2004

Investigation of oxide thickness dependence of Fowler-Nordheim parameter B

Shashank Bharadwaj
University of South Florida

Investigation Of Oxide Thickness Dependence Of Fowler-Nordheim Parameter B

by

Shashank Bharadwaj

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Electrical Engineering
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Dr. Yun Lee Chiou, Ph.D.
Dr. Wilfrido A. Moreno, Ph.D.
Dr. Christos Ferekides, Ph.D.

Date of Approval:
March 25, 2004

Keywords: i-v characteristics, c-v characteristics, nvm, tunneling, scaling

ACKNOWLEDGEMENT

I owe my grateful acknowledgement to Dr. Chiou for his invaluable guidance and constant encouragement over the course of time spent in this research effort. Finally, he poured every inch of the report with painstaking attention to detail and offered extremely useful comments.

My special thanks are due to Professor Ferekides and Professor Moreno for serving on my committee.

I would also like to thank Mehdi Mansouri and Surendra for all their help during the course of this research work. A special thanks to Sharp Laboratories of America, Inc. for providing our lab with the wafers, without which this research work would not have been possible.

Lastly, I would like to put on record my appreciation for the patience, encouragement, and indulgence of my parents, my sister and my wife.

Table of Contents

List of Tables

List of Figures

Investigation of Oxide Thickness Dependence of Fowler-Nordheim Parameter B

Shashank Bharadwaj

ABSTRACT

During recent years the thickness of the gate oxide has been reduced considerably. The progressive miniaturization of devices has caused several phenomena to emerge such as quasi-breakdown, direct tunneling and stress induced leakage currents. Such phenomena significantly modify the performance of the scaled-down MOSFETs. As a part of this research work, an effort has been made to study the performance and characteristics of the thin Gate oxide for MOSFETs and Tunnel Oxide for Floating Gate (FG) MOS devices. The exponential dependence of tunnel current on the oxide-electric field causes some critical problems in process control. A very good process control is therefore required. This can be achieved by finding out the exact value of F-N tunneling parameter. This research work also is an effort of finding an accurate value for parameter B and its dependence on the oxide thickness as the device are scaled down to a level where the probability of Direct Tunneling mechanism gains more prominence.

A fully automated Low Current Measurement workstation with noise tolerance as low as 10^{-15} A was set up as a part of this research. C-V and I-V curves were analyzed to extract, determine and investigate the oxide thickness dependence of F-N parameter B. For oxide thickness in the range10~13 *nm*, the parameter B ranged between 260 and 267. Thus it can be said that it is not sensitive to the change in oxide thickness in this range. However it was noticed that for thickness around 7*nm* wide variety of results were obtained for the Fowler-Nordheim parameter B (B ranged from 260 to 454). This can be attributed to the

enhancement in the leakage current due to the direct tunneling. Hence to have tight control over V_T for a NVM, new algorithms need to be developed for even better process control for oxide thickness in the range of 7 *nm* and below.

Chapter 1

Introduction

Over the past two decades, the steady downscaling of transistor dimensions has been the main stimulus to the growth of silicon integrated circuits (ICs) and the information industry [1,2] . The more an IC is scaled, the higher becomes its packing density, the higher its circuit speed, and the lower its power dissipation [3]. These have been key in the evolutionary progress leading to today's computers and communication systems that offer superior performance, dramatically reduced cost per function, and much-reduced physical size compared to their predecessors.

The prevailing VLSI technology today comprises CMOS devices because of their unique characteristic of negligible standby power, which allows the integration of tens of millions of transistors on a processor chip with only a very small fraction (<1%) of them switching at any given instant. Also it wouldn't be wrong to state that the microelectronics industry owes a great deal of its success to the existence of the thermal oxide of silicon, i.e., silicon dioxide (SiO_2). A thin layer of SiO_2 forms the insulating layer between the control gate and the conducting channel of the transistors used in most modern integrated circuits. As circuits are made denser, all of the dimensions of the transistors are reduced ("scaled") correspondingly. As the CMOS dimension, in particular

the channel length, is scaled to the nanometer regime (<100 nm), however, the electrical barriers in the device begin to lose their insulating properties because of thermal injection and quantum-mechanical tunneling [4]. This results in a rapid rise of the standby power of the chip, placing a limit on the integration level as well as on the switching speed. Also the gate oxide has been scaled to a thickness of only a few atomic layers, where tunneling gives rise to a sharp increase in gate leakage currents.

Therefore to sustain the performance trends of last twenty years new solutions such as high dielectric constant and shallow ultra low resistivity junctions need to be developed.

1.1 Gate Dielectrics

Silicon dioxide has served for more than three decades as the gate insulator responsible for blocking current in insulated gate field-effect transistor channels from the gate electrode in CMOS devices. The reason for the nearly exclusive use of silicon dioxide in this application is that silicon dioxide uniquely possesses the required combination of several properties: good mobility of holes and electrons flowing in silicon at the silicon dioxide interface, ability to keep electronic states (surface states) at this interface low, relatively low trapping rates of holes and electrons, and excellent compatibility with CMOS processing. Since the gate insulator also serves to couple the electric potential from the gate electrode to the channel, the best control of the channel by the gate is obtained when the gate insulator is made as thin as possible

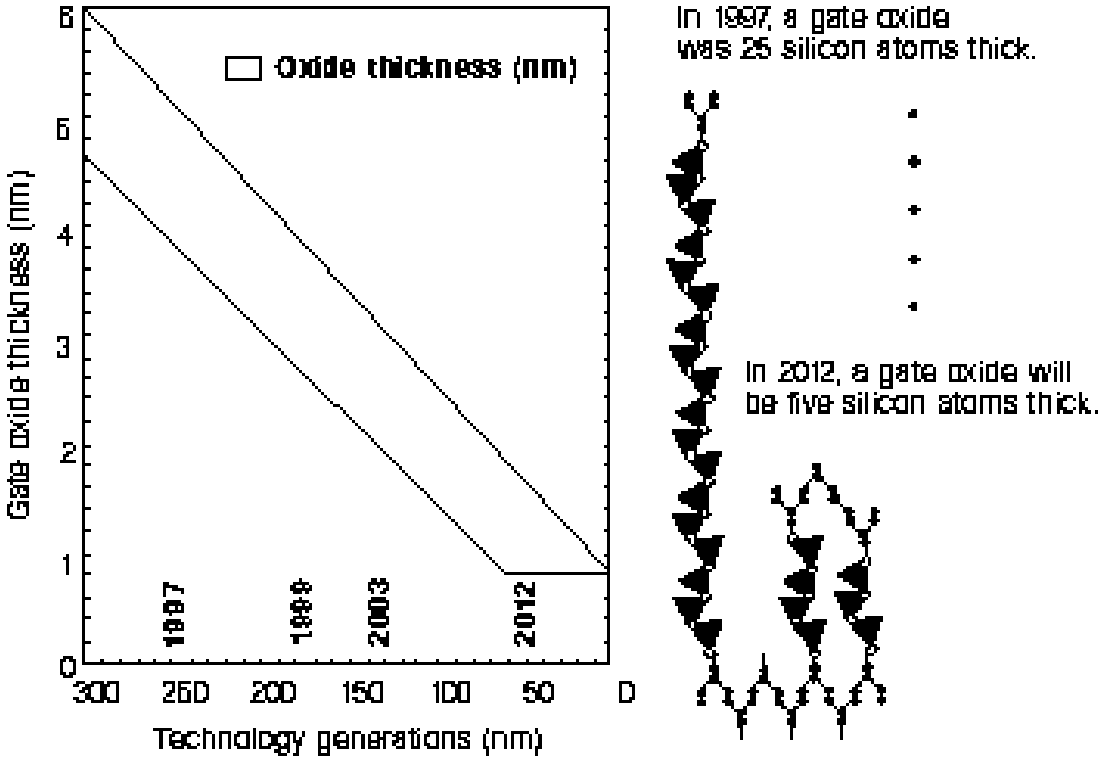

Figure 1.1 Predictions for the Next Few Generations of Chips

Figure 1.1 shows the 1997 prediction for the next few generations of chips [5]. In this plot the generations are defined by a critical device size, which is projected to decrease from 200 nm to 50 nm over the next 12 years. The gate oxide must be reduced in turn, from 25 silicon atoms today to 5 atoms in 2012 to achieve the roadmap goal. Clearly, there must be a limit to this scaling down because the gate-oxide thickness will eventually reach zero. As a matter of fact fundamental limits have been reached for SiO_2 scaling. Increased problems with dopant penetration through ultra thin SiO_2 layers and direct tunneling for ultra-thin oxide films dictate an aggressive exploration for new materials to be used in future gate dielectric applications.

1.2 High K Dielectric

One possible solution is to use an insulating film with a dielectric constant higher than that for SiO_2. A high dielectric constant allows the use of a thicker insulating layer for the gate electrode. There are a wide variety of films with higher k values than SiO_2, ranging from Si_3N_4 with a k value of 7, up to Pb-La-Ti (PLT) with a k value of 1,400. Unfortunately many of these films are not thermodynamically stable on silicon, or are lacking in other properties such as a high breakdown voltage, low defect density, good adhesion, thermal stability, low deposition temperature, ability to be patterned easily and low charge states on silicon. Currently interest seems to be centered on films such as HfO_2 and ZrO_2 with k values of 30-40 and 25 respectively, enabling a 6.4x to 10.3x increase in film thickness for equivalent performance. Intel recently reported over five orders of magnitude reduction of leakage for high-k oxides based on HfO_2 and ZrO_2 versus SiO_2 for equivalent oxide thicknesses. Transistors based on these films showed excellent overall performance. Another film which has shown excellent dielectric property has been La_2O_3 [6]. Although study initiated has been pretty recent, La_2O_3 could well be a good replacement candidate.

Figure 1.2 gives a brief idea about the effort, research and publications put on the different alternate high-k materials in the last two years or so. It is clearly seen that HfO_2 and ZrO_2 have been most popular research materials as a possible replacement for SiO_2.

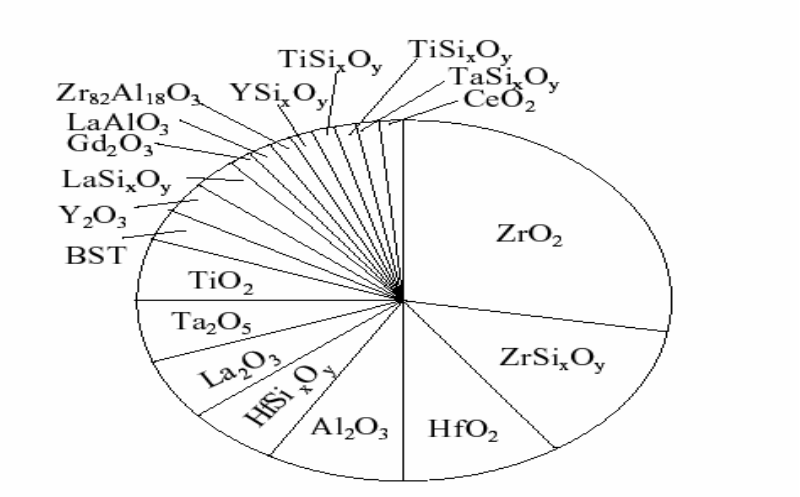

Figure 1.2 Reported High K Materials [7, 8]

1.3 Future Trends in Silicon Technology and Purpose of the Thesis

The guiding principle for the development of high-performance, deep sub-half-micron and sub-quarter- micron, ULSI devices is to increase functions that can be handled in a unit area of the device. Additionally, device reliability must be maintained while enhancing speed performance. During the last decade device physicists, researchers and engineers have been continuously faced with new elements that made the task of MOSFET characterization more and more crucial as well as difficult. The progressive miniaturization of devices has caused several phenomena to emerge such as quasi-breakdown, direct tunneling and stress induced leakage currents. Such phenomena significantly modify the performance of the scaled-down MOSFETs. Therefore, the first effort in this research is to study the performance and characteristics of thin Gate oxide MOSFETs.

The growing demand of high-density NVM (Non Volatile Memory) for the portable computing and telecommunications market has encouraged serious interest in Flash memory with the capability of multilevel storage and low voltage operation [9]. Multilevel storage implies the capability of storing two bits in a single cell. To do so, four different threshold voltages need to be correctly identified in the FG (Floating Gate) transistor. Even if a single bit is stored in a cell, threshold voltage levels need to be accurately determined to prevent errors from creeping during the write and erase operating modes of flash device. This can be achieved by finding out the exact value of F-N tunneling parameter B.

Research work in this direction has been an ongoing process and a wide range of values for parameter B has been published. However, up to the present time, there are no generally accepted values for parameter B. This research work is an effort of finding an accurate value for parameter B and its dependence on the oxide thickness as the device are scaled down to a level where the probability of Direct Tunneling mechanism gains more prominence. Such studies are necessary in order to obtain a better understanding of the dynamics of the writing and erasing in the memory devices, and the leakage current in the gate for MOSFETS.

Including this Introduction, this thesis contains 5 chapters. In chapter 2, the basic structures and operations of MOS devices have been discussed. Importance of scaling is also presented. Chapter 3 deals with the characterization theory and reliability issues. In chapter 4 analysis has been performed and results justifying the research have been

presented. Chapter 5 discusses the conclusions derived and also presents future work in this direction.

Chapter 2

MOS Devices

2.1 Ideal MOS Diode

An ideal MOS diode is defined as follows:

1 The energy difference between the metal work function Φ_m and the semiconductor work function Φ_s is zero.

2 The only charges that can exist in the structure under any biasing conditions are those in the semiconductor and those with the equal but with opposite sign on the metal surface adjacent to the insulator. This is equivalent to oxide charge being zero.

3 The resistivity of the insulator (oxide) is infinite so that there is no carrier transport under dc-bias.

The band diagram for metal, oxide and semiconductor when they are in contact and under zero bias is shown in Figure 2.1. As can be seen at zero applied voltage there is no energy difference between the metal work function Φ_m and the semiconductor work function Φ_s. In other words the band is flat (flat-band condition) when the applied voltage is zero.

If Φms is the work function difference then

$$\Phi_{ms} = \Phi_m - \Phi_s = \Phi_m - \left[\chi + \left(E_c - E_i\right)/q + \Phi_F\right] = 0 \qquad (2.1)$$

for p type, where χ is semiconductor electron affinity, E_c the band gap and Φ_F the potential difference between the Fermi level and intrinsic Fermi level.

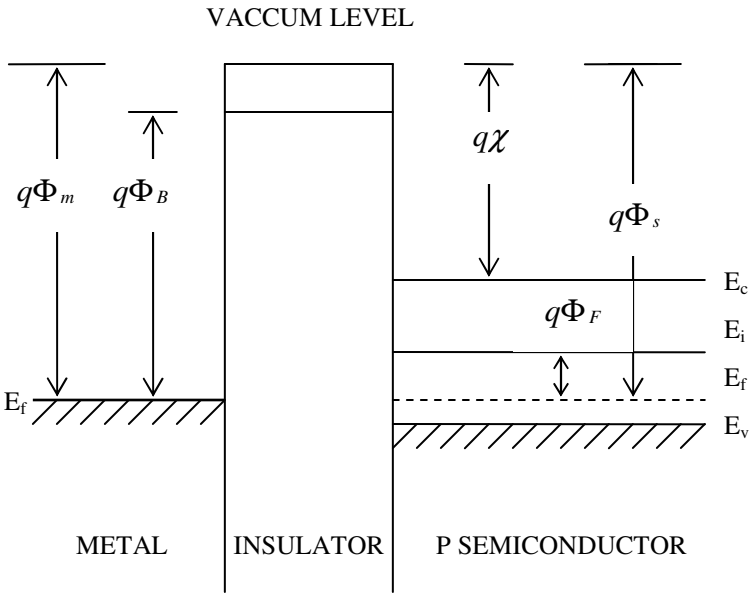

Figure 2.1 Band Diagram for Ideal MOS Diode [10, 11]

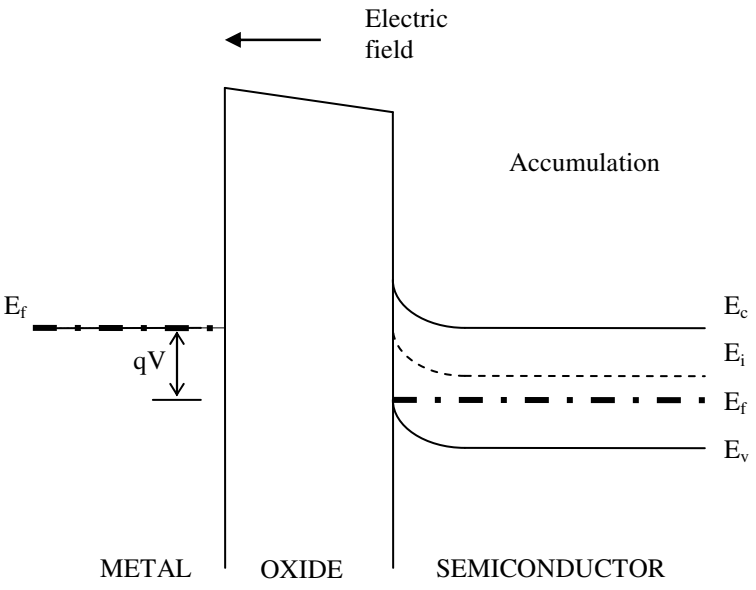

Figure 2.2 Band Diagram of MOS at Negative Voltage ($V_{GS} < 0$)

(Hole Accumulation of the Semiconductor Surface)

When an ideal MOS diode is biased with positive or negative voltages, three cases may exist at the semiconductor surface. Consider p-type semiconductor as shown in Figure

2.2. Now when a negative voltage V is applied to the metal plate, the top of the valence band bends upward and is closer to the Fermi level. Since the carrier density depends on the energy difference (E_F - $E_{V)}$ and Fermi level remains constant because of no current flow in an ideal MOS, this band bending causes accumulation of majority carriers (holes for p-type and electrons for n-type) near the semiconductor surface. This phenomenon is termed as Accumulation process.

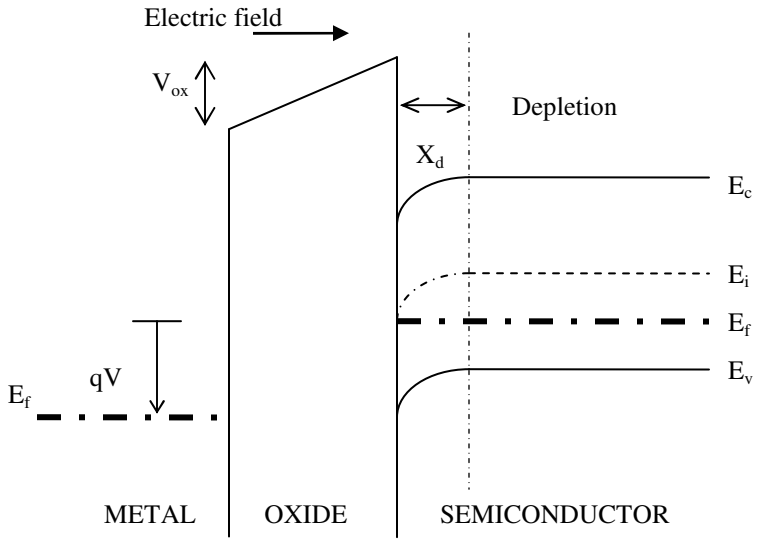

Figure 2.3 Band Diagram of MOS at Positive Voltage ($V_{GS} > 0$)

(Hole Depletion of the Semiconductor Surface)

For $V_{GS} > 0$ (positive voltage across the metal plate), metal Fermi level is lowered by qV relative to its equilibrium position. As a result, the oxide conduction band is again tilted (moving the metal oxide down relative to the semiconductor side). The positive voltage deposits positive charge on the metal and calls for a corresponding net negative charge at

the surface of the semiconductor. Such a negative charge in the P- type material arises from the depletion of holes from the near surface, leaving behind uncompensated ionized acceptors. This is the depletion case as shown in Figure 2.3. If the voltage is further increased in the positive direction, holes are repelled to a large extent. The bands bend even more downward such that the intrinsic level E_i at the surface crosses over the Fermi level as seen in Figure 2.4. At this point electrons become the majority carriers. The surface is thus "inverted" and this process is termed as inversion.

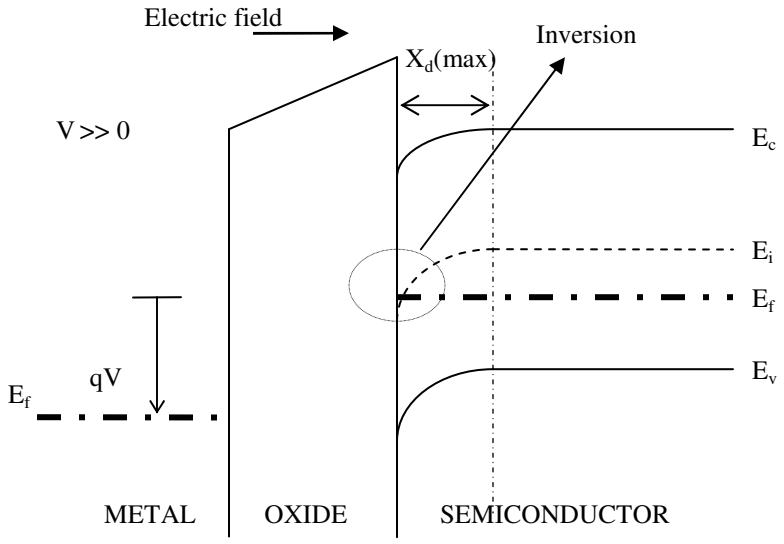

Figure 2.4 Band Diagram of MOS at a Large Positive Voltage

(Hole Depletion of the Semiconductor Surface)

It is should be noted that similar results can be obtained for the n-type semiconductor with the polarity of the applied voltage being reversed.

