ABSTRACT

Title of Dissertation: EXPERIMENTAL STUDY OF BIAS TEMPERATURE INSTABILITY AND PROGRESSIVE BREAKDOWN OF ADVANCED GATE DIELECTRICS

Baozhong Zhu, Doctor of Philosophy, 2005

Directed By: Professor Joseph Bernstein
Department of Mechanical Engineering
Dr. John Suehle
Semiconductor Electronics Division, NIST

With shrinking gate dielectrics, the reliability requirements of semiconductor gate dielectrics become more and more difficult to maintain. New physical mechanisms and phenomena are discovered and new challenges arise.

At the same time, some issues, which have been minor in the past, begin to show bigger impact, such as the Negative Bias Temperature Instability issue. The dynamic NBTI phenomenon was studied with ultrathin SiO_2 and HfO_2 devices. With a dynamic stress condition, the device lifetime can be largely extended due to the reduced NBTI degradation. This reduction is contributed to the annealing of fixed oxide charges during the stress off period. A mathematical model is also established to explain this phenomenon.

With alternative gate dielectrics' introduction, new issues associated with these materials and device structures are also raised. Those issues need to be studied in detail before fully incorporation of new materials. Compared with SiO_2 devices, the NBTI degradation of HfO_2 has a similar trend. However, it is found that they have different

frequency response than the SiO_2 devices. This difference is later found due to the traps inside the gate dielectrics. Detailed studies show that NBTI degradations at dc stress and dynamic stress conditions have different temperature acceleration factors due to the bulk traps. The disappearance of this difference by insetting a detrapping period further proves this observation.

As we enter the ultrathin gate dielectrics regime, the electron tunneling mechanisms behind the gate dielectrics breakdown shift. Consequently, gate dielectrics breakdown mode also shifts from the clear-detected hard breakdown to the noisy soft breakdown. Thus new lifetime extrapolation models are needed. The progressive breakdown of ultrathin SiO_2 is studied by a two-step test methodology. By monitoring the degradation of the progressive breakdown path in terms of the activation energy, the voltage acceleration factor, two kinds of breakdown filaments, the stable one and the unstable one, were studied. The stable filament is found to be a breakdown filament independent of the original breakdown filament, and the unstable filament is the continuing degradation of the original filament.

EXPERIMENTAL STUDY OF BIAS TEMPERATURE INSTABILITY AND PROGRESSIVE BREAKDOWN OF ADVANCED GATE DIELECTRICS

By

Baozhong Zhu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:
- Professor Joseph Bernstein, Chair/Advisor
- Dr. John Suehle, Co-Advisor
- Professor Ichiro Takeuchi
- Professor Carol Smidts
- Professor Ali Mosleh
- Professor Martin Peckerar

© Copyright by
Baozhong Zhu
2005

Dedication

To my parents and my wife

Acknowledgements

At first, I would like to give my sincere thanks to Prof. Joseph Bernstein. He is the one who led me into the field of microelectronics reliability engineering. He has given me the greatest opportunity to explore a whole new world.

At the same time, I would like to thank my co-advisor Dr. John Suehle at National Institute of Standards and Technology. For the past three years, Dr. Suehle has been not only an advisor in the academic study and technical research but also a great friend in life. Most important of all, he gave me a great opportunity to make full use of my practical research abilities.

Special thanks go to Dr. Eric Vogel at NIST for technical discussions and support. Additional thanks would also go to Dr. Dawei Heh at University at Maryland, Dr. Yuan Chen at JPL, Dr. Jinping Han at Infineon, and Dr. Sang-Mo Koo at NIST for their great help.

I would also like to thank my parents for giving me strength and encouragement. Great thanks would also go to my brother and sister; they have been sharing my part of responsibilities to support our whole family.

An accomplishment as this could not be made without the great support of my wife. I appreciate her tolerance of a PhD at home.

Table of Contents

Dedication ii
Acknowledgements iii
Table of Contents iv
List of Figures vi
Chapter 1 Introduction 1
  1.1 Brief history of semiconductor devices 1
  1.2 Scaling of MOS devices 1
  1.3 Reliability requirements of gate dielectrics 2
Chapter 2 MOSFET Basics 5
  2.1 Basics of MOSFETs 5
    2.1.1 I_d-V_g Characteristics 5
    2.1.2 Simplified Device Fabrication 6
  2.2 Device Quality Characterization 8
    2.2.1 Capacitance Voltage Measurement (CV) 8
    2.2.2 The Dielectric Tunneling Current Characterization (IV) 10
    2.2.3 Charge Pumping Measurement (CP) 12
    2.2.4 Subthreshold Swing (SS) 14
Chapter 3 Gate Dielectrics Integrity 19
  3.1 Introduction 19
  3.2 Reliability Engineering Basics 19
    3.2.1 Definitions and Basic Concepts 19
    3.2.2 Common Reliability Functions and Bathtub Curve 20
    3.2.3 Accelerated Testing 22
  3.3 Ultrathin Gate Dielectric Reliability 23
    3.3.1 Physical Mechanisms of Dielectric Breakdown 23
    3.3.2 Dielectrics Stress Methodologies 27
    3.3.3 Lifetime Extrapolation Models 29
  3.4 Statistics and Percolation Theory of Breakdown 32
Chapter 4 Negative Bias Temperature Instability 40
  4.1 Introduction 40
  4.2 Static NBTI 41
    4.2.1 Degradation Phenomena 41
    4.2.2 NBTI Mechanisms 44
    4.2.3 Device Lifetime Extrapolation 46
    4.2.4 NBTI Impact on Circuits 47
  4.3 Dynamic NBTI of Ultrathin SiO_2 48
    4.3.1 Introduction 48
    4.3.2 Devices and Experimental Setup 49
    4.3.3 Results and Discussion 49
    4.3.4 Conclusion 55
  4.4 Dynamic NBTI of HfO_2 Devices 55
    4.4.1 Introduction 55

4.4.2 Device and Experimental Setup ... 56
4.4.3 Results and Discussion ... 57
4.4.4 Conclusion ... 63
Chapter 5 Progressive Breakdown of Ultrathin Gate Dielectrics ... 77
5.1 New Breakdown Phenomena of Advanced Gate Dielectrics ... 77
5.1.1 Hard Breakdown and Soft Breakdown ... 77
5.1.2 Definition of Failure ... 78
5.1.3 New Reliability Methodology ... 78
5.1.4 Impact of Soft Breakdown ... 80
5.1.5 Dielectric Breakdown of HfO_2 ... 81
5.2 Progressive Breakdown of SiO_2 ... 83
5.2.1 Introduction ... 83
5.2.2 Devices and Experimental Setup ... 83
5.2.3 Results and Discussion ... 84
5.2.4 Conclusion ... 89
Chapter 6 Conclusion ... 103
Appendices ... 105

List of Figures

Fig. 1.1. The scaling of device minimum feature size………………………………...…4
Fig. 1.2. The scaling of the equivalent oxide thickness of different applications……4
Fig. 2.1. The three regions of a MOSFET……………………………….……………..17
Fig. 2.2. The CV curve of a MOS capacitor……………………………………………18
Fig. 2.3. The peak current vs. the frequency……………………………………………18
Fig. 3.1. The bathtub curve of the product lifetime……………………………………38
Fig. 3.2. The comparison of 1/E model and *E* model…………….…………………….38
Fig. 3.3. The cell based percolation theory……………………………………………39
Fig. 3.4. The relationship between Weibull slope and gate oxide thickness………39
Fig. 4.1. The electric field at different technology generations…………..…………65
Fig. 4.2. The NBTI and PBTI of nMOSFET and pMOSFET…………….…………65
Fig. 4.3. The band diagram of flat band and inversion conditions of nMOSFETs and pMOSFETs………………………………………………………………………………66
Fig. 4.4. The electrochemical reaction model of NBTI…………...…………………66
Fig. 4.5. The experimental set up for NBTI study………………..……………………67
Fig. 4.6. The stress waveforms for the pulsed NBTI study……………………………67
Fig. 4.7. The frequency dependence of ΔV_{th}……………………………………………68
Fig. 4.8. ΔV_{th} vs. frequency, under negative unipolar bias stresses…………………68
Fig. 4.9. The frequency dependence of ΔI_{on}...………………………………………69
Fig. 4.10. The recovery of I_{on} at different temperatures……………………………69
Fig. 4.11. The activation energies extrapolated from data in Fig. 4.10……………70
Fig. 4.12. ΔV_{th} under positive unipolar stress with different frequencies…………70
Fig. 4.13. The annealing of ΔI_{cp} and ΔV_{th}……………………………………………71
Fig. 4.14. ΔV_{th} under different stress conditions with a 10 kHz frequency…………71
Fig. 4.15. The model fitting of ΔI_{on} with different stress conditions and different oxide thickness……………………………………………………………………………72
Fig. 4.16. The measured time constants related with the stress on and stress off period………………………………………………………………………………………72
Fig. 4.17. Time evolution of ΔV_{th} of HfO_2 devices and the control SiO_2 sample…73
Fig. 4.18. Comparison of the frequency dependence of ΔV_{th} of HfO_2 devices and the control SiO_2 sample…………………………………………………………….........73
Fig. 4.19. Comparison of the frequency dependence of normalized ΔV_{th} of HfO_2 devices and the control SiO_2 sample…………………………………………….…...74
Fig. 4.20. The two charge pumping measurements. ………….……………………..74
Fig. 4.21. The frequency dependence of generations of interface traps and bulk traps ……………………………………………………………………………………………75
Fig. 4.22. Different temperature acceleration factors of charge pumping currents due to interface traps and bulk traps……………………………………………………75

Fig. 4.23. The temperature acceleration of generations of interface traps and bulk traps. ……………………………………………………………………………76
Fig. 4.24. The charge pumping current due to the bulk traps with and without detrapping period………………………………………………………………76
Fig. 4.25. The temperature acceleration of ΔV_{th} of HfO_2 devices and control SiO_2 devices at two stress conditions. ……………………………………………………77
Fig. 5.1. The soft breakdown of gate dielectrics……………………………………91
Fig. 5.2. The schematic structure of the HfO_2 device. ………………………………91
Fig. 5.3. Constant voltage stress of the HfO_2 devices………………………………92
Fig. 5.4. Lifetime distribution of HfO_2 devices……………………………………92
Fig. 5.5. Comparison of the temperature dependence of HfO_2 and SiO_2 devices…93
Fig. 5.6. Comparison of the voltage acceleration of HfO_2 and SiO_2 devices………93
Fig. 5.7. The substrate injection stress of the HfO_2 device…………………………94
Fig. 5.8. The influence of substrate hot electron injection to the device lifetime distribution………………………………………………………………………94
Fig. 5.9. The substrate hot electron injection of the HfO_2 device……………………95
Fig. 5.10. Stable soft breakdown of the gate oxide…………………………………96
Fig. 5.11. Unstable soft breakdown of the gate oxide. ………………………………97
Fig. 5.12. The digital and analog phases……………………………………………98
Fig. 5.13. The correlation between the length of the analog and digital phase of the unstable soft breakdown ……………………………………………………………98
Fig. 5.14. The correlation between the first breakdown and final breakdown of stable soft breakdown………………………………………………………………………99
Fig. 5.15. The correlation between the first breakdown and final breakdown of unstable soft breakdown……………………………………………………………99
Fig. 5.16. The voltage acceleration of the unstable soft breakdown……….………100
Fig. 5.17. The temperature dependence of the unstable soft breakdown…………100
Fig. 5.18. The temperature dependence of the unstable, stable soft breakdown and the first breakdown……………………………………………………………………..101
Fig. 5.19. The voltage acceleration of the residual time for unstable, stable soft breakdown and the first breakdown…………………………………………………101
Fig. 5.20. SHEI test for the unstable breakdown filament…………………………102
Fig. 5.21. SHEI test for the stable breakdown filament……………………………102
Fig. 5.22. The influence of the current compliance to the consequent stress leakage current………………………………………………………………………………103
Fig. 5.23. The influence of first breakdown hardness to the residual time……….103

Chapter 1

Introduction

1.1 Brief history of semiconductor devices

W. Shockley, J. Bardeen and W. H. Brattain invented the bipolar transistor in 1947. After that, great efforts were invested in the field of semiconductors. In 1959, J. Kilby demonstrated the concept of Integrated Circuits (IC). D. Kahng and M. M. Attala fabricated the first Metal Oxide Semiconductor Field Effect Transistor in 1960. These two events provided the basis for the microelectronics industry evolution. Since then, the MOSFET has become the most important part for very large-scale integrated circuits. By 1990, the number of devices manufactured on a chip had grown from 100,100 devices per chip to more than 32 million devices per chip. Thus, the era of ultra large-scale integration has begun.

1.2 Scaling of MOS devices

The increasing of device density has been accompanied by shrinking minimum feature size, which had decreased from 25 μm at 1960s to 0.5 μm at 1990s', and 0.25 μm at year 2000, as shown by Fig. 1.1 [1, 2].

The scaling of device dimensions (channel length, equivalent oxide thickness, and junction depth) largely increased the device density and reduced the transistor cost. Initially, the industry followed the constant voltage scaling. With this scaling mode,

the oxide voltage is kept constant whereas device dimensions are reduced. However, this resulted in large oxide field, which degraded the oxide integrity. Later, a constant electrical field scaling method was used. However, this made integration with different technologies difficult, as they would have different operating voltages. Practically, the industry follows a compromised scaling model of the constant voltage and electrical field models. Further process improvements and new structures are introduced to overcome those difficulties.

Suggested by the International Technology Roadmap for Semiconductors (ITRS) [3], the introduction of alternative gate dielectrics is predicted from 2005 to 2007 depending on the technology application, as shown by Fig. 1.2. Due to the stringent leakage requirements for low standby power devices, the leakage current specifications cannot be met with conventional gate dielectrics in the sub 1.5 nm regime. Thus, the high-*k* dielectrics are likely first introduced in these applications. These alternative dielectrics will be deposited by either a chemical vapor deposition technique [4, 5], physical vapor deposition[6], jet vapor deposition [7] or molecular beam epitaxy [8, 9]. More work and new technologies are still needed.

1.3 Reliability requirements of gate dielectrics

With shrinking gate dielectrics, the reliability requirements become more and more difficult to maintain. New physical mechanisms and phenomena are discovered and new challenges arise. As we enter the ultrathin gate dielectrics regime, the electron tunneling process is dominated by the direct tunneling instead of FN tunneling [10]. The gate dielectrics breakdown mode also shifts from the clear-detected hard breakdown to the noisy soft breakdown [11, 12]. Thus new lifetime

extrapolation models are needed. At the same time, some issues, which have been minor in the past, begin to show bigger impact, such as the Negative Bias Temperature Instability issue. Therefore, new study methods and understandings are required.

With alternative gate dielectrics' introduction, new issues associated with these materials and device structures are also raised. Traps inside the bulk dielectrics and near the interface cause instability to the threshold voltage of MOSFETs and impose new risk to the reliability of devices [13]. New gate dielectric breakdown mechanism has also been suggested [14]. Those issues need to be studied in detail before fully incorporation of new materials.

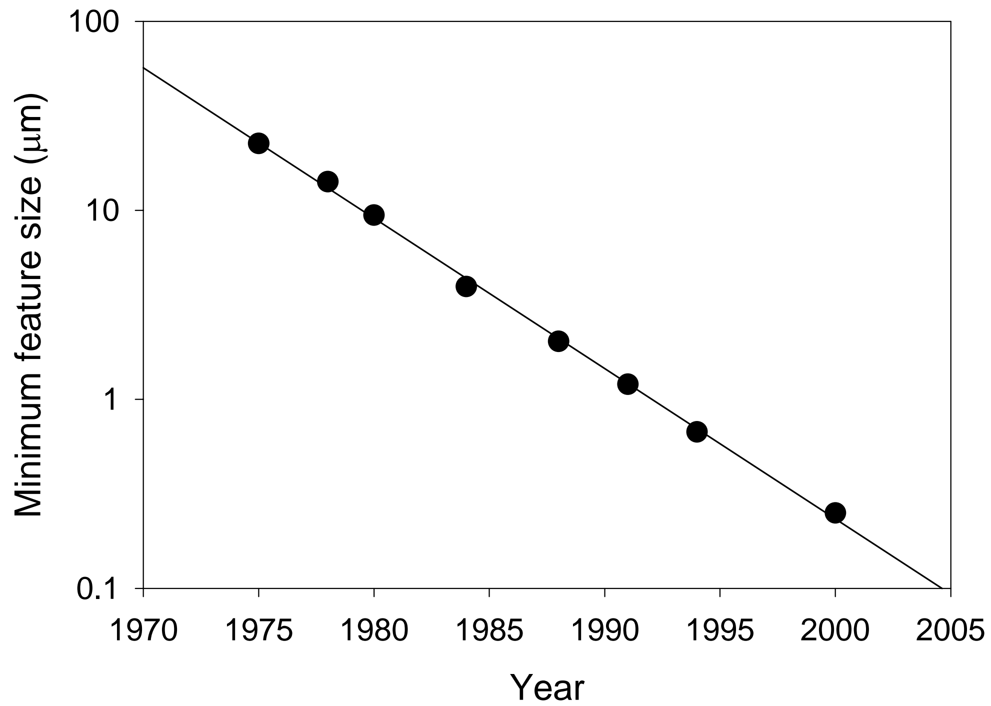


Fig. 1.1. The scaling of device minimum feature size.

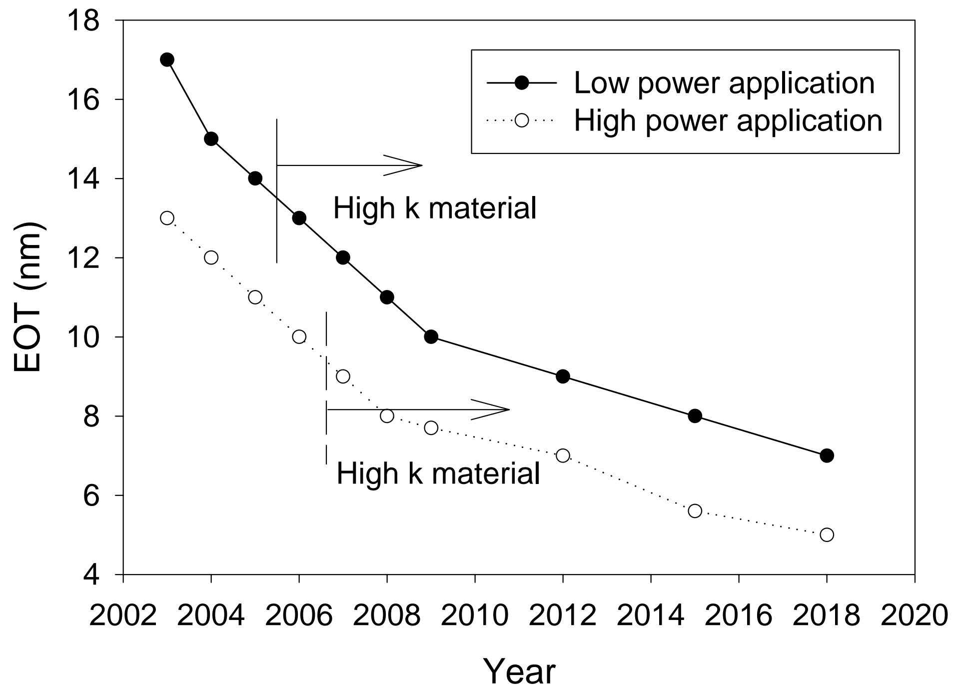


Fig. 1.2. The scaling of the equivalent oxide thickness of different applications.

Chapter 2

MOSFET Basics

2.1 Basics of MOSFETs

The Metal Oxide Semiconductor Field Effect Transistor (MOSFET) is a four terminal field controlled device. The *n*-type MOSFET (nMOSFET) 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. The gate is separated from the semiconductor by the gate oxide. The basic structure of an nMOSFET and the corresponding circuit symbol are shown in Fig. 2.1. The basic device parameters are the channel length (L), which is the distance between the two n^+-p junctions, the channel width (W), the dielectric thickness t_{OX}, the junction depth and the substrate doping density (N_A).

2.1.1 I_d-V_g Characteristics

Many models developed for the simulation of MOSFET's operation. We will briefly describe the MOSFET's operation based on the quadric model, which is relative simple and gives good approximation [15].

