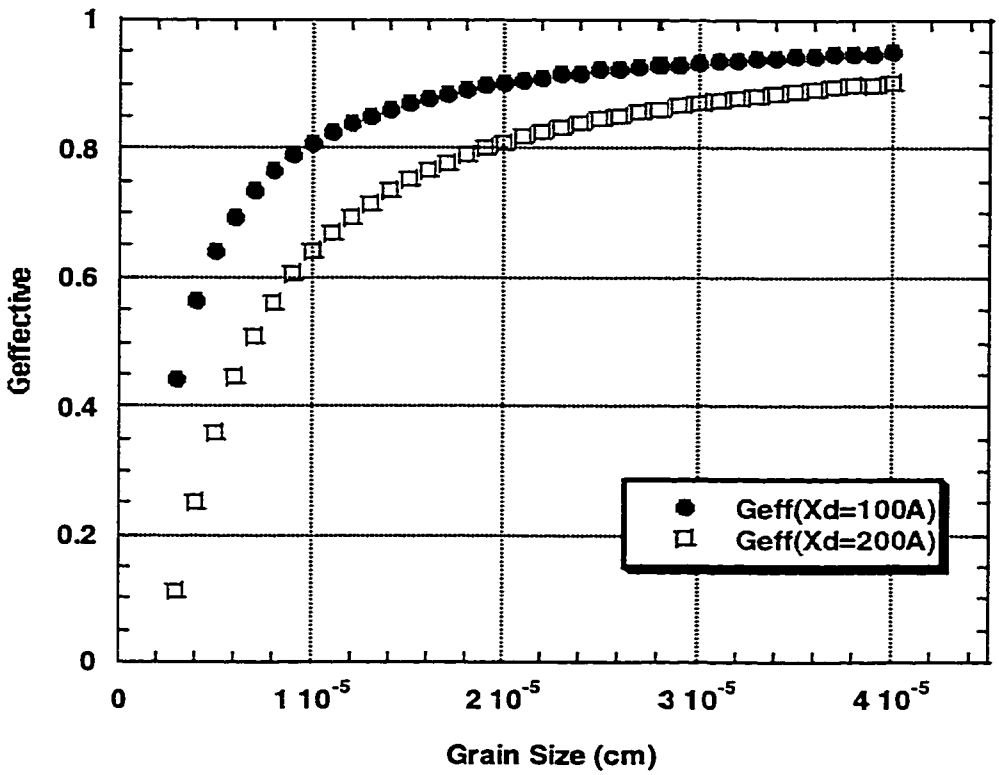


Figure 5.29: Dependence of Geffective on Grain Size

5.4.4 Modeling of Grain Area and Asperity Effects

As has been previously established, the effective floating gate polysilicon grain area (Geff) and field enhancement factor (μ) can be used to explain the fast and normal bit programming voltage characteristics for a 1T-Flash EEPROM device. The formulations of Fowler-Nordheim tunneling injecting field with the incorporation of Geff and enhancement factor μ(Rc) resulted in the threshold voltage equations developed in chapter 2. The final equation 2.67 developed in chapter 2 was solved with all sub-dependent equations and constants. This enabled the generation of the model plots versus experimental data of programming versus time for Fast-programming bits and Normal bits . The established

model was used to demonstrate the effects of grain size on the programming characteristics. As a result the fast bit behavior of process-A can be predicted with a reasonable level of accuracy. As can be seen in Figures discussed, the model developed in chapter 2 (equation 2.67):

$$V_T(t) = \left[\frac{\left(E1 + \frac{\alpha cg Vt0}{X_{tun}} \right)}{1 + \frac{\mu(Rc)\left(E1 + \frac{\alpha cg Vt0}{X_{tun}} \right)}{E_C} \ln\left(1 + \frac{t}{t0} \right)} - E1 \right] * \frac{X_{tun}}{\alpha cg} \quad \text{----------(2.67)}$$

can indeed predict the fast programming bit behavior as a function of the floating gate polysilicon grain size effect on tunneling area (Geff). The programming characteristics are fairly predicted by the model however potential errors which is not providing accurate fit is the electron trapping and how that process is distributed in the tunneling region. Appendix B shows the key parametric inputs and outputs from the model for the 1T-Flash memory device.

5.4.5 Model Discussion

The validity of the compact model and how it compares with experimental results will be discussed in this sub-section. As evident from Figure 5.30, the effect of Geff and its influence on t0 is clear. Since t0 has a direct impact in the effective injecting field Einj(t), the effects of Geff on program threshold voltage in Figure 5.32 and 5.34 can explained. From chapter 2 expressions for t0, as Geff increases t0 decreases hence equation 2.67

predicts a lower threshold voltage. Figure 5.31 reveals an interesting dependence of program Vt on the field enhancement factor (μ) for normal bit (Geff=0.725) and fast bit (Geff=0.880). The observed delta Vt is clearly only due to the grain size factor (Geff). This strong dependence of the field enhancement factor (μ) on program Vt helps to explain the characteristics in Figure 5.33. Similarly from equation 2.67 as the enhancement factor μ(Rc) increases the threshold voltage drops rapidly. This is because the field enhancement is directly proportional to the injecting field hence the effect is more pronounced. Figure 5.34 is a combined simulation for the fast and normal bits taking into account the Geff and the field enhancement factors simultaneously, which is a confirmation that the solutions are self-consistent. The simulations in Figure 5.35 and 5.36 were performed to further the understanding of the contribution of both the grain size factor (Geff) and the asperity related field enhancement factor (μ). From Figure 5.35, its evident that the lowest program Vt that can be achieved occurs when μ=1 and Geff=1 and the Vt=1.58V. This leads to the conclusion that the normal bit can be explained by the grain size factor only and no field enhancement is required. In an attempt to enhance the understanding, the threshold voltage distribution for a normal and fast programming bits in Figure 5.38 and 5.39 has to be analyzed. The delta Vt (i.e. Vtmax – Vtmin) can be reasonably predicted from the model with Geff of 0.767 (Vtmax) and 0.847 (Vtmin). This Geff range corresponds to grain sizes of 1610Ă to 2100Ă. These values are well within the grain size distribution of process-C (normal bits).

Similarly the Vt distribution of process-A in Figure 5.39 can be analyzed. For the fast bit case the distribution is broken into zone-A and zone-B as depicted in the figure. The threshold voltage in zone-B (Vtmax – Vtpeak) can be modeled effectively with the Geff

contribution with Geff of 0.792 (Vtmax) and 0.990 (Vtpeak). This corresponds to grain sizes of 1817Ă and 3980 Ă which is also within the measured grain size distribution for process-A (1800 Ă - 4564 Ă). Considering the threshold voltage range in zone-A (Vtpeak – Vtmin) it is clear that the Geff alone will not be sufficient to account for a program Vtmin of ~0.434V. However by drawing upon the field enhancement factor (μ) the Vt in zone-A can be predicted by model as evident in Figure 5.36. A field enhancement factor of μ(Rc)=2 in conjunction with the Geff =0.880 can accurately predict a fast bit. The delta Vt in zone-A can be predicted with μ(Rc) ranging from 1.5 to 2. Figure 5.39 also reveals that the field enhancement factor alone cannot explain the entire distribution. This is because even with the lowest μ(Rc) =1 and Geff=0.880, Vtmax is ~1.8V. This indicates that the μ(Rc)=1 and a Geff =0.792 combination is required to predict the Vtmax of ~2.25V for the fast bit distribution. Table 5.3 and 5.4 depicts a comprehensive summary of the threshold voltages at Vtmin,Vtmax and Vtpeak for the various Geff and μ(Rc) discussed above.

In summary, the knowledge of fast programming bit origin has been enhanced. The fast bits arise from larger grain floating gate polysilicon, which has increased effective grain tunneling area. This in conjunction with the grain sidewall poly asperities generated after gate etch and post re-oxidation, lead to enhanced Fowler-Nordheim tunneling and thus produce bits with ultra-low program threshold voltages. This work therefore highlights the importance of polysilicon grain size and microstructure engineering during the development of non-volatile Flash EEPROM memories.