2.1.1 Threshold Voltage

The *threshold voltage* of an MOS transistor is the voltage required to turn on the transistor and allow current flow between the drain and source nodes. The transistor is considered to be turned on when an *inversion layer* has been established in the channel. An inversion layer is said to be established when the electrostatic potential between the bulk semiconductor and the oxide-semiconductor surface (called surface potential, Φ_S) is twice the Fermi potential Φ_F. For an nMOS transistor, when $\Phi_S = 2\Phi_F$, the p-type substrate semiconductor acts like an n-type semiconductor at the surface, and thus the transistor is in inversion.

The fundamental equation for MOS threshold voltage, V_T is

$$V_T = 2\Phi_F - \frac{Q_B}{C_{OX}} \tag{2.2}$$

To understand this equation better, certain parameters need to be defined at this stage. The Fermi potential ϕ_F is the difference (in volts) between the intrinsic energy level and the Fermi energy level deep inside the semiconductor substrate (i.e. away from the oxide surface).

$$\phi_F = \frac{E_i - E_F}{q} \tag{2.3}$$

and

$$E_i - E_F = kT \ln\left(\frac{N_A}{n_i}\right) \tag{2.4}$$

where N_A is the substrate impurity concentration. Hence the Fermi potential for an nMOS transistor can be written as:

$$\phi_F = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right) \tag{2.5}$$

When a transistor is in inversion, there is a depletion layer in the device channel (under the gate oxide) which provides a charge that must be balanced by the gate voltage. Thus, the threshold voltage must also account for this depletion charge. This charge in the bulk depletion layer, which is due to ionized impurities, is given by:

$$Q_B = -qN_A X_d = -\sqrt{2q\varepsilon_s N_A (2\phi_F)} \tag{2.6}$$

where X_d is the depletion layer width and is given by

$$X_d = \left[\frac{2\varepsilon_s \phi_s}{qN_A}\right]^{1/2} \tag{2.7}$$

2.1.2 MOS Capacitor under Bias

MOS capacitor consists of a metal-oxide-semiconductor layer structure which forms a voltage dependent capacitor. The most common MOS capacitor structure is shown in Figure 2.5. As seen the metal plate is a heavily doped p^+ poly-silicon layer. The insulating layer is silicon dioxide and the other plate of the capacitor is the semiconductor layer (in this case its an n-type silicon).

The capacitance depends on the voltage applied to the gate with respect to the body (substrate) and the frequency. This dependence is shown in Figure 2.6 (high as well as low frequency). There are roughly three regions of operation separated by two voltages. These are (1) *Accumulation* in which carriers of the same type as the body accumulates at the surface (2) *Depletion* in which the surface is devoid of any carriers leaving only a

space charge or depletion layer, and (3) *Inversion* in which carriers of the opposite type from the body aggregate at the surface to "invert" the conductivity type.

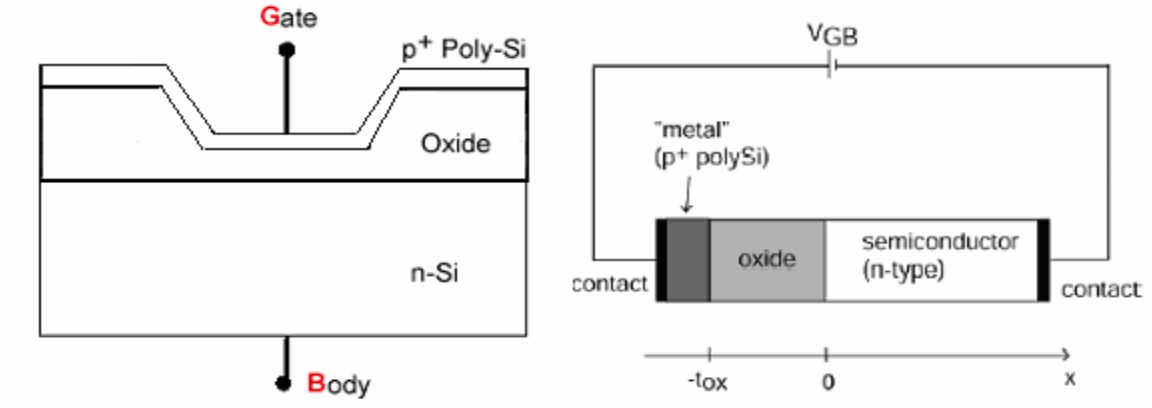

Figure 2.5 The MOS Capacitor Structure [12]

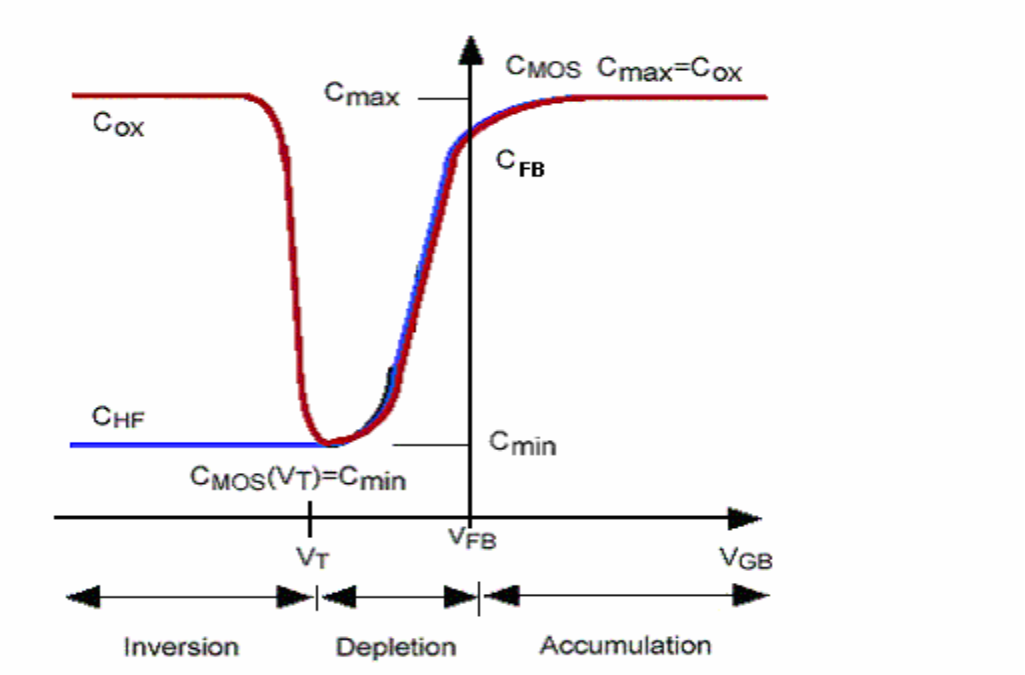

Figure 2.6 C-V Characteristics of an Ideal MOS Capacitor (n type substrate) [12]

The two voltages that demarcate the three regimes are *Flatband Voltage* (V_{FB}) which separates the accumulation regime from the depletion regimes and the *Threshold Voltage*.

2.1.2.1 Surface Accumulation

V_{FB} is the voltage at which there is no charge on the plates of the capacitor and hence there is no electric field across the oxide. Hence it's also called Flatband voltage and is zero for ideal MOS diode. Now when V_{GB} (applied voltage) > V_{FB} a positive charge is induced on the "metal" gate and a corresponding negative charge in the semiconductor. For an n-type silicon electrons accumulate at the surface. This condition is termed as surface accumulation. The charge distribution and equivalent circuit is shown in Figure 2.7. In accumulation the MOS structure appears like a parallel-plate capacitor and the MOS capacitance per unit area is given by

$$C = C_{OX} = \varepsilon_{OX} / t_{OX} \tag{2.8}$$

Where ε_{OX} is the permittivity of the oxide and t_{OX} is the thickness of the oxide.

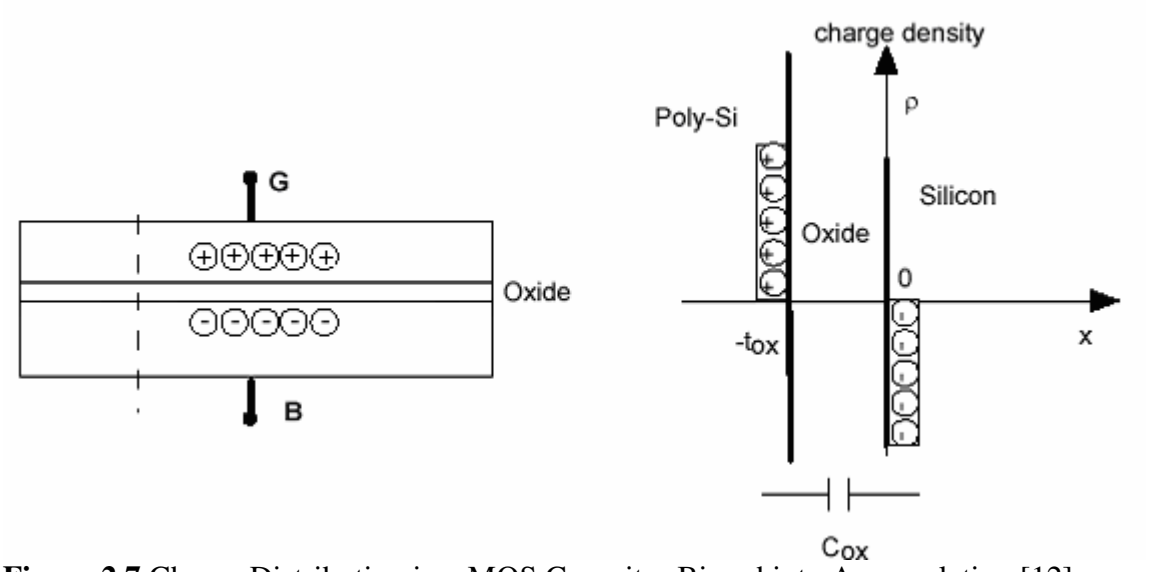

Figure 2.7 Charge Distribution in a MOS Capacitor Biased into Accumulation [12]

2.1.2.2 Surface Depletion ($V_T < V_{GB} < V_{FB}$)

If the applied voltage is less then the Flatband voltage, a negative charge is induced at the interface between the poly-silicon gate and the gate oxide. This leads to a positive charge being induced at the other interface i.e. the oxide/semiconductor interface. As a result the surface of the semiconductor is devoid of any mobile carriers (this being a n-type substrate). This layer at the surface is termed as the depletion layer as it prevents carriers from moving towards the semiconductor-oxide interface. Charge distribution under these circumstances is as shown in Figure 2.8. As seen this layer can be said to behave as an capacitor having a capacitance per unit area (C_D), which depends on V_{GB} and is given by

$$C_D = \varepsilon_{Si} / X_d \tag{2.9}$$

Where ϵ_{si} is the permittivity of the silicon and X_d is the depleted silicon layer thickness. It should be noted from the equation that C_D and X_d are function of the input voltage.

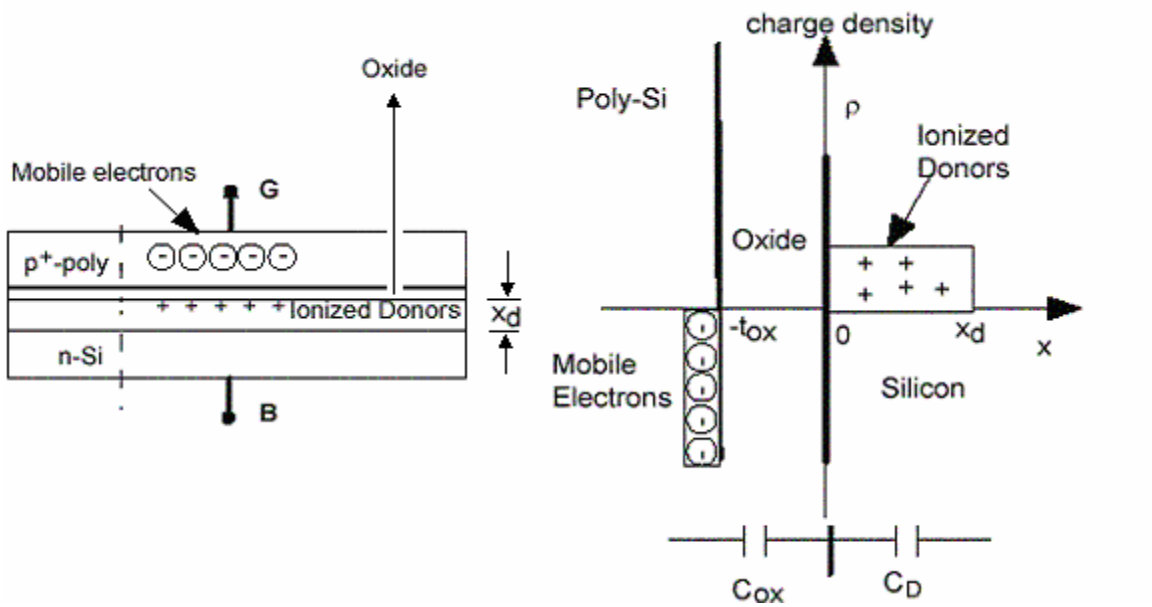

Figure 2.8 Charge Distribution in a MOS Capacitor Biased into Depletion and Equivalent Circuit Diagram [12]

A closer look at the figure will lead to another conclusion that oxide capacitance per unit area (C_{OX}) and the depleted silicon capacitance per unit area (C_D) are connected in series. Thus the capacitance of the MOS structure in the depletion regime is given by:

$$1/C_{MOS}(dep) = 1/C_D + 1/C_{OX} = (C_{OX} + C_D)/C_D * C_{OX} \quad (2.10)$$

2.1.2.3 Surface Inversion ($V_T > V_{GB}$)

Now consider a scenario where the applied gate voltage is lowered below the threshold voltage. The number of electrons decreases as the applied voltage decreases. Correspondingly the number of holes increases. This is owing to the fact that in a MOS capacitor in depletion or inversion, the holes and electrons are generated in the depleted silicon surface region. The holes are attracted to the Si/SiO_2 interface while the electrons are "pushed" into the substrate. However the holes could also come from a p-doped region that is in close proximity to the MOS capacitor such as the source/drain region of a n-MOSFET. Thus at a voltage less then the threshold voltage, V_T, the concentration of the holes at the surface far exceeds the concentration of electrons in the bulk. The conductivity type of the silicon surface is inverted. Figure 2.9 shows the charge distribution of the MOS capacitor in inversion.

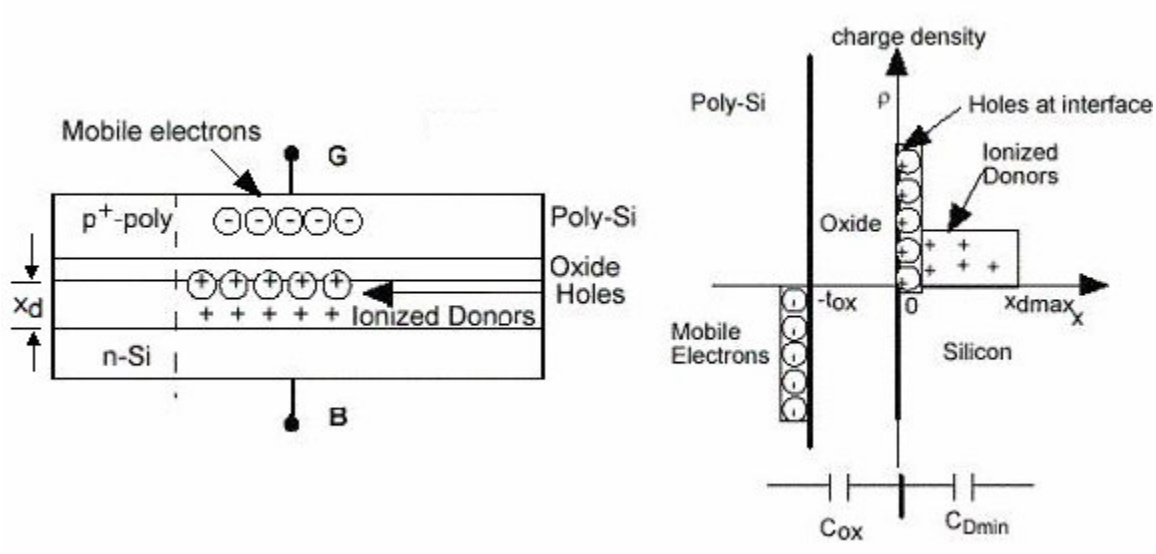

Figure 2.9 Charge Distribution in a MOS Capacitor Biased into Inversion [12]

After surface inversion, the depletion layer thickness reaches a maximum. Thus when the gate voltage is equal to the threshold voltage, the depleted layer capacitance per unit area reaches a minimum C_{Dmin} and likewise the MOS capacitance. This capacitance is given by

$$C_{\min}(inv) = C_{OX}C_{D\min} / C_{D\min} + C_{OX} \tag{2.11}$$

2.1.3 Non Ideal MOS Capacitor

The work function of the gate material Φ_m is not necessarily same as the work function of the semiconductor, and as a result there is a built-in potential across the gate oxide. This built-in potential is equal to the work function difference $\Phi_{ms.}$ A voltage applied to the gate must account for this work function difference before an inversion layer can be established. Thus, for a non ideal MOS Capacitor, threshold voltage, as defined by equation 2.2 changes to

$$V_t = \phi_{MS} + 2\Phi_F - \frac{Q_B}{C_{OX}} - \frac{Q_{OX}}{C_{OX}} \tag{2.12}$$

where Q_{ox} is the oxide charge and is explained in detail below.

The discussion about the ideal MOS capacitor assumed that there were no oxide charges in the oxide region. In a practical MOS capacitor, interface traps and oxide charges exist and affect the ideal MOS characteristics. These charges have energy states in the forbidden silicon band gap region and are able to exchange charges with the silicon in a short time. The interface trap charge Q_{it} charges can possibly be produced by excess silicon, excess oxygen, and impurities. The fixed oxide charges Q_f are located at or near the interface and are immobile under an applied electric field. The oxide trapped charges Q_{ot} can be created by hot-electron injection. These charges are distributed inside the oxide layer. The mobile ionic charges Q_m such as sodium ions, are mobile within the oxide under bias-temperature aging conditions.

The presence of oxide charge causes a voltage drop across the oxide, which is given by

$$\Delta V = \frac{-Q_{ss}}{C_{ox}} \tag{2.13}$$

In equation 2.12 Q_{ss} is the total oxide charge. The entire C-V curve shifts to a more negative value if Q_{ss} is positive.

.

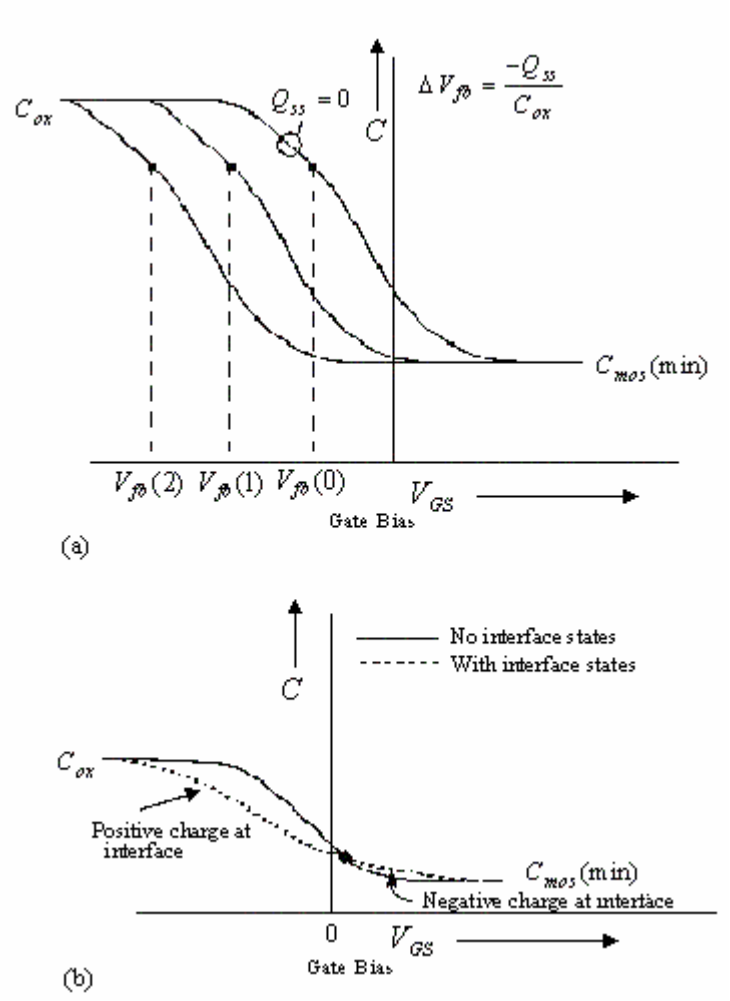

Figure 2.10 (a) A Plot of the High Frequency Capacitance-Voltage of a MOS Capacitor with Different Values of Fixed Oxide Charge. (b) The Effect of the Interface States is to "Smear" out the C-V Curves

A typical shift in the C-V characteristic is depicted in Figure 2.10a. The value of Q_{ss} can be obtained by measuring the shift with respect to the calculated ideal curve. Such measurements are very important for characterizing the quality of MOS devices. The

shift in the C-V curve is due to the fact that Q_m, Q_f and Q_{ot} are constant and independent of voltage.