The operation of a MOSFET can be divided into three regions: cutoff, linear, and saturation, as shown in Fig. 2.1. At the cutoff region, the gate voltage V_g is smaller

than the threshold voltage of the device V_{th}, which is the turn on gate voltage of the device. The drain current I_d is zero.

When the gate voltage increases to $V_g \geq V_{th}$, and $V_g - V_{th} \geq V_d \geq 0$, the device is in the linear region. The drain current Id is given by

$$I_d = \frac{W}{L} \mu_{eff} C_{OX} (V_g - V_{th}) V_d .$$

Here, μ_{eff} is the effective carrier mobility.

The threshold voltage is given by

$$V_{th} = V_{FB} + 2\phi_F + \frac{\sqrt{2\varepsilon_{Si} q N_A (2\phi_F)}}{C_{OX}} .$$

The device turns into the saturation regime when $V_g \geq V_{th}$ and $V_d \geq V_g - V_{th}$. The drain current is given by

$$Idsat = \frac{W}{2L} \mu_{eff} C_{OX} (V_g - V_{th})^2 .$$

2.1.2 Simplified Device Fabrication

A brief description of MOSFET fabrication is given using nMOSFET as an example. The starting material is a p-type, <100> oriented, lightly doped polished silicon wafer. The <100> orientation is preferred over <111> because it has an interface trap density that is about one-tenth that of <111>. The first step is to form the oxide isolation region using LOCOS technology. A thin pad oxide is thermally grown, followed by a silicon nitride deposition. A photoresist mask defines the active device area and boron channel stop nitride layer not covered by the photoresist mask

is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide, field oxide, then the nitride layer is removed and to drive in the boron implant.

The second step is to grow the gate oxide and to adjust the threshold voltage. The composite nitride-oxide layer over the active device area is removed, and a thin gate oxide layer is grown. For an enhancement device, boron ions are implanted in the channel region to increase the threshold voltage to a predetermined value. For a depletion device, arsenic ions are implanted in the channel region to decrease the threshold voltage.

The third step is to form the gate. Polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20 to 30 Ω/sq. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm. For shorter devices, a polycide (a composite layer of metal silicide and polysilicon) can be used as the gate material to reduce the sheet resistance to about 1 Ω/sq.

The fourth step is to form the source and drain. After the gate is patterned, it serves as a mask for the arsenic implantation to form the source and drain, which are self-aligned with respect to the gate.

The last step is metallization. A phosphorus-doped oxide is deposited over the entire wafer and is followed by heating the wafer to give a smooth surface topography. Contact windows are defined and etched. A metal layer, such as aluminum, is then deposited and patterned. The gate contact is usually made outside the active device area to avoid possible damage to the thin gate oxide.

2.2 Device Quality Characterization

2.2.1 Capacitance Voltage Measurement (CV)

Maintaining the quality and reliability of gate oxides is one of the most critical and challenging tasks in any semiconductor fab. Vigorous characterization and monitoring is critical for maintaining gate oxide uniformity and quality across the wafer.

Many electrical techniques have been developed over the years to characterize gate oxide quality. CV test offers a wealth of device and process information, including bulk, interface charges, and many MOS device parameters. CV measurements are typically made on a capacitor, manufactured with the gate oxide grown at the same time under similar circumstances. From the analysis of the CV characteristics, we can get information of the fast interface state density, fixed oxide charge and slow interface state density. The shortage of CV method is the requirement of a large area capacitor and the difficulty to extrapolate the results to the actual device in the product line. The generated CV curve is characterized into three regions, as indicated by Fig. 2.2.

Accumulation region

For a p-substrate MOS capacitor, the accumulation region of the CV curve is observed when negative voltages are applied to the gate. The negative polarity causes majority carriers to be attracted toward the gate. Because the oxide is a good insulator, these holes accumulate at the oxide/substrate interface.

A CV test measures the oxide capacitance in the strong accumulation region; the CV curve slope is essentially flat. Therefore, the oxide thickness can be extracted

from the oxide capacitance. However, the CV curve for a very thin oxide often does not saturate to a flat slope. In that case, the measured oxide capacitance differs from the true oxide capacitance.

Depletion region

nMOS capacitor differs from a parallel-plate capacitor as the gate voltage moves toward positive values. Roughly, the following occurs:

1) The positive gate electrostatically repels holes from the oxide/substrate interface.

2) A carrier-depleted area forms beneath the oxide, creating a depletion region.

As a result, the high frequency CV analyzer measures two capacitances in series, the oxide capacitance and the depletion capacitance.

As the gate voltage becomes more positive, the following occurs:

1) The depletion region penetrates more deeply into the substrate.

2) The depletion capacitance becomes smaller, and the total measured capacitance becomes smaller consequently.

Inversion region

As the gate voltage increases beyond the threshold voltage, carrier generation and recombination move toward the carrier generation. The positive gate voltage both generates electron-hole pairs and attracts electrons, which are minority carriers, towards the gate. Again, because the oxide is a good insulator, these minority carriers accumulate at the oxide/substrate interface. Thus forms an inversion layer. Above a certain positive gate voltage, most of the available minority carriers are in the

inversion layer, and further gate voltage increases do not further deplete the semiconductor. Thus, the depletion region reaches a maximum depth.

However, the minority carrier generation is much slower than the MHz frequency range of high frequency CV measurements [16]. Therefore, once the depletion region reaches a maximum depth, the capacitance measured by the high frequency CV (HFCV) is still based on the majority carrier position and distribution. Thus, the capacitance measured by HFCV is the oxide capacitance in series with the maximum depletion capacitance, which is often referred to as the minimum capacitance.

When measuring the high frequency capacitance while sweeping the gate voltage "quickly", deep depletion occurs [17]. Quickly means that the gate voltage must be changed fast enough so that the structure is not in thermal equilibrium. One then observes that, when ramping the voltage from flatband to threshold and beyond, the inversion layer is not or only partially formed. This occurs since the generation of minority carriers cannot keep up with the amount needed to form the full inversion layer.

2.2.2 The Dielectric Tunneling Current Characterization (IV)

As the oxide thickness continues scaling, the reliability issues as well as the physical mechanisms behind also change. One of the new phenomena of ultrathin gate oxides is the electron tunneling mechanism changes from the Fowler-Nordheim (FN) tunneling to the direct tunneling, as shown in Fig. 2.3. FN tunneling is a quantum mechanical tunneling process. The electrons will inject into the conduction band of dielectric through a triangular barrier.

The tunneling density is given by

$$J_{FN} = AE_{OX}^2 \exp\left(-\frac{B}{E_{OX}}\right).$$

Here

$$A = \frac{q^3}{16\pi^2 \eta \phi_B}, \quad B = \frac{4(2m_{OX})^{1/2}\phi_B^{3/2}}{3q\eta}.$$

Direct tunneling is also a quantum mechanical tunneling process. It occurs when electrons tunnel through the gate oxide region directly from the gate to the channel region. The direct tunneling current density is given by

$$J_{DT} = \frac{AE_{OX}^2}{\left[1-\left(\frac{\phi_B - qV_{OX}}{\phi_B}\right)^{1/2}\right]^2} \exp\left[-\frac{B}{E_{OX}}\frac{\phi_B^{3/2}-(\phi_B - qV_{OX})^{3/2}}{\phi_B^{3/2}}\right]$$

Here A and B have the same definition as in the FN tunneling equation.

For thick oxides, which are larger than 5.0 nm, the FN tunneling is the dominating transport mechanism. However, when with the scaling of gate oxides, the conduction mechanism is dominated by the direct tunneling, where the electrons inject into the conduction band of anode through a trapezoidal barrier. This is true especially with ultrathin oxides with V_{OX} less than ϕ_B, which is about 3.0 V [10, 18, 19].

Another important tunneling process is trap assisted tunneling [20-22]. Trap assisted tunneling occurs when electrons tunnel through the oxide into traps and then from the traps into the silicon substrate.

Besides those three tunneling processes, there are other factors that can influence the behavior of the tunneling current. One of them occurs due to the gate and drain overlap region [23]. The tunneling process through the valence band is also important

especially in ultrathin gate oxides [10]. Electron hopping, field emission and the Poole Frenkel emission are other factors that need consideration [16].

2.2.3 Charge Pumping Measurement (CP)

The Si/SiO_2 interface in MOSFETs plays an important role in determining the device parameters and affects reliability and lifetime of the device. Thus it is important to measure and characterize the interface properties in order to understand the origin and physical properties of the interface states in a MOS system.

Depending on the distance from the interface, interface states can be divided as fast states and slow states. The fast states are those locating at or near the Si/SiO_2 interface and therefore they can readily exchange charge with the Si substrate by capturing or emitting charge carriers. The amount of charge captured by the fast states is a function of the surface potential and therefore a function of the gate voltage. The slow states are located in the oxide within a short distance from the interface. The charge carriers can tunnel through the oxide. Thus the tunneling time constant is the time constant for the slow state to charge up or discharge. The tunneling time constant increases exponentially with the tunneling distance.

The charge pumping method uses a measurable substrate current due to the recombined trapped charge induced by repeatedly pulsing the gate from accumulation to inversion to characterize interface states in MOSFETs. This method does not rely on admittance measurements, and is well suited for use on small transistors. A mean interface trap density representative of values between flatband and threshold is determined according to measured charge pumping current as a function of gate pulse frequency. A distribution of interface trap density between flatband and threshold can

also be obtained by measuring charge pumping current as a function of gate pulse rise and fall times.

The main advantage of the charge pumping measurement is that the measurements are done on the actual transistor without needing a separate test device. The results of this technique are relatively more precise and can be used to obtain quick results in a simple experiment or detailed results with sophisticated modifications. The sensitivity of this technique permits the measurement on small geometry transistors present in modern VLSI technology.

A model describing the behavior of both slow and fast states has been developed [24-27]. From Shockley-Read-Hall (SRH) statistics, the occupancy factor, F, of a trap as a function of time can be described as

$$\frac{dF}{dt} = c_n(E) - F\left[c_n(E) + c_p(E)\right]$$

Where c_n is the capture rate of electrons and c_p is the capture rate of holes. In this equation, the emission of electrons and holes are neglected because capture processes dominate the charge pumping.

The capture rate of electrons and holes is given by $c_n = n_s \sigma_n v_{th}$ and $c_p = p_s \sigma_p v_{th}$. Here, n_s and p_s are surface concentration of electrons and holes, σ_n and σ_p are capture cross section of electrons and holes, v_{th} is the thermal velocity.

For fast states, the filling fraction, which represents a state's change in occupancy factor during one gate pulse period, is defined as

$$\Delta F_{FS} = F_{\max} - F_{\min} = \frac{\left(1 - e^{-\frac{c_n}{2f}}\right)\left(1 - e^{-\frac{c_p}{2f}}\right)}{1 - e^{-\frac{c_n}{2f} - \frac{c_p}{2f}}},$$

Where $F_{\max}$ and $F_{\min}$ are the maximum and minimum value of F during one gate pulse, and f is the gate pulse frequency.

The charge pumping current is obtained as

$$I_{cp}(f) = qAf \int_{E_{low}}^{E_{high}} N_{FS}(E) \Delta F_{FS} dE,$$

Here A is the area of the device, E_{high} and E_{low} are the maximum and minimum positions in the energy of the Fermi level within the bandgap during one gate pulse. NFS is the density of the fast states.

Apparently, the measured charge pumping current will be roughly a linear function of the signal frequency *f*, as proved by Fig. 2.3.

2.2.4 Subthreshold Swing (SS)

Another simpler and more convenient interface traps assessment method is the subthreshold swing evaluation [28]. The subthreshold swing, S, is defined as the gate voltage required to vary the subthreshold current I_d by one decade.

$$S = \frac{\partial V_g}{\partial \log I_d}.$$

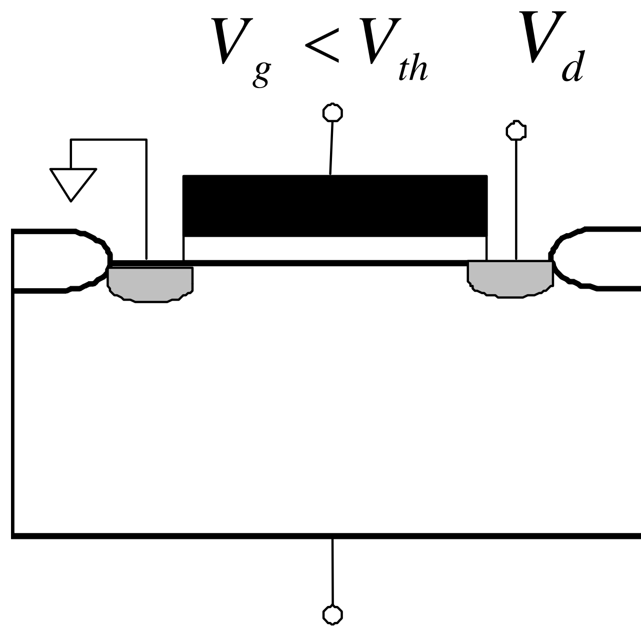
It can be further simplified as [29]:

$$S = \frac{kT}{q} \ln(10)(1 + C_d / C_i)$$

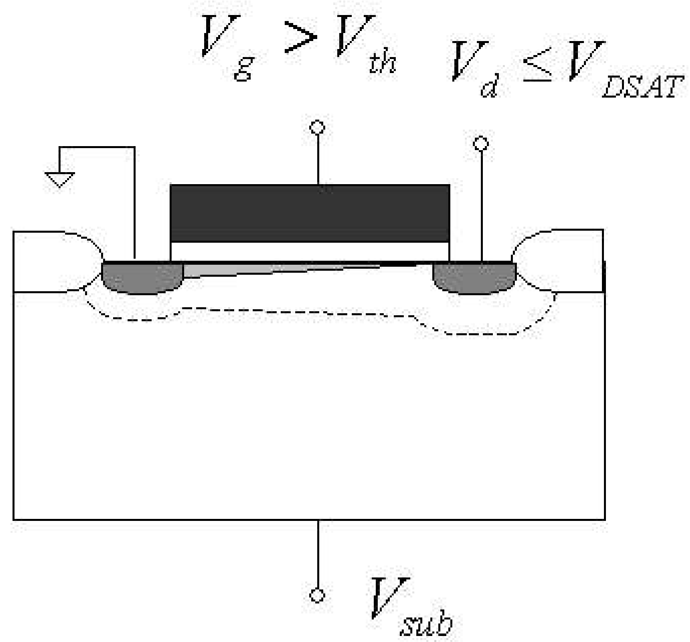
If there is a large interface state density D_{it}, an interface state capacitance $C_{it} = qD_{it}$ is in parallel with C_d. Consequently, C_d in the above equation will increase to $C_d + C_{it}$.

Thus, the change of subthreshold swing ΔS is a direct evidence of the change of interface state density [16]:

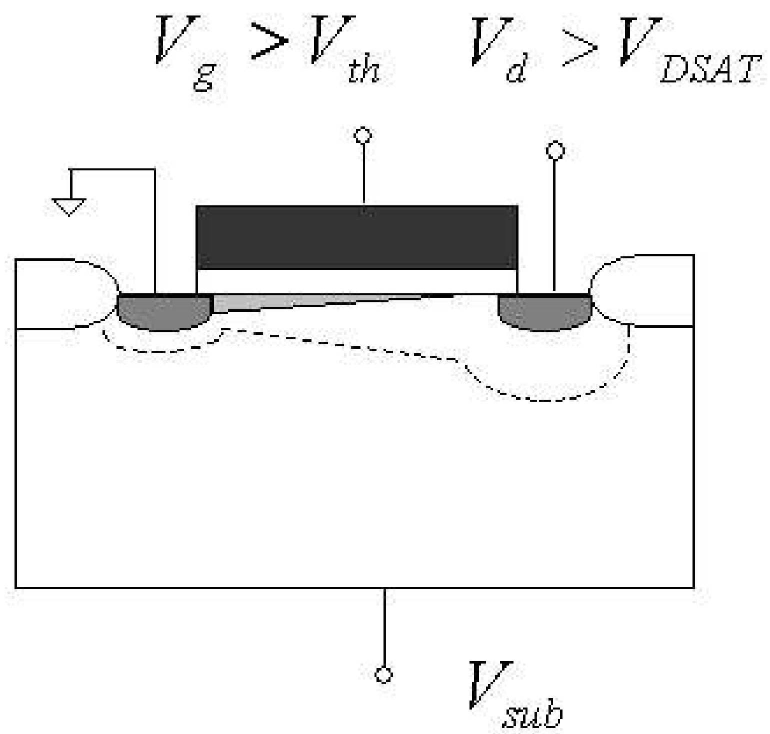
$$\Delta S = \frac{kT}{q} \ln(10) \frac{qD_{it}}{C_i}.$$



Cutoff Region



Linear Region



Saturation Region

Fig. 2.1. The three regions of a MOSFET.

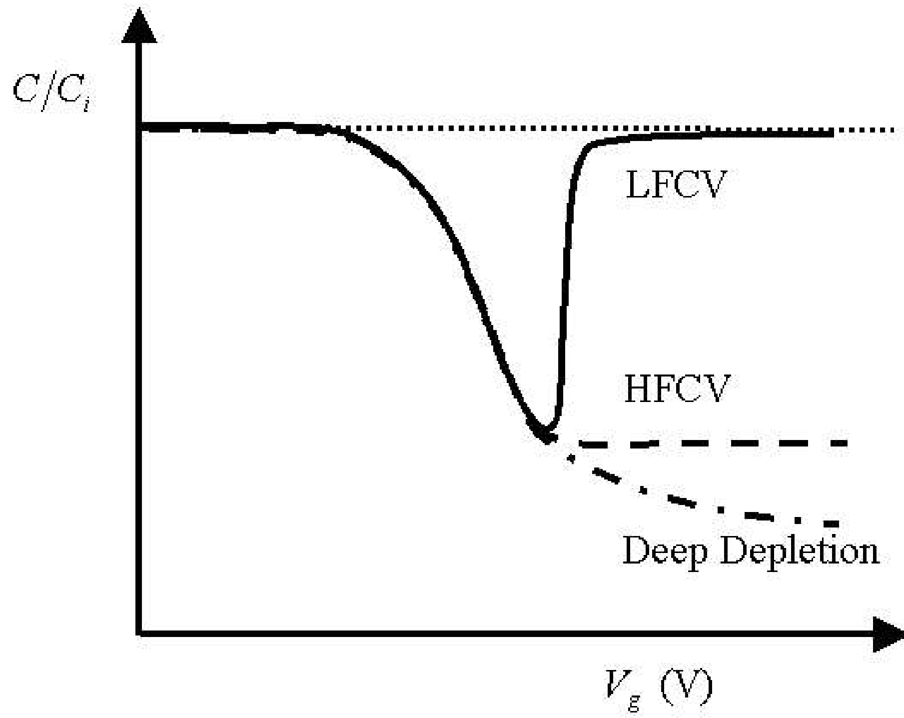


Fig. 2.2. The CV curve of a MOS capacitor.