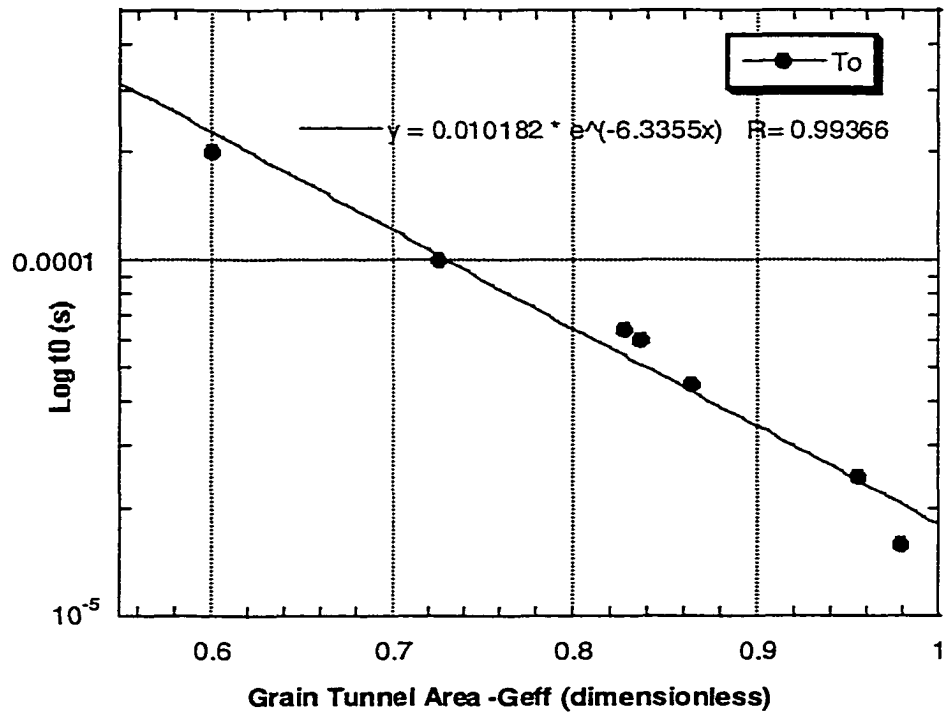


Figure 5.30: Effects of Grain Size Factor (Geff) on Turn-on Characteristic Time (t0)