The interface charge has a somewhat different effect on the C-V characteristics. In an ideal system, there are no electron states allowed in the bandgap of a semiconductor. However, since the $Si - SiO_2$ interface is not ideal, a certain density of interface states are produced that lie in the bandgap region. In contrast to the fixed charge, the electrons can flow into and out of these interface states. However, the electron motion is dependent upon the position of the fermi level. The nature of the interface states is characterized as either "acceptor-like" or "donor-like". An acceptor state is neutral if it is unoccupied, which occurs if the fermi level is below the state. An acceptor state is negatively charged if it is occupied, which occurs if the fermi level is above the state. A donor state is neutral if it is occupied, which occurs if the fermi level is above the state. A donor state becomes positively charged if it is unoccupied, which occurs if the fermi level is below the state. As a result, when the position of the fermi level is altered, the charge at the interface changes and the voltage drops. When the interface charge is positive the C-V curve shifts towards the negative voltage. When the interface charge is negative the curve shifts towards the positive voltages. The amount of shift is voltage dependent and is depicted schematically in Figure 2.10b. The curve is "smeared out" due to the presence of interface states.

2.2 MOSFET

2.2.1 Basic Structure and Principle of Operation of the MOSFET

A Metal-Oxide-Semiconductor Field-Effect-Transistor (MOSFET) is a four terminal device. The n-channel version of the device and the corresponding circuit symbol are illustrated in Figure 2.11. As shown it consists of a source and a drain, two highly conducting n-type semiconductor regions which are isolated from the p-type substrate by reversed-biased p-n diodes. A metal (or poly-crystalline) gate covers the region between source and drain, but is separated from the semiconductor by the gate oxide.

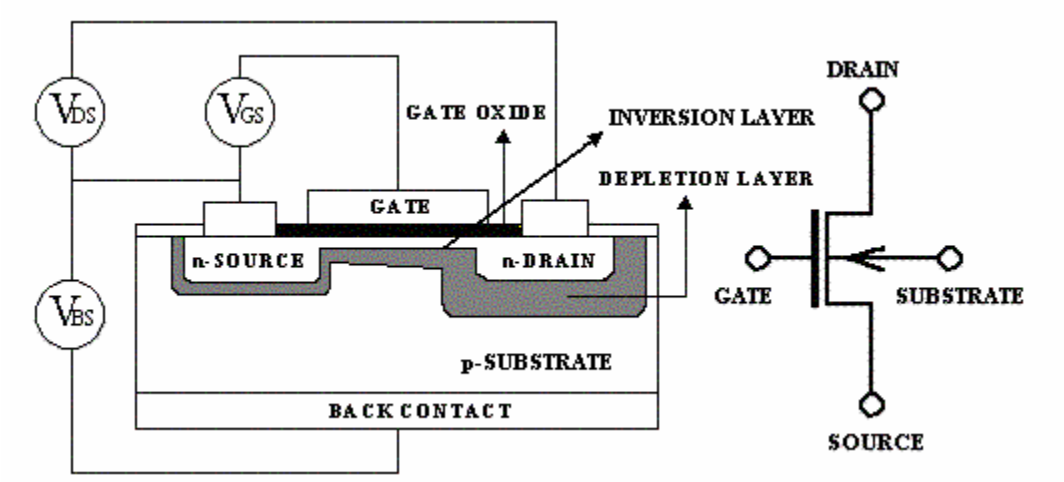

Figure 2.11 Cross-Section and Circuit Symbol of an n-channel (MOSFET)

As can be seen, the source and drain regions are identical when no voltage is applied. It is the applied voltages, which determine which n-type region provides the electrons and becomes the source, while the other n-type region collects the electrons and becomes the drain. The voltages applied to the drain and gate electrode as well as to the substrate by means of a back contact are referred to the source potential.

A conceptually similar structure was first proposed and patented by Lilienfeld and Heil [13] in 1930. The main technological problem was the control and reduction of the surface states at the interface between the oxide and the semiconductor.

Initially it was only possible to deplete an existing n-type channel by applying a negative voltage to the gate. Such devices have a conducting channel between source and drain even when no gate voltage is applied and are called "depletion-mode" devices. In contrast there are other set of devices, which do not have a conducting channel unless a positive voltage is applied. Such devices are referred to as "enhancement-mode" devices. The electrons at the oxide-semiconductor interface are concentrated in a thin (~10 nm thick) "inversion" layer. By now, most MOSFETs are "enhancement-mode" devices.

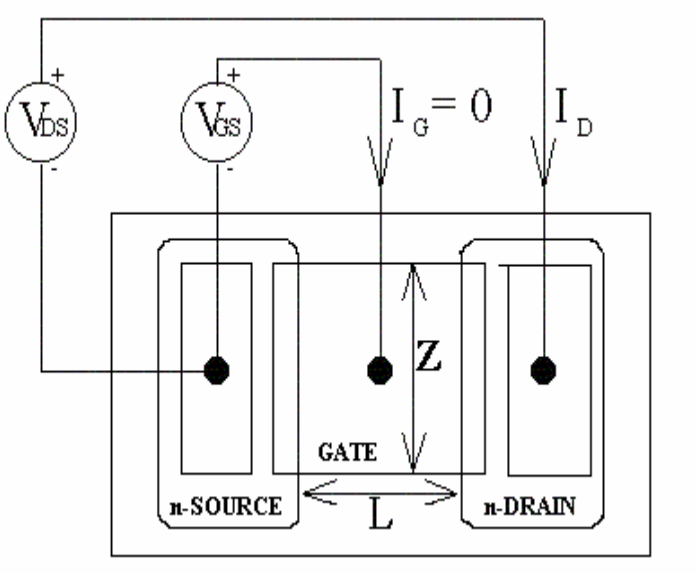

Figure 2.12 Top View of MOSFET

A top view of the same MOSFET is shown in Figure 2.12, where the gate length, *L*, and gate width, Z, are identified. Note that the gate length does not equal the physical dimension of the gate, but rather the distance between the source and drain regions

underneath the gate. The overlap between the gate and the source/drain region is required to ensure that the inversion layer forms a continuous conducting path between the source and drain region. Typically this overlap is made as small as possible in order to minimize its parasitic capacitance. The flow of electrons from the source to the drain is controlled by the voltage applied to the gate. A positive voltage applied to the gate attracts electrons to the interface between the gate dielectric and the semiconductor. These electrons form a conducting channel between the source and the drain called the inversion layer. No gate current is required to maintain the inversion layer at the interface since the gate oxide blocks any carrier flow. The net result is that the current between drain and source is controlled by the voltage, which is applied to the gate.

2.2.2 Current-Voltage Characteristics of the MOSFET

The electrical function of a MOSFET is to carry current between its source and drain terminal depending on the applied gate voltage. Any proper analysis of the MOSFET requires that charge flow in the device be studied in three dimensions. Several models have been proposed to find the characteristics of the device. The simplest model is discussed here. This model assumes that the mobility of the electron is constant and independent of the electric field. This is a good assumption for the long channel MOSFET. The condition for the n-MOSFET to be in the cut off region, where the current is zero, is when $V_G - V_S < V_T$. When the MOSFET is in the cut off region the source and drain are electrically isolated with respect to one another. When on, a bias is applied between the source and the drain, current flows in the channel near the Si-SiO_2 interface. The charge density in the channel is controlled by the gate bias as well as the

source-drain bias. Therefore, the gate bias can modulate the current flow in the channel. In order to find the current in the device, a constant electron mobility has been assumed. The current is assumed to be due to carrier drift and is given by

$$I_D = Q_n \mu_n \frac{dV_s}{dx} Z \tag{2.14}$$

where Q_n is the charge due to electrons within the inversion layer and is the difference between the total surface charge Q_s and Q_B after strong inversion.

The current, I_D is assumed constant throughout the cross-section of the channel. Thus

$$I_D dx = Q_n \mu_n Z dV_s \tag{2.15}$$

The integration of equation 2.14 throughout the length of the device gives

$$I_D = \frac{\mu_n Z C_{ox}}{L} \left[\left\{ V_{GS} - V_T - \frac{V_{DS}}{2} \right\} \right] V_{DS} \tag{2.16}$$

Equation (14) is valid for the device before pinch off. In the case when the drain bias, V_{DS} is much less than $V_{GS} - V_T$, the current equation is given by

$$I_D = \frac{\mu_n Z C_{ox}}{L} [(V_{GS} - V_T) V_{DS}] \tag{2.17}$$

For very small values of drain bias, the current increases linearly with drain bias because of the $\frac{\mu_n Z C_{ox}}{L}$ factor in equation 2.17. Therefore, it is defined as the linear region.

As the drain voltage is increased, the voltage across the oxide decreases near the drain, which forces Q_n to decrease. As a result, the channel becomes pinched off at the drain end and the current saturates. The saturation condition is given approximately by

$$V_{DS}(sat) = V_{GS} - V_T \tag{2.18}$$

and

$$I_D(sat) = \frac{\mu_n Z C_{ox}}{2L}\left[(V_{GS} - V_T)^2\right] \quad . \tag{2.19}$$

The drain current at saturation remains essentially constant for larger values of drain voltage.

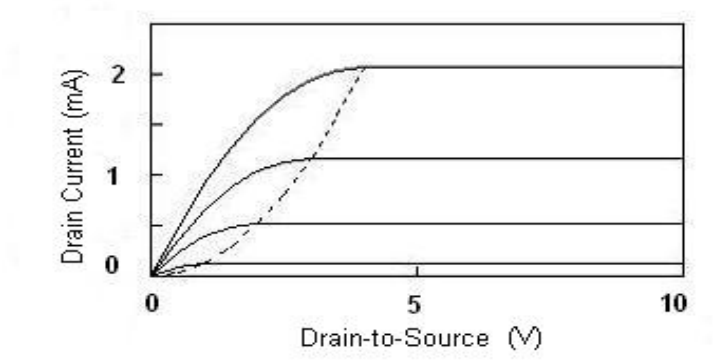

Figure 2.13 Current-Voltage Characteristics of a MOSFET

Figure 2.13 depicts the typical I-V characteristics of a MOSFET [14]. The dotted line separates the two main regions of operation, which are the linear and saturation regions.

2.2.3 Importance of MOSFET

One of the major advantages of this type of FET in semiconductor industry is that it amplifies electrical signals. The current gain is made possible owing to the fact that no gate current is required to maintain inversion layer. The device therefore has an infinite current gain in DC. The voltage gain of the MOSFET is caused by the fact that the current saturates at higher drain-source voltages, so that a small drain current variation can cause a large drain voltage variation. In addition MOSFET devices exhibit excellent switching characteristics owing to different region of operations. MOSFET can also be

used as a resistor. Resistor characteristics are obtained by permanently biasing the gate terminal for conduction. The ratio of the source-drain voltage to the channel current then determines the value of resistance. A particularly useful device for digital applications is a combination of n-channel and p-channel MOS transistors on adjacent regions of the chip. Such a semiconductor device is termed as Complementary MOS (or CMOS). The basic circuit is an inverter. The beauty of this circuit is that one of the devices is turned in either condition (input 1 or input 0). As a result CMOS structures greatly reduces power consumption. Hence it has found wide spread acceptance in the design of high density complex digital circuits.

2.3 Non Volatile Memory (NVM)

Nonvolatile Memory is a device into which the data can be programmed electrically and which will retain data even in the absence of a power supply. The basic operating principle of the nonvolatile memory devices is the storage of charge in the gate insulator of a MOSFET. This has led to the subdivision of devices into two groups namely the MIOS (Metal Insulator Oxide Semiconductor) and the FGMOS (Floating Gate Metal Oxide Semiconductor). This research work concentrates on FGMOS which is most commonly used as compared to MIOS in flash memories.

The FGMOSFET is the basic building block of the flash memory cell [9, 15]. In this device the charges are injected from the silicon across the first insulator and stored in the insulator-oxide interface of the floating gate of the FGMOS. Fowler Nordheim tunneling techniques can be used to write as well as erase data.

2.3.1 Structure of a Typical Floating Gate Device

The schematic cross section of an FG device is presented in Figure 2.14. The top electrode, also known as the control gate (word line), has a provision for direct electrical access. Below the control gate is the floating gate, which is capacitively coupled to the control gate and the underlying silicon. Generally, the floating gate is separated from the

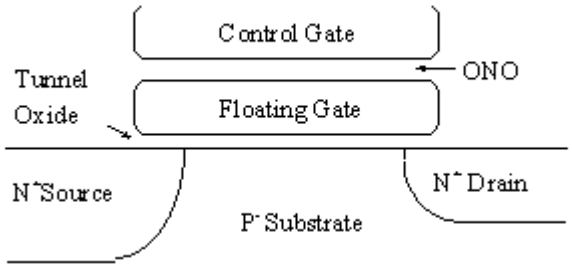

Figure 2.14 Basic Structure of a Typical FG Device

control gate by a stacked Oxide-Nitride-Oxide(ONO) dielectric. Electrical access to the floating gate is only through capacitors.

2.3.2 Operation Conditions

Table 1 summarizes the bias conditions, which identify the three operational modes of an Industry Standard cell. The Program (Write) and Erase and Read operation modes are discussed in the following chapters.

Table 2.1 Source, Drain and Control Gate Biases for Operation of a Typical Flash Call

	SOURCE	DRAIN	CONTROL GATE
PROGRAM	GND	V_{dd}	V_{pp}
ERASE	V_{cc}	Float	GND
READ	GND	V_{read}	V_{cc}

2.3.2.1 Write Operation

Flash and EPROM implement hot electron injection to place charge on the floating gate during a WRITE. Figure 2.15 shows the cell bias condition during Program operation During a WRITE, a high programming voltage (V_{pp}) is placed on the control gate. This

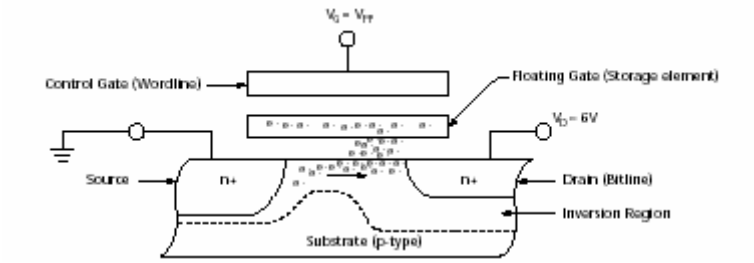

Figure 2.15 Flash Cell During a Write Operation [16]

forces an inversion region to form in the p-type substrate. The drain voltage is increased to approximately half the control gate voltage while the source is grounded (0 volts), increasing the voltage drop between the drain and source as shown in Figure 2.15. With the inversion region formed, the current between drain and source increases. The resulting high electron flow from source to drain increases the kinetic energy of the electrons. This causes the electrons to gain enough energy to overcome the oxide barrier and collect on the floating gate. After the WRITE is completed, the negative charge on the floating gate raises the cell's threshold voltage (V_T) above the wordline logic 1 voltage. When a written cell's wordline is brought to logic 1 during a READ, the cell will not turn on. The sense amps detect and amplify the cell current and output a "0" for a written cell.

2.3.2.2 Erase Operation

Flash employs Fowler-Nordheim tunneling to remove charge from the floating gate to bring it to the erased state. Using high-voltage source erase, the source is brought to a high voltage (VPP), the control gate grounded (0 volts) and the drain left unconnected as shown in Figure 2.16. The large positive voltage on the source, as compared to the floating gate, attracts the negatively charged electrons from the floating gate to the source through the thin oxide. Because the drain is not connected, the ERASE function is a much lower current-per-cell operation than a WRITE that uses hot electron injection. After the ERASE is completed, the lack of charge on the floating gate lowers the cell's VT below the wordline logic 1 voltage. When an erased cell's wordline is brought to a logic 1 during a READ, the transistor will turn on and conduct more current than a written cell.

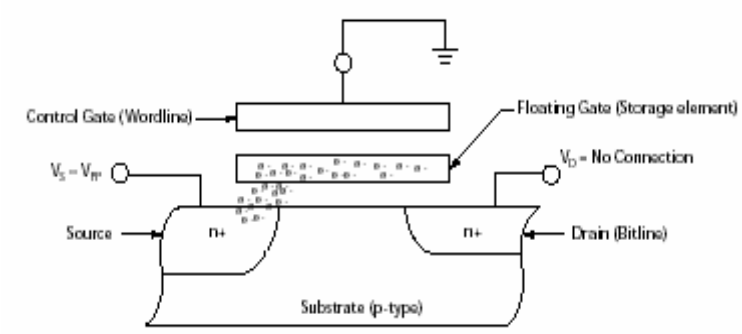

Figure 2.16 Erase Operation Employing Fowler Nordheim Tunneling [16]

Some flash devices use Fowler-Nordheim tunneling for WRITEs as well as ERASEs.

2.4 MOS Device Scaling

Early transistors were crude by today's standards, having channel lengths of greater then 10um, gate oxide thickness greater then 100 *nm* and operating at 12 V. As a result integrated circuits in those times had around 10000 MOS transistors only and operated at frequencies less then 1 MHz. Today, however, MOS transistors have channel lengths less then 0.1um, gate oxide thickness less then 10 nm and operate at 3.3V. Integrated circuits today use more then 3 million transistors and the operating frequency is more then 100 MHz. Table 1 shows the past trends in transistor scaling [17]:

Table 2.2 Historical Trends in Scaling

	1977	1980	1983	1986	1989	1992
Gate Length (um)	3	2	1.5	1.1	0.7	0.6
Channel length (um)	2	1.5	1.2	0.9	0.6	0.4
Gate Oxide (nm)	70	40	25	25	20	15
Junction Depth (um)	0.6	0.4	0.3	0.25	0.20	0.15
Gate Delay (ps)	800	350	250	200	160	90

MOS technology dimensions are being aggressively scaled to such a plateau where device performance must be accessed against fundamental limits. The exponential growth in the last three decades has been largely driven by technological innovations and have enabled a steady reduction in MOSFET dimensions. However, moving further, scaling will not be as rampant as it has been in the past. This is owing to the fact that when the dimensions of a MOSFET are scaled down, both the voltage level and the gate-oxide thickness must also be reduced. Figure 2.17 shows the scaling trend of power-supply voltage (V_{dd}), threshold voltage (V_T), and gate-oxide thickness (t_{ox}) as a function of CMOS channel length [2]. As seen that the power-supply voltage has not been decreasing at a

rate proportional to the channel length. This means that the field has been gradually rising over the generations between 1-µm and 0.1-µm channel lengths. The effect of this increase in electric field is not felt because thinner oxides are still reliable at such high electric field, thus allowing operation at the reduced but nonscaled supply voltages

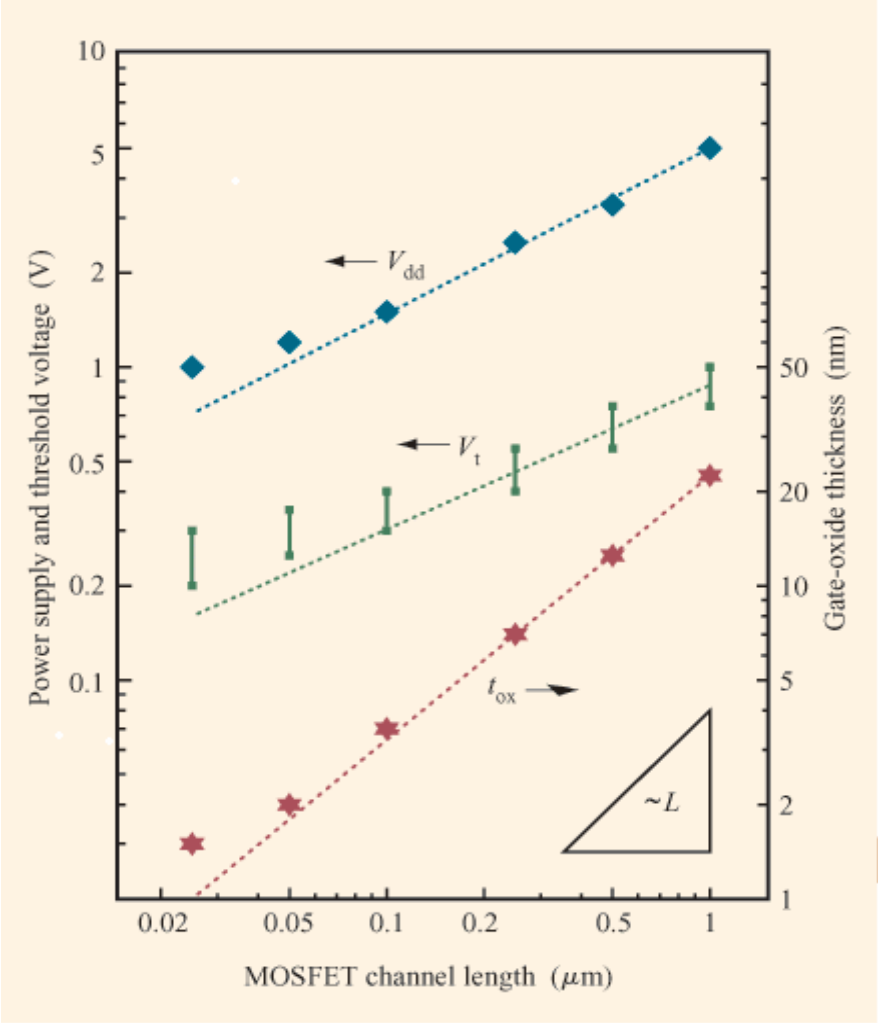

Figure 2.17 Past Trends in Scaling [2]

Therefore it can be said that a point will be reached in the near future at which further gains from scaling of traditional planar CMOS devices will be very difficult, limited by

leakage due to direct tunneling and switching power considerations. Further expansion in the specialization of device structures and design points will proceed over the next few generations of CMOS, with increased emphasis on new materials and structures.

As of now the Tunnel oxide in NVM has been scaled down to 8 *nm*.