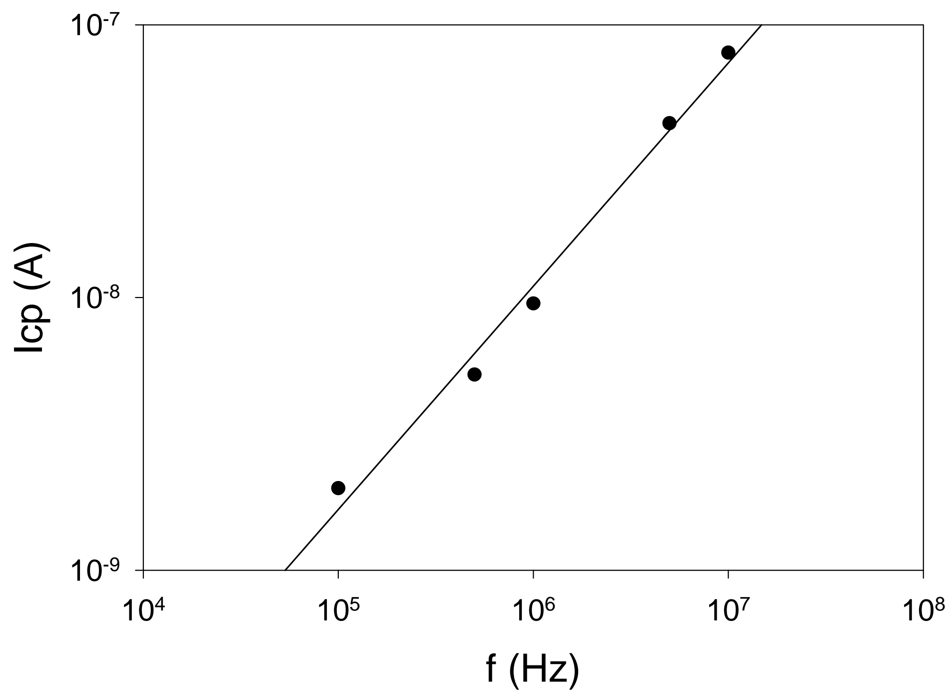


Fig. 2.3. The peak current vs. the frequency

Chapter 3

Gate Dielectrics Integrity

3.1 Introduction

Driven by the higher performance and circuit density, the scaling of microelectronics continues with gate leakage current and the reliability of gate dielectrics as two critical limiting factors. Due to the characteristics of reliability engineering, it is crucial to use correct physical models, statistics and acceleration factors to extrapolate the device lifetime from the accelerated tests. Thus, accelerated tests and lifetime extrapolations based on those tests must be robust, valid, and performed on statistically significant sample sizes.

3.2 Reliability Engineering Basics

3.2.1 Definitions and Basic Concepts

Reliability: The probability that a component will perform a specific function under specific conditions for a specific period of time.

Thus, the objectives of reliability engineering are to meet/exceed customer expectations, eliminate reliability concerns prior to qualifications and volume ramp, and allow more aggressive performance without increasing risk of failure.

Several functions are defined before the study of dielectrics reliability. Those functions are commonly used in the field of statistics as well as reliability engineering [30, 31].

Reliability function $R(t)$ is the fraction of population that survives until time t, it is also known as the survival function.

Failure probability function $F(t)$ is the fraction of population that fails before time t. Related with the reliability function, we have $F(t) = 1 - R(t)$.

Probability density function $f(t)$ mathematically describes the number of devices failing between a time interval dt, as shown by $f(t) = \frac{dF(t)}{dt}$.

The failure rate $\lambda(t)$ is defined as the ratio of probability that the failure occurs in this interval divided by the interval length Δt:

$$\lambda(t) = \frac{R(t) - R(t + \Delta t)}{\Delta t R(t)}.$$

Hazard rate or instantaneous failure rate $h(t)$ describes the rate at which a unit is expected to fail, given that it has survived time t.

$$h(t) = \lim_{\Delta t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t R(t)} = \frac{1}{R(t)} \frac{-dR(t)}{dt} = \frac{f(t)}{R(t)}.$$

The failure rate $\lambda(t)$ and hazard rate $h(t)$ are mathematically different, however, they are often used interchangeably in conventional reliability engineering.

3.2.2 Common Reliability Functions and Bathtub Curve

Exponential distribution: $f(t) = \lambda e^{-\lambda t}$, λ is the failure rate.

Weibull distribution: $f(t)=\left(\frac{\beta}{\eta}\right)\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}$, β is the Weibull slope and shape parameter, η is the characteristic lifetime.

Lognormal distribution: $f(t)=\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Ln(t)-Ln(\mu)}{\sigma}\right)^{2}}$, σ is the shape parameter and μ is the median time-to-fail.

Reliability engineers often describe the lifetime of a population of products using a graphical representation called the bathtub curve, as shown in Fig. 3.1 [30, 32]. The bathtub curve consists of three periods: an infant mortality period with a decreasing failure rate followed by a normal life period (also known as "useful life") with a low, relatively constant failure rate and concluding with a wear-out period that exhibits an increasing failure rate. The bathtub curve failure rate curve is really the addition of two kinds of failure rates: the extrinsic failure rates and the intrinsic failure rates.

The bathtub curve does not depict the failure rate of a single item, but describes the relative failure rate of an entire population of products over time. Some individual units will fail relatively early (infant mortality failures), others will last until wear-out, and some will fail during the relatively long period typically called normal life. Failures during infant mortality are highly undesirable and are always caused by defects and blunders: material defects, design blunders, errors in assembly, etc. Normal life failures are normally considered to be random cases of "stress exceeding strength." However, many failures often considered normal life failures are actually infant mortality failures. Wear-out is a fact of life due to fatigue or depletion of materials. A product's useful life is limited by its shortest-lived component. A product

manufacturer must assure that all specified materials are adequate to function through the intended product life.

Note that the bathtub curve is typically used as a visual model to illustrate the three key periods of product failure and not calibrated to depict a graph of the expected behavior for a particular product family. It is rare to have enough short-term and long-term failure information to actually model a population of products with a calibrated bathtub curve.

3.2.3 Accelerated Testing

Why do we need the accelerated testing?

During the technology development, it is hard to quantify the bathtub curve as the time until a failure occurs is over a period of months to years. To solve this problem, we use accelerated testing to simulate the filed use reliability degradation mechanisms. The validity of accelerated testing is based on two assumptions:

(1) The failure mechanisms can only be quantified if they are observed.

(2) Accelerated testing is a "clock multiplier" to observe failures in a reasonable time.

The following table shows the common variables and their respective affecting fields in the semiconductor industry [31].

Variables	Affecting fields
Voltage	Dielectrics, transistor
Temperature	Dielectrics, transistor, interconnect
Current	Interconnect
Humidity	Packaging
Mechanical stress/vibration	Packaging

The time-to-failure modeling with accelerated testing is an exact application of accelerated tests. With the help of accelerated testing, the time-to-failure (*TTF*) of a sample can be modeled as

$$TTF = A * \xi^{-N} * e^{-\frac{E_A}{k_B T}}.$$

Here ξ is the generalized stress, which brings about an irreversible change in the material's properties and causes fails, E_A is the Arrhenius activation energy, *A* is a process dependent coefficient, and *N* is the stress dependent coefficient.

3.3 Ultrathin Gate Dielectric Reliability

3.3.1 Physical Mechanisms of Dielectric Breakdown

There are two categories of oxide failures: the extrinsic and intrinsic failures. The extrinsic failures occur as a result of defects in the oxide or problems associated with the fabrication of the oxide. Those failures tend to occur early in the life of the oxide. They also have a different distribution than intrinsic failures. This can result in a bimodal distribution if the poor quality oxides are included in the accelerated tests. Intrinsic failures are failures due to breakdown in the defect-free oxide. Intrinsic failures determine the true lifetime of a high quality oxide.

The exact physical mechanism of oxide breakdown is still not known [33]. The general idea is that a driving force such as the applied voltage or the tunneling electron current creates defects in the gate oxide. The effects accumulate with time and eventually reach a critical density and trigger a sudden loss of the dielectric properties. Thus the defect generation is critical to understand the mechanisms of

oxide breakdown. Currently, there are three main theories established for the defect generation, the anode hole injection model, the anode hydrogen release model, and the voltage driven model.

Anode hole injection model (AHI)

This model was established on the evidence that tunneling electrons must be the driving force for wearout and breakdown in ultrathin SiO_2[34, 35]. According to this model, the oxide breakdown is caused by holes injected from the anode. Electrons injected from the cathode into the oxide cause impact ionization events generating holes. Those holes inject back from the anode and trapped in the oxide near the cathode. They will distort the band diagram and increase the electric field near the cathode, and consequently enhance the electron tunneling current. In the direct tunneling regime, which is the dominating tunneling process with ultrathin gate oxides, the energy of the tunneling electrons is proportional to the applied gate voltage [36]. When considering a physical process for defect creation in thin oxides, it is important to consider the energy required for such a process to occur. At energies below those required for anode impact ionization, anode hole injection was believed to take place via surface plasma process [36, 37]. This process requires electrons with energies greater than 7 eV for hot hole injection and subsequent trapping.

The AHI model was questioned when the origin of substrate current could be linked to physical mechanisms other than tunneling holes. These mechanisms include generation recombination processes in the substrate and photo excitation processes due to photons generated by energetic electrons in the gate region[38, 39]. However, recent experimental evidence and modeling have demonstrated that AHI can indeed

be operative in ultrathin oxides at low gate voltages. It is showed that bulk defect generation is increased significantly and Q_{bd} decreased for devices with lightly doped n^+ polysilicon gate electrodes[10].

Anode hydrogen release model (AHR)

There is evidence for a defect generation mechanism involving the release of atomic hydrogen from the anode by energetic tunneling electrons [40]. It has been know that hydrogen can induce a number of defects in silicon dioxide films by internationally exposing MOS devices to hydrogen [41-43]. A trap creation process attributed to the release of atomic from the Si/ SiO_2 interface has been shown to have a threshold voltage of approximately 5.0 V [40]. This process has been shown to continue at the voltage as low as 1.2 V, which includes the regime of circuit operating voltages. There is evidence for hydrogen involvement in defect generation and breakdown. It was shown that exposure of bare SiO_2 films to atomic hydrogen radicals, in the absence of any electric field, will produce electrically active defects essentially identical to those produced by electrical stress or radiation.

The primary argument against the hydrogen release model for oxide breakdown is the observation that Q_{bd} does not appear to improve if an isotope of hydrogen is used to passivate the Si/SiO_2 interface [44]. It has been reported that deuterated oxide film has suppressed hydrogen desorption from the interface with silicon and consequently improved immunity to interfacial trap generation due to hot carrier injection.

E model

Different from the AHI and AHR models, which are current driven, the *E* model suggests that the electrical field drives the defect generation and current flowing through the oxide plays a second role at most. This model considers the interaction of the electrical field with the dipole moments associated with oxygen vacancies in the SiO_2. The dipole and electric field interaction reduces the activation energy for the Si-Si bond breakage. The oxygen vacancy is the breakdown precursor. The *E* model assumes that the tunneling currents play no role in breakdown and the acceleration factor is constant independent of field electric field or gate voltage. The activation energy required to break the bond is reduced by the dipolar energy, leading to a quantitative prediction for the field dependence of the activation energy for dielectric breakdown, which agrees well with experiments. The distribution of energies of the weak bonds could account for a wide range of observations of the temperature and field dependence of SiO_2 breakdown times, since the defect which dominates the breakdown process may change depending on stress conditions [45].

The *E* model was hypothesized based on an electric field driven defect generation process that yielded the same observed dependence of oxide lifetime on electric field [46]. This model known as the *E* model indicated that the applied electric field interacts with the weak Si-Si bonds associated with oxygen vacancies in the amorphous SiO_2 film. The applied electric field eventually breaks the weak bond and creates a permanent defect or trap. This defect is referred to as the E' center, which is a structure in the SiO_2, having an unpaired electron localized on a silicon atom, which

is bonded to three oxygen atoms. Tunneling electrons are not necessary in the E model to create defects.

The E model has attained widespread acceptance. However, the exponential dependence on field is not proof of the validity of the particular physical model. As the SHEI experiment pointed out that the charge to breakdown Q_{bd} is strongly dependent on the substrate bias, even though the oxide field is fixed. Therefore Q_{bd} is correlated with the electron energy not the electric field.

3.3.2 Dielectrics Stress Methodologies

In order to predict the lifetime of devices correctly, proper accelerated stress tests are needed. Depending on the process technologies and test goals, three stress methods are commonly used in accessing the oxide reliability [47].

Constant Voltage Stress (CVS)

The constant voltage stress (CVS) is popularly used during the sample stresses. In this method, a constant stress voltage is applied to the device gate or substrate, and the current vs. time (I-t) characteristics are recorded. The oxide breakdown is characterized by a sudden increase of the current (hard breakdown), or when the gate current noise exceeds a specified current noise level, which corresponds to a soft breakdown event (soft breakdown). This breakdown is characterized by distribution parameters, thus a sufficient sample size is required to determine those parameters accurately. At a fixed stress voltage, T_{bd} and Q_{bd} decrease with oxide thickness; therefore it is used often with ultrathin oxides tests.

Ramp I-V

In a ramp I-V the gate voltage is increased rapidly from zero until a sudden increase in the gate leakage current is measured. In this measurement, the breakdown is monitored as a function of gate bias from which a breakdown field is determined. In principle different failure modes can be detected with this method. This method is beneficial for extrinsic distribution, and improves resolution.

Constant Current Stress (CCS)

Constant current stress (CCS) is another popular stress method. As stated by the name, a constant current is applied to the device, and the voltage vs. time characteristics (V-t) is recorded. It is found that the charge to breakdown, Q_{bd} is stress current dependent, and for a fixed stress current, T_{bd} and Q_{bd} increase strongly for ultrathin oxides. Therefore, this method is often used in thicker oxides.

Q_{bd} vs. T_{bd}

Q_{bd} and T_{bd} are two breakdown parameters critical to the gate oxide breakdown study. Q_{bd}, defined as the time integrated current density that flows through the oxide until breakdown occurs is a physically meaningful quantity. However, the quantity of interest for an electronic component is the failure rate derived from the lifetime or time-to-breakdown, T_{bd}. For a constant voltage stress, $\int_0^{T_{BD}} Jdt = Q_{BD}$, where J is the instantaneous value of the current density [48]. For ultrathin oxides, the current is nearly constant until breakdown (in remarkable contrast to thicker oxides), therefore $T_{BD} = \frac{Q_{BD}}{J}$. Therefore T_{bd} and Q_{bd} will yield comparable result with ultrathin oxide breakdown.

3.3.3 Lifetime Extrapolation Models

E model

E model is an empirical model for breakdown developed by observing the electric field dependence of TDDB data [49, 50]. When the logarithm of the time-to-failure was plotted against applied electric field, a straight line was observed, i.e. $t_{BD} \propto e^{-\gamma E}$, where E is the electric field and γ is the electric field acceleration factor. Acceleration parameters for the electric field and temperature could be extracted from the model to allow extrapolation of oxide lifetime from accelerated stress conditions.

1/E model

The 1/E model is a popular breakdown model with ultrathin oxides. According to this model, a fraction of electrons entering the anode have sufficient energy to create hot holes that are injected back into oxide. The injected holes will generate defects in the oxide, and breakdown occurs when a critical hole fluency is reached. According to the 1/E model, the acceleration factor increases at smaller E_{OX}.

1/E model was proposed based on anode hole injection [35, 51]. In this case, tunneling electrons (based on FN tunneling) transferred energy to holes in the anode, where they are injected into the oxide film. An earlier version of the model was based on a feedback mechanism where injected holes became trapped and modified the oxide potential barrier to enhance additional electron injection. Eventually, the positive feedback process caused a current runaway leading to breakdown. In this case, the time-to-failure was proportional to the inverse of applied electric field, i.e. $t_{BD} \propto e^{\beta E}$, where β is the electric field acceleration factor. The reciprocal field dependence is a consequence of the FN tunneling current, which is the driving force

for the defect generation. At higher voltages (~9 V), holes can be generated in the oxide by impact ionization processes and subsequently trapped, leading to breakdown.

Both the *E* model and the 1/E model assumed a temperature dependence of the form $t_{BD} \propto e^{\frac{E_a}{kT}}$; where E_a is referred to as the thermal activation energy, *k* is Boltzmann constant, and T is the absolute temperature. E_a is due to different physical processes depending on the model assumed. In the *E* model, Ea is related to the enthalpy of activation for oxide breakdown or it is related to the temperature dependence of the hole generation coefficient and the slope of the FN tunneling characteristic in the 1/E model [52].

The Unification of E and 1/E model

The prediction of E model and 1/E model is quite different at the low electric field. However, this cannot be verified by direct test since T_{bd} is extremely long with such low fields. In recent years, there have been efforts to unify these two models. First, it is shown that the field enhanced bond breaking and hole generation and trapping process coexist and degrade oxide. At high fields 1/E and at low fields *E* model prevails [53]. Another research pointed out that the dominant carriers were different at different fields. At high fields, the holes are the dominant factor and electrons dominate the low fields [54]. Similar results showed that E and 1/E are competing processes, the dominating mechanism depends on precursor mix and hole current density [55]. The molecular physics based complementary model also indicates that field and current induced oxide degradation occur simultaneously if one reaction dominates [45].

Voltage driven model

The validity of an electric field driven model becomes questionable after the observation of results from substrate hot electron injection experiments and the thickness and polarity dependence of ultrathin gate oxide breakdown. A thermochemical wearout process should only be related to the magnitude of the electric field and the length of time it is applied to the gate, not to the amount of charge injected. The substrate hot electron injection (SHEI) experiment shows that T_{bd} is proportional to the inverse of the indicted current density [23]. Other experiments were conducted where changing the polysilicon gate doping while maintaining a fixed gate voltage can modulate the electric field [10, 56]. The results show that the lifetime is nearly independent of polysilicon doping, which is contradictory to the electric field driven model.

Even though the voltage driven model is accepted as the correct model for the ultrathin gate dielectrics, there is a still question that how to extrapolate the device lifetime based on the accelerated tests.

At first an exponential law was suggested as

$$T_{bd} \propto \exp(\gamma V_g)$$

Here γ is the stress voltage acceleration factor.

However, it is found that γ is not constant for different V_g, even for the same oxide thickness [57]. The measured T_{bd} and Q_{bd} at lower voltages are much higher than expected from the exponential law extrapolation based on high voltage data and there is a systematic concave line up of the data, leading to higher γ at lower voltages [57, 58].

To solve this problem, an empirical power law is established as

$$T_{bd} \propto V_g^{-n}$$

Based on the experimental data, n~44 is extrapolated. However, it is difficult to explain with a simple physical process. Recently, the multiple hydrogen release model is used to explain the power law dependence [59, 60]. However, it is still too early to draw conclusions.

3.4 Statistics and Percolation Theory of Breakdown

Projection of dielectric lifetime of a product from data collected by stressing test devices under accelerated test conditions requires a number of correct models and assumptions. The voltage and temperature must be accurately extrapolated from the accelerated test conditions. Scaling from the area of the test structure to that of the product must be performed with the proper model. Finally, the choice of the failure distribution must be correct, especially, since the 100 ppm level is usually specified requiring accurate estimation of the low failure rate tails.

Lognormal statistics have been used for some time to fit reliability data from accelerated life tests to induce various failure mechanisms such as electromigration and oxide breakdown [61-64]. There has been discussion concerning the validity of lognormal statistics for thin oxide breakdown. Although lognormal statistics may fit failure data over a limited sample set, it has been demonstrated that the Weibull distribution more accurately fits large samples of TDDB failure data, particularly, at lower failure rates. More importantly, lognormal statistics does not predict the observed area dependence of T_{bd} for ultrathin gate oxides. Furthermore, the Weibull

shape parameter β is experimentally observed to decrease as the oxide thickness is decreased. The model based on percolation theory has been proposed to explain the oxide thickness dependence of β and how it relates to N_{bd}, the number of defects at breakdown.

According to the percolation theory, the defects are generated randomly throughout the gate oxide film. As the defects generation continues, the defect density increases. When the defect density reaches a critical value N_{bd}, the defects will connect each other and finally can connect the both edges of the oxide film, which leads to the breakdown of the oxide. If this is true, the thinner oxide will have a lower critical defect density, and a larger area device will have a larger probability to have a conduction path with the same oxide thickness and defect density. In this way, the area dependence of the breakdown can also be explained.

The N_{bd} will decrease with decreasing oxide thickness. As the oxide thickness scaling down, one defect will be enough to create the breakdown path, which means that N_{bd} becomes constant instead of decreasing with thinner oxide.

The statistics of gate oxide breakdown are described using the Weibull distribution $F(x) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$, where F is the cumulative failure probability; x is either the charge or time to breakdown. The characteristic life α is the 63.2% lifetime, and β is the slope parameter or Weibull slope. Plotting $W \equiv Ln[-Ln(1-F)]$ against $Ln(x)$ yields a straight line with slope β.

The gate oxide failure is a weakest-link type of problem because failure of the whole chip is defined by the failure of the first individual device, and a device fails if

any small portion of the gate area of the device breakdown. From statistics, if the probability of any one unit failing is p then the probability of any one of N independent units failing is

$$F = 1 - (1 - p)^N ,$$

So that

$$Ln[-Ln(1-F)] = Ln(N) - Ln[-Ln(1-p)].$$

As predicted by the percolation model, the value of N_{bd} will decrease with thinner oxide thickness, showing by the wider failure distribution and smaller β.