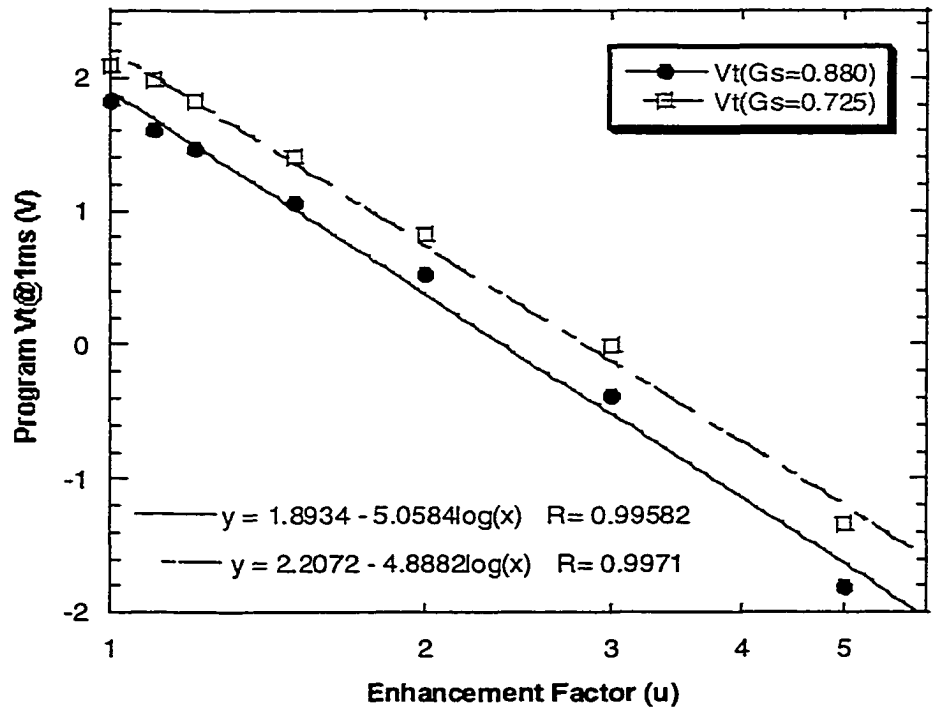


Figure 5.31: Analysis of Field Enhancement factor (u) for Normal and Fast Bits

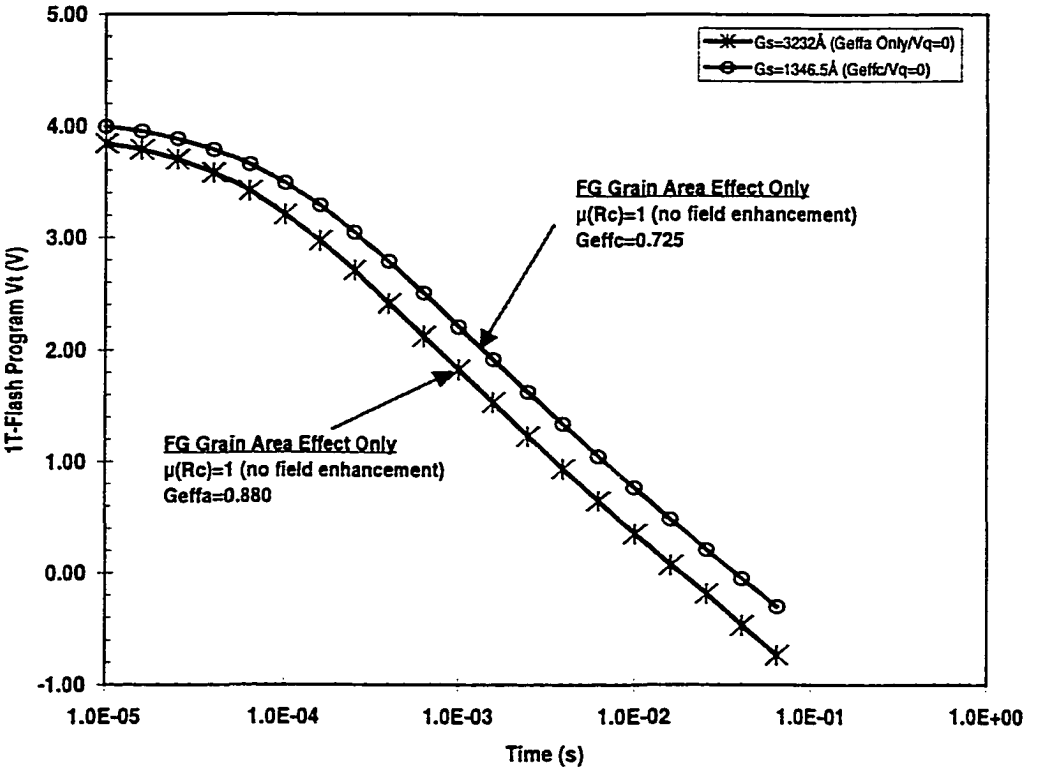


Figure 5.32: Modeling of Program Vt versus Time with Area Factor (Geff) Only

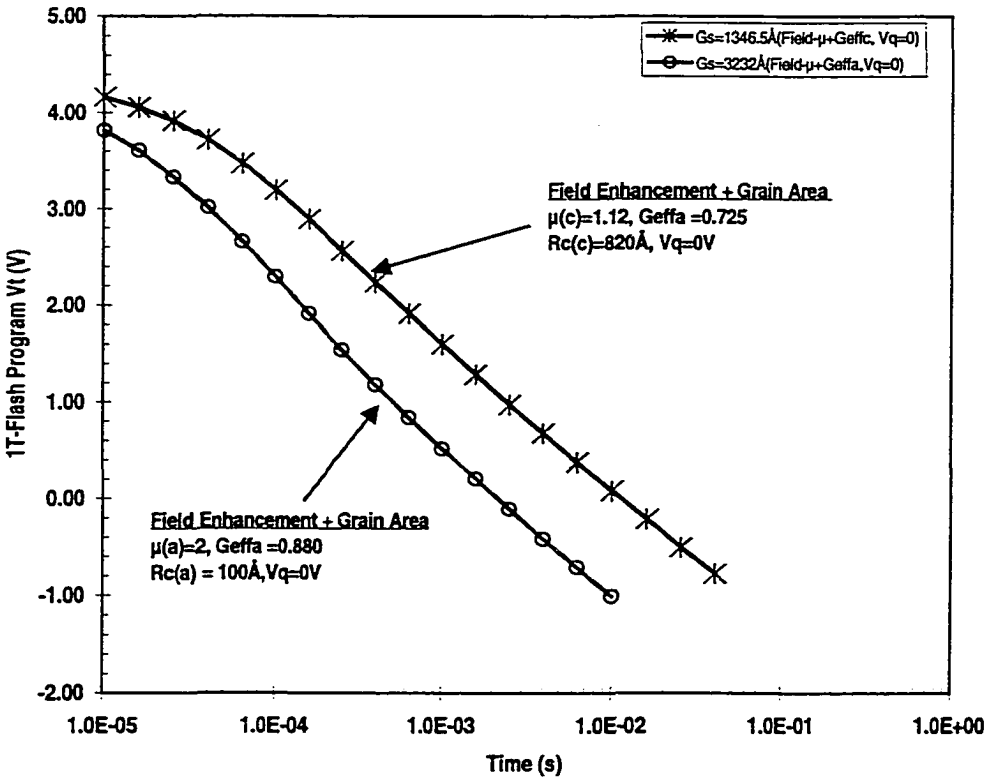


Figure 5.33: Model Program Vt vs. Time with Field (μ) and Grain Area (Geff) Factors

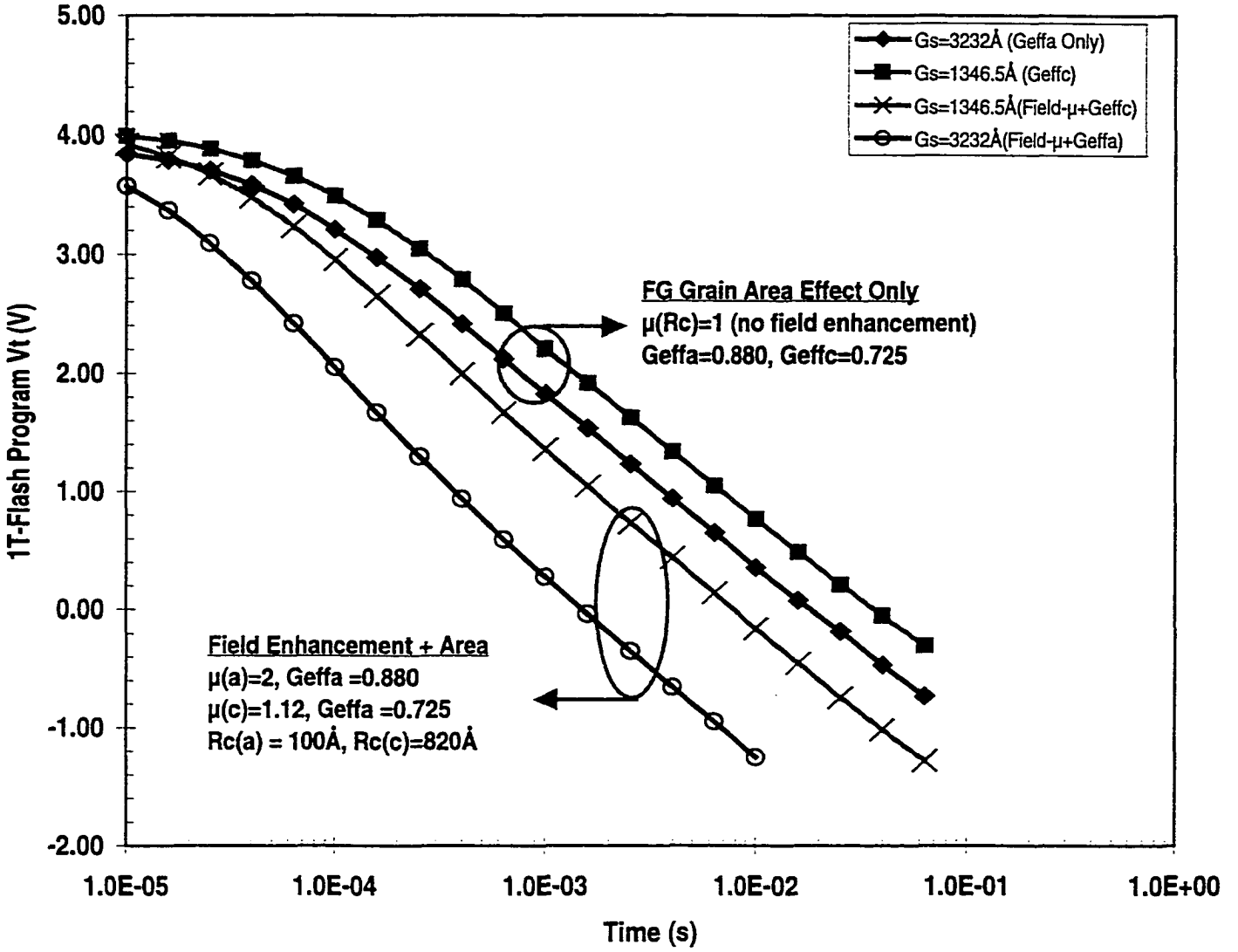


Figure 5.34: Modeling of Program Vt with Area (Geff) and Field (u) Factors Simultaneously

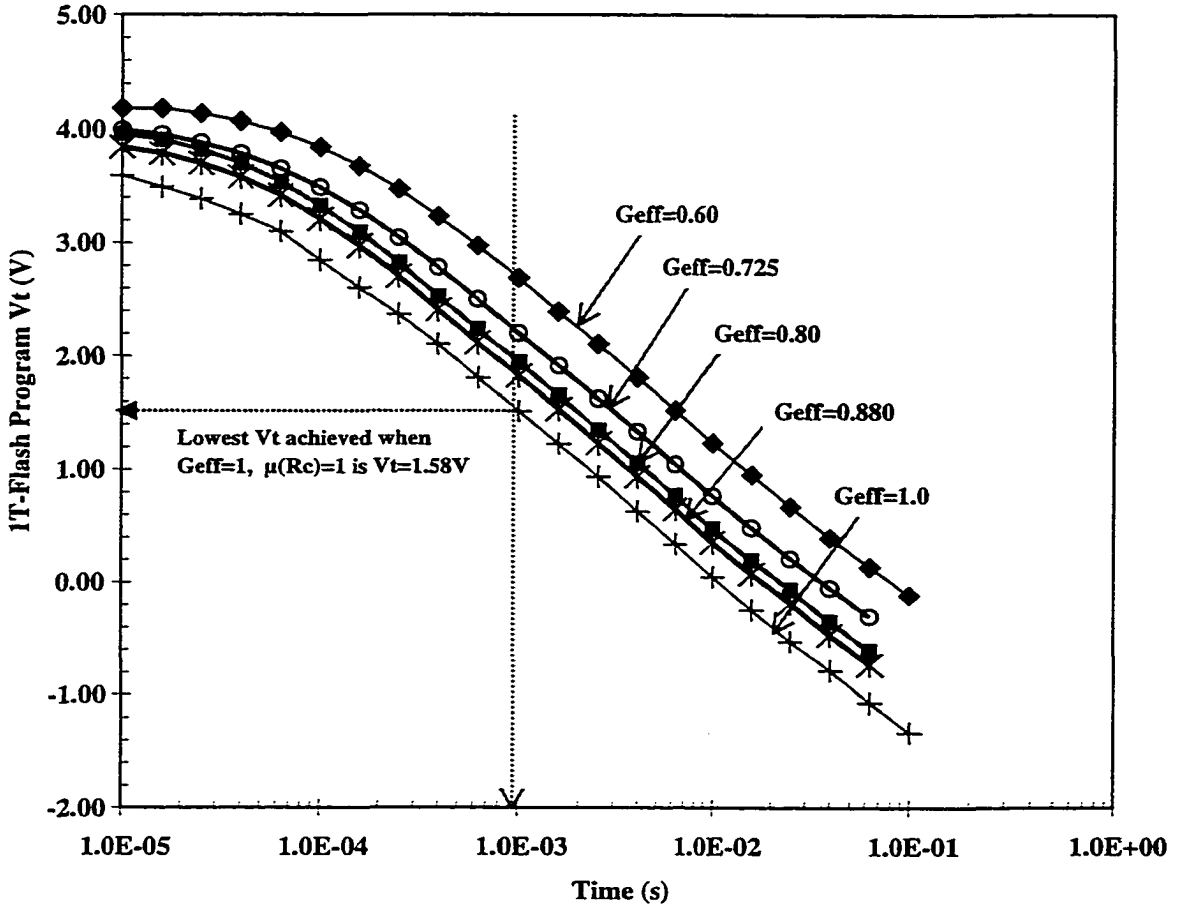


Figure 5.35: Modeling of Program Vt Family of Curves versus Grain Area factor (Geff)