Chapter 3

Characterization of Thin Gate Oxide

3.1 Introduction

The microelectronics industry owes a great deal of its success to the existence of the thermal oxide of silicon, i.e., silicon dioxide (SiO_2). A thin layer of SiO_2 forms the insulating layer between the control gate and the conducting channel of the transistors used in most modern integrated circuits. Although no insulator is ideal, amorphous SiO_2, with a bulk resistivity of ~10^{15} Ω-cm and dielectric breakdown strength of ~10^7 V/cm, is one of the best nature has provided [3]. Its usefulness in silicon technology is due to the fact that films of this material can be grown by thermal oxidation of clean silicon surfaces, and that the resulting films can be made with sufficiently low densities of charge traps. In addition, the energy barrier (conduction-band discontinuity) between Si and SiO_2 is 3.1 eV, only about 1 eV lower than the Si/vacuum barrier. Silicon dioxide uniquely possesses the required combination of several properties: good mobility of holes and electrons flowing in silicon at the silicon dioxide interface, ability to keep electronic states (surface states) at this interface low and excellent compatibility with CMOS processing.

Continuous device scaling requires the continued reduction of the gate dielectric thickness. This needs to be done to keep short-channel effects under control. Therefore

gate oxide thickness is reduced almost in the same proportion as the channel length. This is necessary in order for the gate to retain more control over the channel than the drain. For CMOS devices with channel lengths of 100 nm or less, an oxide thickness less then 3 *nm* is needed [18]. This thickness comprises only a few layers of atoms. Oxide film this small no longer acts as a perfect barrier. A leakage current, dependent on the applied voltage and the corresponding electric field, exists. This phenomenon can be attributed to the Quantum-mechanical tunneling [19]. As the silicon dioxide scaling reaches its fundamental limits, direct tunneling effects are more felt now then ever before.

3.2 Current Conduction Mechanism

The lifetime of a particular gate oxide thickness is determined by the total amount of charge that flows through the gate oxide by leakage current. Ideally, an oxide doesn't allow charge to pass through. However this is not the case when the electric fields become high and the oxide becomes thin. There are several mechanisms that allow charge to pass through the oxide. For oxide layers thicker than 6 nm, the leakage current mechanism is explained by Fowler-Nordheim electron tunneling in MOS structures [20]. On the other hand, the tunneling in the ultra-thin oxide layers (2-5nm), as already examined by Maserjian in 1974 [21], has been termed as direct tunneling.

3.2.1 Thermionic Emission

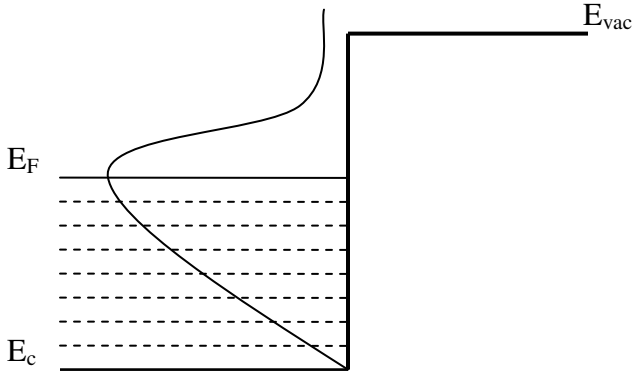

Figure 3.1 Thermionic Emission

In a conducting material, the electrons are governed by Fermi-Dirac statistics. The baseline electron energy is the Fermi Energy, and at low temperatures the electrons all exist at or below this level. As temperature is increased, so the distribution function for the electrons develops a high energy 'tail'. Some of these electrons have sufficient energy to pass over the surface potential barrier between the material and the vacuum. This process of increasing the temperature of a bulk material to increase the number of electrons which can leave the material is called thermionic emission. The thermionic emission has also been observed in metal-oxide structures.

3.2.2 Fowler-Nordheim Tunneling

Fowler Nordheim tunneling is a quantum mechanical tunneling process where the electrons can penetrate through the oxide barrier into the conduction band of the oxide (Figure 3.2). Fowler-Nordheim tunneling is dependent on the voltage across the gate

oxide and increases exponentially with the applied voltage. It can occur in most any gate oxide, provided the voltage is sufficient for electrons to tunnel through the barrier. Figure 3.3 shows the F-N tunneling mechanism for a p+poly-Si/ SiO_2/p-Si structure.

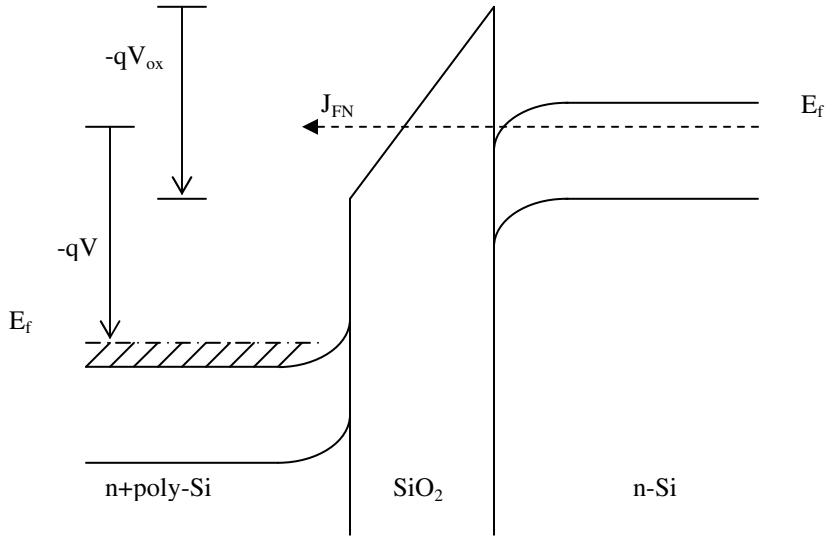

Figure 3.2 Energy Band Diagram for an n+poly-Si/SiO_2/n-Si Structure in Case of Fowler Nordheim Tunneling

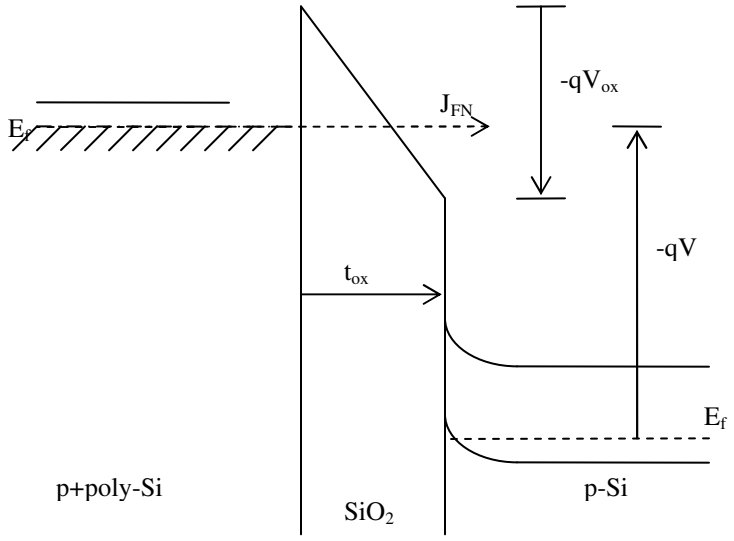

Figure 3.3 Energy Band Diagram for a p+poly-Si/SiO_2/p-Si Structure in Case of Fowler Nordheim Tunneling

3.2.3 Direct Tunneling

Direct tunneling is also a quantum mechanical tunneling process as shown in Figure 3.4. Direct tunneling is a phenomenon that is important to understand in ultrathin oxides. It occurs when electrons tunnel through the gate oxide region directly from the gate to the channel region and vice versa. Direct tunneling is dependent on the thickness of the gate region; it increases exponentially as the thickness of the oxide decreases. Direct tunneling is relatively independent of the electric field across the gate oxide.

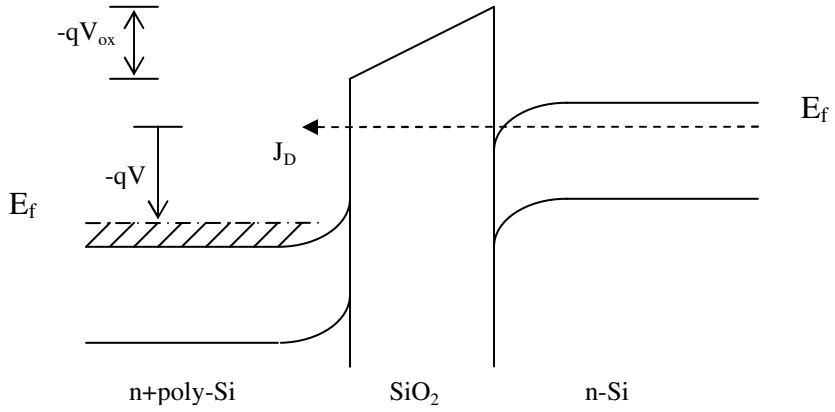

Figure 3.4 Energy Band Diagram for an n+poly-Si/SiO_2/n-Si Structure in Case of Direct Tunneling

The difference between Fowler-Nordheim tunneling (FNT) and direct tunneling (DT) of electrons is determined by the shape of the tunnel barrier. If the oxide tunnel barrier is triangular, FNT occurs while DT takes place through trapezoidal barrier. The transition from FNT to DT occurs if the difference in potential energy over the oxide layer qVox becomes smaller than the barrier height at the injecting interface.

3.2.4 Analytical Expressions for the Tunnel Currents

For a simplified metal-semiconductor or heterojunction or a band to band transition can be approximated by a triangular potential barrier. For a triangular potential barrier as shown Figure 3.5 the tunneling probability T_t can be given by WKB approximation (Wentzel-Kramer-Brillouim method) as shown below:

$$T_t = \exp\left[-2\int_{x_1}^{x_2} |k(x)|dx\right] \tag{3.1}$$

where $|k(x)|$ is the absolute value of the wave vector of the carrier in the barrier, and x_1 and x_2 are the classical turning points as shown in the figure below:

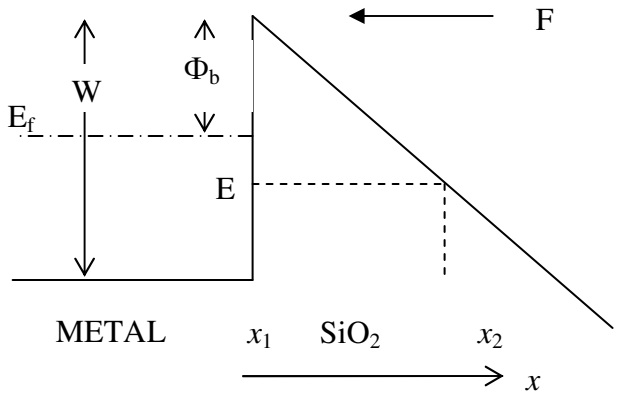

Figure 3.5 Triangular Potential Barrier

The value of k can be found from the conservation of energy and momentum which is justified only in the tunneling process and is given by

$$\frac{h^2k^2}{2m} = [U - E] \tag{3.2}$$

where E is the electron energy and U the potential energy and is given by

$$U = W - qFX \qquad \text{for } x > 0 \tag{3.3}$$

where F is the electric field. Replacing this value of U in equation 3.2 gives the value for k. By substituting this value of k in equation 3.1 tunneling probability can be rewritten as

$$T_t = \exp\left[-\frac{4}{3}\left(2m/h^2\right)^{1/2} \frac{(W-E)^{3/2}}{F} \right] \tag{3.4}$$

The tunneling current density J_{FN} in a three dimensional lattice from the n type Si accumulation layer to poly-Si gate or from the gate electrode to the accumulation layer of p-Si can be obtained from the equation below

$$J = q\int\int\int \frac{2}{h^3} T_t V_x dP_x dP_y dP_z \tag{3.5}$$

where V_x is the velocity due to electrons in the x direction.

Replacing the tunneling probability T_t as described in equation 3.4 in 3.5 it can be shown that

$$J_{FN} = AF^2 \exp\left[-\frac{B}{F} \right] \tag{3.6}$$

where,

$A = \dfrac{q^3}{16\pi^2 h \phi_b}$ and $B = \dfrac{4}{3}\dfrac{(2m_{ox})^{1/2}}{qh}\phi_b^{3/2}$

These are two basic parameters for the gate oxide tunnel current calculation and the parameters A and B depend on the tunnel barrier height Φ_b and the effective mass of the tunneling electron m_{ox} and can be derived from a F-N plot of 1n(J_{FN} / F*F) vs. 1/F. The slope of this straight line gives B while A is determined from the intercept.

It is clear from these relationships that operating a CMOS device at voltages greater than foundry specification results in an exponential increase in the amount of oxide current. Once electrons have breached the oxide potential barrier they are accelerated through the oxide by the electric field which is determined by the applied voltage and the oxide thickness. Charge accelerated in the gate oxide achieves greatest energy at the oxide-silicon interface presuming there have been no collisions in transit. At the end of its travel through the oxide, it deposits its energy at the oxide-silicon interface.

Making these equations as the basis of the analytical part of this research work, results are presented in the following chapter.

In the case of direct tunneling the current density, where the potential barrier takes a rectangular shape, is given by [22]

$$J_D = \frac{AF^2}{\left[1-\left(\frac{\phi_s - qV_{ox}}{\phi_s}\right)^{1/2}\right]^2} \exp\left[-\frac{B}{F}\frac{\phi_s^{3/2} - (\phi_s - qV_{ox})^{3/2}}{\phi_s^{3/2}}\right] * \left\{1-\exp\left[-\frac{3}{2}\frac{\phi_s^{1/2} - (\phi_s - V_{ox})^{1/2}}{\phi_s^{3/2}} E_f\right]\right\} \tag{3.7}$$

3.3 Dielectric Breakdown

The reliability of SiO_2 in microelectronics, i.e., the ability of a thin film of this material to retain its insulating properties while subjected to high electric fields for many years, has always been a concern and has been the subject of numerous publications over the last 35–40 years, ever since the realization that SiO_2 could be used as an insulating and passivating layer in silicon-based transistors [23, 24]. The gate leakage current causes

increased power consumption and may affect device and circuit functionality. Thus, the leakage current imposes a practical limit on oxide thickness. The current flowing through an ultrathin gate oxide is not merely a nuisance parasitic, but also causes reliability problems by leading to long-term parameter shifts (wear out) and eventually to oxide breakdown

3.3.1 Time Dependent Dielectric Breakdown (TDDB)

The time dependent nature of the breakdown, or TDDB, is one of the most heavily researched failure mechanisms in the semiconductor reliability community. Since the 1960s, researchers have struggled to understand the nature of how an oxide degrades over time [25]. By definition, TDDB is wear-out of the insulating properties of silicon dioxide in the CMOS gate leading to the formation of a conducting path through the oxide to the substrate. With a conducting path between the gate and the substrate, it is no longer possible to control current flow between the drain and source by means of the gate electric field. TDDB lifetime is strongly affected by the number of defects in the gate oxide produced during wafer fabrication. Therefore, foundries strive to produce an ultra-clean oxide in their process to maximize the TDDB lifetime. Even if a foundry could produce a perfectly defect free oxide, TDDB would remain a concern for ASIC (Application Specific Integrated Circuit) designers. TDDB occurs at all gate voltage bias conditions. The goal of the foundry is to trade off gate oxide thickness with operating voltage specifications to achieve both speed and lifetime targets for the technology. The

lifetime of a particular gate oxide thickness is determined by the total amount of charge that flows through the gate oxide by leakage current.

3.3.2 Dielectric Breakdown Models

The essential elements of our present understanding of oxide wear out and breakdown are illustrated in Figure 3.6. The figure represents outline of the mechanism of defect generation leading to breakdown in SiO_2. The gate voltage causes energetic electrons to flow across the oxide, leading to a buildup of damage in the form of microscopic defects which act as electron traps. When sufficient defect density is reached, the oxide no longer acts as a good insulator.

Electrons flowing across the oxide trigger several processes depending on their energy, which is determined by the gate voltage for thin oxides. At least three defect-generation mechanisms have been identified: impact ionization, anode hole injection and trap creation process attributed to hydrogen release from the anode. These models form the basis of the different analysis and tests Reliabilty engineers perform to check device performance.

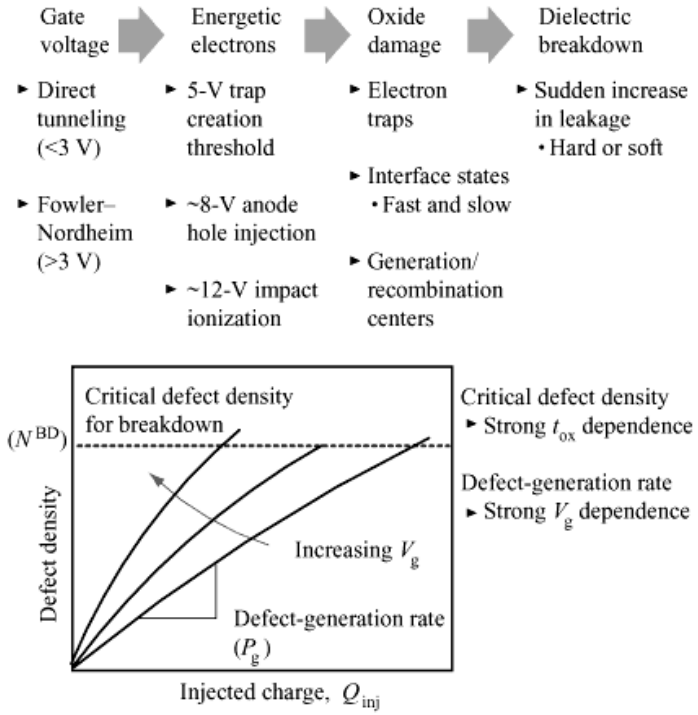

Figure 3.6 Outline of the Mechanism of Defect Generation Leading to Breakdown of SiO_2 [26]

3.3.2.1 Bandgap Ionization Models

DiMaria (IBM Microelectronics) developed the Bandgap Ionization in the early 1990s [27]. He postulated that when the energy of an electron approaches the energy of a bandgap (~ 9eV), electron-hole pairs can be generated that cause defects in the oxide. This process occurs in dielectrics thicker than ~ 20nm at fields higher than ~ 7 MV/cm, and is associated with the high-energy tails of the electron distribution. Although this model was popular in the early to mid 1990s, it is not in wide use today. Soon afterward, researchers began reporting data on ultra thin oxides that could not be reconciled with the

Bandgap Ionization model. While the model is valid for thicker oxides, it is not as valid for ultra thin oxides.

3.3.2.2 Anode Hole Injection Model

Schuegraf and Hu developed the Classic Anode Hole Injection Model in the mid 1990s [28]. The Classic Anode Hole Injection Model, commonly called the 1/E model, was the source of some controversy in the late 1990s as researchers sided either it or with the Thermo chemical (E) model. Schuegraf and Hu postulated that a fraction of the electrons entering the anode have enough energy to create a "hot" hole which can tunnel back into the oxide. These holes then in turn create defects in the oxide. Their equation modeling the process predicts a 1/E dependency, hence the name 1/E (see equation below).

Some researchers have suggested based on both theoretical and experimental evidence that significant injection and trapping of anode holes does not occur until the electron obtains a sufficiently high energy (~7.6 eV) in the anode. Recent studies using improved models for impact ionization suggest that anode hole injection can occur at low energies and that the dependence of anode hole induced breakdown on voltage is linear. However, there are no known studies conclusively showing that anode injected holes at low voltage are trapped in the oxide and lead to breakdown.

3.3.2.3 Hydrogen Release Model

Di Maria and his colleagues developed The Hydrogen Release Model after physically observing the behavior and properties of the oxides during and after stress [29]. DiMaria (and others) observed hydrogen release and buildup after stressing various oxide

structures. He postulated that energetic electrons create oxide damage by interacting with the oxide lattice itself or with a secondary species such as hydrogen. Based on data taken at IBM he determined that hydrogen release requires electrons with energy levels of at least 5 eV in the anode, and 2 eV in the oxide.

3.4 Reliability

Reliability engineers also study the statistics associated with the oxide breakdown. There have been many models proposed but the most commonly models referred to for analysis are Percolation or Spheres model of breakdown [30] and Weibull distribution [31]. A detailed explanation of these models and reliability projections is beyond the scope of this research work.

Breakdown is generally understood to be the result of a gradual and predictable buildup of defects such as electron traps in the oxide, but the precise point at which breakdown occurs is statistically distributed so that only statistical averages can be predicted. The rate of defect generation in the oxide is proportional to the current density flowing through it, and therefore the reliability margin for gate-oxide breakdown has been drastically reduced as a consequence of device scaling. At present, predicted reliability "limits" for the gate-oxide thickness range from less than 1.5 nm to 2.8 nm. Dielectric breakdown occurs by two mechanisms: extrinsic or intrinsic failure. Extrinsic failures occur with a decreasing failure rate over time and are caused by process defects such as metallic, organic, or other contaminants on the crystalline silicon surface, and by surface roughness. Most extrinsic failures occur early in the lifetime of a device and can cause

significant reliability problems as the number of dielectric failures exceed a predefined specification. Failure rates generated extrinsically can be reduced by applying a burn-in or voltage screening process to the devices. The failures are accelerated at elevated voltages and temperatures. These procedures screen out dielectric defects before the product is shipped, thereby containing early life failure rates within product specifications. Minimizing extrinsic defects in a process and reducing dielectric fails early in the life cycle of a product is important for processing ultrathin gate dielectrics. It is also critical that reliability testing detect systematic and/or random types of process-induced defects that may contribute to an increase in number of dielectric failures. In full-scale product manufacturing, in-line monitoring of both yield and reliability ensures the tight process control of defects to minimize the device and/or chip failures. The second type of failure for a thin gate dielectric, sometimes referred to as wear-out, is intrinsic to the material. Intrinsic failures occur with an increasing failure rate over time and are usually caused by an inherent imperfection in the dielectric material. It is essential that these fails do not occur during the intended useful lifetime of the device when it is operating under specified conditions. The intrinsic failure of ultrathin gate dielectric has been the subject of intensive research. The process of intrinsic degradation in gate dielectric begins with trap creation and formation of interface states as the device gate dielectric is stressed at elevated voltages and temperatures. The creation of these defects continues with the injected current (and therefore time) until the defect density reaches a critical value, after which dielectric breakdown occurs. Parameters that affect when and

how the dielectric breakdown occurs include the applied stress voltage, temperature, dielectric thickness, device dielectric area, and intrinsic dielectric lifetime.