The Weibull plot thus has the extremely useful property that if the area is increased by a factor N then the curve shifts vertically by $Ln(N)$. If the desired low failure rate is F_{chip} over the product lifetime T_{life} for the total gate area A_{ox} on the chip, this is equivalent to a higher failure rate F_{test} in time T_{test} on the test structures with area A_{test}. Then we get

$$\frac{T_{life}}{T_{test}} = \left(\frac{A_{test}}{A_{ox}}\right)^{\frac{1}{\beta}} \left(\frac{Ln(1-F_{chip})}{Ln(1-F_{test})}\right)^{\frac{1}{\beta}} \approx \left(\frac{A_{test}}{A_{ox}}\frac{F_{chip}}{F_{test}}\right)^{\frac{1}{\beta}}.$$

The assumption in the above equation is that β is independent of V_g. This equation is used to scale measured breakdown times to the expected product lifetime, or equivalently to estimate the chip failure rate from test structure measurements. Since F_{chip} is less than F_{test}, and A_{test} is typically less than A_{ox}, then T_{test} is larger than T_{life}. So it is always necessary to use accelerated stress conditions. Thus it is important to understand the voltage dependence and is also reason so much attention is paid to the physical model for trap generation and breakdown.

Another key characteristic of the percolation theory is the relationship between the Weibull distribution slope β and the oxide thickness t_{ox}. With a thicker oxide, β becomes lager. Many papers are published to establish this relationship [48, 65, 66]. With the help of Fig. 3.3, a simple model can be used to explain this.

The bulk oxide can be considered as a lattice structure with a characteristic length a_0. Thus if the oxide area is A, the number of column is $N = A/a_0^2$. With the similar definition, the number of cells in one column is $n = t_{OX}/a_0$.

Assume λ is the fraction of defective cells, the cell cumulative failure distribution is then given as

$$F_{cell} = \lambda .$$

If all the cells in one column are fail, the column will fail. Thus the failure distribution of the column is

$$F_{cell} = \lambda^n .$$

Since the oxide breakdown is a weakest-link issue, the device failure distribution and the Weibit, W are given as

$$F = 1 - \left(1 - \lambda^n\right)^N$$

$$W = Ln\left[-Ln(1-F)\right] = Ln\left[-NLn\left(1-\lambda^n\right)\right]$$

Since λ usually is much less than 1, the above equation can be further simplified as

$$W = Ln\left[-NLn\left(1-\lambda^n\right)\right] \approx Ln(N) + nLn(\lambda).$$

Compared with the original model of the percolation theory, n is corresponding to the Weibull distribution slope β. Thus the relationship between the oxide thickness t_{ox} and β is established as

$$\beta = \frac{t_{OX}}{a_0} \tag{a}$$

In this way, as the oxide thickness continues shrinking, β will approaches to unit t_{ox} as is equal or less than a_0.

Fig. 3.4 shows the published data of the Weibull slope β as a function of oxide thickness [48, 59, 67-70]. There are two characteristics of this figure. First, the Weibull slope β has a linear relationship with the oxide thickness if the oxide thickness is thicker than about 2 nm, which can be explained well by equation (a). Second, as the oxide thickness continues shrinking, β approaches unit. The reason for this is also very clear since one defect would cause the failure of the entire gate oxide, as the defect size is equal or smaller than the oxide thickness.

It should be mentioned that although the percolation theory can predict the statistical behavior of breakdown well, the parameters extrapolated from it might not correspond to the real physical mechanisms. For example, the defect size extrapolated form it may not be the real physical dimensions. As we know, the defect generation inside the oxide is not uniform. There are more defects near the interface and source and drain region. However, percolation theory and the Weibull distribution are more reasonable and intuitive compared with other theories. At the same time, even though this theory was developed for the SiO_2, it is still applicable to the other gate dielectrics. However, when the high-*k* dielectrics are used, an interfacial layer is

always inserted between the silicon substrate and the gate bulk oxide. Thus, even for the same structure, the Weibull slope will be quite different depends on the breakdown occurs at the bulk oxide or the interfacial layer. At the same time, the critical oxide thickness when β approaches unit will also be different, depending on the exact physical structures of the defect. Further study in this area is still needed[71, 72].

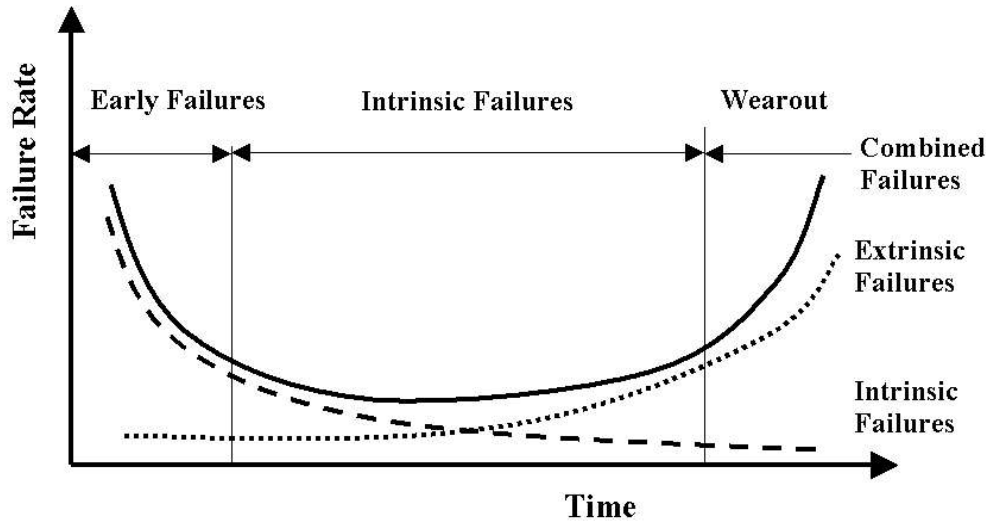


Fig. 3.1. The bathtub curve of the product lifetime.

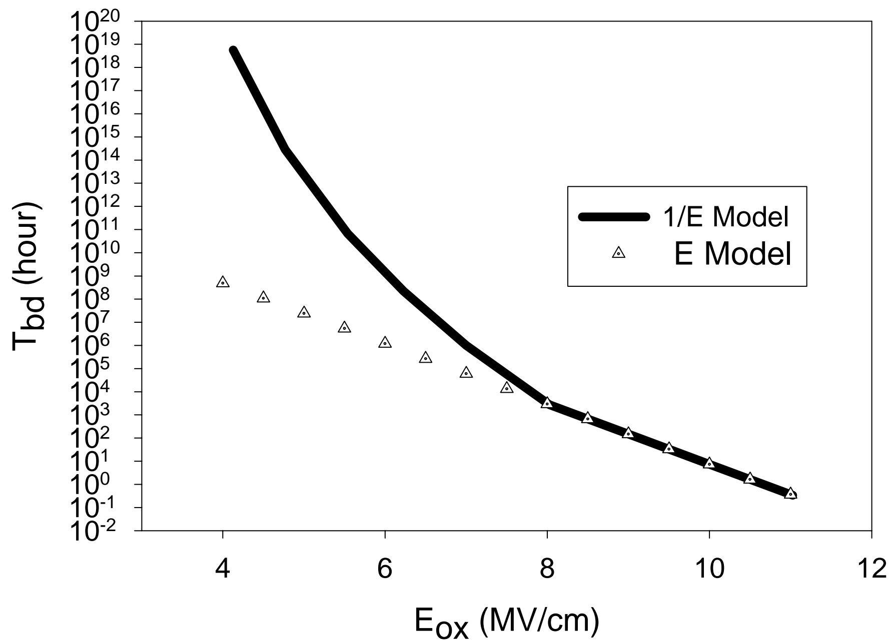


Fig. 3.2. The comparison of 1/E model and *E* model.

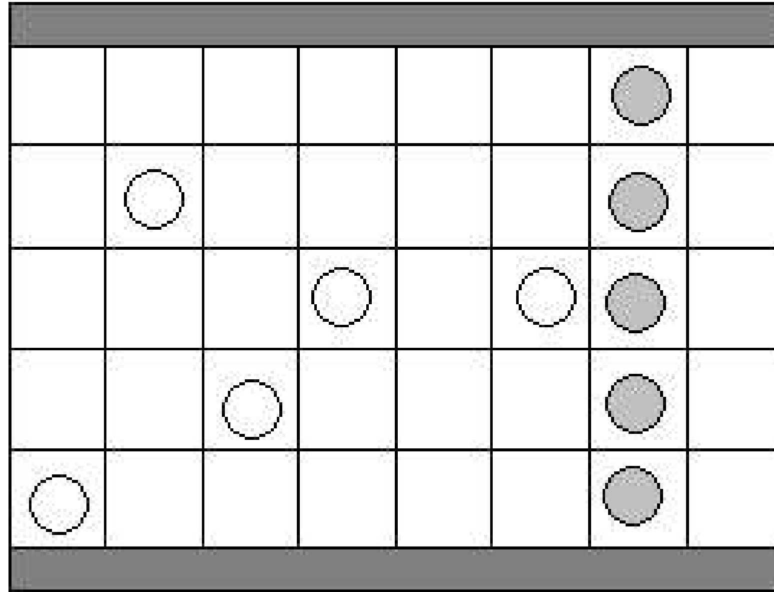


Fig. 3.3. The cell based percolation theory.

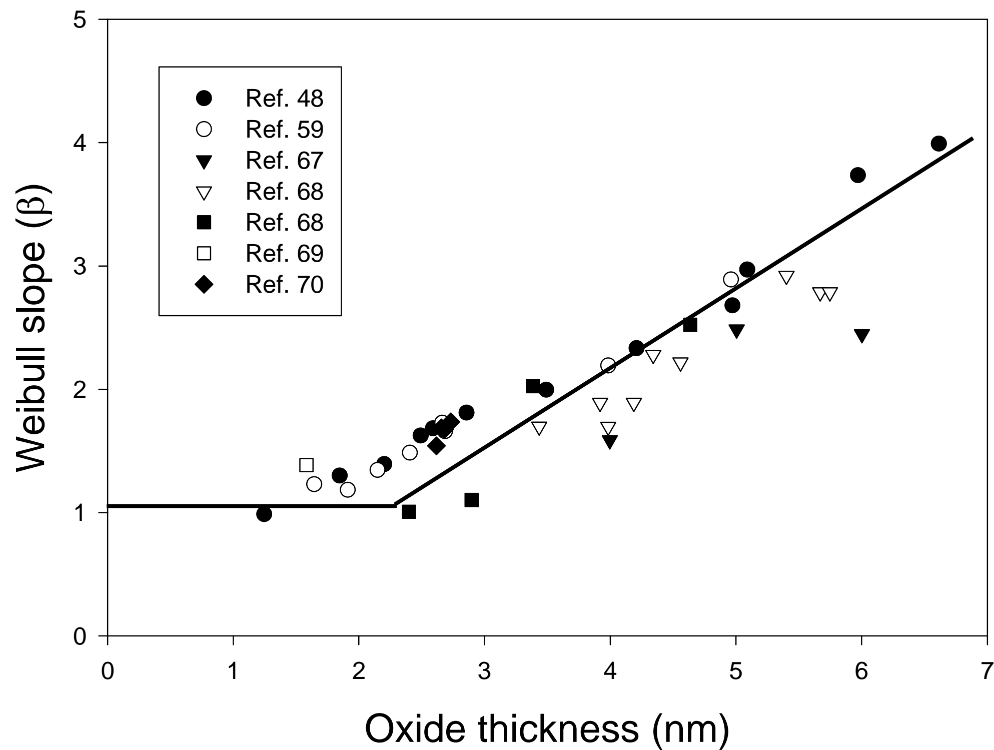


Fig. 3.4. The relation ship between Weibull slope and the oxide thickness.

Chapter 4:

Negative Bias Temperature Instability

4.1 Introduction

In the modern semiconductor industry, statistical physics and the physics of understanding how complex systems interact will become increasingly important especially as we enter the era of megascale integration and SOC IC production. This is especially true for variance and defect interactions that become magnified when investigating the electrical output characteristics as device geometry shrinks [73].

For the next generation of SOC and IC products, there is a new barrage of challenge for us to overcome[74, 75]. Those challenges have critical impact on product yield, product reliability, chip testability/performance prediction, and understanding of the process integration. Among the key problems affecting the semiconductor industry are Negative Bias Temperature Instability (NBTI), gate oxide leakage current, power consumption, etc. As devices are scaled and the density on chip increases, the probability of a circuit encountering lethal accelerated NBTI degradation increases. Hence, the impact of NBTI on yield is expected to increase as SOC complexity and integration increase.

4.2 Static NBTI

4.2.1 Degradation Phenomena

pMOSFETs suffer NBTI when stressed with negative gate voltages at elevated temperatures. The typical stress condition of NBTI is 100-250 ^{o}C and less than 6 MV/cm oxide electrical fields. A higher electrical field will lead to hot carrier degradation. This stress condition is typical during the burn in. However, it is also approached in high performance ICs during the routine operation. Fig. 4.1 shows the trend in electric fields for CMOS circuits, showing that oxide electrical fields of the magnitude to generate NBTI are typical in today's circuits[76].

The main characteristics of NBTI are decreases of drain current I_{on} and transconductance g_m, and the absolute off current I_{off}, and threshold voltage V_{th} increase. The threshold voltage V_{th} and flatband voltage V_{fb} of a MOSFET are given by (1) and (2) respectively [15].

$$V_{th} = V_{FB} - 2\phi_F - \frac{|Q_B|}{C_{ox}}$$

$$V_{fb} = \phi_{MS} - \frac{Q_f}{C_{ox}} - \frac{Q_{it}(\phi_s)}{C_{ox}}$$

The fixed oxide charge Q_f and the interface-trapped charge density Q_{it} are the two factors determine the threshold voltage shift. Positive increases in them will lead to the negative threshold voltage shift.

$$\Delta V_{th} = -\frac{\Delta Q_{it}(\phi_s)}{C_{OX}} - \frac{\Delta Q_f}{C_{OX}}$$

During the NBTI degradation, the threshold voltage will shift to the negative direction, thus either the interface traps or the fixed oxide charges will change during the NBTI stress.

The simplest form of the driving current I_{on} and transconductance g_m of a MOSFET are given as

$$I_{on} = \frac{W}{2L}\mu_{eff}C_{ox}\left(V_g - V_{th}\right)^2$$

$$g_m = \frac{W}{L}\mu_{eff}C_{ox}\left(V_g - V_{th}\right)$$

As shown by those two equations, the parameters leading to I_{on} and g_m degradation are the threshold voltage shift and the mobility μ_{eff} change. The mobility degradation mainly comes from the interface trap generation, leading to additional surface-related scattering.

NBTI mainly occurs to pMOSFETs and appears to be negligible for positive gate voltage and for either positive or negative gate voltages in nMOSFETs as shown in Fig. 4.2.

The reason of this phenomenon can be well explained by the interface traps behavior of nMOSFETs and pMOSFETs. An interface trap in MOSFETs is an interface trivalent Si atom with an unsaturated valence electron at the SiO_2/Si interface, denoted by $Si_3 \equiv Si^{\bullet}$. The $\equiv$ represents three complete bonds to other Si atoms and the * represents the fourth, unpaired electron in a dangling bond.

Interface traps are electrically active defects with an energy distribution throughout the Si band gap. Acted as generation/recombination centers, they contribute to low frequency noise, leakage current, and reduced mobility, drain

current, and transconductance. Once occupied by electrons or holes, the device threshold voltage will shift by equation (1).

Interface traps act as acceptor-like or donor-like depending on their relative position to the band gap [77]. Acceptor-like interface traps are neutral when they are empty and negative when occupied by electrons. Donor-like interface traps are positive when they are empty and neutral when occupied by electrons. Depending on its relative position to the Fermi energy level, an interface trap will be occupied by an electron or empty.

As shown by Fig. 4.3, at flatband condition electrons occupy the states in the lower half of the band gap, and those traps are neutral. Those between mid gap and the Fermi energy are negatively charged since they are occupied acceptor states. Those above the Fermi level are unoccupied acceptors states and they are neutral. For a pMOSFET device, the fraction of interface traps between mid gap and the Fermi level is unoccupied donors, when the device is in inverted state. Those interface traps are positively charged, thus leading to the negative threshold voltage shift in a pMOSFET.

Since the interface traps are acceptors in the upper half of the band gap and donors in the lower half of the band gap, they affect the threshold voltage shift in nMOSFET and pMOSFET differently. At flatband condition, the interface trap charge in nMOSFET is positive and in pMOSFET is negative. At inversion condition, the interface trap charge is negative in nMOSFET and positive in pMOSFET. However, the fixed oxide charge is positive, thus at inversion nMOSFET: $Q_f - Q_{it}$, pMOSFET: $Q_f + Q_{it}$. Hence the pMOSFET is more vulnerable to NBTI stress [78].

4.2.2 NBTI Mechanisms

After first been studied by Deal [79], NBTI has been a hot spot and many models for the physical mechanisms have been established. Among those models, holes injected into the oxide, tunneling electrons and electrochemical reactions were the main subjects.

Hole trapping model

The hole trapping model is based on avalanche hole injection measurements on unstressed MOS capacitors and the NBTI tests[80-82]. This model proposes that the negative midgap voltage shift, which is believed to be a measure of the change of positive oxide charge without the contribution from interface states, is due to the filling of intrinsic hole traps. All the positive charge generated by preceding negative bias stress can be removed by the positive bias stress. But the exact mechanism for hole injection into the oxide is still unknown.

Thermally assisted electron tunneling model

The thermally assisted electron tunneling model was established by Breed[83, 84]. According to this model, the neutral or positive centers, which cause the charge trapping, are located near the interface in the oxide. Under negative bias stress, the centers are excited. The electrons in the excited states then tunnel into empty states of the conduction band of the silicon. This process is a thermally assisted tunneling process.

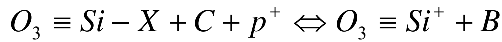
Reaction-Diffusion model

Several authors proposed the electrochemical reaction model, or the Reaction-Diffusion model, which has been accepted by many researchers in recent years [85-

88]. This model explains the NBTI effect in terms of electrochemical reactions. Fig. 4.4 shows the schematic diagram of this model.

Basically, there are two reactions in this model.

$$Si_3 \equiv Si - X + A \Leftrightarrow Si_3 \equiv Si^* + B$$



$$O_3 \equiv Si - X + C + p^+ \Leftrightarrow O_3 \equiv Si^+ + B$$

$$B_{interface} \Leftrightarrow B_{bulk} \quad \text{(Diffusion)}$$

Here, Si3≡Si-X is a trivalent silicon, a defect near the interface. X, A, B, and C are unknown neutral species. Si3≡Si• is a neutral trivalent silicon backboned by three other silicon atoms, functioning as an interface trap near the interface. O3≡Si+ is a positively charged species, serving as a positive charge in the oxide.

The first reaction accounts for the interface trap generation. The trivalent silicon reacts with A and produces the interface trap and B. There are different definitions for A and B. Some suggest A as a hydrogen ion and B a hydrogen atom. But there are also some suggestions in different ways. The second reaction is for the fixed oxide charge generation reaction. There are a number of suggestions for the C species as well. The diffusion of the product B would produce the phenomenon of NBTI. There are also some other variant electrochemical reaction models, which explain the NBTI effect with one electrochemical reaction. It is clear that regardless of mechanism, holes certainly play an important role in the NBTI effect.

4.2.3 Device Lifetime Extrapolation

In order to precisely predict the lifetime of the device experienced NBTI degradation, a correct model is needed. Generally, a threshold voltage shift ΔV_{th} caused by NBTI can be expressed as

$$\Delta V_{th} = Af_1(t)f_2(V_g)\exp\left(\frac{-E_a}{kT}\right).$$

Here f_1 and f_2 are functions account for the time dependence and gate voltage dependence.

Based on the physical mechanisms and experimental data, several models for the time dependence have been suggested.

Logarithmic time dependence

$$\Delta V_{th} = ALog(t)$$

This model was established on the ideal of charge trapping, wherein carriers tunnel into existing traps[79]. According to this model, the NBTI is field accelerated, and there is little or no temperature activation. The saturation behavior is due to the finite trap density. There is significant deviation at long time when using this model. However, it is frequently observed in recent high-*k* gate dielectrics experiments.