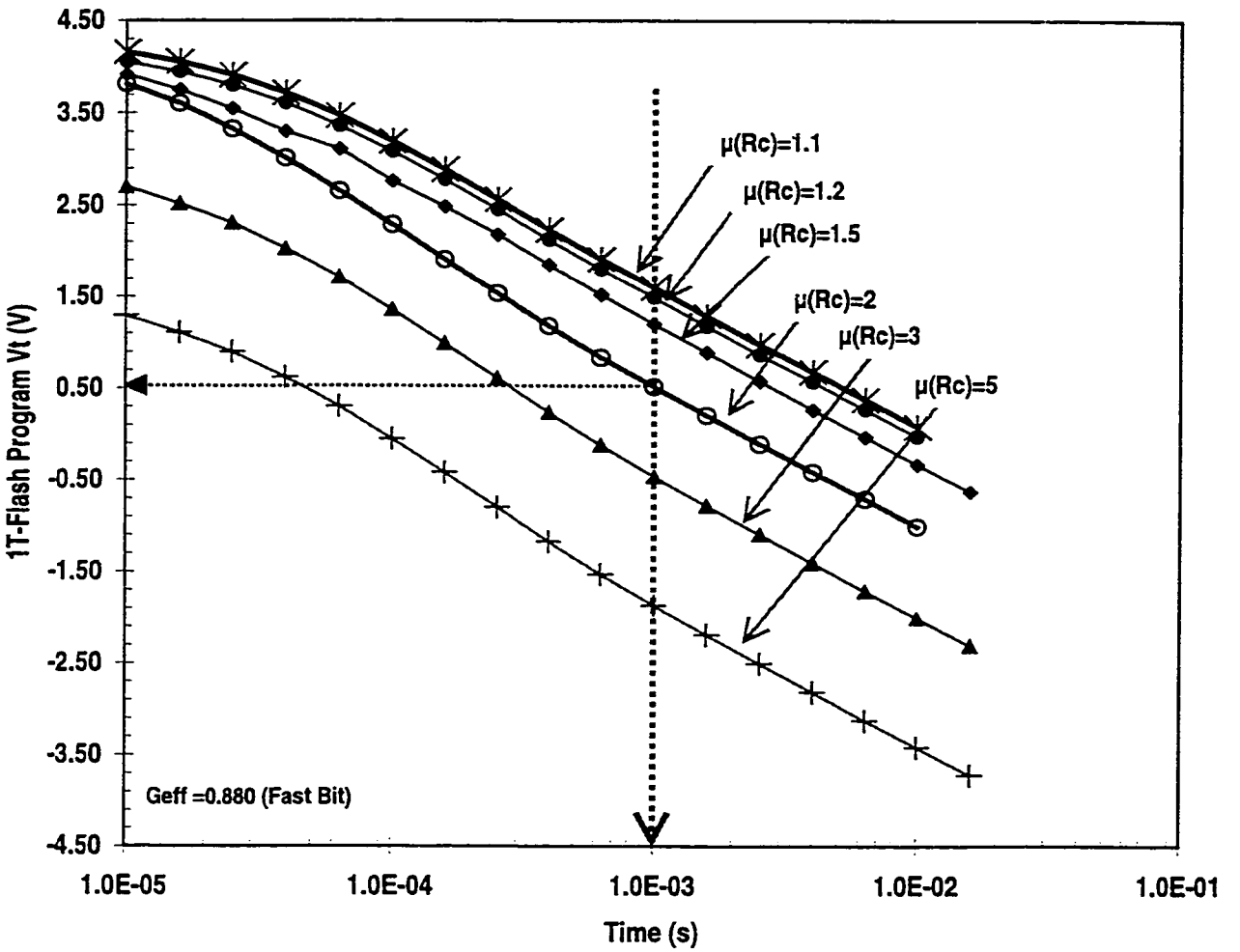


Figure 5.36: Modeling of Program Vt Family of Curves versus Enhancement Factors (u)