3.4.1 Reliabilty Projections

Figure 3.7 shows various oxide-reliability projections from different research groups . Plotted in this figure is the maximum allowable voltage that can be applied to the total gate area on a chip, such that no more than a specified failure rate will result. The failure rate in this case is defined as the fraction of chips which will experience one or more oxide breakdown events. As can be ascertained from this figure, there remains considerable disagreement among laboratories regarding this reliability limit for SiO_2. The various projections in this figure are compared to the latest international roadmap for the operating voltage and physical oxide thickness. Note that various laboratories may use different methods to determine the oxide thickness, which can produce results varying by as much as 0.4 nm. As a rule of thumb, a factor of 10 change in any reliability criterion (e.g., lifetime, failure rate, device area) results in ~0.1 nm change in the minimum t_{ox} for a given supply voltage.

It is clear that there is no agreement at the present time regarding the correct oxide-reliability extrapolation. While all of the estimates for the minimum t_{ox} are rather closely clustered in the 1.5–2.5-nm range, this is precisely the range of interest for the industry roadmap, and it represents a large variation in device performance and circuit density. Thus, the situation is complicated by the fact that neither a worst-case or a best-case outlook can be adopted safely. The worst-case scenario appears to be unacceptable from a

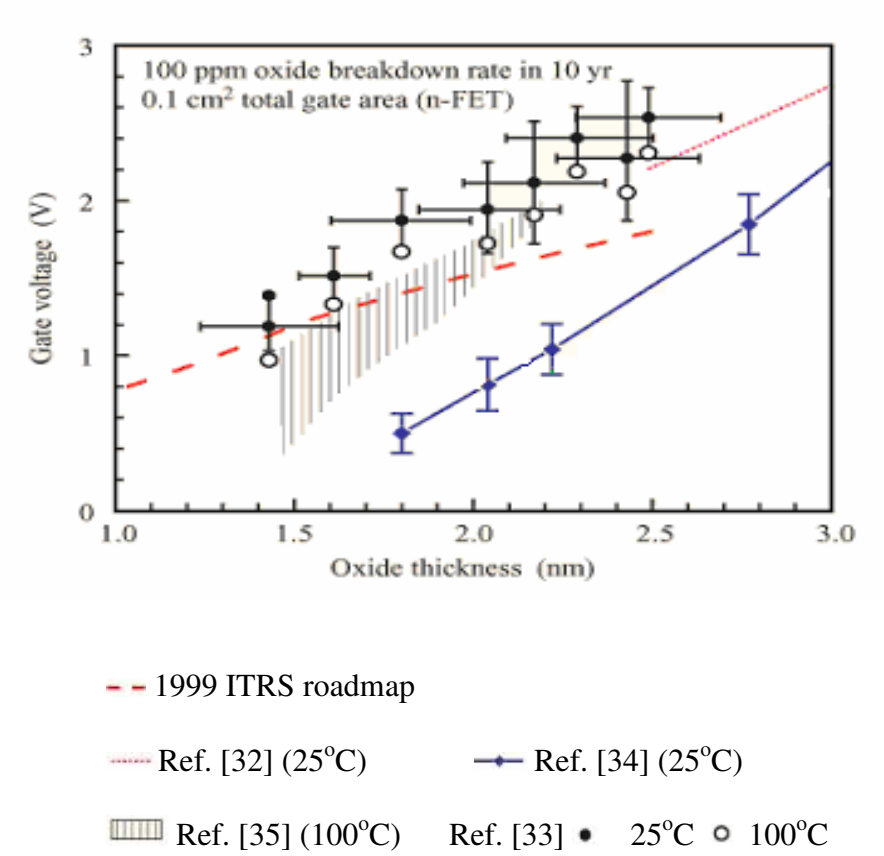

Figure 3.7 Oxide Reliability Projections by Different Labs

technological and economic perspective, and the best-case scenario is risky because even in this case there are no margin for error. Thus, it appears that the traditional method of assessing oxide reliability, by observing the breakdown time during an accelerated stress experiment and extrapolating this to operation conditions using physical or empirical models, may no longer be adequate to convincingly ensure the long-term reliability of products based on CMOS circuits. It has therefore become necessary to look in more detail at the nature of the breakdown event and the behavior of devices and circuits after oxide failure.

Chapter 4

Determination and Extraction of Parameter B

4.1 Introduction

The Fowler Nordheim tunneling-injection mechanism is widely used in NVM, particularly in EEPROM. There are three main reasons for this choice: first, tunneling is a pure electrical mechanism; second, the involved current level is quite low and thus allows the internal generation of supply voltages needed for all operations; third, it allows one to obtain the time to program (1 ms) 12 orders of magnitude shorter than retention time (10 y), which is a fundamental request for all NVM technologies.

On the other hand, the exponential dependence of tunnel current on the oxide-electric field causes some critical problems of process control because, for example, a very small variation of oxide thickness among the cells in a memory array produces a great difference in programming or erasing currents, thus spreading the threshold voltage distribution in both logical states [9]. A very good process control is therefore required. Moreover, the tunneling currents may become important in device reliability at low fields either in the case of bad-quality tunnel oxides or when thin oxides are stressed many times at high voltages. In fact, bad quality oxides are rich of interface and bulk traps, and trap-assisted tunneling is made possible since the equivalent barrier height seen by electrons is reduced and tunneling requires a much lower oxide field than 10 MV/cm.

Thus in-depth analysis is required to achieve the required threshold voltage window so that no errors creep in. If this is not the case then a desired WRITE program may result in ERASE and vice versa.

4.2 Importance of Parameter B

The parameter B plays an important role in the operation of writing and erasing in semiconductor nonvolatile MOS memory devices such as flash memory and the gate current in MOSFETS. The electron effective mass in the dielectrics and the potential barrier height which control the F-N tunneling current in these devices can be found from the extracted parameter B. Therefore, such studies are necessary in order to obtain a better understanding of the effect of novel gate materials on the F-N tunneling current, the dynamics of the writing and erasing in the memory devices, and the leakage current in the gate for MOSFETS. Consequently improvement in the programming speed of memory devices and the reliability of the gate oxide can be achieved [36].

However up to the present time, there are no generally accepted values for parameter B. If one assumes B = 260 MV/cm and ε_0 = 10 MV/cm, a -5% uncertainty of B will result in 370% uncertainty in the current density. This research work is an effort in the direction of finding a value of B which is as accurate as possible.

4.3 Sample Fabrication

The fabrication of the oxide of a MOSFET is one of the critical steps when fabricating MOSFETs. This is due to the need for an ideal oxide-semiconductor interface with low

surface-state density and also because of the extremely thin oxides that are currently used for sub-micron MOSFETs. Two techniques are commonly used to form silicon dioxide; one involves the oxidation of the silicon yielding a thermal oxide and the other relies on the deposition of SiO_2 using a chemical vapor-deposition process. Former technique has been used to fabricate wafers used in this research. Gate oxides were thermally grown in a dry O_2 ambient at a temperature in the range of 900-950 ^oC. Aluminum was deposited on the front and the back of the wafer. Resistive evaporation and wet etching of the aluminum were used to minimize damage to the gate oxide. The resistivity of wafer was 1 ohm/cm. There was no post-metallization annealing in order to minimize reactions between the aluminum and the gate oxides. After they were cleaned using standard cleaning process and HF dipped, thermal oxide of sizes in the range of 7 *nm* ~ 13 *nm* were grown.

4.4 Measurement Setup

Measuring leakage currents on semiconductor wafers and within devices has always been a challenge for analytical probing equipment, as well as the supporting instrumentation. With market demands for higher density ICs, lower power, improved reliability, and features like programmability, semiconductor makers engineer processes to push leakages and device "off currents" ever lower. When these currents are less than a picoamp, direct measurement can be very difficult. The set up used during the course of this research has provided some really good data which were later used for analysis

purpose. A brief overview of the set up has been provided. Details have been presented in the appendix A and B.

The equipments used in this setup are: personal computer, Keithley 617 Electrometer [37], Keithley 230 Voltmeter[38], Boonton Capacitance meter, Micromanipulator 7000 series probe station. The functions of all the devices and equipments that have been used in the setup are discussed below.

The setup requires an IBM compatible computer with LabVIEW software installed on it and IEEE-488 interface bus that carries data between computer and other equipment in the setup. LabVIEW is a program development application that uses a graphical programming language to create program in block diagram form. LabVIEW like any other programming language has an extensive library of functions for any programming task. It includes libraries for data acquisition, serial instrument control, data analysis, data presentation and data storage. A program had been written in labVIEW to control the flow of data in the setup. A virtual model of Kiethley 617 electrometer and 230 voltmeter, with all the functionalities, had been developed to automate the whole process. Details have been provided in appendix B.

4.4.1 Overview of Equipment

The Keithley model 230 is a programmable voltage source with full range voltage from 199.95 mV to 101 V. The three selectable current limits allow the user to tailor the current limit from 2mA to 20mA to 100mA. The 100 memory locations allow up to 100 storage locations for programming source, I-limit and dwell time.

The Keithley Model 617 Programmable Electrometer is a highly sensitive instrument designed to measure voltage, current, charge, and resistance. The measuring range of the Model 617 is between 100mV and 200V for voltage measurements, 0.lfA and 20mA in the current mode, l0fC and 200C in the coulombs mode and 0.1 ohms to 200 Mega ohms in the resistance mode. The very high input impedance and extremely low input offset current allow accurate measurement in situations where many other instruments would have detrimental effects on the circuit being measured. A 4% digit display and standard IEEE-488 interface give the user easy access to instrument data.

Micromanipulator Probe station is an ultra stable station with coarse and fine wafer stage movement to provide faster wafer movement as well as submicron resolution. It is designed for high density, small geometry probing. Our Probe station is a 7000 series analytical probing station with X-Y stage movement by fine manual and two-speed motor controls. It consists of a high magnification microscope, with long working distance objectives, a refined microscope translation, with one inch travel in the X and Y directions, a pneumatic platen lift back that allows one inch lift, a motorized stage drive, with two speed operation, a pneumatic platen, to incorporate vacuum base manipulators and a stage position, that allows up to six inch travel. The noise level during the course of this research work had been reduced to 10^{-15}A.

4.4.2 Lab Setup for I-V Measurements

The setup used for the IV measurements during the course of this thesis has been shown in Figure 4.1. The kiethley 230 voltage source is fed voltage range with desired delay

using the LABVIEW software through GPIB interface. The voltage range could be negative or positive depending upon the device under test. This action triggers the voltage source meter. The voltage range, with desired delay, gets applied onto the MOS capacitor placed on the shielded probe station through a probe. Corresponding current values are displayed on the Kiethley 617 Electrometer and subsequently stored as an excel file on the computer. Generally the voltage range is selected such that the current ranges from picoamps to microamps. Device under test is assumed broken when current readings extend into microamps.

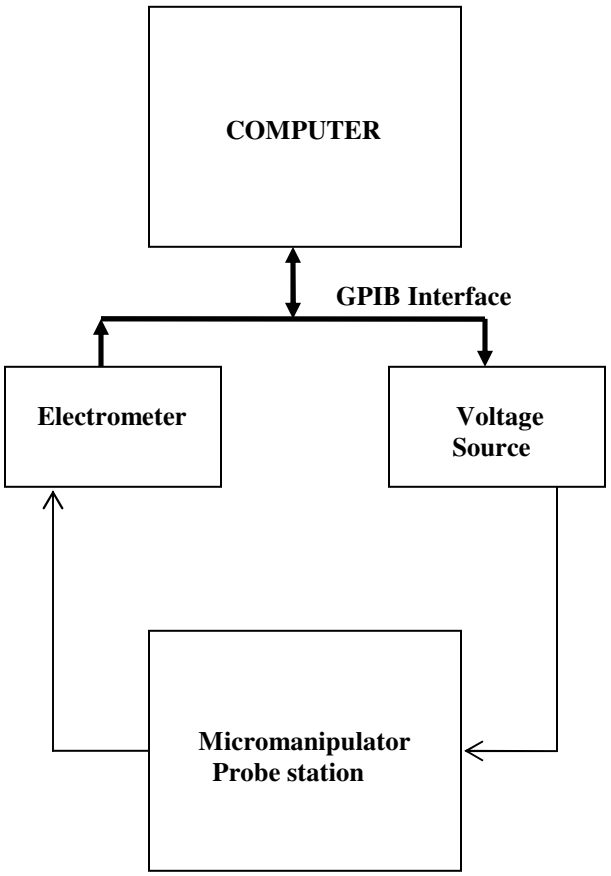

Figure 4.1 I-V Measurement Setup

4.4.3 Lab Setup for C-V Measurements

Figure 4.2 shows the setup for C-V measurements in the lab. As shown the Device Under test (DUT) is connected to the Capacitance meter. Now there is no interface between the capacitance meter and the computer. So to automate the flow the output from the meter is fed to Kiethley 617 Electrometer which in turn is connected to the computer as already shown in the setup for I-V measurements. The results are stored in the excel datasheet. Similarly the desired input voltage source is fed to the Capacitance meter through the kiethley Voltage Source. Thus the process of recording data for analysis has been made fully automated.

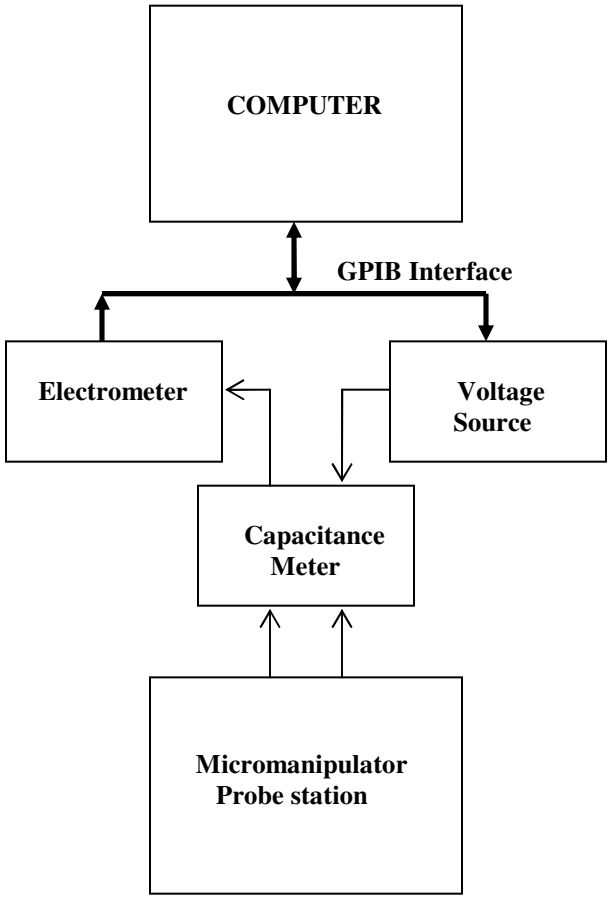

Figure 4.2 C-V Measurement Setup

4.5 Analysis

Before going into the analysis it is of prime importance to validate the wafers, otherwise any type of analysis would not yield the desired result. One of the simplest tests to see if wafer is useful for analysis is to check for current uniformity. This test was performed by measuring current on dies of different area on the wafer at constant voltage. As already discussed the cell dimensions are 10^{-4} cm^2, $4*10^{-4}$ cm^2 and $16*10^{-4}$ cm^2.

In this research two different voltages viz 5V and 8V were applied and current measured for the different areas at each voltage. This test was performed on both the p-type as well as n-type wafer. Figure 4.3 and Figure 4.4 gives a plot between the measured current in Amps versus the Area at a constant applied voltage of 5V. As seen, the curve is linear which implies that the current density is constant and doesn't vary with increase in area.

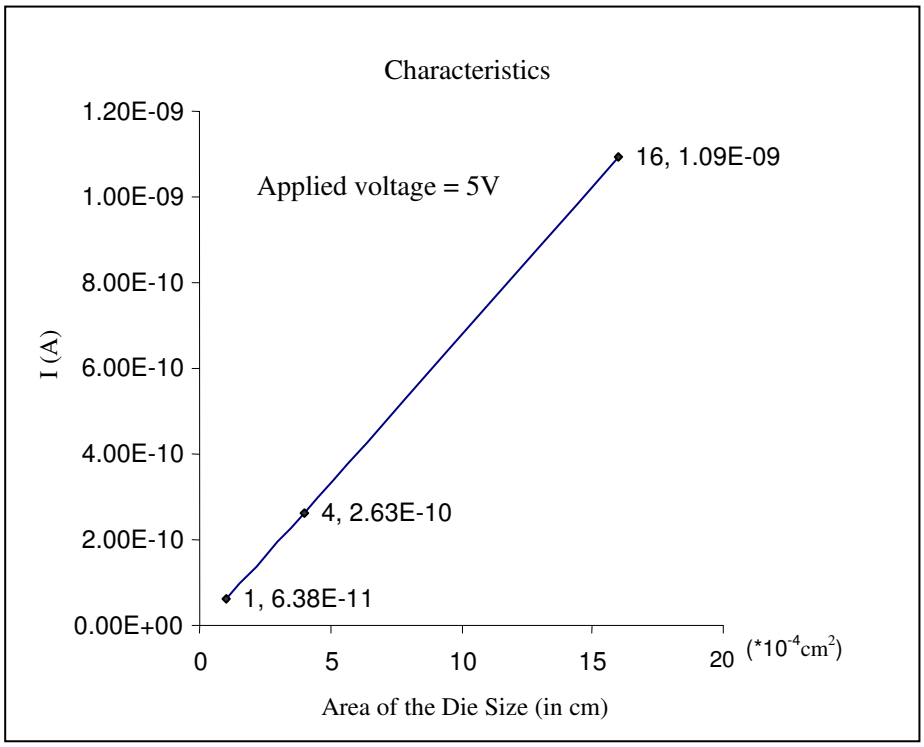

Figure 4.3 Plot of Current versus Area for a 7*nm* n-type Device

The plots between the measured current versus the area at constant input voltage of 8V are not presented. But current was found to be uniform. However the slopes were different as compared to the ones shown in Figures 4.3 and 4.4.This can be attributed to the fact that the current density depends on the electric field, which in turn changes according to the applied voltage.

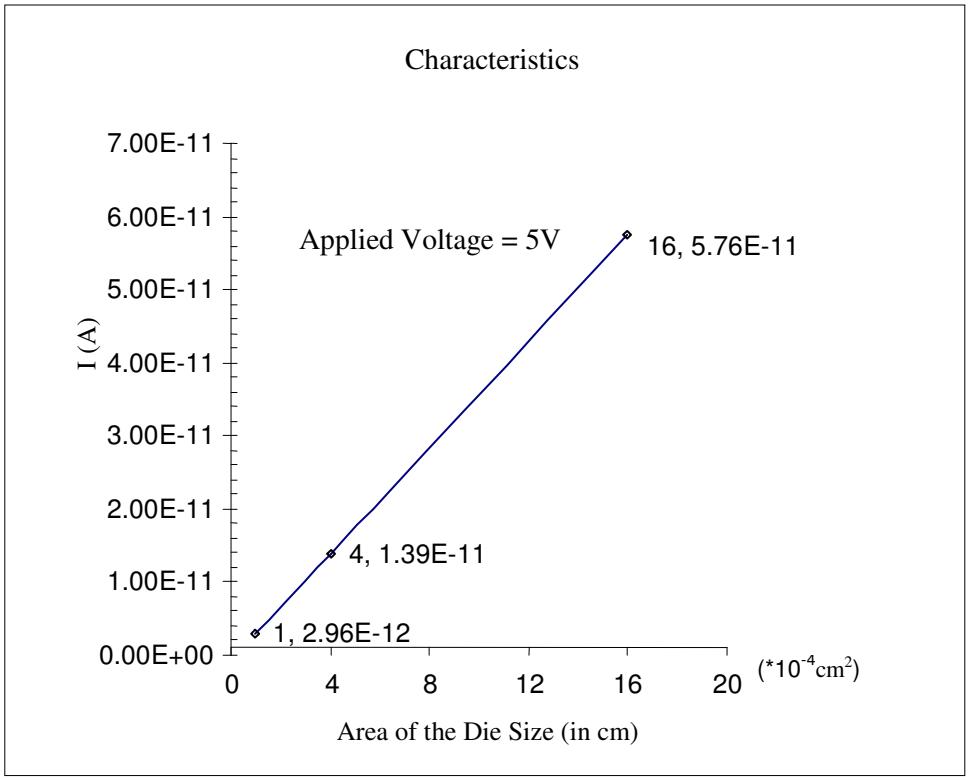

Figure 4.4 Plot of Current versus Area for a 7*nm* p-type Device

Based on the above conclusions, it was deducted that the wafers were good for further analysis and measurements.

4.5.1 Time Dependent Dielectric Breakdown (TDDB)

Dielectric breakdown refers to the destruction of a dielectric layer, usually as a result of excessive potential difference or voltage across it. It is usually manifested as a short or leakage at the point of breakdown.