Exponential time dependence

$$\Delta V_{th} = A\exp\left(\frac{t}{\tau}\right)$$

$$\Delta V_{th} = A\exp\left(\frac{t}{\tau_1}\right) + B\exp\left(\frac{t}{\tau_2}\right)$$

A single exponential time dependence model was established at first based on the first order reaction, which was limited by the hole concentration[85]. Different from

the logarithmic time dependence, this model suggested the temperature activation from the reactions. Later, a two-exponential model was further suggested[89, 90].

Power law time dependence

$\Delta V_{th} = At^{n}$

The power law model was based on the reaction diffusion mechanism. According to this model, the hydrogen profile determines the time dependency[86, 87]. The temperature dependence arises from the reaction and diffusion processes and the saturation behavior comes from the diffusion barrier or the available Si-H bonds. Compared with the other two models, this model has the most observed features and is widely accepted.

4.2.4 NBTI Impact on Circuits

In CMOS circuits, NBTI leads to timing shifts and potential circuit failure due to increased spreads in signal arrival in logic circuits. Asymmetric degradation in timing paths can lead to non-functionality of sensitive logic circuits and hence leads to product field failures.

In the analog mixed signal circuit, NBTI is a concern because V_{th} shifts are a major reliability issue, especially in matching applications where circuit operation may force matched transistors into asymmetrical bias conditions resulting in a significant asymmetric stress induced mismatch [91]. Even though the analog design techniques can avoid some of these issues and minimize the NBTI stress conditions, this will increase the expense of design complexity and potential performance tradeoffs. There are have been reports of the NBTI effects on the current mirrors and operational amplifiers [92].

NBTI stress induced degradation in digital device saturation drive current leads to significant timing issues. The induced threshold voltage shift will lead to reduced current and frequency degradation of ring oscillators and reduced standard random access memory noise margin[92]. For digital circuits, NBTI can affect the FPGA performance [93], the ring oscillator, SRAM static noise margin [92] and microprocessors [94].

4.3 Dynamic NBTI of Ultrathin SiO_2

4.3.1 Introduction

Recent studies have shown that a fraction of NBTI degradation can be annealed if the applied gate voltage is removed [95-98]. In the real circuits, the electronic device experienced more the pulsed stresses rather than the static stresses. Thus, to accurately assess the NBTI degradation of devices under dynamic stress conditions becomes an important task.

Since there are two factors influencing the NBTI degradation, the interface traps and the fixed oxide charge, the reduction of either of them will cause the relaxation of NBTI degradation. Based on the Reaction-Diffusion model, the dynamic NBTI was explained by the generation and annealing of interface traps [99]. However, there are also other reports suggested the fixed oxide charge was the main factor, and the interface trap density remained constant with different stress conditions [100].

In the following paragraphs, the Dynamic NBTI degradation of ultrathin SiO_2 was studied at first. The NBTI degradation of HfO_2 device is at the second part.

4.3.2 Devices and Experimental Setup

Surface channel dual-gate pMOSFETs with 2.0 nm and 5.0 nm gate oxide thickness are used in this study. The channel width is 10.0 μm, and the length is 0.25 μm.

Constant voltage stresses are applied to the gate under dc and unipolar-pulsed stresses with source, drain, and substrate grounded as shown in Fig. 4.5. For the pulsed stress, the amplitude is identical to the dc stress voltage during the pulse "on" phase and was zero during the pulse "off" phase. The duty cycle is maintained at 50 %. The stress was periodically interrupted, and the drain saturation current I_{on} was measured immediately.

4.3.3 Results and Discussion

Briefly speaking, during the NBTI stress, the threshold voltage shift ΔV_{th} is caused by the fixed oxide charges and the interface trapped charges, as shown by:

$$\Delta V_{th} = -\frac{\Delta Q_{it}(\phi_s)}{C_{OX}} - \frac{\Delta Q_f}{C_{OX}}.$$

Here ϕ_s is the surface potential, and C_{OX} is the oxide capacitance per unit area. V_{th} is extrapolated from the linear region of the drain current vs. gate voltage (I_d-V_g) measurement [16].

Fig. 4.7 shows the ΔV_{th} under negative unipolar voltage stress with pulse repetition frequencies up to 100 kHz. According to this figure, ΔV_{th} exhibits significant frequency dependence. ΔV_{th} decreases as the frequency increases, and

with a frequency of 500 kHz or higher, ΔV_{th} is only about half of that under dc stress condition, which is also observed by the other researchers [95].

Drive current, I_{on}, is an important factor in determining the speed of CMOS circuits. I_{on} and V_{th} have similar shit behavior with the NBTI stress, which is explained by the simplest relation between them:

$$I_{on} = \frac{W}{2L}\mu C_{OX}\left(V_G - V_{th}\right)^2.$$

Here, W and L are the device channel width and length respectively, μ is the effective mobility of minority carries. However, I_{on} and V_{th} are not simply related by square root since there is also mobility degradation during the NBTI stress, which is due to the generation of the interface states. Compared with V_{th}, I_{on} shows the similar dependence on frequency, as shown in Fig. 4.9.

One of the possible explanations for the frequency dependence is that the damage caused by NBTI during the stress on period is reduced or annealed during the stress off period.

We monitored the recovery of I_{on} in detail as a means to evaluate the damage caused by NBTI after the stress termination, as shown in Fig. 4.10. Although the monitoring of V_{th} is more preferable, we cannot conduct that test as fast as I_{on}, since I_d-V_g measurement takes a longer time. As pointed out by some reports, the time between the measurements is critical for the study of NBTI. The data in this plot were taken every 1 s after the stress terminated.

Apparently, the I_{on} recovery data at each test could not fit a simple straight line, which implies the existence of multiple mechanisms. This is further confirmed by the

extrapolated activation energy using the data in Fig. 4.10 as shown in Fig. 4.11. A chemical reaction process is related with unique activation energy, which is the minimal energy needed.

Obviously, the activation energy associated with I_{on} recovery is much higher right after the NBTI stress terminates and saturates later, which suggests the existence of at least two different processes. The first process has higher activation energy, has faster annealing rate, and happens earlier. The second process has lower activation energy, has slower annealing rate and dominates the later period. The first process may happen so fast that we cannot get the exact value due to the equipment time response limit. The observed decreasing period should be the transient period from the first process to the second one.

Since both reactions (1) and (2) can be reversible, the recovery of NBTI could be attributed to the annealing of either fixed oxide charges or interface traps, or a combination of the two.

At first, the annealing of interface traps is discussed. It has been reported that, different from NBTI (which includes both interface trap generation and positive charge formation), the positive bias temperature instability (PBTI) only results in interface trap generation, and the final saturation level is fixed by the number of defects in the device [101]. Thus, the study of the frequency dependence of PBTI will reveal whether there is annealing of interface traps, as shown in Fig. 4.12. This shows that there is almost no frequency dependence of V_{th} with PBTI, at least in the frequency range from dc to 10 kHz. This suggests that the annealing of PBTI, or interface traps during the stress off period within each pulse, is much smaller or

negligible. The interface traps generated by PBTI and NBTI should be the same as identical devices are tested.

Additional tests are performed to study the annealing of interface traps during the stress off period as shown by Fig. 4.13.

In this figure, we compare the annealed fraction of ΔI_{cp} and ΔV_{th}. They are defined as:

$$\Delta V_{th(t)} = \frac{V_{th(t)} - V_{th(1000)}}{V_{th(1000)} - V_{th(0)}} \times 100\%$$

$$\Delta I_{cp(t)} = \frac{I_{cp(t)} - I_{cp(1000)}}{I_{cp(1000)} - I_{cp(0)}} \times 100\%$$

Here, $V_{th(1000)}$ is the threshold voltage at 1000 s, at which the stress is turned off, $V_{th(t)}$ is the threshold voltage measured at time *t*, and $V_{th(0)}$ is the initial threshold voltage for a fresh device. I_{cp} is defined in a similar way.

If the annealing is mainly associated with interface traps, the annealing of ΔI_{cp} will be comparable to that of ΔV_{th}. However, we see the annealing of ΔV_{th} is much larger, thus the annealing of fixed oxide charges is dominating. We believe the earlier recovery shown in Fig. 4.8 is related with the annealing of fixed oxide charge, which is the product mainly from a chemical reaction at the interface, while the later process is because of the annealing of interface traps, as it is associated with diffusion process, which happens at a lower rate.

It has been reported that fixed oxide charges could be reduced or detrapped with a positive gate bias following the NBTI stress [102]. Based on reaction (1), one would argue interface traps would show similar behavior. One provable test is the bipolar

stress test in which the device is stressed with negative gate bias and positive gate bias one after another. If both fixed oxide charges and interface traps reduce with an alternative gate bias stress, the resulting ΔV_{th} will reduce consequently. However, Fig. 4.14 shows that ΔV_{th} under bipolar stress is larger. This means that there is not only no reduction but also the amount of increase of V_{th} induced by increased interface traps during the positive bias period could be larger than the reduced amount caused by the reduction of fixed oxide charges. This may further suggest that alternative gate bias stress will probably not reduce the degradation caused by NBTI.

As discussed earlier, ΔV_{th} consists of the change of the fixed oxide charge and interface traps. Thus the reduction of ΔV_{th} could only come from the reduction or annealing of fixed oxide charge. Recent studies have also confirmed that, during the NBTI stress, the interface traps density is almost independent of the gate voltage frequency [102].

Following this argument, we can establish a model to explain the frequency dependence of NBTI. The parameter shift, such as V_{th} shift, could be expressed as:

$$\Delta V_{th(f)} = \frac{1}{C_{OX}}\left[\Delta Q_{it(dc)} + \Delta Q_{f(f)}\right] \qquad (3)$$

$\Delta V_{th(f)}$ is V_{th} shift with unipolar stress, $\Delta Q_{it(dc)}$ is the interface traps change with dc stress, and $\Delta Q_{f(f)}$ is the fixed oxide charge changes with unipolar stress and is a function of gate voltage frequency.

For a unipolar stress condition, the fixed oxide charge density will increase during the stress on period (t_{on}) and decrease during the stress off period (t_{off}), as the

generation of fixed oxide charge can be reversible. By assuming the first order kinetics, we can create the following equations.

During the nth stress off period:

$$Q_{\min}^{n} = Q_{\max}^{n} e^{-\frac{t_{off}}{\tau_{on}}},\ t_{on} + t_{off} = \frac{1}{f}.$$

During the n+1th stress on period:

$$Q_{\max}^{n+1} = Q_{\min}^{n} + (Q_{f(t)} - Q_{\min}^{n})(1 - e^{-\frac{t_{on}}{\tau_{off}}}).$$

Here $Q_{\max}$ and $Q_{\min}$ are the upper trace and lower trace of $Q_{f(f)}$ respectively, f is the frequency of the gate voltage, and τ_{on} and τ_{off} are time constants for the stress on and off period respectively.

After a certain time, the system will be in a quasi-equilibrium condition, that is $Q_{\max}^{n} = Q_{\max}^{n+1}$. Assuming $Q_{\max}$ to be $Q_{f(f)}$, we have:

$$Q_{f(f)} = \frac{Q_{f(dc)}\left(1 - e^{-\frac{t_{on}}{\tau_{on}}}\right)}{1 - e^{-\frac{t_{off}}{\tau_{off}}} e^{-\frac{t_{on}}{\tau_{on}}}}.$$

To simplify the problem by assuming there is no mobility degradation during the NBTI stress, we can directly use this model to fit the ΔI_{on} data we get at different frequencies, as in Fig. 4.15. All the data in this figure are normalized to the dc stress data.

By this model, we find the parameter shift caused by NBTI will be largely reduced in the unipolar stress conditions. And this kind of reduction will almost saturate at higher repetition frequencies. This can also explain the reason some

reports did not observe the independence at higher frequencies [95]. The fit parameters are 0.4 s and 1 s for τ_{on} and τ_{off}, respectively.

In order to determine experimentally the time constants τ_{on} and τ_{off}, we conduct two tests. In these tests, the pulse on (off) period is set to be 5 ms, and the pulse off (on) period varies from 2 μs up to 10^4 ms. The total effective (DC) stress time is 1000 s for both tests. ΔV_{th} is plotted against the pulse frequency as in Fig. 4.16. Based on these tests, τ_{on} will be at the order of 100 ms, and τ_{off} will be at least one order higher than τ_{on}, which is consistent with our model.

4.3.4 Conclusion

The root course of the NBTI reduction with unipolar stress was discussed. It was found that the annealing of NBTI under unipolar stress is mainly due to the annealing of fixed oxide charge. The annealing of interface traps is a much slower process compared with that of fixed oxide charges. A model based on the fixed oxide charge generation and annealing was established to explain the reduced NBTI under the unipolar stress conditions.

4.4 Dynamic NBTI of HfO_2 Devices

4.4.1 Introduction

As the scaling trend of CMOS goes on, new gate materials other than silicon dioxide are inevitably needed [2]. HfO_2 has been studied extensively and believed to a promising candidate [103, 104]. Before integrated into circuit, the reliability of the devices made with HfO_2 need to be characterized. One of the important problems

with the CMOS devices is the NBTI phenomenon, which influences the threshold voltage (V_{th}) and driving current (I_{on}) of pMOSFETs [76]. Compared with SiO_2, with which high quality gate oxide can be achieved routinely, there exist more defects and traps inside HfO_2. It has been reported these bulk traps influence the V_{th} stability, thus the device performance [13, 105]. Several papers have been published on the NBTI of HfO_2 devices [90]. In those papers, the interface traps were evaluated using the charge pumping measurement, or the dc current-voltage method. However, the contribution of bulk traps to NBTI degradation needs further study. In this work, we evaluated the NBTI degradation of HfO_2 devices at different pulse bias stress conditions and compared them with SiO_2 control samples. The contribution of bulk traps are found to be important in NBTI study of HfO_2 devices and will influence the device lifetime extrapolation.

4.4.2 Device and Experimental Setup

The devices used in this work are HfO_2/Hf silicate hybrid stacks pMOSFETs. The gate dielectrics were fabricated by metal organic chemical vapor deposition (MOCVD), post deposition annealing (PDA) with NH_3 at 700 ^{o}C for 60 s. The equivalent oxide thickness (EOT) is 1.7 nm. Between the HfO_2 bulk oxide and silicon substrate, there is an interfacial layer of SiO_2 with a thickness of 1.0 nm. For comparative study, a SiO_2 control wafer with an EOT of 2.0 nm is used. The channel width is 10 μm and the channel length 0.25 μm. Constant voltage stresses (CVS) are applied to the gate with dc and unipolar pulse stresses while the source, the drain, and the substrate are grounded. For the pulsed stress, the amplitude is identical to the dc

stress voltage during the pulse “on” phase and was zero during the pulse “off” phase, and the duty cycle is maintained at 50 %. Unless specifically pointed out, the devices tested in this study are stressed with a –2.0 V dc or unipolar stress voltage. All the devices are stressed with equivalent time. The stress is periodically interrupted and the monitoring tests are performed immediately.

4.4.3 Results and Discussion

The Frequency Dependence

The NBTI degradation is first investigated by measuring the ΔV_{th} as a function of stress time, gate voltage and ambient temperature. V_{th} is determined using the constant current method [28]. It was found that when stress by dc gate voltages, ΔV_{th} of HfO_2 shows the similar trend as SiO_2 devices, except with different parameters, this is also observed by other reports [90]. A typical ΔV_{th} time evolution is shown in Fig. 4.17.

The frequency dependence of NBTI degradation of SiO_2 devices has been widely reported and several models have also been proposed to explain this behavior [100, 106]. For the HfO_2 devices in this study, we observed similar behavior, as shown in Fig. 4.18.

In this figure, we compared the ΔV_{th} of HfO_2 and SiO_2 devices. For better understanding, Fig. 4.19 shows the normalized ΔV_{th} data with the dc stress condition set to be 1. Compared with SiO_2 control devices, HfO_2 devices show several noticeable features.

First, the ΔV_{th} of HfO_2 is much larger than SiO_2 devices as shown in Fig. 4.18. The oxide thickness difference itself cannot cause such a larger difference. This is must because of the intrinsic characteristics difference of the gate oxide, as we know there are larger interface trap density and bulk defects in HfO_2 gate oxide.

Second, the ΔV_{th} of both devices are significantly reduced at higher stress frequencies. With a 10 MHz gate voltage stress, the ΔV_{th} is only around 50% of that caused by dc stress for both devices. This suggests there are time constants associated with the generation and relaxation of positive oxide charges and/or interface traps. It is believed that the injected holes generate the oxide traps [102]. With a higher frequency, i.e. a shorter stress on time, fewer oxide traps are generated. Possibly, the models proposed to explain this behavior of SiO_2 can also be adopted for HfO_2 devices.

At last, more importantly, the frequency dependence of HfO_2 and SiO_2 devices at middle frequencies are different. For example, ΔV_{th} of a 1 KHz stress voltage is about 80 % of that caused by dc stress with HfO_2 devices. However, for the SiO_2 devices, this ratio is about 60 %. In a separate study, we have demonstrated that for the SiO_2 devices, this dependence is found to be independent of gate oxide thickness [100]. Therefore, we believe the bulk traps inside the gate oxide could be one of reasonable explanations.

The Contribution of Bulk Traps

Besides the positive oxide charges, the generation of oxide traps can cause reduced drain driving current, increased threshold voltage. It is believed that there is a large quantity of bulk traps inside the high-*k* dielectrics [13, 105, 107]. The charge

pumping measurement is widely used to detect the trap density at and near the interface between the oxide and substrate. Since there are two interfacial layers in the high-*k* dielectric stacks, the conventional charge pumping measurement, in which the amplitude is fixed (FACP), cannot provide accurate information. In order to distinguish the bulk traps from the interface traps, a method called variable amplitude charge pumping (VACP) measurement was applied recently [13, 105]. A simplified setup is shown in Fig. 4.20. In our tests, we used both of these two methods to monitor the interface traps and bulk traps generation.

With the conventional charge pumping measurement, the interface traps generation is given by (1).

$$\Delta N_{it(t)} = \frac{\Delta I_{cp1}}{qAf} = \frac{\left(I_{cp1(t)} - I_{cp1(0)}\right)}{qAf} \quad (1)$$

Here $I_{cp1(t)}$ is the peak charge pumping current measured from substrate after stress time t, q is the electronic charge, A is the device area, and f is frequency. The charge pumping current measured from the amplitude variable charge pumping method is believed to be originated from the total oxide trap density including the bulk traps and the interface states. With a similar definition, this trap density increase is given by (2).

$$\Delta N_{trap(t)} = \frac{\Delta I_{cp2}}{qAf} = \frac{\left(I_{cp2(t)} - I_{cp2(0)}\right)}{qAf} \quad (2)$$

$I_{cp2(t)}$ is the charge pumping current at a constant charge pumping amplitude after stress time t. Note this trap density is the sum of interface trap density and bulk trap

density. Thus, the bulk trap density generation is given by the total trap density generation subtracting the interface state generation, as (3).

$$\Delta N_{bulk} = \frac{\Delta I_{bulk}}{qAf} = \frac{\Delta I_{cp2} - \Delta I_{cp1}}{qAf} \qquad (3)$$

In this way, by measuring the charge pumping current change we can directly get the information of interface traps generation and bulk trap generation.

We performed both charge pumping measurements on the HfO_2 devices and the control SiO_2 samples. For the HfO_2 devices, we observed the distinct generation of bulk traps in addition to the interface traps with stress time. But there is almost no bulk traps generation from SiO_2 devices during the stress, which is similar to other reports [5, 6]. In addition to the dc stress, the same measurements were taken with different stress conditions. Fig. 4.21 shows the interface traps and bulk traps generations as a function of stress frequencies.

Evidently, the generation of bulk traps also depends on the pulsed stress repetition frequency. At dc stress conditions, the increased charge pumping current due to the bulk traps is comparable to that due to the interface traps. However, at higher frequencies, the bulk trap generation is about one order lower than the interface trap generation. This means that the contribution of bulk traps to the NBTI degradation at high stress frequencies is negligible even though it accounts a larger portion at the dc stress condition. This could be also the reason why the frequency dependence of ΔV_{th} is distorted compared with the SiO_2 control samples, since there is almost no bulk trap generation in the SiO_2 devices. The exact reason for the lower generation of bulk trap at higher stress frequencies is still unclear. We suspect that those bulk traps are caused by the injected holes. However, compared with interface traps, these traps are

deep inside the oxide, thus it is more difficult for holes to reach those positions and create bulk traps when stressed with higher frequencies.