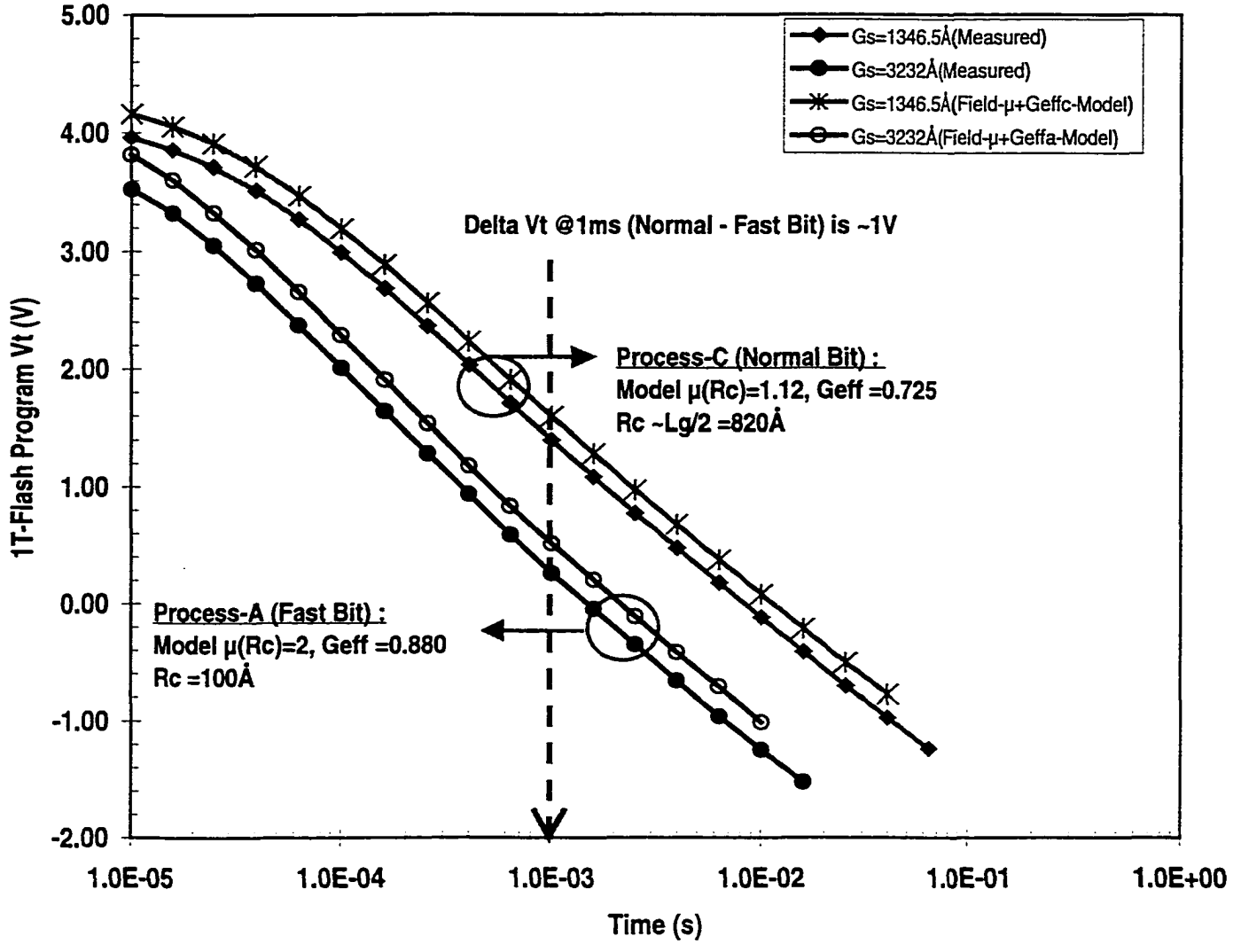


Figure 5.37: Program Vt Model versus Experimental Data for Fast and Normal Bits

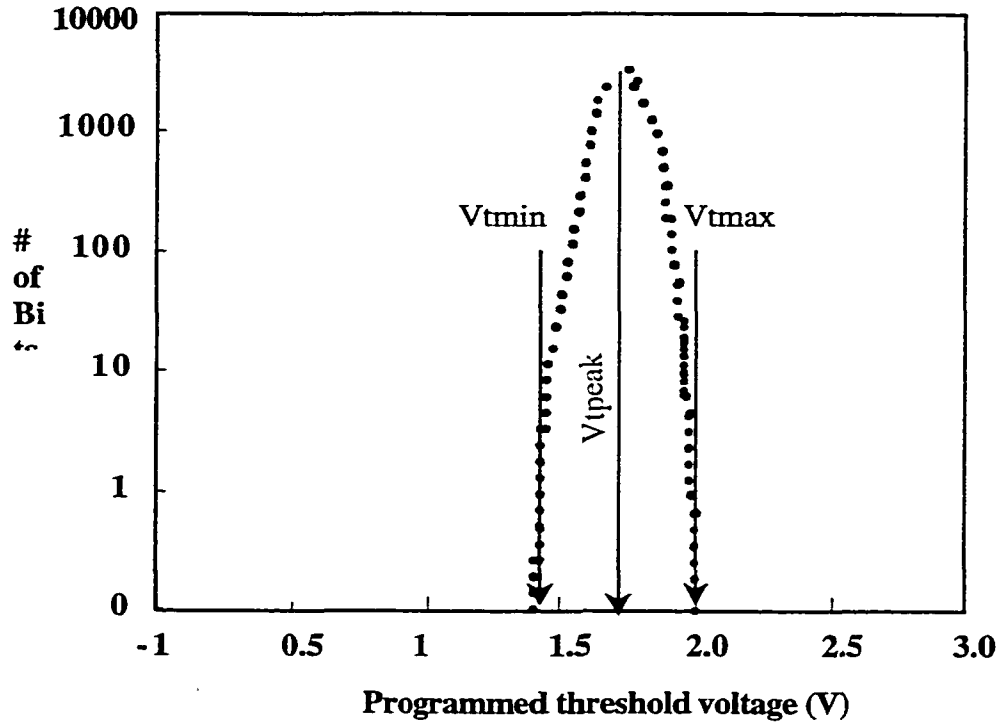


Figure 5.38: Experimental Data of 1T-Flash Program Vt Distribution for Normal Bit

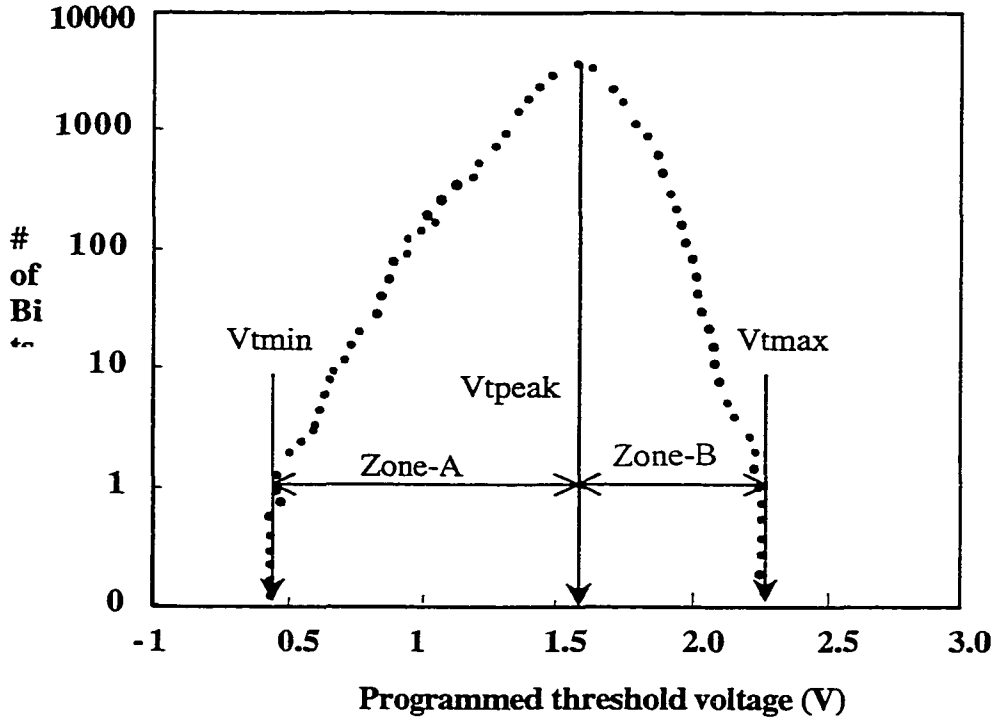


Figure 5.39: Experimental Data of 1T-Flash Program Vt Distribution for Fast Bit

Variable	Vtmin	Vtpeak	Vtmax
Vt (V)	1.4	1.75	2
Geff	0.847	0.8	0.767
Lgrain (A)	2510	1894	1610
μ(Rc)	1	1	1

Table 5.3: Summary of Vt as a function of Model Parameters Geff and μ(Rc) for Normal Bit

Variable	Vtmin	Vtpeak	Vtmax
Vt (V)	0.434	1.6	2.25
Geff	0.88	0.821	0.792
Lgrain (A)	3232	2200	1817
μ(Rc)	2	1	1

Table 5.4: Summary of Vt as a function of Model Parameters Geff and μ(Rc) for a Fast Bit

Chapter 6

Summary of Research and Future Work

This dissertation has addressed several aspects of the performance degradation in 1T-Flash non-volatile memory devices. In particular, the physics of Fowler-Nordheim tunneling of electrons from a floating gate and the effects of the floating gate polysilicon morphology on the memory threshold voltages. The following is a summary of the key results of this research.

6.1 Summary

First, the physical mechanisms that govern the device operation and leads to degradation of the 1T-Flash device performance have been described. They include Fowler-Nordheim tunneling of electrons, band-to-band tunneling of energetic holes. These mechanisms usually lead to the generation of electrically active defects that degrade the device performance.

A number of electrical measurement techniques have been used to detect and characterize the electrical behavior of 1T-Flash EEPROMs. Measurements such as subthreshold measurement, charge pumping and the gate-induced drain leakage (GIDL), endurance cycling and data retention bakes have been used to probe the reliability of 1T-Flash EEPROMs.

Simulation tools were employed during the device design phase of the research, which proved vital to the implementation of the boron halo flash drain implant. This device

design was very crucial in achieving the desired short-channel margins of 100nm, which allowed the simultaneous optimization of leakage and programming speed.

A electric field model or simplified capacitive network model was introduced in chapter 2 to predict the floating gate potential of a 1T-Flash EEPROM cell which cannot be directly determined from the electrical measurements. The capacitive network model used for the extraction of the floating gate potential is based on the capacitive coupling coefficients that relate the floating gate potential to the terminal voltages and the floating gate charge, and yield results comparable to the exact electric field model.

1T-Flash devices were fabricated with the final device design parameters obtained from the supreme simulations. These devices were then used in both the wafer level characterization and package level reliability characterization.

New classes of 1T-Flash fast-programming bits have been reported during this research. Fast and Normal bits are quantified and characterized by examining bitcell programming and erase stability, before and after UV-erase and Data Retention. To further understand the role of FG poly grain morphology effects on tunneling currents, various FG - poly deposition and doping techniques were examined, which were further correlated to the memory programmed V_t distribution of bits in a 2MBit Flash EEPROM array. A number of these bits were found to be insensitive to annealing and do not exhibit the type of erratic behavior previously reported [11]. We report the effects of Flash "channel" programming, or severe gate disturb, on the threshold voltage of fast or over-erased bits. Experiments were performed to establish that this class of fast bits are non-erratic and remain fast after 250°C bake. These fast bits exhibit identical sub-threshold characteristics similar to that of a normal

bit after UV-erase, thus establishing that the initial charge stored on the floating gate is the same for both normal and fast bits.