Since SiO2 is a very common dielectric material, its breakdown mechanism has been understood over the years. SiO2 breakdown is believed to be due to charge injection, and may be broken down into 2 stages. During the first stage, current starts to flow through the oxide as a result of the voltage applied across it. High field/high current regions are then formed as charges are trapped in the oxide. Eventually, these abnormal regions reach stage 2, a critical point wherein the oxide heats up and allows a greater current flow. This results in an electrical and thermal runaway that quickly leads to the physical destruction of the oxide. In this research work, Time-Dependent-Dielectric Breakdown (TDDB) of *Al* gate with 7nm gate oxide thickness (for both n-type and p-type wafers) were measured and analyzed under constant voltage stress.

The time-to-breakdown for a dielectric is voltage dependent. Figure 4.5 shows the TDDB for a 7 *nm* thick gate oxide for an n-type device when subjected to constant voltages of 10.5v, 11.5v and 12.0v. When V=10.5v it takes approximately 20 minutes to breakdown where when V=11.5v and V=12.5v it takes approximately 14 and 6 minutes respectively. Figure 4.6 shows the TDDB for a 7 *nm* thick gate oxide for a p-type device when subjected to constant voltages of 10.5v, 11.5v and 12.0v. When V=10.5v it takes approximately 24 minutes to breakdown where when V=11.5v and V=12.5v it takes approximately 19 and 8 minutes respectively.

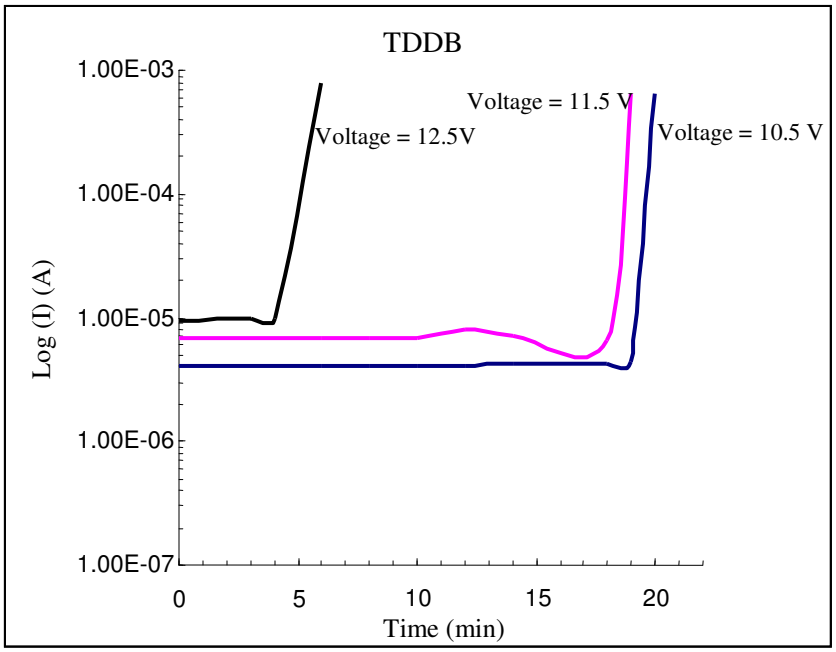

Figure 4.5 Time Dependent Dielectric Breakdown Behavior of a 7*nm* Oxide Thickness n-type Device

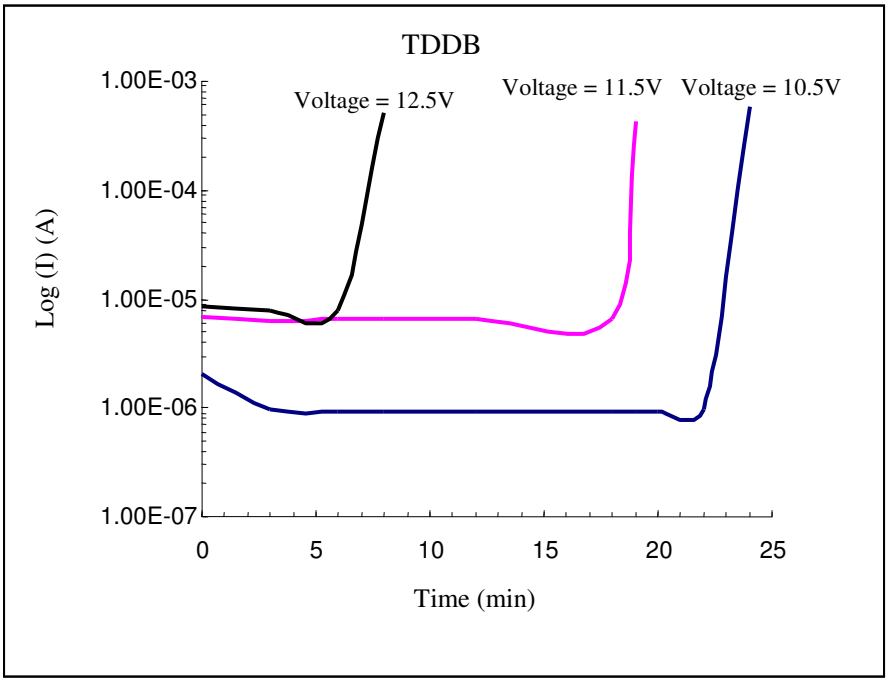

Figure 4.6 Time Dependent Dielectric Breakdown Behavior of a 7*nm* Oxide Thickness p-type Device

4.5.2 F-N Region

For a relatively thick thermal oxide, no detectable current flows until the occurrence of Fowler-Nordheim (F-N) tunneling, which occurs with the presence of an electric field above 6-7 Mv/cm. Analysis have been carried out in the following sections to determine parameter B in the F-N region.

4.5.2.1 C-V Characteristics

MOS capacitance-voltage measurement is one of the most common processes monitoring diagnostics employed in device manufacturing. A vast amount of information can be derived from this simple test. First and foremost a look at the Capacitance versus voltage plot gives an idea about the Device Under Test (DUT). The applied voltage range was from -5V to 5V to sweep the DUT from depletion to accumulation and retrace. Hysteresis was noticed during retrace. Cells where the hysteresis was prominent (more trap charges) were marked not good for further analysis. Only the characteristics similar to the plots shown in figures 4.7 (for n-type) and 4.8 (for p-type) were analyzed further. This was another way of validating the cells in addition to checking the uniformity of current (as already discussed in the previous section).

It is very important became imperative to find out the exact value for oxide thickness because the parameter B is sensitive to the uncertainty of the oxide thickness [36]. This uncertainty in B can be given by the following expression:

$$B = B_0 [1/ (1+t)] \qquad (4.1)$$

where t is the uncertainty of the oxide thickness. Thus as seen the uncertainty in the parameter B is almost proportional to the uncertainty of the oxide thickness.

The wafers used in this research had oxide thickness with the nominal value of 7 *nm*. Now this is not an accurate indication of the oxide thickness. This value is just the average for all the cells. An accurate value for parameter B requires an accurate value for oxide thickness. This was achieved, in this research work, by analyzing the C-V data in accumulation.

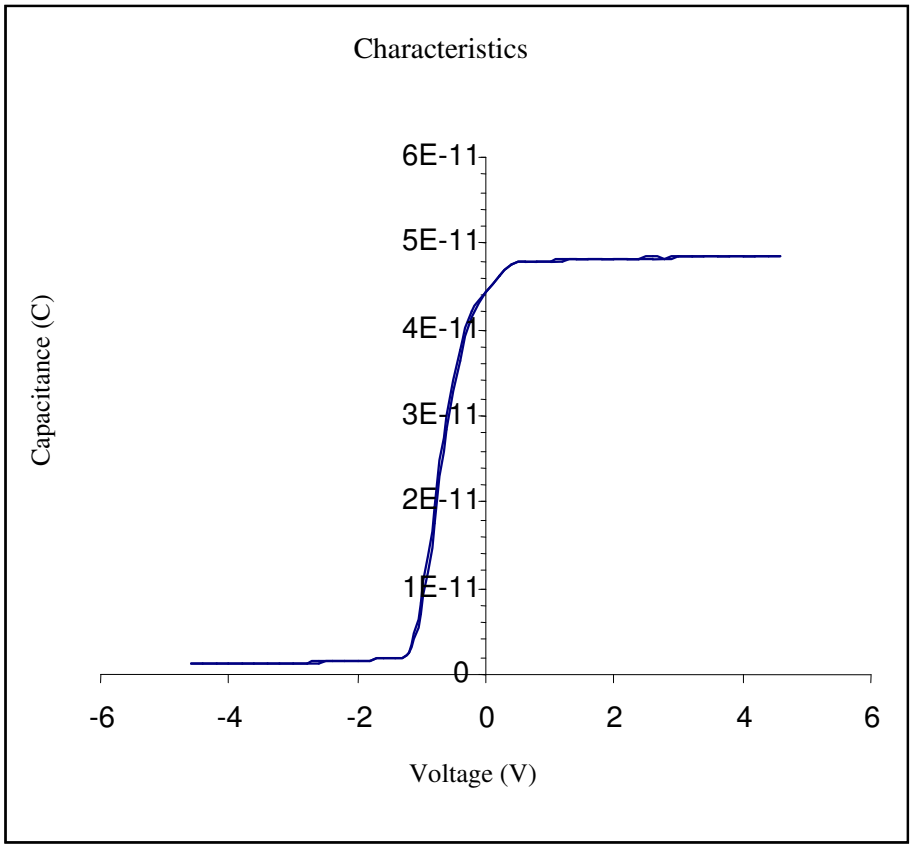

Figure 4.7 C-V Characteristics of a 7 *nm* n-type Device

Theoretically Oxide thickness is given by

$$t_{ox} = \left(\frac{A * \varepsilon_{ox}}{C_{ox}} \right) \tag{4.2}$$

where C_{ox} is the capacitance in accumulation, A is the area of the cell under investigation and ϵ_{ox} is the permittivity of oxide. The value of unknown parameter C_{ox} was found by looking at the C-V curve and noting down the value for Capacitance when the cell was in accumulation.

Thus an accurate value of gate oxide thickness was determined. In the course of this research it was found out that oxide thickness ranged from 6.76 *nm* to 7.167nm. The plot shown in Figure 4.7 is for n-type wafer and the oxide thickness for the same is 7.134*nm*.

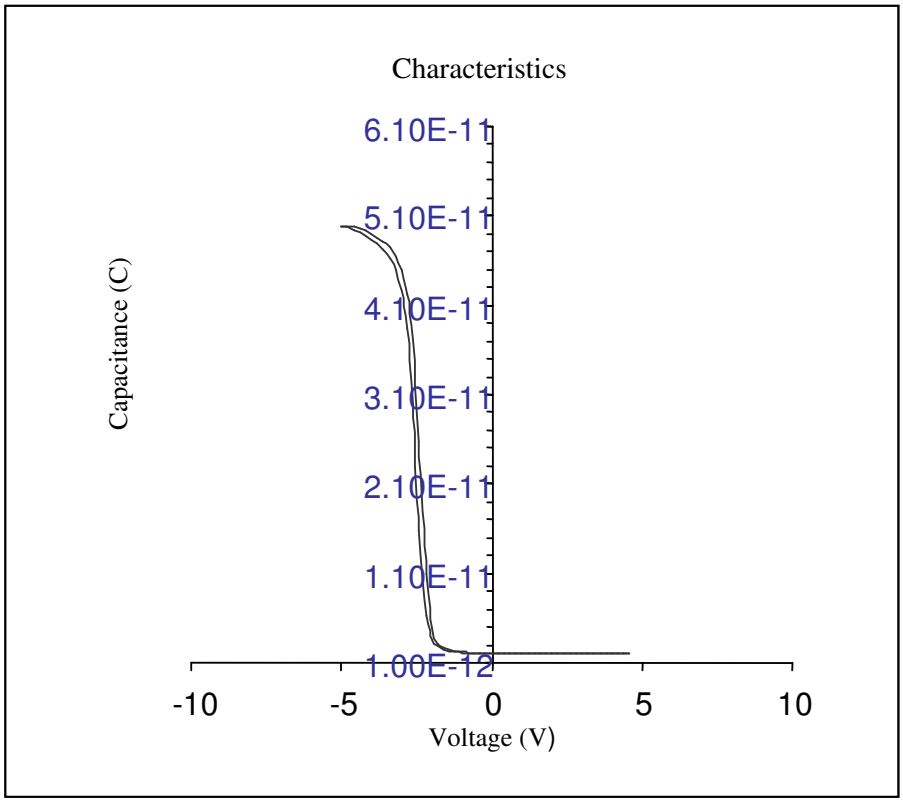

Figure 4.8 C-V Characteristics of a 7 *nm* p-type Device

The Plot shown in figure 4.8 is for p-type wafer and the thickness for the same is 6.99 *nm*.

Next step in order to find an accurate value for parameter B was to find flat band voltage V_{FB}. Ideally V_{FB} should be zero. But because there is a work function difference and a finite value for oxide charge, V_{FB} has a value other then zero. This value was found out relying on C-V data. To find V_{FB} it was important to find flat band capacitance which was given by

$$C_{FB} = \left[\left(\frac{C_i * C_d}{C_i + C_d}\right)\right] * A \tag{4.3}$$

where C_d is the depletion capacitance at flatband and is given by

$$C_d = \epsilon_{si} / L_d \tag{4.4}$$

where L_d is the Debye length .

Once C_{FB} was determined, the corresponding flat band voltage, V_{FB} was determined from the C-V curve. After analyzing lots of C-V data, it was found out that for the samples used in this research V_{FB} ranged from -0.90V to -1.05 V for n type device and from -1.80 V to -2.05 V for p-type. The flat band voltage, V_{FB} for the figure 4.7 was determined to be -1.00 V and for figure 4.8 it was -2.00V.

4.5.2.2 I-V Characteristics

A typical I-V curve can be divided into three distinct regions. At lower voltage for a given oxide thickness, the behavior is explained by Direct tunneling. As the voltage increases so does the electric field. This region where the current increases sharply with

the corresponding increase in the electric field is termed as F-N region. As this research concentrates on finding the F-N parameter B, concentration was limited to the region. The third region is the dielectric breakdown. This is accompanied by a very large increase in current value.

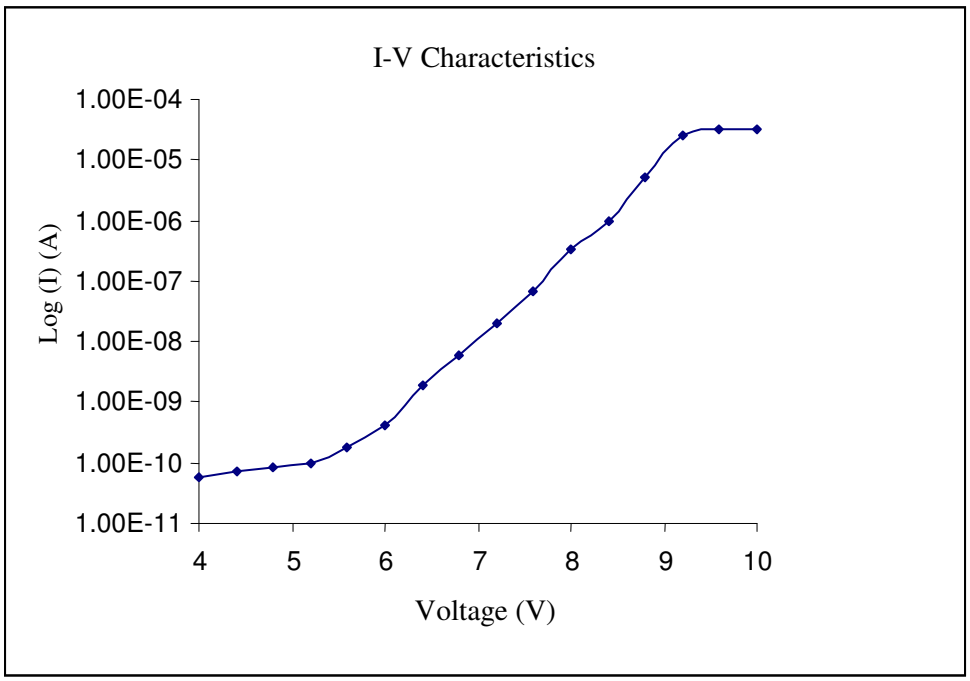

Figure 4.9 I-V Characteristic for a 7 *nm* n-type Device

Figure 4.9 shows I-V characteristics of an n-type device with 7 *nm* oxide thickness, biased for various voltage values.

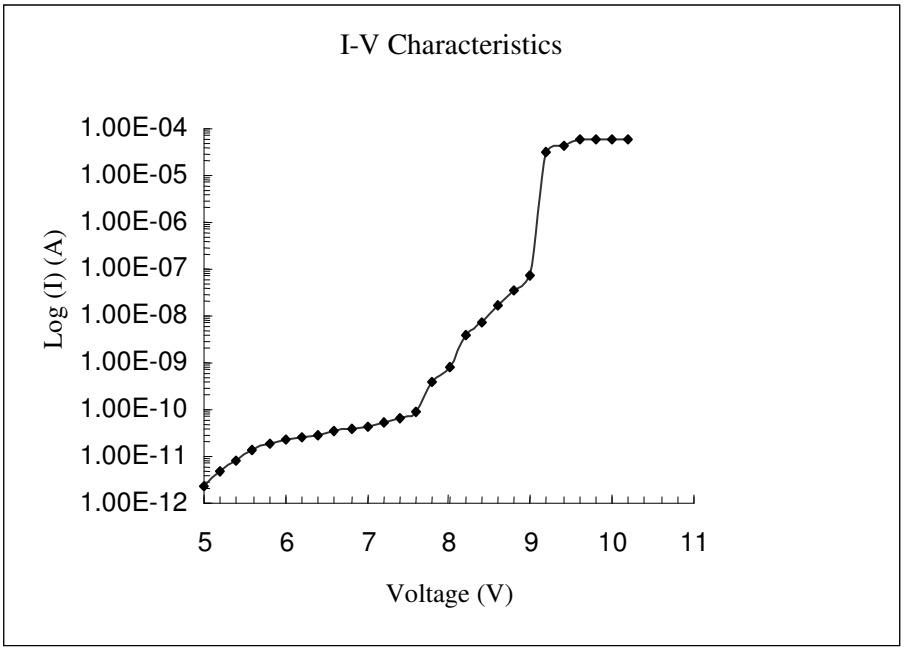

Figure 4.10 I-V Characteristic for a 7 *nm* p-type Device

Figure 4.10 shows I-V characteristics of an p-type device with 7 *nm* oxide thickness, biased for various voltage values.

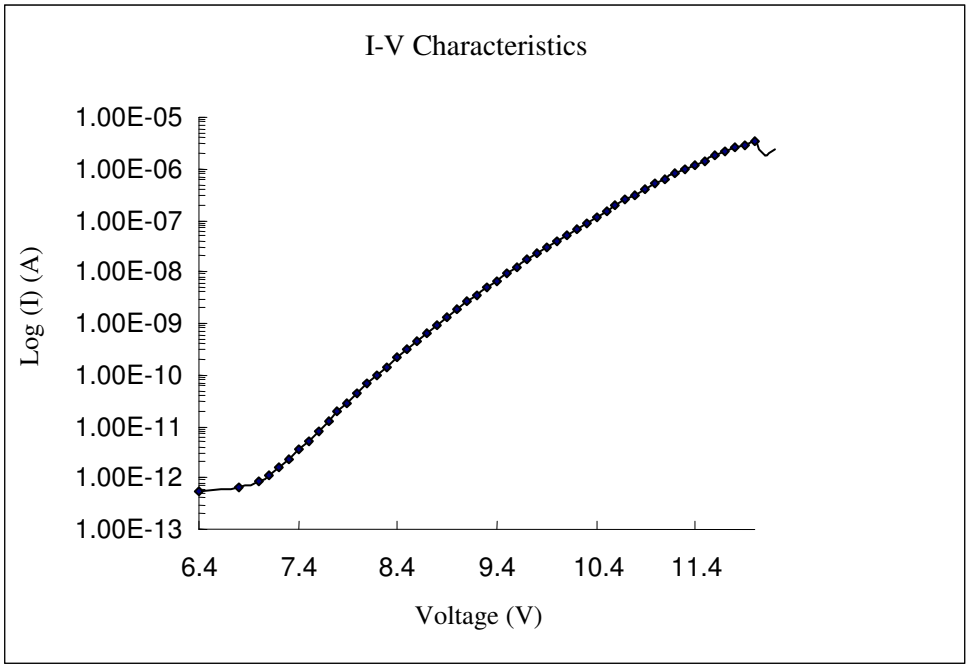

Figure 4.11 I-V Characteristic for a 13 *nm* n-type Device

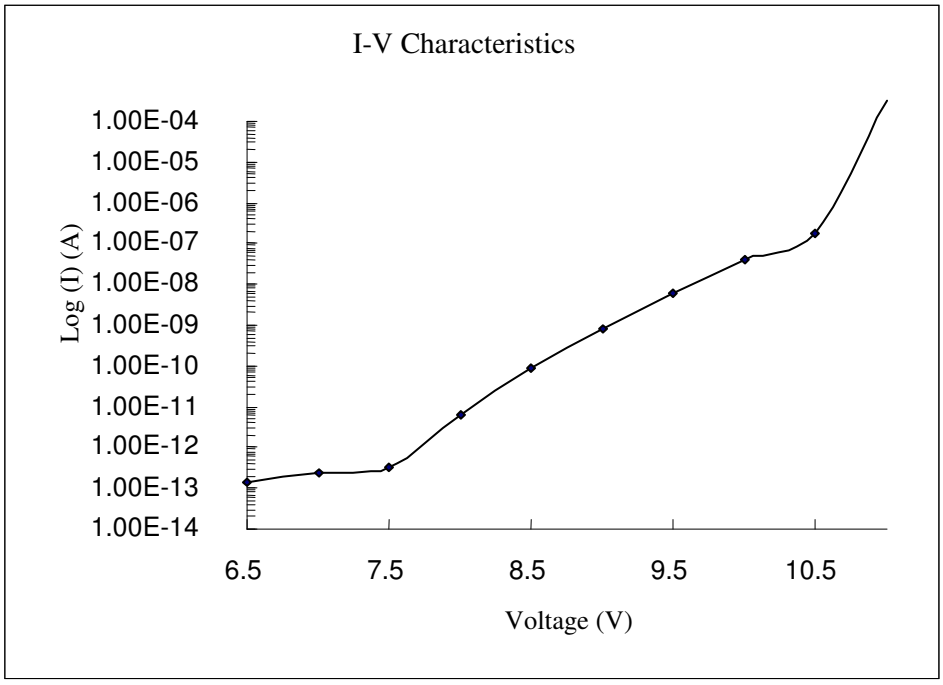

Figure 4.12 I-V Characteristic for a 10 *nm* p-type Device

Figures 4.11 and 4.12 show I-V characteristics of both n-type and p-type devices (with 13 *nm* and 10 *nm* oxide thickness respectively), biased for various voltage values.