In addition to the different response to pulsed stress frequencies, we found the generations of bulk traps and interface traps also have different temperature dependences as shown in Fig. 4.22. For the interface traps generation, the activation energy is 0.11 eV; however, the bulk traps generation has a much higher activation energy, which is 0.24 eV. This suggests the interface traps generation of HfO_2 devices is less sensitive to the ambient temperature than bulk trap generation.

To further study the influence of bulk traps NBTI degradation, we investigated the ΔV_{th} at different temperatures. In this test, we especially compared the dc and 10 MHz stress conditions. As we know, there are two factors that influence ΔV_{th}, the interface traps and trapped oxide charges. The oxide charge generation, which is mainly due to hole injection, is almost independent of temperature. Thus the activation energy extrapolated from the ΔV_{th} mainly reflects the temperature dependence of interface traps. As shown in Fig. 4.23, the temperature dependences of SiO_2 devices at dc stress and 10 MHz stress are quite similar with activation energies of 0.14 eV and 0.13 eV respectively. However, the situation for HfO_2 devices is quite different. Even though the data don't fit well to a straight line, the activation energy extrapolated from dc stress is quite different from the 10 MHz stress. According to our experimental data, under dc stress conditions, the activation energy of ΔV_{th} is 0.11 eV, while it is only 0.05 eV for the 10MHz stress. This clearly demonstrates the influence of bulk traps at different stress conditions, since we know from the previous

test that the interface traps are less sensitive to stress temperature than bulk traps, and there are much less bulk traps generated with a 10 MHz gate voltage stress.

In order to verify the influence of bulk traps, we add a 1.0 V, 10 s detrapping period before each I_d-V_g and charge pumping measurement. According to our tests, this detrapping period will not stress the device. Thus any further difference should not be caused by the stress during this period. We found that the charge pumping current due to the bulk traps almost disappeared after this detrapping period, as shown in Fig. 4.24, indicating that there was no more contribution from the bulk traps to the degradation of the device. In other words, the trapped charges observed during the NBTI stress are due to preexisting bulk traps and no additional bulk traps are generated during the NBTI stress.

Using this method, we did the similar temperature acceleration tests on the HfO_2 devices and the control SiO_2 samples. In addition to the annealing of the bulk traps, the injected holes could also be neutralized by this detrapping period. As shown in Fig. 4.25, after the detrapping period, the activation energies extrapolated under the dc stress and 10 MHz stress conditions are almost the same, which is similar to the control SiO_2 samples. Since the bulk traps and injected holes are already annualized during the detrapping period, these activation energies actually reflected the temperature sensitiveness of the interface traps. This could be the reason for the similar activation energies for two stress conditions of SiO_2 and HfO_2 respectively. In the HfO_2 devices, there is a 1.0 nm SiO_2 interfacial layer, thus we would expect similar interface traps. However, since the thickness of the interfacial layer is so small, the interface traps could have a structure of $Hf_3{\equiv}Hf\bullet$ in addition to the $Si_3{\equiv}Si\bullet$

structure. This can be the reason for the different activation energies between HfO_2 devices and the control SiO_2 samples.

4.4.4 Conclusion

The contribution of gate oxide bulk traps to NBTI degradation of HfO_2 devices is evaluated and compared with SiO_2 control samples. The ΔV_{th} and oxide trap generation of HfO_2 and SiO_2 devices show the dependence of gate voltage frequencies. Besides interface traps, bulk traps also influence the NBTI degradation of HfO_2, and this influence is also dependent on the stress conditions. At dc stress conditions, the bulk trap generation is comparable to that of interface traps, and thus plays an important role for the ΔV_{th}, however it becomes negligible at higher frequency stresses. In addition to the different frequency response, the bulk trap generation also shows different temperature dependence than the interface traps. We found that interface traps generation has much higher activation energy than bulk traps, which also explain the different temperature dependences of ΔV_{th} at dc and high frequency stress conditions. We should be cautious when extrapolating the device lifetime based on the experimental data especially using the temperature acceleration tests, as the extrapolation will be quite different under the dc and pulsed stresses.

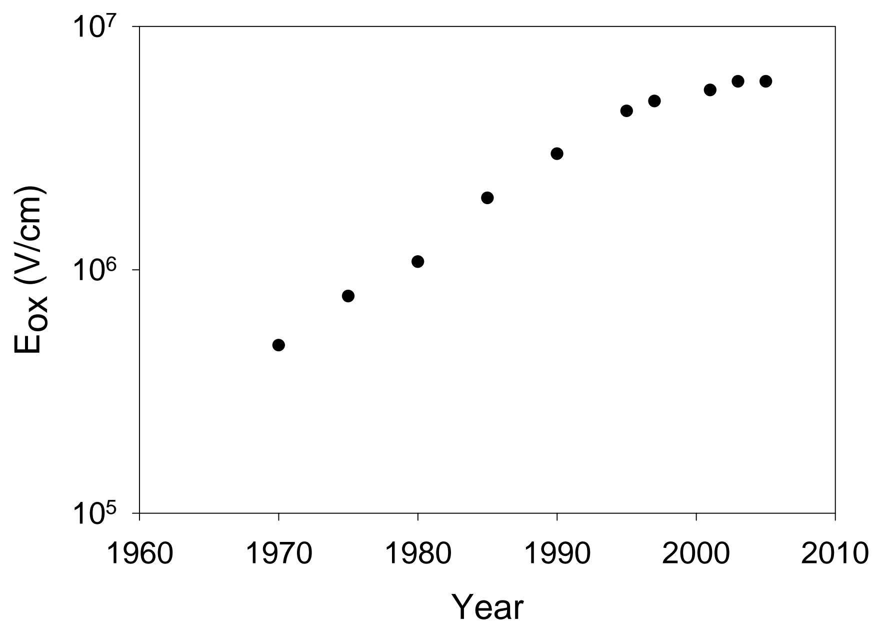


Fig. 4.1. The electric field at different technology generations.

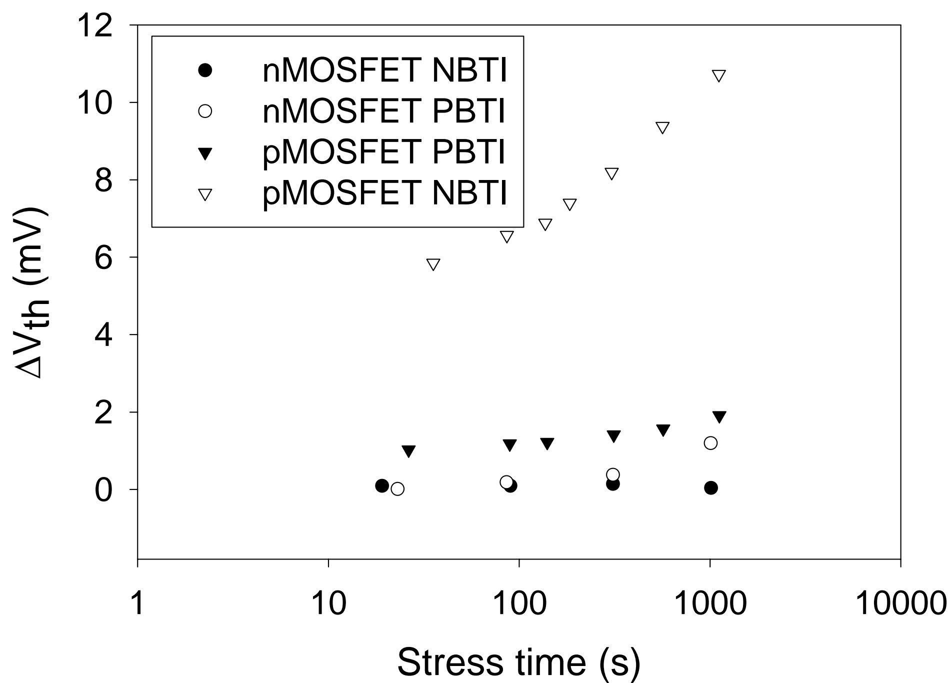


Fig. 4.2. The NBTI and PBTI of nMOSFET and pMOSFET.

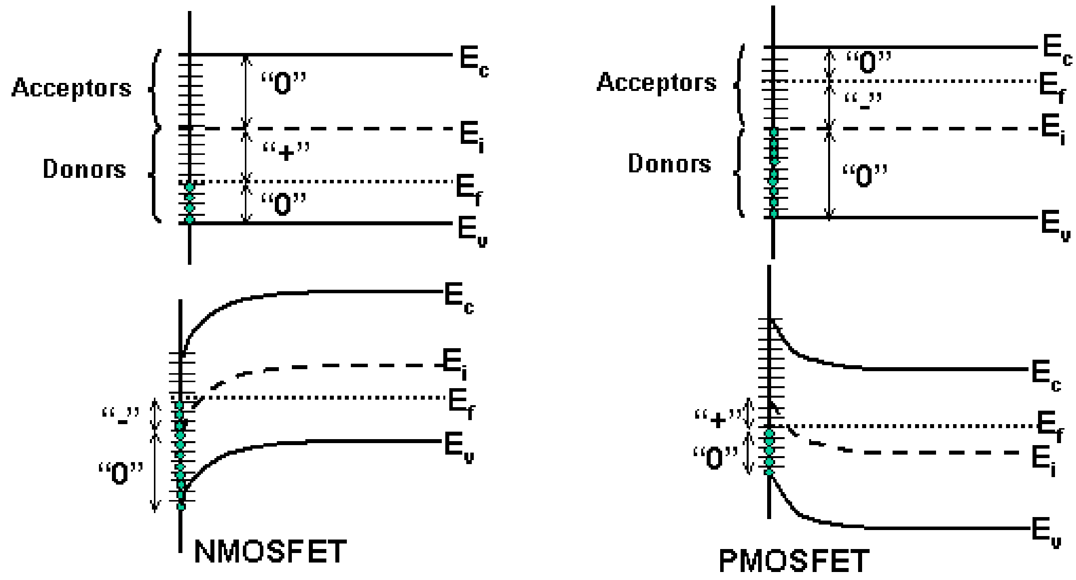


Fig. 4.3. The band diagram of flat band (upper) and inversion condition (lower) of nMOSFET and pMOSFET.

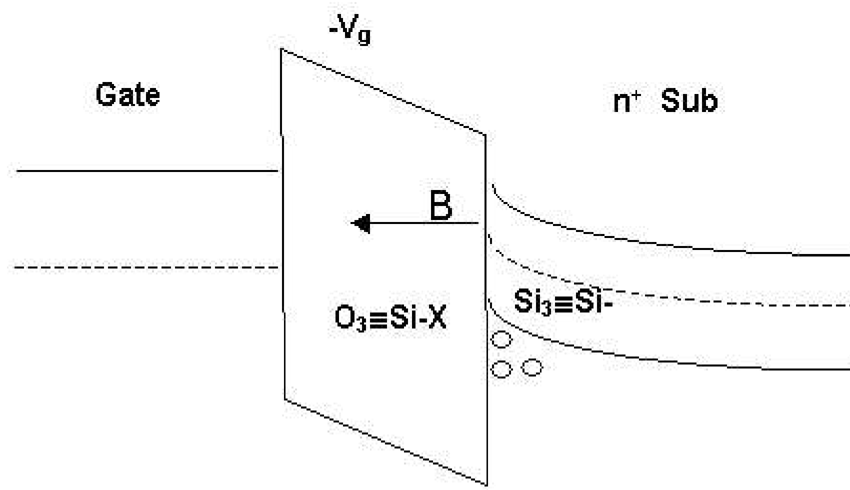


Fig. 4.4. The electrochemical reaction model of NBTI.

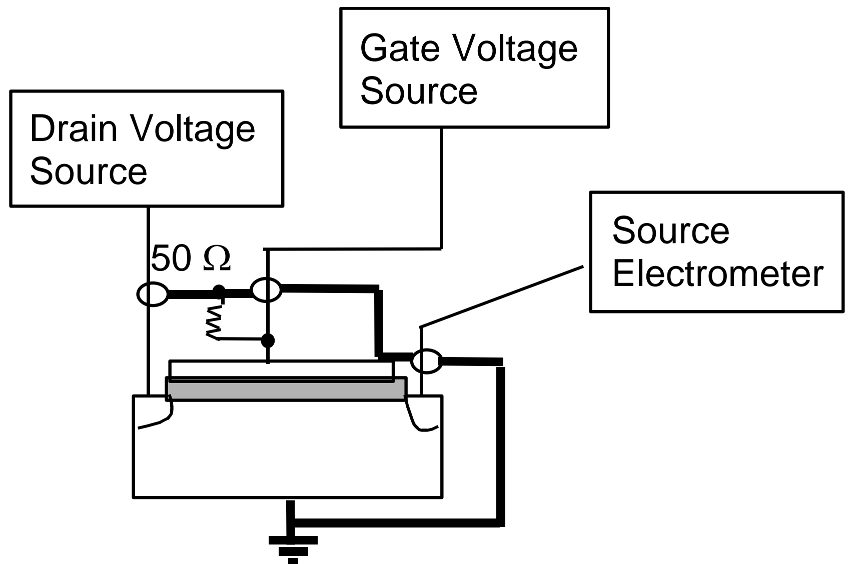


Fig. 4.5. The experimental set up for NBTI study.

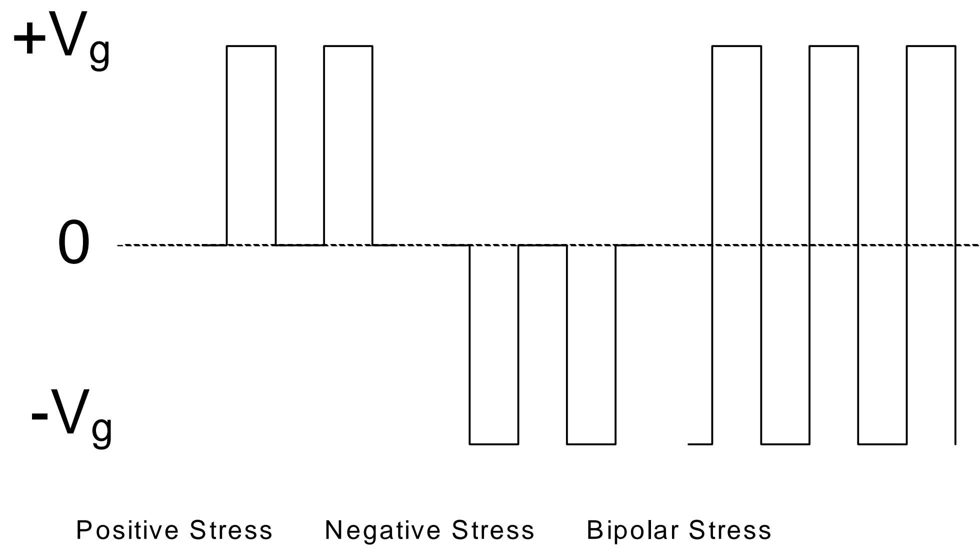


Fig. 4.6. The stress waveforms for the pulsed NBTI study.

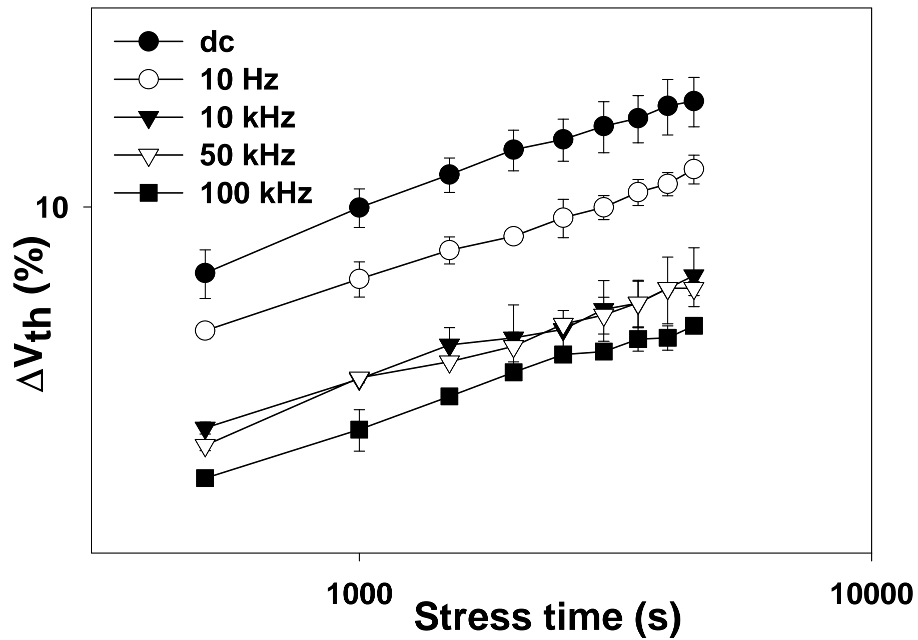


Fig. 4.7. The frequency dependence of ΔV_{th}.
Stress condition: Vstress=-3.0 V, T=150 °C.

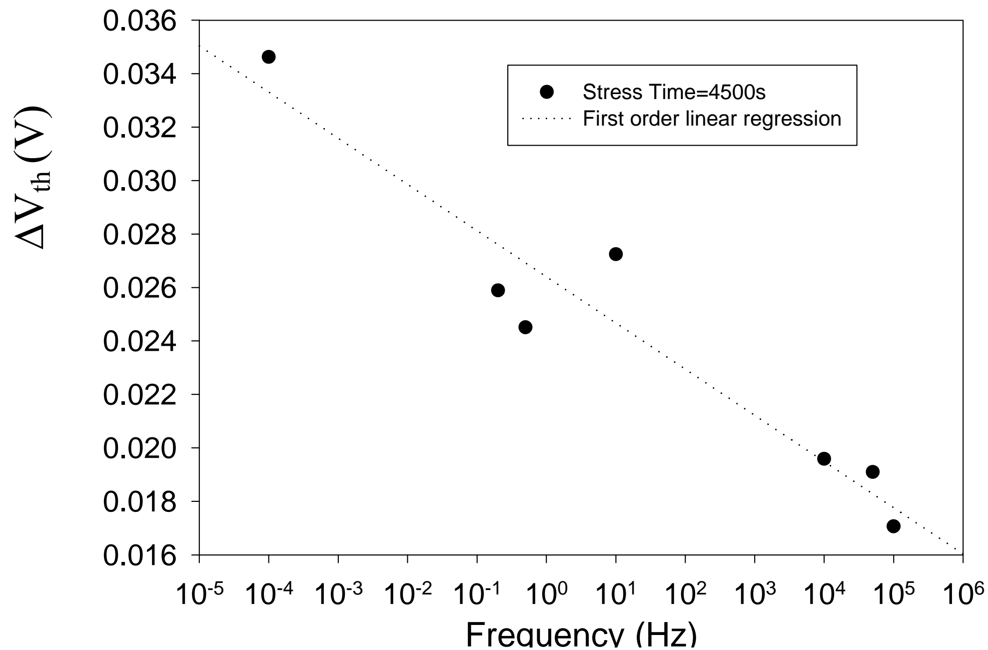


Fig. 4.8. ΔV_{th} vs. frequency, under negative unipolar bias stresses,
Stress condition: V_{stress}=-3.0 V, T=150 °C.