Experiments were conducted to further investigate the possible causes of the fast programming bits. "Channel" programming or severe gate disturb was used to modulate the memory V_t state of the fast bits. It was discovered that the fast programming bit appears normal if the programming takes place over the entire channel region instead of the FG-drain junction overlap region. Further measurements of low B-B tunneling leakage at the flash drain junction indicates that the enhanced F-N tunneling is influenced at the FG-tunnel oxide interface and not by band-to-band tunneling of energetic holes. To further this understanding, MOS capacitors were fabricated using the flash floating gate as the cathode on tunnel oxide. The 1500Å FG poly was deposited and doped in three ways: 1) sandwich of un-doped plus phosphorus in-situ doped polysilicon, 2) un-doped followed by an Arsenic implant and 3) in-situ doped polysilicon film. From AFM measurements, the un-doped plus in-situ doped stack for the FG poly exhibited the smoothest morphology with the smallest RMS value of 1.06 which is a measure of the FG poly surface roughness. This was also confirmed with TEM grain size characterization as shown in chapter 5. The average grain size is largest for process - A (implanted) FG-poly and smallest for process-C (un-doped plus in-situ doped) FG-stack, correlated well with the AFM surface roughness characterization and the corresponding TEM micrographs.

We reported here that after a few cycles ($N > 50$), the programmed V_t of the fast-programming bits asymptotically increases to a higher steady state value where it remains, even after 6000 P/E endurance cycles. Although the observed V_t spread of the fast bit is larger than the normal bit, there is no evidence of an erratic V_t behavior.

In this research the experimental data showed convincingly that the fast - programming bit behavior could be modulated with the Floating gate polysilicon microstructure. As a results the modeling of the fast bit programming characteristics was focus on the incorporation of polysilicon physics into the memory V_t programming transient characteristics, which was then benchmarked against the experimental results. Besides the use of the capacitor model and I-V characteristics of the floating gate device for extraction and quantification of the coupling coefficients, they can also aid in the modeling of the memory bitcell behavior. A comparison of measures and calculated characteristics reveals the validity of the assumptions made concerning the physical mechanisms governing the programming and erase behavior of both the fast-programming bits and the normal bits.

In summary, the knowledge of fast programming bit origin has been enhanced. The fast bits arise from larger grain floating gate polysilicon, which has increased effective grain tunneling area. This in conjunction with the grain sidewall poly asperities generated after gate etch and post re-oxidation, lead to enhanced Fowler-Nordheim tunneling and thus produce bits with ultra-low program threshold voltages. A field enhancement factor of $\mu(Rc)=2$ in conjunction with the Geff =0.880 can accurately predict a fast bit. The delta Vt in zone-A (Vtpeak –Vtmin) can be predicted with $\mu(Rc)$ ranging from 1.5 to 2. Figure 5.39 also reveals that the field enhancement factor alone cannot explain the entire distribution. This is because even with the lowest $\mu(Rc) =1$ Vtmax is ~1.8V which indicates that the $\mu(Rc)=1$ and a Geff =0.880 combination is required to predict the Vtmax of ~2.25V for the fast bit distribution. This work therefore highlights the importance of polysilicon grain size and microstructure engineering during the development of non-volatile Flash EEPROM memories.

6.2 Future Work

The major reliability and endurance problems in the Flash EEPROM cell involve a charge gain or loss process, which depends heavily on the quality of the tunnel oxide. There are several processing techniques [38,39,40] which have shown to improve the quality of tunnel oxides, including a highly precise nitridation method, and fluorination. These techniques should be evaluated for nonvolatile memory applications.

Although this research has discussed that hole trapping could cause random enhancement of gate current through band-to-band tunneling, it is not yet clear whether the enhancement is caused by holes trapped in the bulk states or in the near-interface states. The defects created by Fowler-Nordheim tunneling needs to be examined in more detail in terms of their effect on the field enhancement. In addition, the oxide damage arising from erase which is over the entire channel should be examined to determine its effects on cell failures.

The modeling of all device operations of Flash EEPROM is very crucial to the development of embedded Flash cores for consumer applications. Although the model developed in the research provides a good first order model that incorporates the floating gate polysilicon physics and field enhancement factor ($\mu(Rc)$), future work should attempt to expand on it by incorporating the grain boundary trap density physics. This should include all classes of traps especially the amphoteric traps which has the potential effects of causing some low level memory threshold voltage instability.

References

[1] IEEE Non-Volatile Semiconductor Memory Workshop, CA., 1995.

[2] D. Khang and S.M. Sze. " A Floating Gate and its Application to Memory Devices". Bell Sys. Tech. Journal, 46:1283, 1967.

[3] H.A.R. Wegener et al. "The Variable Threshold Transistor, A New Electrically Alterable Nondestructive Read –Only Storage Device". IEDM, Washington D.C., 1967.

[4] Takeda et al; IEEE Transactions on Electron Devices ED-29 (4) April 1982, pg. 611-618.

[5] Simon Tam et al; IEEE Transaction on Electron Devices ED-29, 1982, pg. 1740

[6] N. Mielke et al. "Reliability Comparison of FLOTOX and Textured Polysilicon EEPROMs". IRPS Tech. Dig., page 85, 1987.

[7] R. E. Shinner et al. " Characterization and Screen of Si02 defects in EEPROM Structures". IRPS Tech. Dig. Page 248, 1983.

[8] N. Mielke et al. "Reliability Comparison of FLOTOX and Textured Polysilicon EEPROMs". 8th NVSM Workshop Vail Co., 1986.

[9] J. Lee, B. Euzent, N. Boruta and C. Jenq. "Reliability Aspects of A Floating Gate EEPROM". IRPS, pages 11-16, 1981.

[10] Wu et al; IEEE IRPS, 1990, pg. 145.

[11] T.C. Ong, A. Fazio, N. Mielke, S. Pan, N. Righos, G. Atwood and S. Lai, "Erratic erase in ETOX™ Flash memory array", *Symposium of VLSI Tech.*, 1993, pp. 83-84.

[12] S.K. Lai and V.K. Dham. Comparison and trends in today's dominant E2 technologies. IEDM Tech. Dig., pages 580-583, December 1986.

[13] J. Frenkel. "On pre-breakdown phenomena in insulators and electron semiconductors. Phys. Rev, 54:647, 1938.

[14] M. Lenzlinger, E.H. Snow. "Fowler-Nordheim Tunneling into thermally grown SiO2". Journal of Applied Physics, vol. 40:278, January 1969.

[15] Schiff. Quantum Mechanics pages 268-279.

[16] R. Shirota, T. Endoh, M. Momodomi, R. Nakayama, S. Inoue, R. Kirisawa and F. Masuoka. "An Accurate Model of sub-breakdown due to band-to-band tunneling and its application. Pages 26-29, 1988.

[17] Chi Chang, Jih Lien. "Corner-field induced Drain Leakage in thin oxide MOSFETs". Pages 714-717, 1987

[18] K. Kurimoto, Y. Odake and S. Odanaka. Drain Leakage Current Characteristics due to band-to-band Tunneling". IEEE-IEDM Tech. Dig., pages 621-624, December 1989.

[19] Mariko Takayanagi and Shuichi Iwahuchi. "Theory of band-to-band tunneling under nonuniform electric fields for sub-breakdown Leakage Currents. IEEE Trans. Electron Devices, Vol 38:pages 1425-1431, June 1991.

[20] I. Nedev, A. Asenov and E. Stefanov. "Experimental Study and Modeling of band-to-band Tunneling Leakage Current in thin-oxide MOSFETs. Solid State Electronics, Vol 34: pages 1401-1408, 1991.

[21] Kirk Prall, Wayne I. Kinney, Jon Marco. "Characterization and Suppression of Drain Coupling in Submicron EPROM cell. IEEE Trans. Electron Devices, ED-34, Vol 12: page 2463, 1987.

[22] M. Wada, S. Mimura, H. Nihira, H. Iuzuke. "Limiting factors for programming EPROM of reduced dimensions. IEDM Tech. Dig., page 38, 1980.