4.5.2.3 F-N Slope Analysis and Experimental Results

In the last three decade the F-N tunneling in the MOS capacitor has been studied extensively by many researchers. According to the simplest model and as already shown in chapter 3, the tunneling current density J is given by

$$J_{FN} = AF^2 \exp\left[-\frac{B}{F}\right] \tag{4.5}$$

where

$$A = \frac{q^3}{16\pi^2 h\phi_s} \tag{4.6}$$

$$B = \frac{4}{3}\frac{(2m_{ox})^{1/2}}{qh}\phi_b^{3/2} \tag{4.7}$$

According to the model, the plot of $\log J/F^2$ versus $1/F$ (also called the F-N plot) should be linear with a slope of B. To find the slope (parameter B), current density (J) and electric field (F) needs to be determined first. This can be achieved by first determining the F-N region in an I-V curve and then finding the corresponding current density and electric field using the following equations. The current density, J is given by

$$J = I / A \tag{4.8}$$

and the Electric field is given by

$$F = (V - V_{FB}) / t_{ox} \tag{4.9}$$

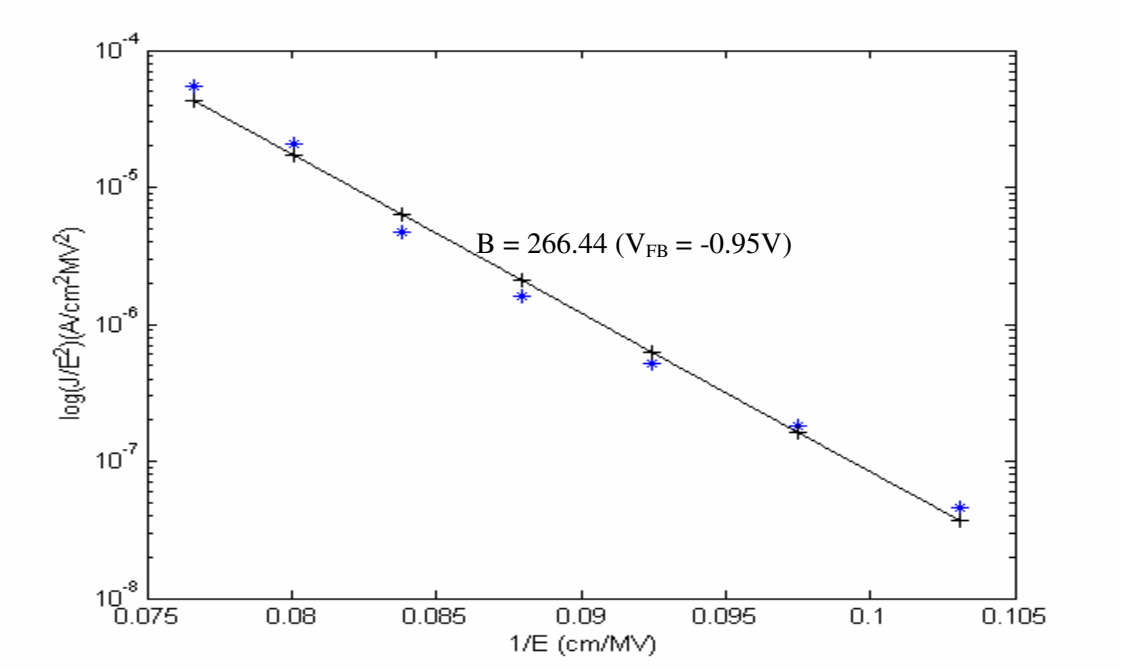

Figure 4.13 Fowler-Nordheim Plot for 7*nm* n-type Device

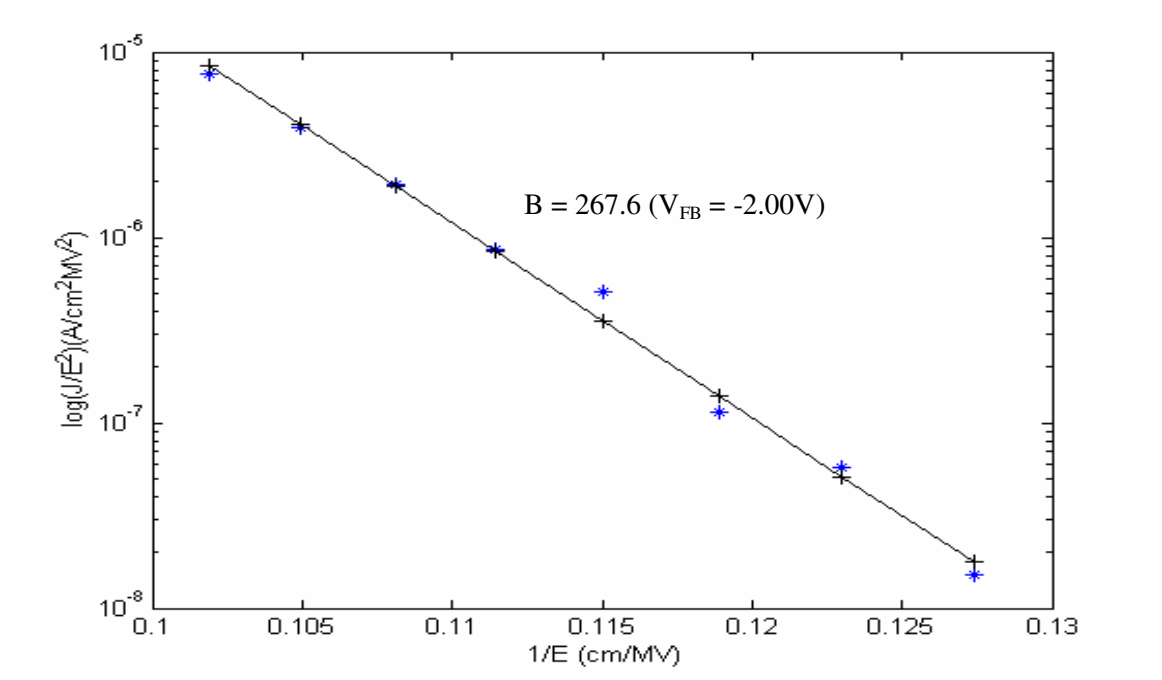

Figure 4.14 Fowler-Nordheim Plot for 7*nm* p-type Device

Figure 4.13 shows the corresponding F-N plot for the F-N region shown in Figure 4.9. This region was approximated to lie between voltages 6.0V and 8.4V. The value of parameter B extracted from the F-N plot was 266.44. Similarly for Figure 4.10, the F-N region was approximated to lie between 7.6V and 9.0V. Figure 4.14 is the F-N plot for the same. The slope of this plot (Parameter B) is 267. Figure 4.15 shows the corresponding F-N plot for the F-N region shown in Figure 4.11. This region was approximated to lie between voltages 7.0V and 11.0V. The value of parameter B extracted from the F-N plot was 263.11. Similarly for Figure 4.12, the F-N region was approximated to lie between 7.5V and 10.5V. Figure 4.16 is the F-N plot for the same. The slope of this plot (Parameter B) is 260.18.

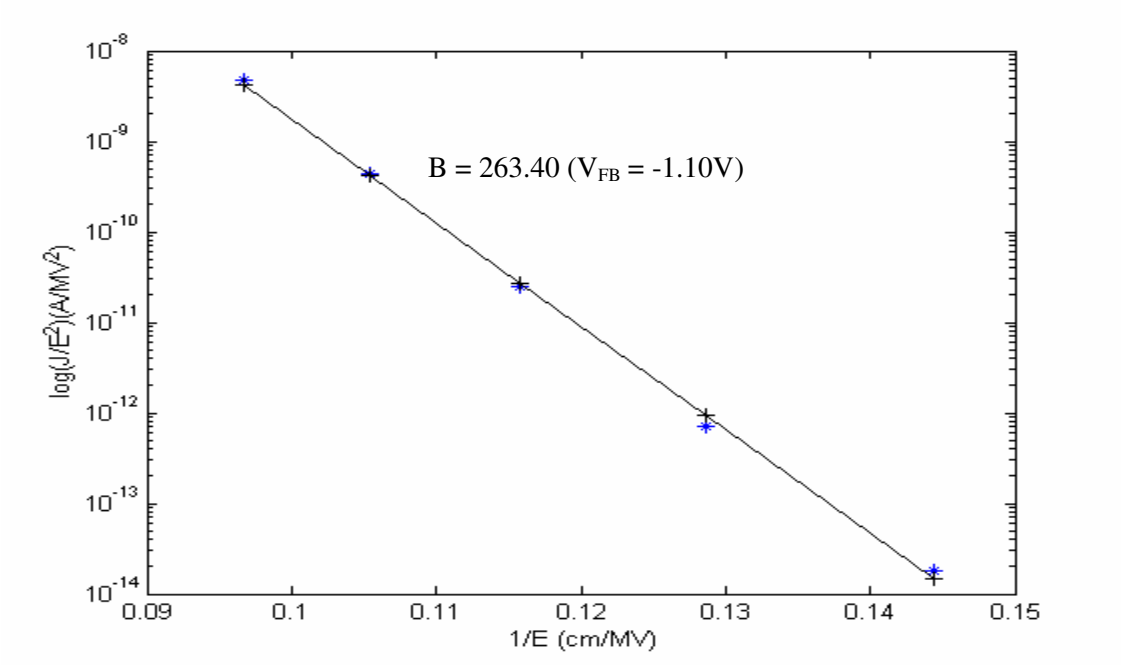

Figure 4.15 Fowler-Nordheim Plot for 13*nm* n-type Device

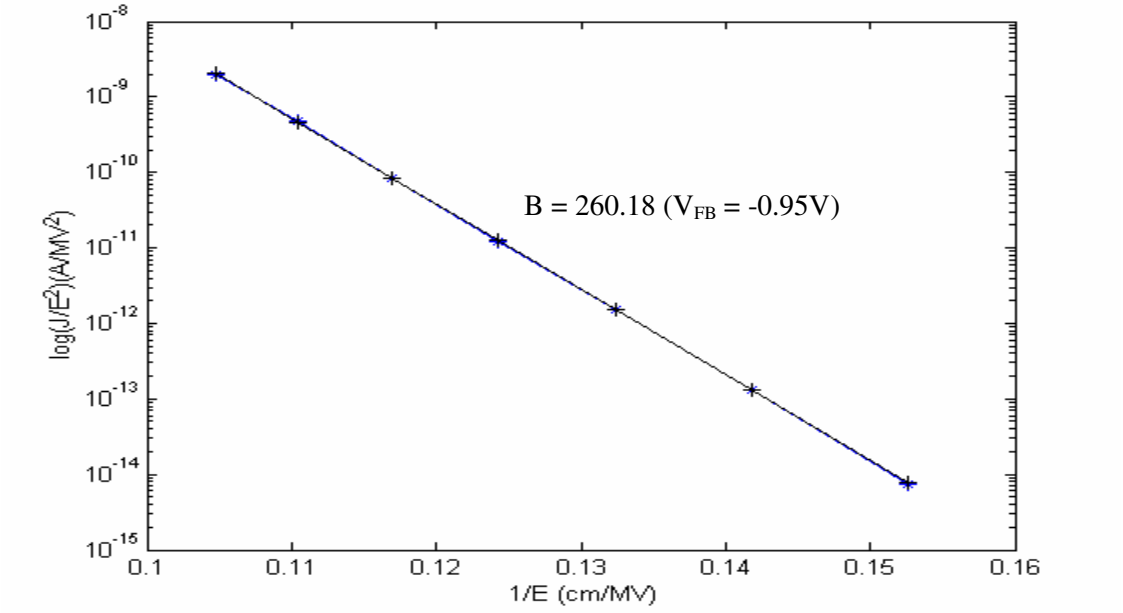

Figure 4.16 Fowler-Nordheim Plot for 10*nm* p-type Device

Chapter 5

Summary and Conclusion

As a part of this research work, one of the first step undertaken was to set up a highly reliable Test Station. The whole set up was then automated to remove human error angle. Improvisation of the software and the Test Station is an ongoing process in the Lab. Current Uniformity was tested for different areas on each wafer. It was observed that though the current increased with increase in area, current density remained more or less constant. Hysteresis in the C-V was observed to be in the range which indicated minimal oxide charge effects.

F-N parameter B is important for determining reliability issues in MOSFET and threshold voltage control in NVM. As reported in [36], the uncertainty of the parameter B is almost proportional to that of the oxide thickness. Thus the nominal value of oxide thickness wasn't used for analysis. Instead an effort was made to calculate accurate value of oxide thickness from the capacitance in accumulation. C-V data was collected and analyzed to determine flat band voltage.

Due to continuous scaling, fundamental limits have been reached for SiO_2 scaling. This can be attributed to the change in oxide thickness behavior due to increased problems with dopant penetration through ultra thin SiO_2 layers and direct tunneling mechanism. This research work concentrated on devices with oxide thickness varying from 7 *nm* to 13

nm. For oxide thickness in the range 10~13 *nm*, the parameter B is not sensitive to the change in oxide thickness. The values of B ranged from 260 to 267 for devices with varying thickness. This result was close to the value of B (258) obtained in [36]. As this sensitivity of parameter B increased with decreasing oxide thickness, most of the focus was on devices with 7 *nm* oxide thickness.

Flat-Band voltage was also extracted from this data. It was observed that for n-type device, the Flat-Band voltage varied from -0.90V to -1.10V and for p-type it varied from -1.80V to -2.10V. It was noticed that for thickness around 7*nm* wide variety of results were obtained for the Fowler-Nordheim parameter B. In other words, the slope variation observed was more than the other thickness ranges. Researchers have agreed that below 5 *nm* thickness range, Fowler-Nordheim Tunneling is no longer dominant. The conduction is explained by Direct Tunneling. However based on the results which exhibited wide range for parameter B, it was concluded that the effects of Direct Tunneling could already be felt for 7 *nm* oxide thickness.

For NVM, the difference in the threshold voltage after writing and erasing, the so-called memory window, is one of the major concerns in the design of device. This difference is given by

$$\Delta V_T = Q / C_2 \tag{5.1}$$

where Q is the charge in the floating gate of the device and C_2 is the capacitance between the control gate and the floating gate. As can be seen from equation 5.1, change in threshold voltage is sensitive to the charge in the floating gate, which in turn depends on the current density. This current density, which depends upon the barrier height, is given

by equation 4.5. As can be seen, F-N parameter B is exponential term in the equation. Hence it won't be wrong to deduct that it is the most dominant factor in determining current density.

The variation of B for oxide thickness in the range of 7 *nm* is attributed to the enhancement in the leakage current due to the direct tunneling. The direct tunneling depends upon the barrier height and oxide thickness. Hence to have tight control over V_T for a NVM, new algorithms need to be developed for even better process control for oxide thickness in the range of 7 *nm* and below.

References

[1] M. Bohr, S.U. Ahmed, L. Brigham, R. Chau, R.Gasser, R. Green, W. Hargrove, E. Lee, R. Natter, S.Thompson, K. Weldon and S. Yang, IEDM TechnicalDigest, 1994, p. 273

[2] M. Bohr, S.S. Ahmed, S.U. Ahmed, M. Bost, T.Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S.Sivakumar, S. Thompson, J. Tsai, and S. Yang, IEDM Technical Digest, 1996, p. 847

[3] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," IEEE J. Solid-State Circuits SC-9, 256-268 (1974)

[4] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS Scaling into the Nanometer Regime," Proc. IEEE 85, 486 –504 (1997)

[5] Semiconductor Industry Association National Technology Roadmap for Semiconductors 1997 (Sematech Inc., Austin, Texas 1997); http://www.sematech.org

[6] S. Ohmi, C. Kobayashi, E. Tokumitsu, H. Ishiwara and H. Iwai, Ext. Abs. SSDM, Tokyo, Japan, pp. 496-497, 2001

[7] M. Takayanagi, S. Takagi, and Y. Toyoshima, Proc. IRPS, Orlando, FL, pp.380-385, 2001

[8] J. R. Hauser, IEDM Short Course, Washington DC, December, 1999

[9] P. Pavan, R. Bez, P. Olivo,and E. Zanoni, "Flash Memory Cells—An Overview," Proc. IEEE 85,1248-1262(1997)

[10] H.K. Henisch, Rectifying Semiconductor Contacts, Clarendon, Oxford, 1957

[11] Physics of Semiconductor Devices, Second edition, S. M. Sze, Wiley & Sons, 1981, Chapter 8

[12] Massachusetts Institute of Technology, “Microelectronics Processing Technology”, Fall 2001

[13] J.E. Lilienfeld, U.S. Patent 1,745,175 (1930) and O. Heil, British Patent 439,457 (1935)

[14] Bart Van Zeghbroeck, “Principle of Semiconductor Devices”, 1999 http://ece-www.colorado.edu/~bart/ecen5355/newbook/

[15] A. Chimenton, P. Pellati and P. Olivo, “Overerase Phenomena: An Insight In to Flash Memory Reliability”, Proc. IEEE 91, NO. 4, April 2003

[16] Micron Technology, Inc., “Boot Block Flash memory”, TN-28-01, FT01.p65-Rev. 12/99

[17] S. Thompson, P. Packan and M. Bohr, " MOS Scaling: Transistor Challenges for the 21st Century," Intel Technology Journal (1998)

[18] Y. Taur, “CMOS design near the limit of scaling”, IBM Journal of Research & Development, Vol. 46, Numbers 2/3, 2002

[19] S.-H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, “Quantum-Mechanical Modeling of Electron Tunneling Current from the Inversion Layer of Ultra-Thin-Oxide nMOSFET's”, IEEE Electron Device Lett. 18, 209–211 (1997)

[20] M. Lenzlinger and E. H. Snow, “Fowler–Nordheim Tunneling into Thermally Grown SiO2,” J. Appl. Phys. 40, 278–283 (1969)

[21] Maserjian, "Tunneling in Thin MOS Structures," J. Vac. Sci. Technol. 11, No. 6, 996-1003 (1974)

[22] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe and M. M. Heyns IMEC, Kapeldreer 75, B-3001 Leuven and Department of Solid State Science, University of Ghent, Krijgslaan 281-S1, B-9000 Ghent, Belgium

[23] J. Frosch and L. Derick, “Surface Protection and Selective Masking During Diffusion in Silicon,” Proc. Electrochem. Soc., p. 547 (1957)

[24] D. Khang, “Silicon–Silicon Dioxide Surface Device,” (Bell Laboratories technical memorandum issued on January 16, 1961), Semiconductor Devices: Pioneering Papers, S. M. Sze, Ed., World Scientific Press, Singapore, 1991, pp. 583–596

[25] D. L. Crook, “Method of Determining Reliability Screens for Time Dependent Reliability Breakdown,” Proceedings of the International Reliability Physics Symposium, p.1, (1979)

[26] J. H. Stathis, “Physical and Predictive Models of Ultra Thin Oxide Reliability in CMOS Devices and Circuits,” IEEE Trans. Device & Mater. Reliabil. 1, 43–59 (2001)

[27] D. J. DiMaria, E. Cartier, and D. Arnold, “Impact Ionization, Trap Creation, Degradation, and Breakdown in Silicon Dioxide Films on Silicon,” J. Appl. Phys. 73, 3367–3384 (1993)

[28] C. Schuegraf and C. Hu, “Metal-Oxide Semiconductor Field-Effect-Transistor Substrate Current During Fowler–Nordheim Tunneling Stress and Silicon Dioxide Reliability,” J. Appl. Phys. 76, 3695–3700 (1994)

[29] D. J. DiMaria and J. W. Stasiak, “Trap Creation in Silicon Dioxide Produced by Hot Electrons,” J. Appl. Phys. 65, 2342–2356 (1989)

[30] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, “New Insights in the Relation Between Electron Trap Generation and the Statistical Properties of Oxide Breakdown,” IEEE Transactions on Electron Devices, Vol. 45, No. 4, pp. 904-911, (1998)

[31] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, and H. E. Maes, “A Consistent Model for the Thickness Dependence of Intrinsic Breakdown in Ultra-Thin Oxides,” IEDM Tech. Digest, p. 866 (1995)

[32] C. Hu, “Gate Oxide Scaling Limits and Projection,” IEDM Tech. Digest, pp. 319–322 (1996)

[33] M. L. Green, E. P. Gusev, R. Degraeve, and E. L. Garfunkel, “Ultrathin (<4nm) SiO2 and Si-O-N Gate Dielectric Layers for Silicon Microelectronics: Understanding the Processing, Structure, and Physical and Electrical Limits,” J. Appl. Phys. 90, 2057–2121 (2001)

[34] J. H. Stathis and D. J. DiMaria, “Reliability Projection for Ultra-Thin Oxides at Low Voltage,” IEDM Tech. Digest, pp. 167–170 (1998)

[35] J. H. Stathis, A. Vayshenker, P. R. Varekamp, E. Y. Wu, C. Montrose, J. McKenna, D. J. DiMaria, L.-K. Han, E. Cartier, R. A. Wachnik, and B. P. Linder, “Breakdown Measurements of Ultra-Thin SiO2 at Low Voltage,” Symposium on VLSI Technology, Digest of Technical Papers, 2000, pp. 94–95

[36] Y.L. Chiou, J.P. Gambino and M. Mohammad, " Determination of the Fowler-Nordheim tunneling parameters from the Fowler-Nordheim plot," Solid-State Electronics 45 (2001) 1787-1791

[37] Instruction Manual Model 617 Programmable Electrometer, Keithley Instrument, Inc., 1984

[38] Instruction Manual Model 230 Programmable Voltage source, Keithley Instrument, Inc., 1988

Appendices

Appendix A: Factors Affecting Testing on Probe Station

Following are the most common problems encountered probing at very low current levels.