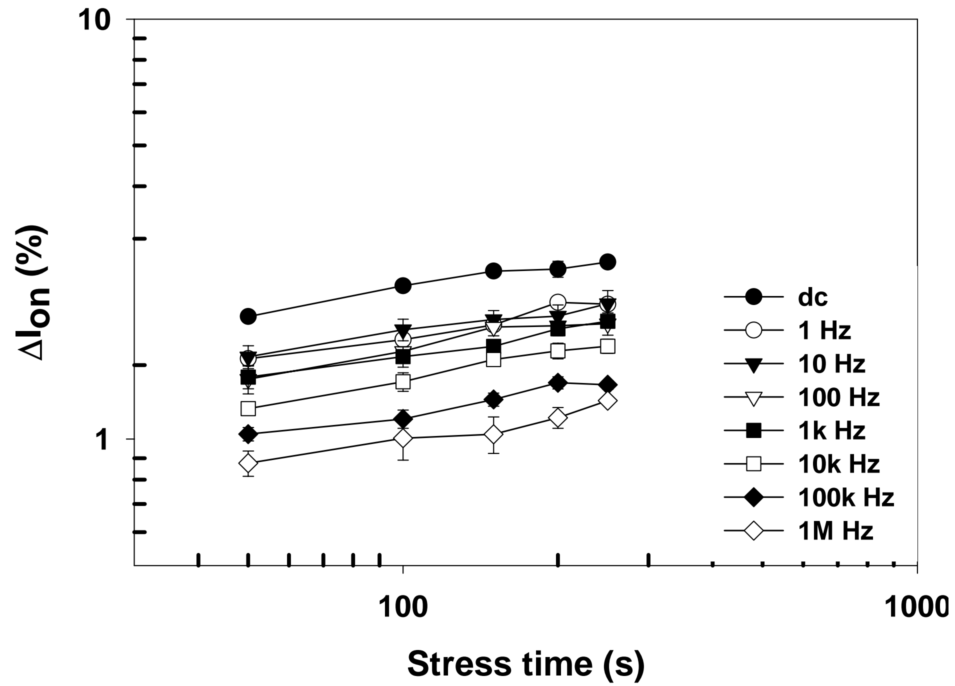


Fig. 4.9. The frequency dependence of ΔI_{on}.
Stress condition: V_{stress}=-3.0 V, T=150 ^{o}C.

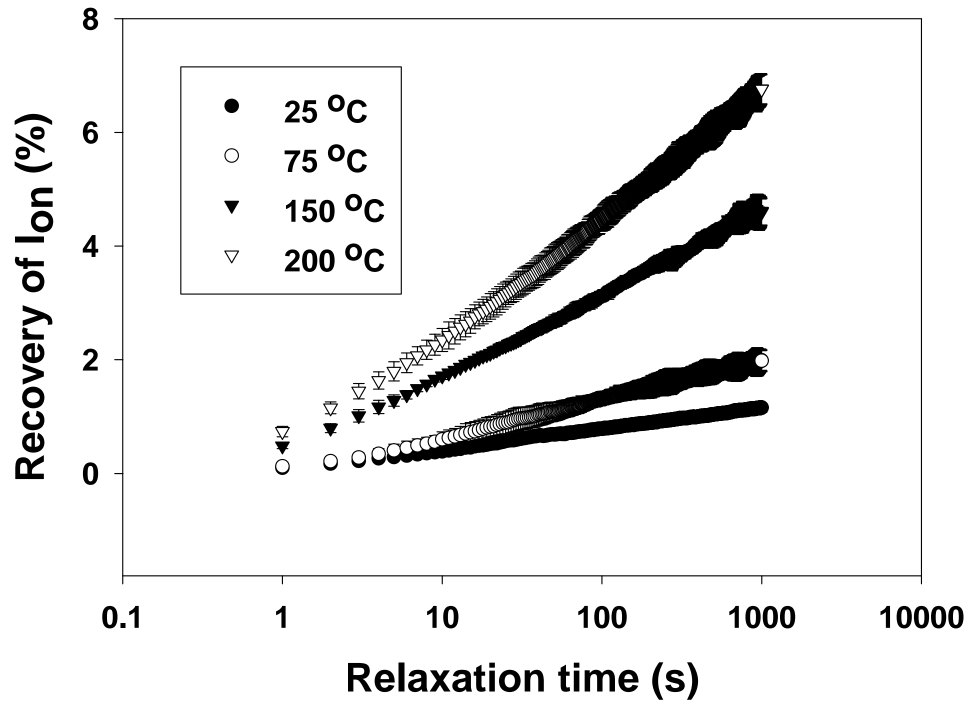


Fig. 4.10. The recovery of I_{on} at different temperatures.
Stress condition: V_{stress}=-2.0 V, total stress time =1000 s.

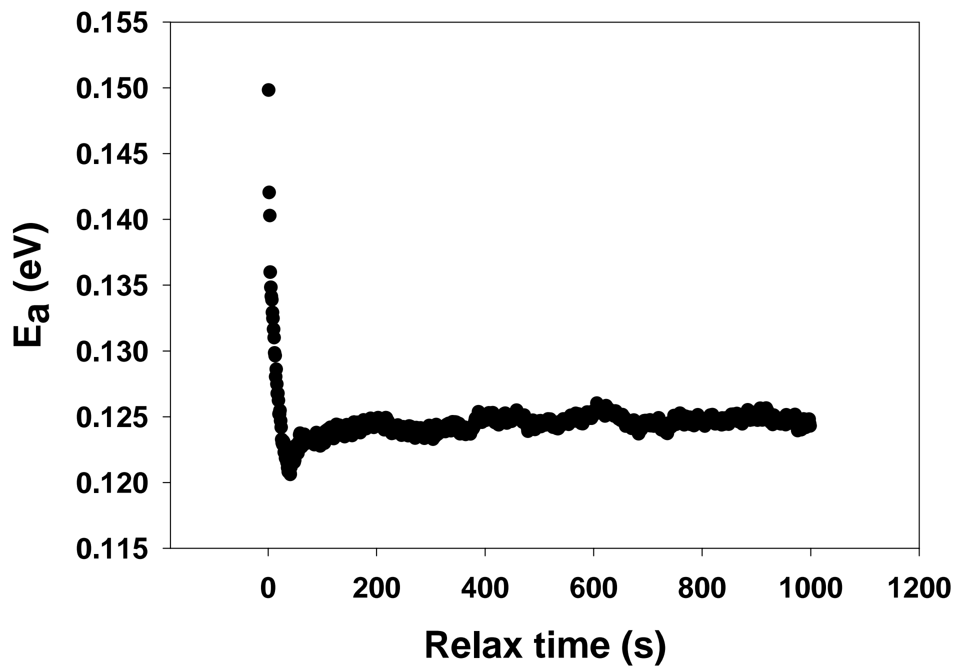


Fig. 4.11. The activation energies extrapolated from data in Fig. 4.10. Two stages are observed.

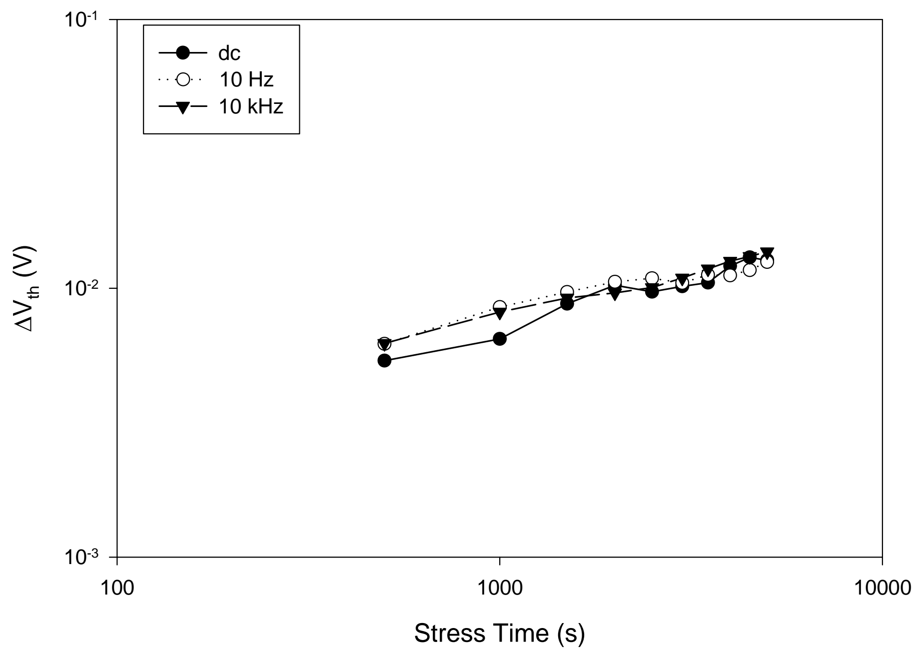


Fig. 4.12. ΔV_{th} under positive unipolar stress conditions with different frequencies, V_{stress}=+3.0 V, T=150 ^{o}C.

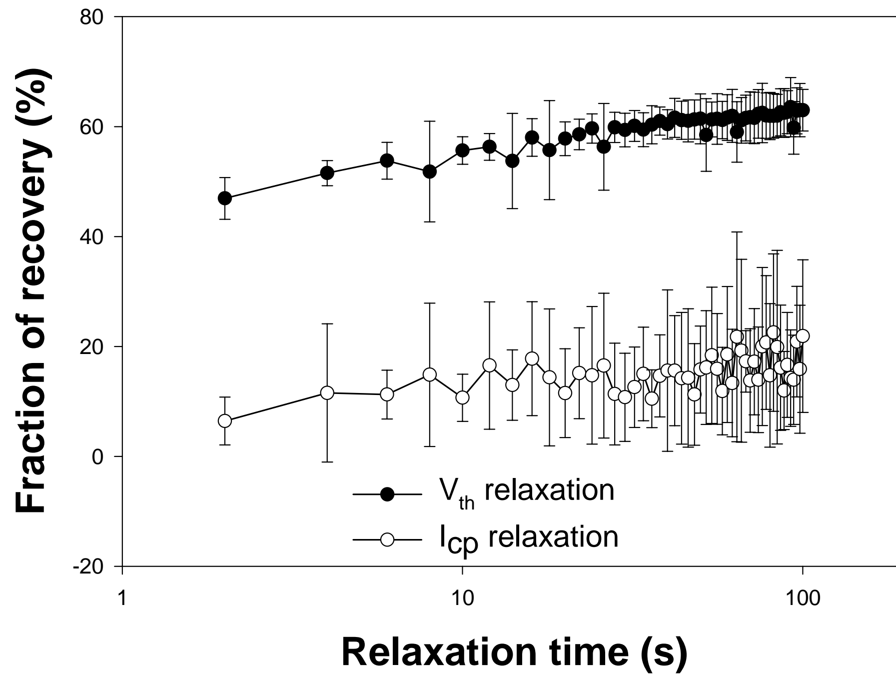


Fig. 4.13. The annealing of ΔI_{cp} and ΔV_{th}.
Stress condition: V_g=-3 V, T=75 °C, total stress time =1000 s.

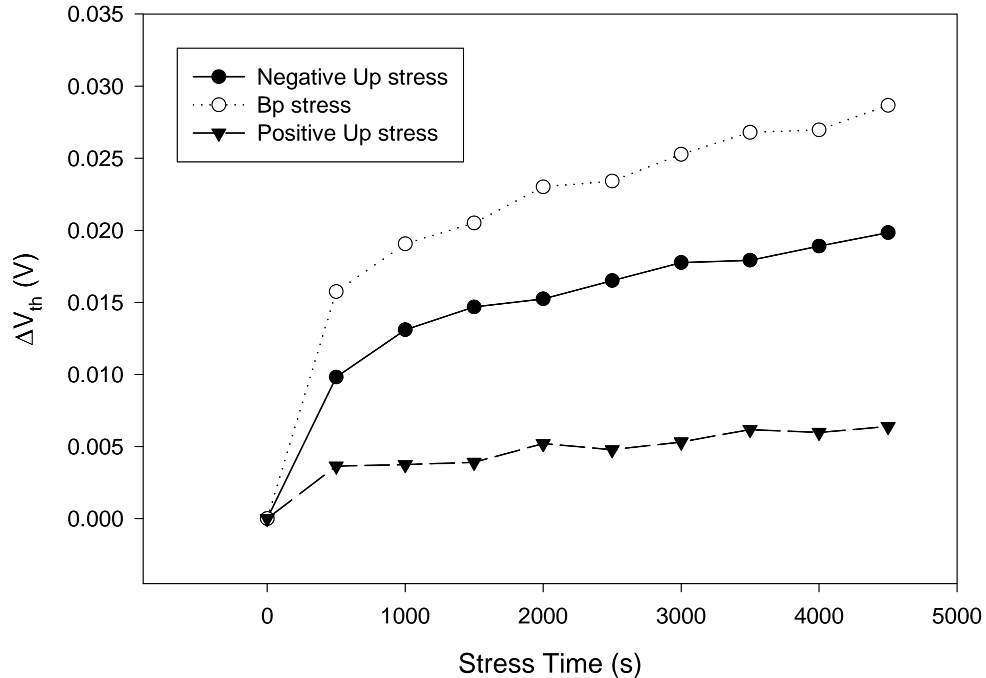


Fig. 4.14. ΔV_{th} under different bias stress conditions with a 10 kHz frequency, V_{stress}=+3.0 / -3.0 V, T=150 °C.

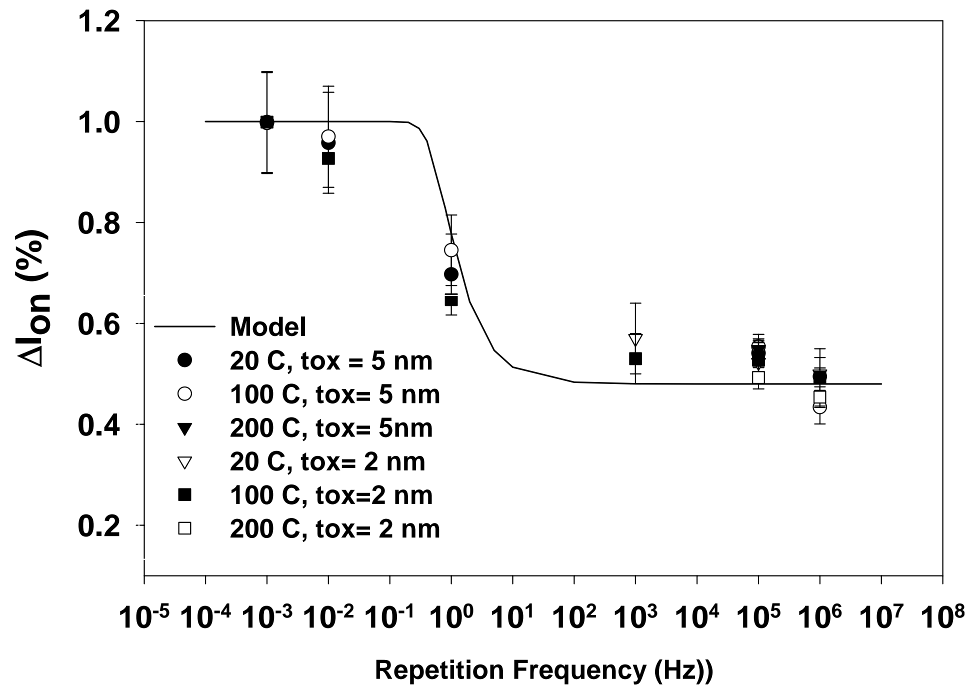


Fig. 4.15. The model fitting of ΔI_{on} with different stress conditions and different oxide thickness.

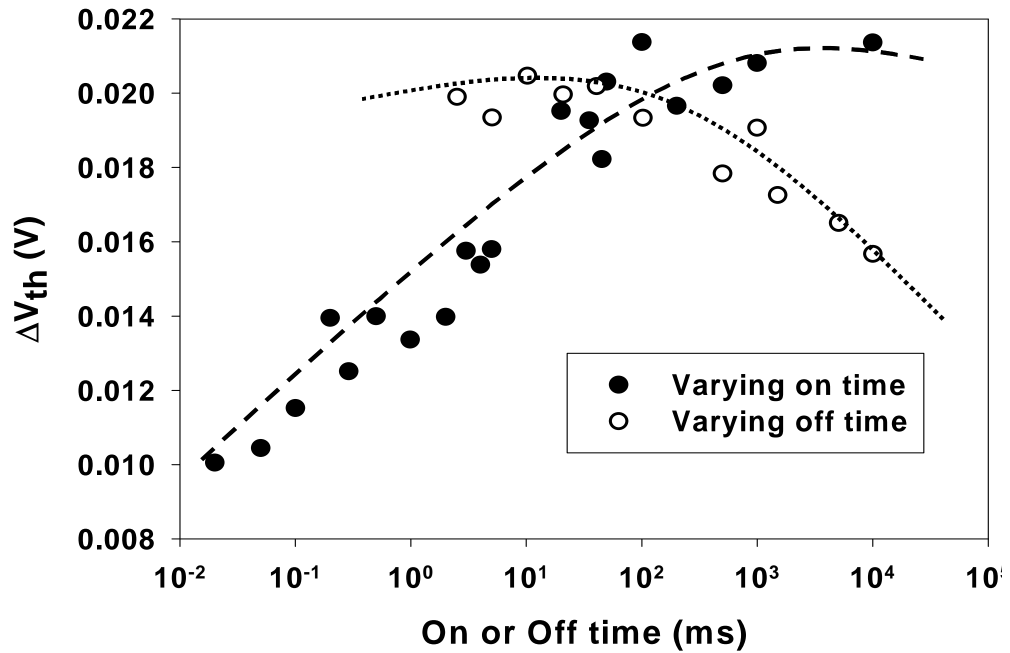


Fig. 4.16. The measured time constants related with the stress on and stress off period.

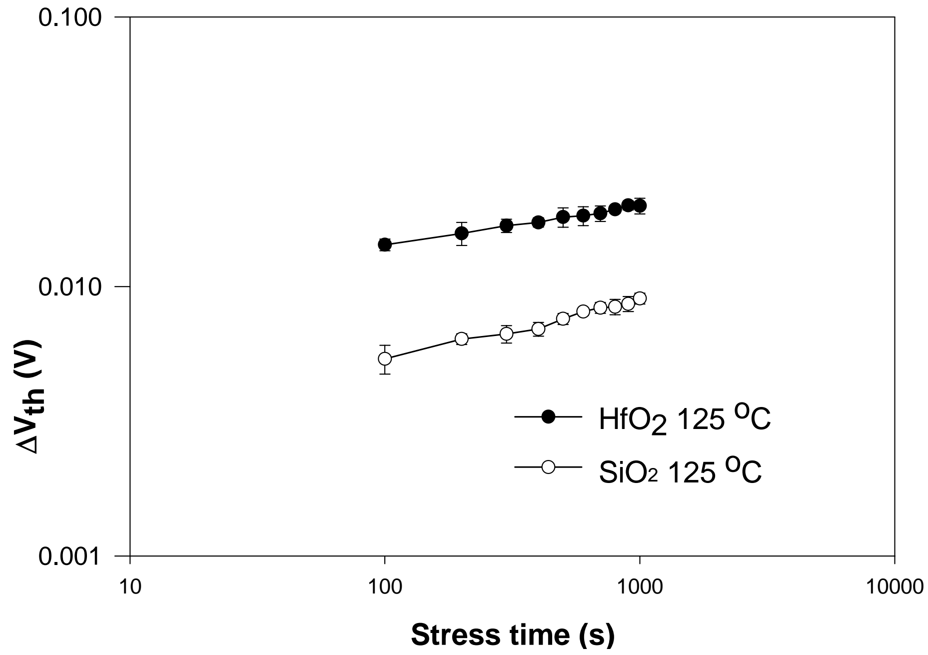


Fig. 4.17. The time evolution of ΔV_{th} of HfO_2 devices and the control SiO_2 samples.