[23] A. Kolodny, S.T.K. Nieh, B. Eitan, J. Shappir. "Analysis and Modeling of the Floating gate EEPROM cells. IEEE Trans. Electron Devices, ED-33, Vol. 6: pages 835, June 1986.

[24] J. S. Brugler and P. G. A. Jespers. "Charge Pumping in MOS Devices", IEEE T Elec. Devices. Vol. ED-16,no. 3:297, 1969.

[25] G. Groeseneken, H.E. Maes, N. Beltran and R. F. De Keersmaecker. "A Reliable Approach to Charge Pumping Measurements in MOS Transistors". IEEE Trans Elect. Dev. ED-31: 42, 1984

[26] W.L. Tzeng. "A New Charge Pumping Method of Measuring Si-Si02 Interface States". J. Appl. Phys., Vol. 62, N0. 2:591, 1987.

[27] J. G. Simmons and L.S. Wei. "Theory of Dynamic Charge Current and Capacitance Characteristics in MIS Systems containing Distributed Surface Traps". Solid State Electronics, 16:53-66, 1973.

[28] R.R. A. Siergiej. " Quantization Effects on Interface Traps Modeling and Techniques to Characterize highly doped Oxy-Nitride and pure oxide Silicon Field Effect Transistors". Ph.D. Thesis, Lehigh University, Bethlehem, PA., 1992.

[29] S.S. Roth, W. Ray, C. Mazure, and H.C. Kirsch, "Polysilicon encapsulated local oxidation," *IEEE Electron Device Lett.*, vol. 12, p. 92, 1991.

[30] Conn-Luan Tran. "The Scaling of Submicron CMOS Devices". Ph.D. Thesis, Lehigh University, 1990.

[31] H. Onoda et al, " A Novel Cell Structure for a 3 Volt Operation, Sector Erase Flash

Memory", IEDM Technical Digest, p. 599-602, December 1992.

[32] S. Mori, Y.Y. Araki, M. Sato, H. Meguro, H. Tsunoda, E. Kamiyia, K. Yoshikawa, N. Arai and E. Sakagami, "Thickness scaling limitations factors of ONO interpoly dielectrics for nonvolatile memory devices", *IEEE Trans. Electron Devices*, vol. 43, p. 47, Jan 1996.

[33] V. Kynett, M. Fandrich, J. Anderson. P. Dix, O. Jungroth, J. Kreifels, R. Lodenquai, B. Vajdic, S. Wells, M. Winston and L. Yang, "A 90ns one million erase/program cycle 1-Mbit flash memory", IEEE J. Solid-State Circuits, Vol. 24, p. 1259, 1989.

[34] J. Van Houdt, D. Wellekens, L. Faraone, L. Haspeslagh, L. Deferm, G. Groeseneken, H.E. Maes, "A 5V -compatible flash EEPROM cell with microsecond programming time for embedded memory applications", IEEE Trans. Comp. Hybrids Manuf. technology, vol.17 p. 380 1994.

[35] G. Groeseneken, H.E. Maes, N. Beltran and R.F. De Keersmaecker, "A reliable approach to charge-pumping measurements in MOS transistors", IEEE Trans. Electron Devices, vol. 31, p 42, 1994.

[36] P. Heremans, J. Witters, G. Groeseneken and H.E. Maes, "Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation", IEEE Trans. Electron Devices vol. 36, p 1318, 1989.

[37] U. Sharma, R. Moazzami, P. Tobin, Y. Okada, S. Cheng and J. Yeargain. "Vertically scaled, highly reliable EEPROM devices with ultra-thin oxy-nitride films prepared by RTP in N20/02 Ambient". IEEE IEDM Tech. Dig., page 461, December 1992.

[38] T. Hori, T. Ohozone, Y. Odake and J. Hirase. "A MOSFET with Si-Implanted Gate SiO2 Insulator for Nonvolatile Memory Applications". IEEE IEDM Tech. Dig., page 469, December, 1992.

[39] H. Fukuda, A. Uchiyama, T. Kuramochi, T. Hayashi, T. Iwabuchi, T. Ono and T. Takayashiki. "High-Performance scaled Flash-type EEPROMs with heavily Oxy-Nitrided Tunnel Oxide Films". IEEE IEDM Tech. Dig., page 465, December 1992.

[40] M. Kuhn. "A Quasi-Static Technique for MOS C-V and Surface State Measurements". Solid State Electronics, Vol. 13: 873, 1970.

[41] A.K. Agarwal. " A Study of Traps in the Metal-Insulator Semiconductor (MIS) System with the Three Terminal Gated-Diode Structure". Ph.D. Thesis, Lehigh University, 1984.

[42] F. R. Libsch. Physics, Technology and Electrical Aspects of Scaled MONOS/SONOS Devices for Low Voltage Non Volatile Semiconductor Memory (NVSMs) PhD thesis, Lehigh University, 1989.

[43] A. Roy and M.H. White. "A New Approach to Study Electron and Hole Charge Separation at the Semiconductor-Insulator Interface." IEEE Trans. Elect. Dev., Vol. 37, No.6:pp. 1504, 1990.

[44] D. Frohman-Bentchkowsky and M. Lenslinger. "Charge Transport and Storage in Metal-Nitride-Oxide-Silicon (MNOS) Structures." J. Appl. Phys, Vol. 40, No8:3307, 1969.

[45] C. T. Sah, T.H. Ning and L.L. Tscholpp. "The scattering of electrons by surface oxide charges and by lattice vibrations at the silicon-silicon dioxide interface". Surface Sci., Vol. 32:561-575, 1972.

[46] S.M. Sze. Physics of Semiconductor Devices. Wiley, 2nd edition, 1981.

[47] S. Yoon. "Reliability of Thin Gate Insulators for Submicron MOSFET and EEPROM Device". PhD. Thesis, Lehigh University, 1992.

[48] J. Levison et al; Journal of Applied Physics, Vol 53, page 1193, 1982.

[49] J.G. Fossum et al; IEEE Electron Device Letters ED-30, No. 8, 1983.

[50] M.K. Hatalis and D.W. Greve, Journal of Applied Physics Vol 63 (7), 1988.

[51] M.K. Hatalis and J.H. Kung, Electrochemical Society Meeting, 1991

[52] I-Wei Wu, W.B. Jackson, A.G. Lewis and A. Chiang, IEEE Electron Device Letters, Vol. 12, No. 4, 1991.

[53] R. Reif and J.E. Knott, Electron Device Letters, Vol. 17 pages 586-588, 1981.

[54] M.K. Hatalis and D.W. Greve, IEEE Electron Device Letters, EDL-8, No. 8 pages 361-364, 1987.

[55] Alan G. Lewis et al; IEDM Technical Digest, page 575, 1991.

[56] S. Ikeda et al; IEDM Technical Digest, page 469, 1990.

[57] T. Kamins, "Polycrystalline silicon for Integrated Circuit Application", page 57.

[58] W.A. Brown and T.I. Kamins, "Solid State Technology", Vol 22, No. 51, pages 51-57, 1979.

[59] M.S. Rodder, PhD Thesis, MIT, 1987.

[60] K.R. Olasupo and M.K. Hatalis, Spring MRS Symposium, 1994.

[61] J. Frenkel, Physics Review Vol. 54, page 657, 1938.

[62] J.G. Fossum, A. Ortiz-conde, H. Shichijo and S.K. Banerjee, IEEE Trans. Electron Devices, Vol. ED-32, N0. 9, pages 1878-1884, 1985.

[63] I. Lundstrom and C. Svensson, Journal of Applied Physics, Vol. 43, No. 12, pages 5045-5047, 1972.

[64] Y. Hu, Ph.D. Dissertation, Lehigh University , 1992

[65] K. Kobayashi et al., " A high-speed parallel sensing architecture for Flash EEPROMs", IEEE J. of Solid-State Circuits, vol 25, no.1 pg 79-83, 1990.

[66] A. Roy, F. Libsch and M. White, "Electron Tunneling from Polysilicon Asperities into Polyoxides", Journal of Solid-State Electronics, vol32., No. 8 pg 655-659, 1989.

[67] M. Offenberg et al., "Role of surface passivation in the integrated processing of MOS structures, "Symposium of VLSI Technology Dig. Tech. Paper pg. 117, 1990.