A.1 Random Noise Introduction

This category includes such problems as EM noise transmitted through shielding, charge coupled noise resulting from activity around the test set up and triboelectric effects induced through probe cables. The data tends to show large distortion which are however not periodic in nature. The best way to take care of this problem is by correcting Shielding problems, selecting proper cable to be used in testing, proper mounting and strain relief to avoid triboelectric effects.

A.2 Periodic Noise Introduction

Some sources can cause periodic noise in the test data. These sources produce data that is repeated throughout the test and may appear in multiple periodic forms. The short noise could be caused by a illuminator power supply fan resting on the prober vibration isolation table while the larger noise is caused by improper tuning (air pressure or balancing) of the vibration table producing low frequency vibration effects.

A.3 Noise Coupled Through Measurement Equipment

Poor tester and prober grounding or a poorly isolated probe will allow electrical noise from power supplies or external circuits to enter the probing environment and be coupled to the measurement.

A.4 Probe Contact

Many times noisy data can result from a bad probe contact. This shows on the graph as distortions at lower current levels and may disappear completely at the higher levels. Such a probe should be replaced or a probe with lower resistance tip properties should be used.

A.5 Leaky Probes

Leaky probes distort data by offsetting it. This is a serious problem as leakage paths in most high performance probes are capacitative in nature and so they charge, changing the characteristics over time. Leaky probes can also couple current to good probes in close proximity. Thus it can be said that one bad probe can make all others look leaky as well. However the solution is easy. Generally all leaky probes are easily spotted with a voltage sweep.

A.6 Factor Affecting the Low Level Environment Current Testing

A number of factors may effect low-level measurement. Such factors and their cure have been described in brief below:

A.6.1 Common Impedance

Common impedance is an impedance or current path shared by a source of noise and the receiving instrument. The noise source causes the current to flow through the common impedance. If this common impedance is in parallel with the instrument's input, a noise voltage will be coupled into the measured signal. If the instrument is measuring current,

then the current flowing through the common impedance will algebraically sum with the signal current at the connecting node.

A.6.1.1 Preventive Measure

Current paths between different circuit functions need to be isolated. Also unintended contact between the ground conductors that reside in different parts of the circuit should be removed.

A.6.2 Magnetically Coupled Noise

Noise is introduced due to magnetic coupling, which may be due to motors, transformers and facility wiring.

A.6.2.1 Preventive Measures

The best way to overcome such a noise is to remove unnecessary source of electromagnetic energy in the area of test circuit.

A.6.3 Incidental Capacitive Coupling

Noise current can be coupled to the measurement circuit via stray capacitance. Stray capacitance is inversely proportional to the distance between conductors. If the circuit under test has high impedance, it will see a noise voltage. If the impedance is low, a current will be developed.

A.6.3.1 Preventive Measures

A distance should be maintained between unshielded noise sources and the test circuit. Separation of two wires by a distance which is more than 40 diameter will provide good electrostatic isolation. Active guard shields will also provide good protection.

A.6.4 Charge Transfer

If a charge body is brought in proximity to the test circuit, that charge will attract or repel charge in the test circuit and cause momentary electron flow. A human body is atypical charge vehicle.

A.6.4.1 Preventive Measure

Limited activity in the area of test circuit measurement should be carried out.

A.6.5 Light

Semiconductor devices are generally light sensitive. This is especially true for MOSFET devices that rely on minority carriers for conduction. At femto ampere levels, the current induced by exposure to light will obliterate many measurements.

A.6.5.1 Preventive Measures

A light tight enclosure (LTE) should be used around the test circuit.

A.6.6 Intrinsic Noise Sources

At low fA levels noise is produced due to the coaxial and tri axial cables. This noise may result in flow of unwanted current. Also contacts between dissimilar metals, crimp joints may develop temperature sensitive Seebeck voltages, which is nothing but the thermal noise.

A.6.6.1 Preventive Measures

All the cables involved in the test must be allowed to rest. Anchoring the cable is another way of reducing noise. It should be made sure that contacts are made between similar metals to check thermal noise.

A.6.7 Leakage Current

Insulation is not 100% leak free. So insulation used in cabling, connectors and probes etc. will permit a degree of leakage that can again hamper the measurements.

A.6.7.1 Preventive Measures

An active guarded shield on all cables should be used to keep the differential voltage at a very low level.

A.7 Work Done on the Probe Station in Noise and Reliability Lab

The probe station was received from Lucent Labs. One of the air valves in one of the legs of the station was leaking. This meant that the lift was not as was required. This air valve

was replaced. Also as there was no separate vacuum line available in the lab, a motor was brought that could create vacuum to hold the sample. A cover for the station was made so that it could be isolated from the noisy environment around.

Appendix B: LabVIEW and Integrated Circuit Testing

B.1 Introduction

The program was developed in LabVIEW to test integrated circuits. The program was able to turn on a Keithley 230 Programmable Voltage Source and record the current measured current measured from a Keithley 617 Programmable Electrometer. The measured data was then graphed over time.

LabVIEW is a graphical programming language used by many people for data acquisition and for controlling laboratory instruments. LabVIEW uses graphical block diagrams instead of conventional code to write programs. The diagrams are complied by LabVIEW into machine code for the computer to process. LabVIEW programs are called virtual instruments, or VI's. Each VI consists of a front panel and a block diagram. The front panel is the graphical user interface of the VI. This is the screen that the user will see when they run the program. The front panel usually contains knobs, push buttons, graphs, and other graphics for the user to interface with.

The block diagram is the VI's source code. The source code is written in the graphical programming language called "G". The block diagram usually contains sub VI's (subroutines), built-in functions, and program execution commands. The graphical commands were wired together to show how data flows through the program. Each component in the front panel was represented in the block diagram so that data can flow between the program and the user.

B.2 Front panel of the VI

Figure B.1 illustrates the front panel of the program. From the front panel the user can decide how to test the integrated circuit. The controls on the left side of the front panel command the Keithley 617 Electrometer and the control panel on the right side, commands the Keithley 230 Voltage Source. The following pages will describe the main components of the front panel of the LabVIEW software written during the course of this research.

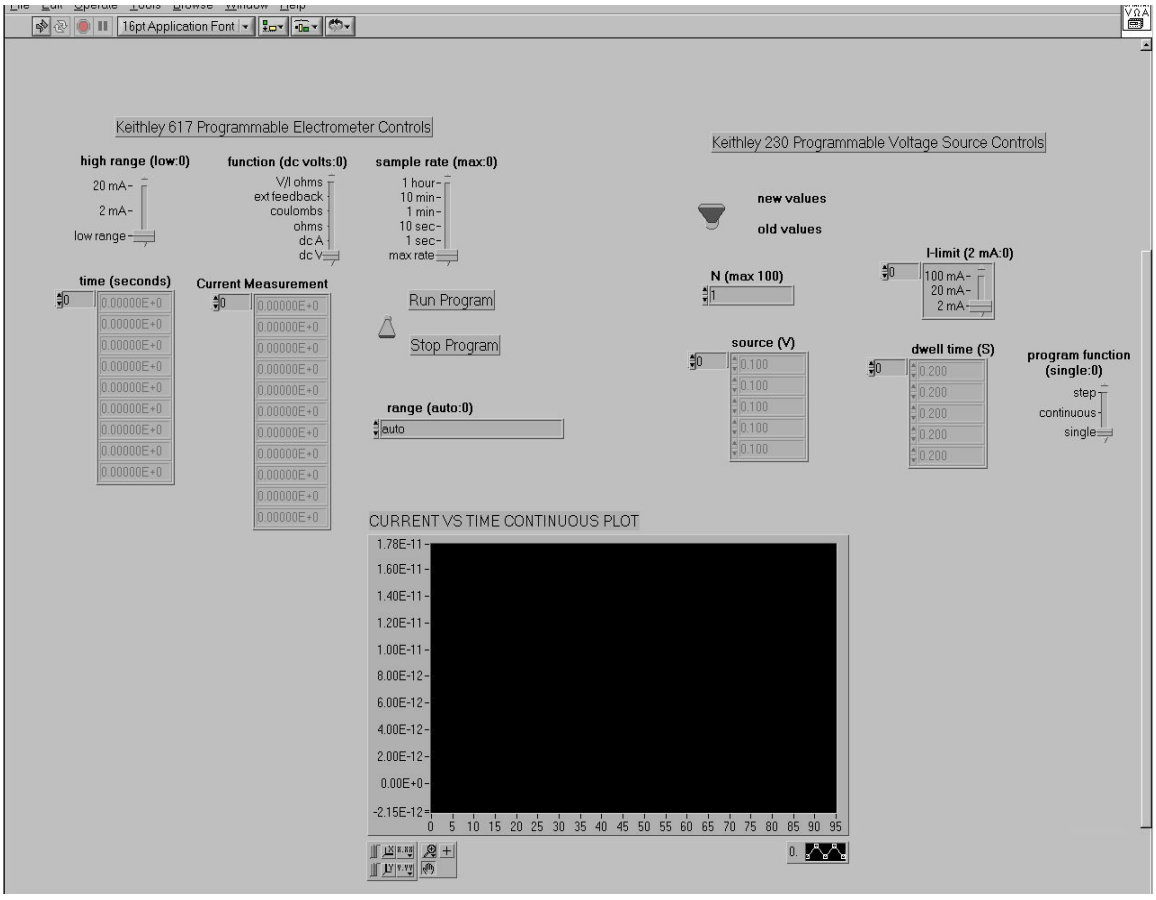

Figure B.1 Front Panel of VI

B.3 Block diagram of the VI

Figure B.2 illustrates part of the Block Diagram (source code) for the program. Notice that the Block Diagram has been placed in what appears to be a frame of film. This is called Sequence Structure. This program has four frames, numbered 0-3. Only one frame is usually visible at a time. To see the other frames user can click on the arrows at the top of the frame. Figure 2 shows the first frame in our Block Diagram. In this frame both the Keithley 617 and the Keithley 230 are turned on and programmed according to the user's inputs in the front panel.

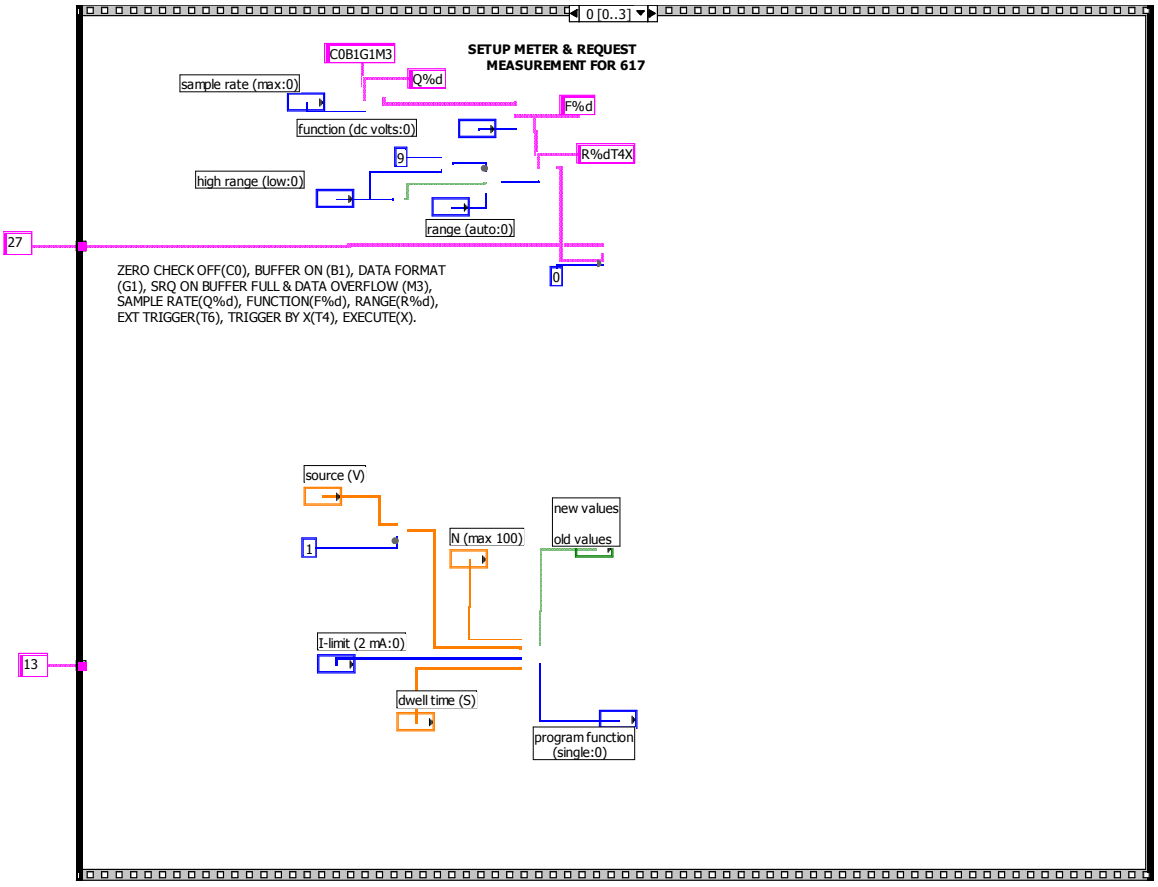

Figure B.2 Frame 0 of the Block Diagram

Appendix B: (Continued)

Figure B.3 below shows the second frame in our Block Diagram. In this frame the Electrometer takes the upto 100 sample, records the time the samples were taken, lists the results in the tables in the front panel, and graphs the results.

The program stays in this frame until the user moves the switch to the "Stop Program" position. When the switch has been pressed the program stops taking data and continues to the third frame of the Block Diagram.

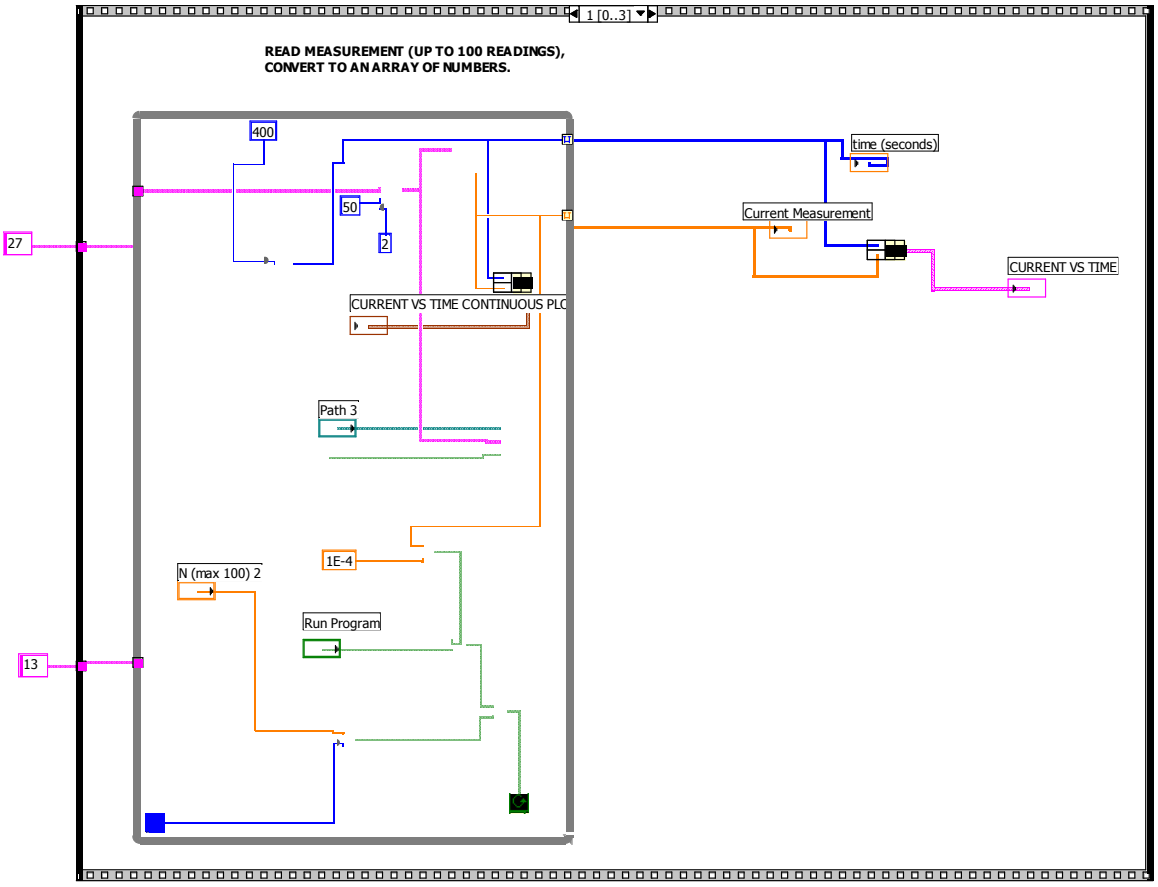

Figure B.3 Frame 1 of the Block Diagram

Figure B.4 shows frame 3 of the block diagram. These frames clear all data from the Keithley 617 so it can be used again. The memory of the Keithley 617 must be erased before new data points can be taken.

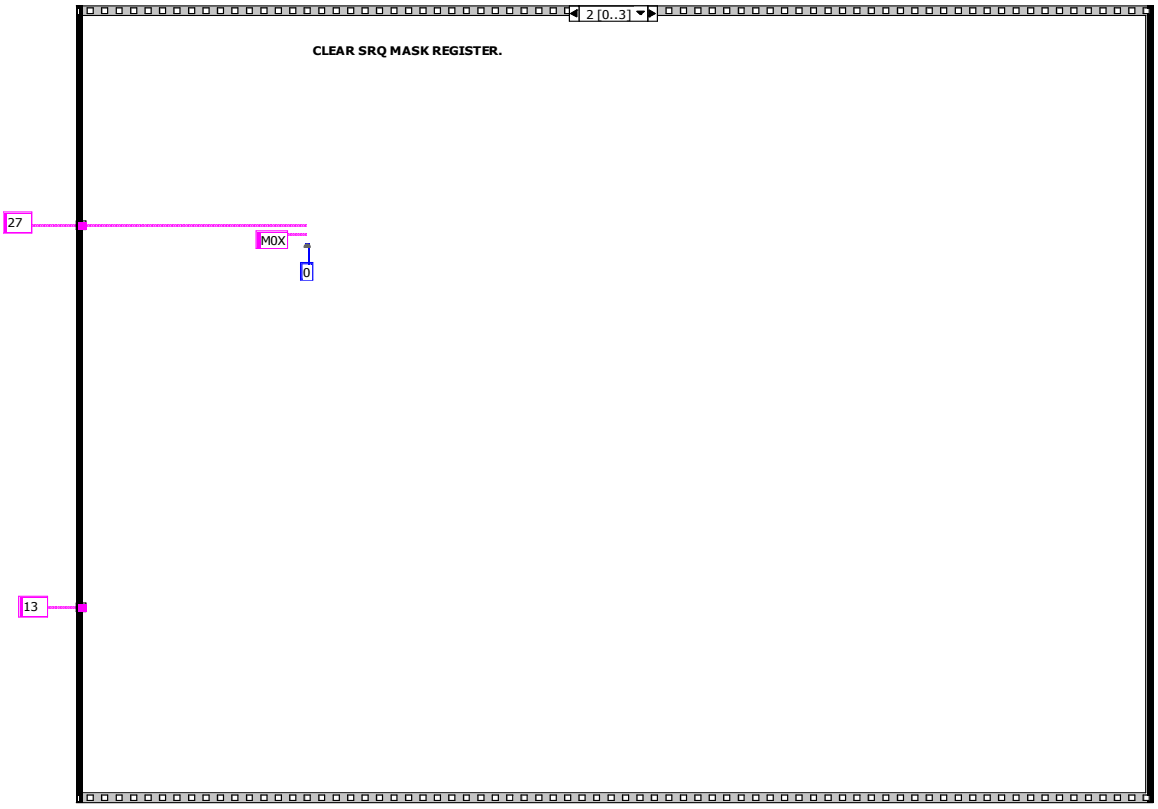

Figure B.4 Frame 3 of the Block Diagram

B.4 How to run the program

The following procedure will take you through step-by-step how to obtain a current versus time plot from LabVIEW.

1. Turn on LabVIEW.
2. Open the file IV.vi

3. Change the sample rate to 10 second.
4. Turn the “Run Program” switch to the on position
5. Change N to 100. This will command the Keithley 617 Electrometer to take 100 samples.
6. Enter the input voltage range.
7. Set the dwell times for each input voltage.
8. Move the Function Vertical Slide to dcA. This tells the Electrometer to take current measurement.
9. Set the program function to continuous.
10. Push the “Run Program” button in the upper left corner of the LabVIEW window.
11. Watch the data light on the Electrometer. When the data light begins to blink the Electrometer is no longer recording data. (Overflow has occurred.)
12. To stop the program turn the “Run Program” switch to the off position. After a few seconds the graph in the bottom left corner of the window will appear with the current versus time plot.