APPENDIX A

Fabrication Flow for 1T-Flash NVSM Transistor

scribe wafers	
nitride densification anneal	
active photo	
active nitride etch	
field oxidation	
active oxide/nitride etch	
sacrificial oxidation	400A
HV/Isolated P-well Photo	
HV/Isolated P-well Photo	Boron 5E12 400KeV
HV/Isolated P-well Implant	Phosphorus 5E12 2.5MeV
HV NMOS Vt Adjust Implant	BF2 6E12 50KeV
HV N-well Photo	
HV N-well Implant	Phosphorus 7E12 600KeV
HV N-well Implant	Phosphorus 5E12 50KeV
sacrificial oxide etch	500A
tunnel oxide	100A
1500A in-situ doped poly1	
poly1 photo	
poly1 etch	
ONO bottom oxide	130A
ONO nitride	180A nitride
ONO photo	
ONO etch	
Pre-HV Gate Oxide Sacrificial Oxide	250A 900C
Pre-HV Gate Oxide Sacrificial Oxide Etch	250A
HV gate oxidation	400A
poly2 deposition	2000A poly
gate poly ARC deposition	
self-aligned poly stack photo	
self-aligned poly stack etch	
self-aligned source/drain photo	
self-aligned source/drain etch	7000A Oxide
gate poly photo	
gate poly etch	
Sidewall oxidation	50A, 850°C
flash EEPROM drain implant photo	
flash EEPROM drain implant	Phosphorus or Arsenic
flash EEPROM drain implant resist strip	
gate spacer deposition	1000A nitride
gate spacer etch	
N-channel S/D implant photo	
N-channel S/D implant	As, 5.0e15, 55 keV

N-channel LDD implant	P, 2.0e15, 45 keV
N-channel S/D implant resist strip	
N-channel S/D anneal	RTA at 1097C
P-channel S/D implant photo	
P-channel S/D implant	B, 5e15, 25 keV
P-channel S/D implant photo resist strip	
P-channel S/D anneal	RTA
S/D screen oxide etch	200A oxide
Ti deposition	700A Ti
first Ti silicide anneal	
salicide selective Ti etch	
Piranha clean	
second Ti silicide anneal	
ILD0 PETEOS deposition	2000A PETEOS
ILD0 BPSG deposition	9000A BPSG
ILD0 BPSG polish	6000A oxide remaining in field
contact photo	
contact etch in AMT5000	
contact glue deposition	
contact glue anneal	
contact plug tungsten deposition	6000A tungsten
contact plug tungsten polish	
metal1 deposition	
metal1 photo	
metal1 etch	
ILD1 PETEOS deposition	2000A PETEOS
ILD1 O3-TEOS deposition	3000A TEOS
ILD1 PETEOS deposition	10000A PETEOS
ILD1 PETEOS polish	6000A oxide remaining over metal1
via1 photo	
via1 etch	15000A PETEOS
via1 glue deposition	
via1 plug tungsten deposition	6000A tungsten
via1 plug tungsten polish	
metal2 deposition	
metal2 photo	
metal2 etch	
metal anneal	
passivation deposition	7000A oxynitride
passivation photo	P09
passivation etch with low power in AMAT 5000	7000A oxynitride

Appendix B

Simple Floating Gate Parameter Calculator
(For Establishing Modeling constants)

Given Values

Item	Value	Unit	Description
Vtfg	0.60	V	Threshold measured from the floating gate
Vtcgp	1.50	V	Programmed threshold measured from the control gate
Vtcge	4.30	V	Erased threshold measured from the control gate
Vgr	3.30	V	Control gate voltage during read
Vdr	1.00	V	Drain gate voltage during read
Vgp	-9.00	V	Control gate voltage during program
Vdp	5.00	V	Drain gate voltage during program
Vge	9.00	V	Control gate voltage during erase
Vde	0.00	V	Drain gate voltage during erase
Tfg	9.50E-07	cm	Tunnel oxide thickness
Tono	1.75E-06	cm	Interpoly ONO thickness (effective Tox)
Tfld	3.00E-05	cm	Average field oxide thickness over which poly 1 overlaps
Tpoly1	1.50E-05	cm	Poly 1 thickness
Weff	4.50E-05	cm	Effective control gate channel width
L	5.50E-05	cm	Physical control gate length
L2	4.50E-05	cm	Any additional poly stack width (in addition to L) on field
Lp1fld	1.70E-04	cm	Total poly 1 overlap of field oxide with poly stack width "L"
Lp1fld2	1.00E-04	cm	Total poly 1 overlap of field oxide with poly stack width "L2"
Lovrlp	0.00E+00	cm	Total poly 2 overlap of poly 1 edge
Δd	1.20E-05	cm	Underlap of the drain under the floating gate
Δs	5.00E-06	cm	Underlap of the source under the floating gate
er	3.90	-	Relative permmitivity of oxide
eo	8.85E-14	F/cm	Permmitivity of free space

Calculated Parameters

Item	Value	Unit	Description
ag	0.753	-	Gate coupling ratio
ad	0.046	-	Drain coupling ratio
Aono	1.63E-08	cm^2	Poly 2/poly 1 overlap area (including poly 1 side wall)
Cono	3.22E-15	F	Poly 2/poly 1 capacitance
Cfg	6.22E-16	F	Poly 1/channel capacitance
Cd	1.96E-16	F	Poly 1/drain capacitance
Cs	8.18E-17	F	Poly 1/source capacitance
Cfld	1.59E-16	F	Poly 1/substrate capacitance (over field oxide)
Cf	1.06E-15	F	Total floating gate capacitance (Cfg + Cd + Cs + Cfld)
Ct	4.28E-15	F	Total capacitance
Qfgp	-2.46E-15	C	Charge on the floating gate in a programmed state
Qfge	-1.15E-14	C	Charge on the floating gate in an erased state
ΔQfg	9.02E-15	C	Change in floating gate charge for a change in logic state
Vfgpo	-0.57	V	Floating gate voltage after program w/all nodes at ground
Vfgeo	-2.68	V	Floating gate voltage after erase w/all nodes at ground
Vfgp	-9.23	V	Peak floating gate voltage during program (before programming begins)
Vfge	6.20	V	Peak floating gate voltage during erase (before erasing begins)
Vfgr	1.95	V	Floating voltage during read
Vtuv	0.74	V	Natural (charge neutral) threshold measured from the control gate
Efgpo	-0.60	MV/cm	Electric field across the tunnel oxide after program w/all nodes at ground
Efgeo	-2.82	MV/cm	Electric field across the tunnel oxide after erase w/all nodes at ground
Efgp	-14.97	MV/cm	Peak electric field across the tunnel oxide during program (drain)
Efgpx	-12.76	MV/cm	Electric field across the tunnel oxide at the end of program (drain)
Efge	6.52	MV/cm	Peak electric field across the tunnel oxide during erase
Efgex	4.31	MV/cm	Electric field across the tunnel oxide at the end of erase
Eonopo	-0.33	MV/cm	Electric field across the ONO after program w/all nodes at ground
Eonoeo	-1.53	MV/cm	Electric field across the ONO after erase w/all nodes at ground
Eonop	0.13	MV/cm	Peak electric field across the ONO during program
Eonopx	-4.07	MV/cm	Electric field across the ONO at the end of program
Eonoe	1.60	MV/cm	Peak electric field across the ONO during erase
Eonoex	2.34	MV/cm	Electric field across the ONO at the end of erase

Vitae

Franklin D. Nkansah received a Bachelors of Science degree from Kutztown University of Pennsylvania in December 1988 with a Major in Engineering Physics. He begun his Solid State career in AT&T Bell Laboratories in Allentown PA in the area CMOS Process Integration, and completed his Master of Science degree at Lehigh University in May 1993 with a Major in Electrical Engineering. Since July 1993 he has been engaged in research/development in Motorola Semiconductor Products Sector in Austin Texas, where he worked on various SRAM, Logic and BiCMOS technologies from 0.4um to 0.18um technology generations. Franklin Nkansah is currently a Device Section Manager leading an advanced technology development team of 5 Ph.D's and 4 Technicians to develop and prototype advanced 0.18um Wireless DSP Baseband Processors for Motorola. He has authored and co-authored 8 external publications and over 30 internal publications and presentations.