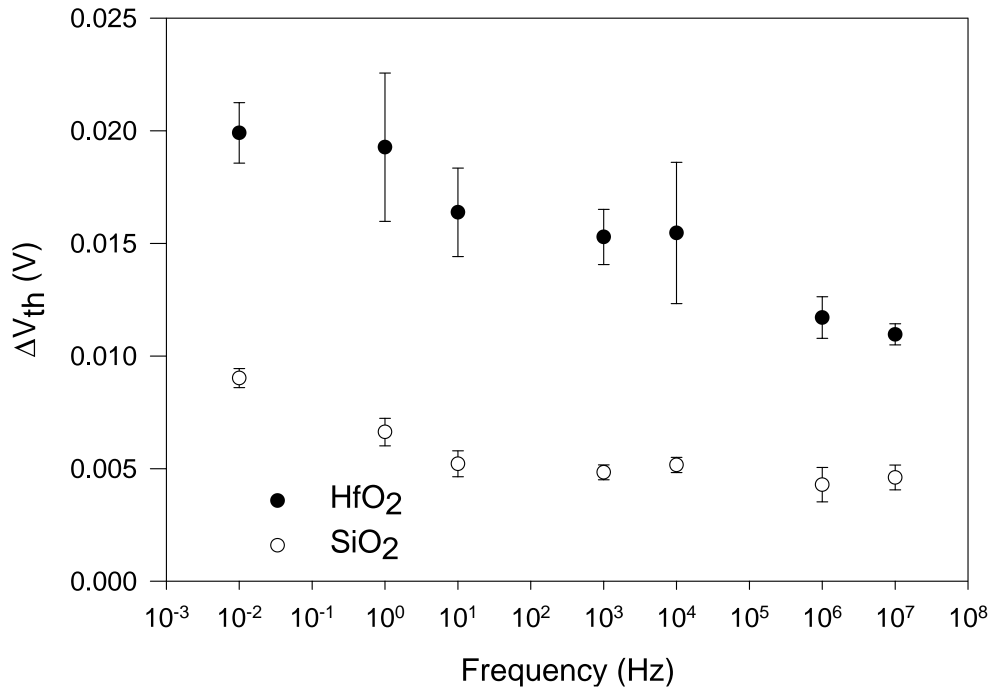


Fig. 4.18. Comparison of the frequency dependence of ΔV_{th} of HfO_2 devices and the control SiO2 sample. Total stress time is 1000 s, stress temperature is 125 ^{o}C. For better view, the frequency is set to be 0.01 Hz for the dc strees condition.

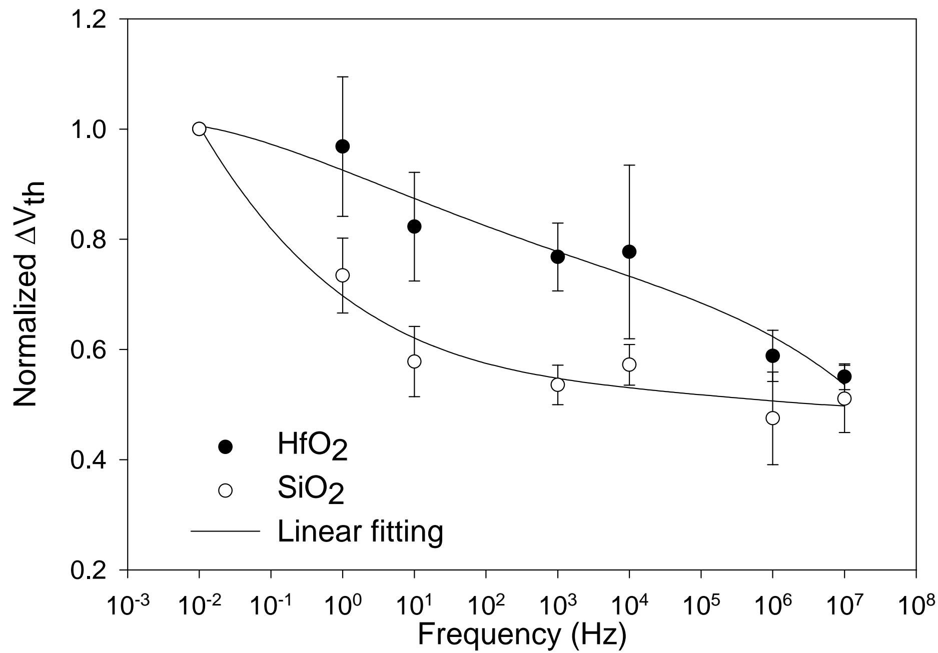


Fig. 4.19. Comparison of the frequency dependence of normalized ΔV_{th} of HfO2 devices and the control SiO2 sample. The data are taken from Fig. 4.18. The line fitting is only for better view.
For better view, the frequency is set to be 0.01 Hz for the dc stress condition.

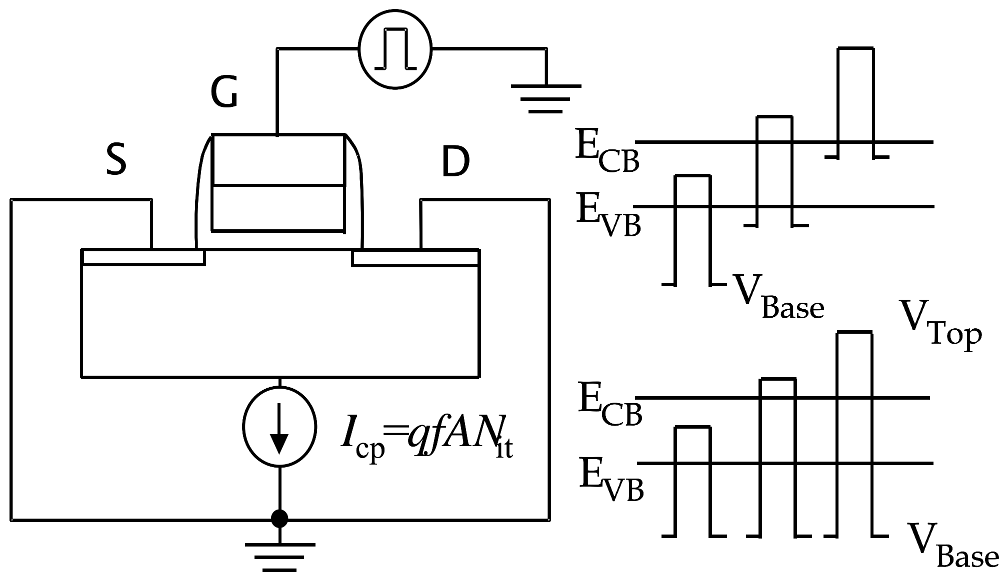


Fig. 4.20. The two charge pumping measurements. In FACP, the V_{Base} is swept; in VACP, the V_{Top} is swept.

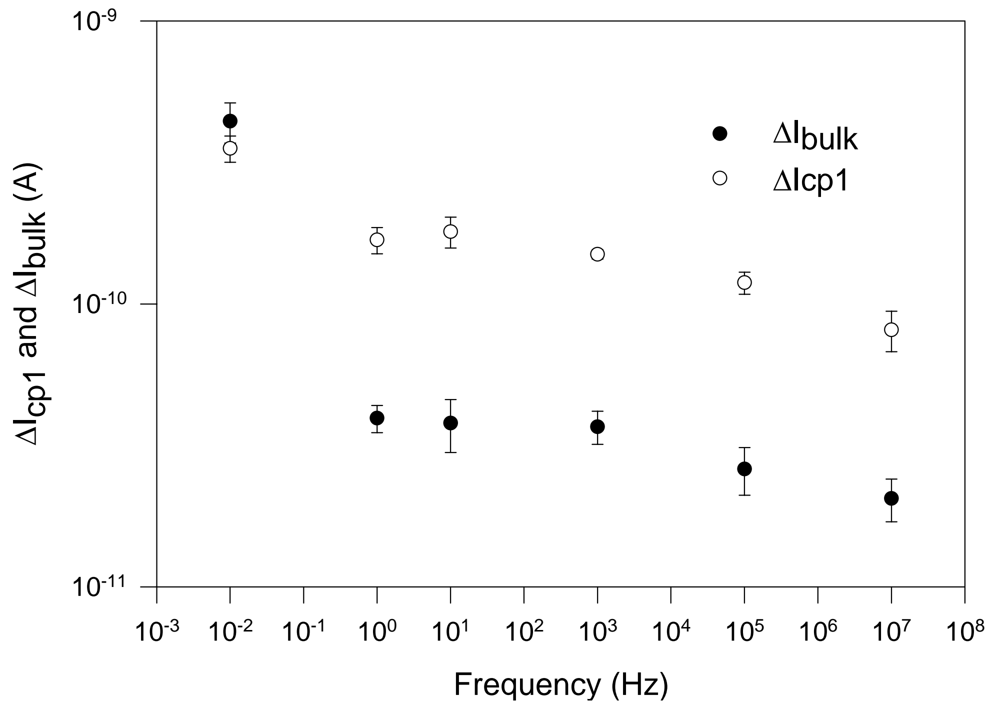


Fig. 4.21. The frequency dependence of generations of interface traps and bulk traps. Total stress time is 1000 s, stress temperature is 125 ^{o}C. For better view, the frequency is set to be 0.01 Hz for the dc stress condition.

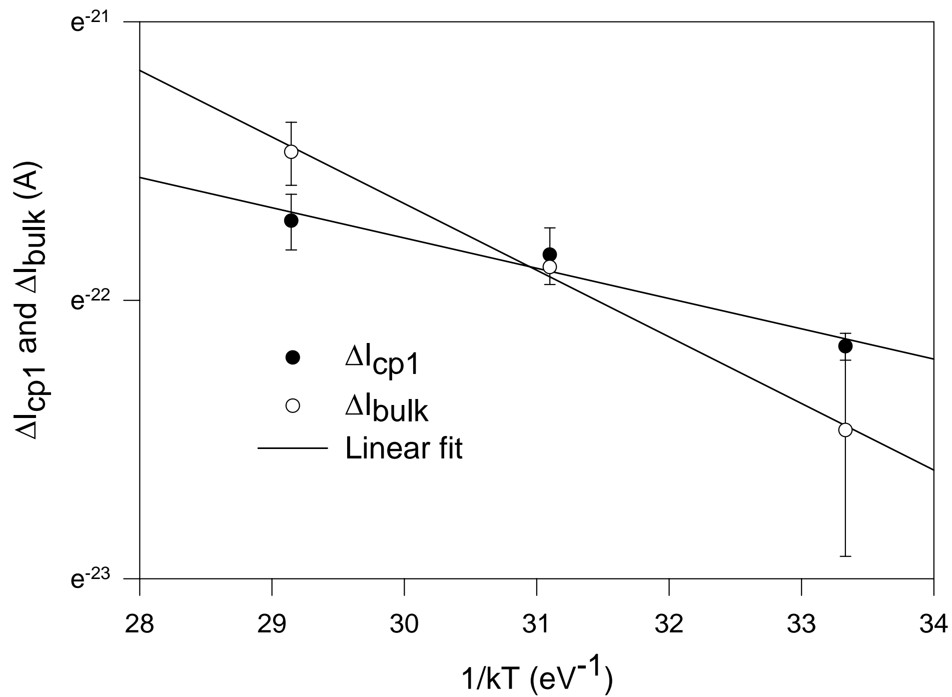


Fig. 4.22. Different temperature acceleration factors of charge pumping currents due to interface traps and bulk traps.

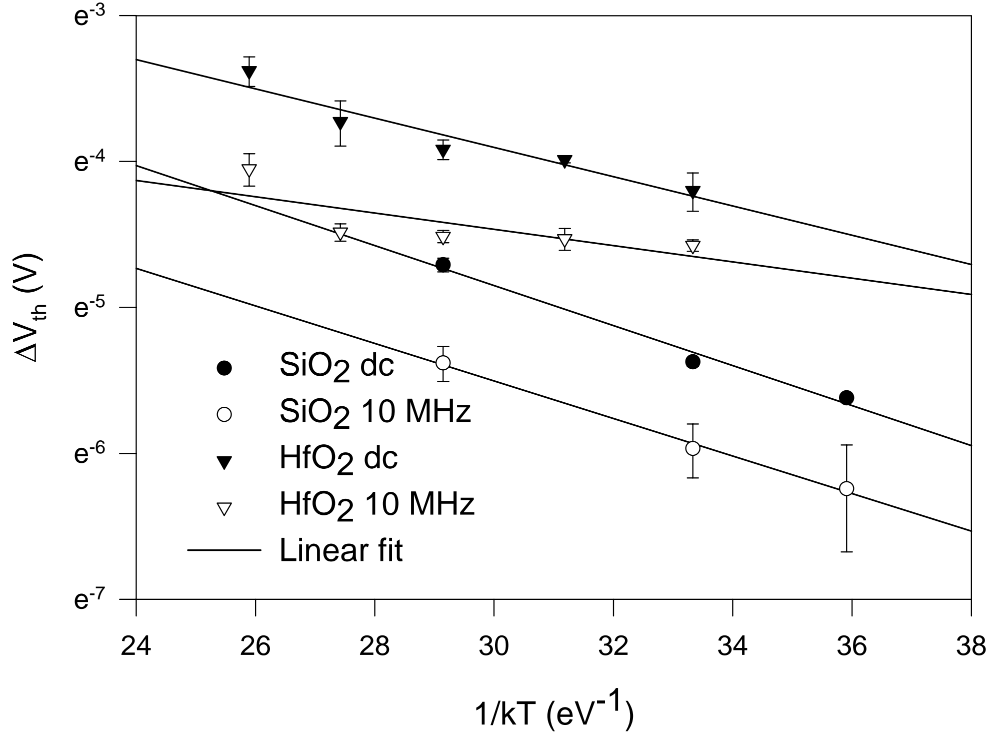


Fig. 4.23. The temperature acceleration of generations of interface traps and bulk traps. Total stress time is 1000 s, stress temperature is 125 ^{o}C.

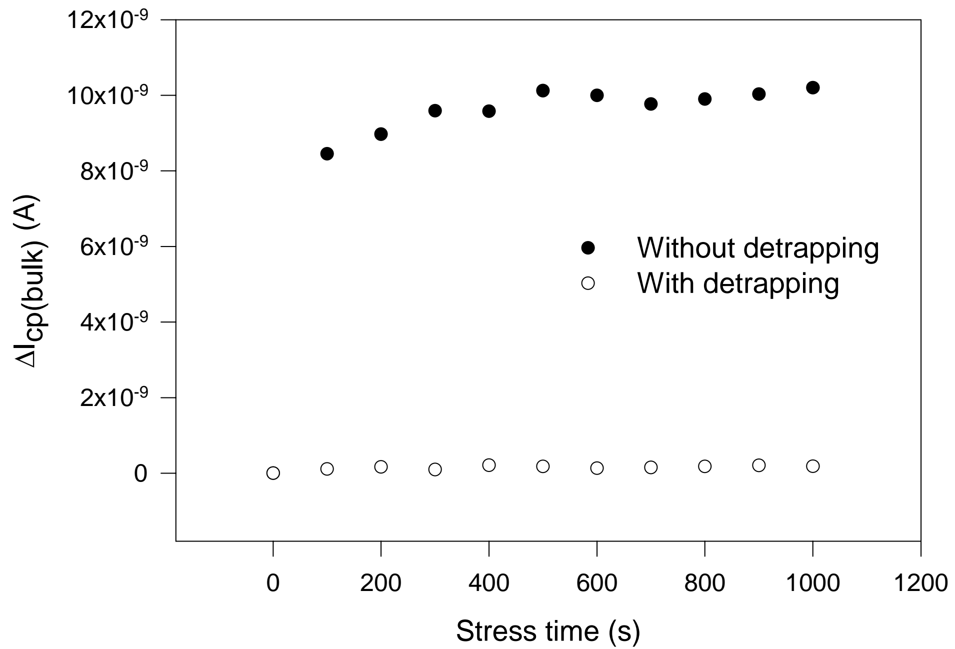


Fig. 4.24. The charge pumping current due to the bulk traps with and without detrapping period.

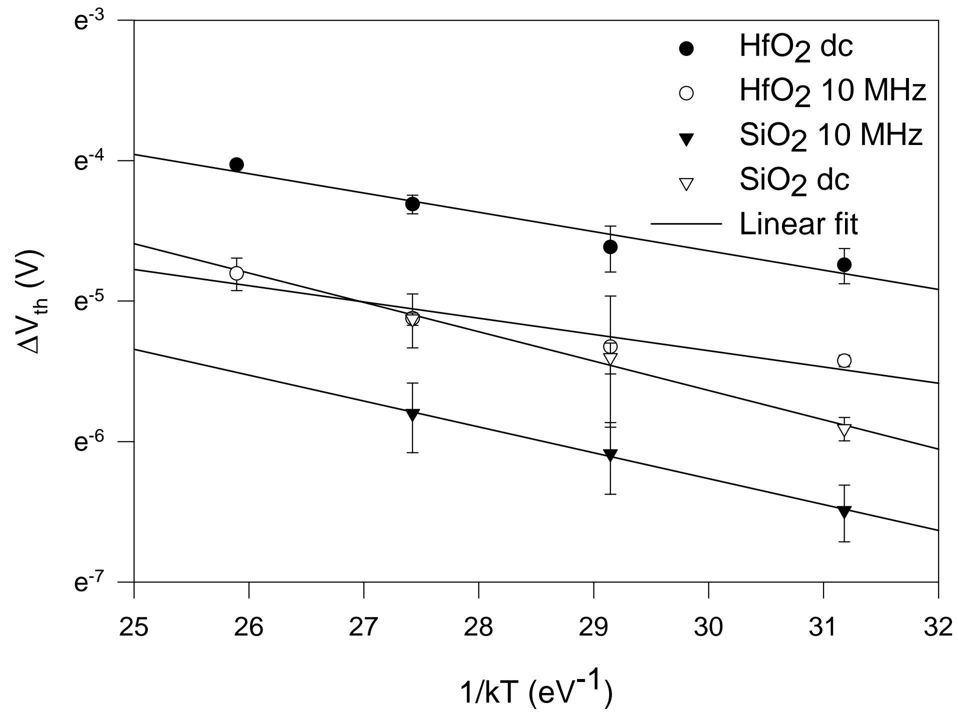


Fig. 4.25. The temperature acceleration of ΔV_{th} of HfO_2 devices and control SiO_2 devices at two stress conditions. The stress condition is the same with Fig. 4.24, except the addition of a detrapping period before each measurement.

Chapter 5

Progressive Breakdown of Ultrathin Gate Dielectrics

5.1 New Breakdown Phenomena of Advanced Gate Dielectrics

5.1.1 Hard Breakdown and Soft Breakdown

Dielectric breakdown has been defined as sudden loss of the insulating properties of the dielectric leading to a large leakage current increase through a localized spot caused by ohmic conduction, which is called hard breakdown. When stressed by a constant voltage, the manifestation of a hard breakdown event is the sudden increase of the leakage current. When stressed by a constant current source, the hard breakdown has a sudden change in the stress voltage. Thus, a hard breakdown event is easier to detect and define.

As the semiconductor device enters into the ultrathin gate dielectric regime, the breakdown phenomenon also shifts from the hard breakdown to the soft breakdown [11, 108-111]. The soft breakdown in the dielectric only causes a current increase and it may be tolerable from the device point of view and may not be considered as a

circuit failure. The soft breakdown is also associated with an increase in the noise of the stress current, which can be used as a breakdown monitor.

5.1.2 Definition of Failure

The importance of the definition of a failure event is evident. However, with the noisy characteristics of soft breakdown, this definition is not as easy as the hard breakdown.

One of the popular definitions is using the noise level of the leakage current as a signature [112-114]. A calculation example is given as follows using the methodology recommended by JEDAC.

Let $I_{g(n)}$ be the nth gate current data, then the noise can be calculated as:

$$N(n) = \sqrt{\frac{\left(I_{g(n-4)}^2 + I_{g(n-3)}^2 + I_{g(n-2)}^2 + I_{g(n-1)}^2 + I_{g(n)}^2\right) - \frac{\left(I_{g(n-4)} + I_{g(n-3)} + I_{g(n-2)} + I_{g(n-1)} + I_{g(n)}\right)}{5}}{4}}$$

As we can see from Fig. 5.1, after stressed 8200 s, the noise level increased at least one order, which clearly indicates the soft breakdown event.

However, as we will see later, that the sudden increasing of the current noise may not be a good criterion for the final breakdown, which indicates the failure of the device and the whole circuit. In our following tests, we define an absolute current level of 200 μA as the final breakdown.

5.1.3 New Reliability Methodology

There are several methodology models about soft breakdown. They are different as different soft breakdown mechanisms are adopted.

The first claims that the soft breakdown and hard breakdown are competing events with different origins [115, 116]. They have different statistical distributions, voltage accelerations and activation energies.

The second model assumes that the soft breakdown and hard breakdown have a common failure mechanism [117]. The soft breakdown and hard breakdown statistically coincide, indicating they have the same defect generation and breakdown triggering process. Based on this model, a prevalence ratio methodology is used to predict the soft breakdown probability. In a breakdown experiment, the first breakdown of each sample is defined as soft breakdown or hard breakdown. The prevalence ratio (α_{SBD}) is the fraction of samples that first experience soft breakdown, given by

$$\alpha_{SBD} = \frac{n_{SBD}}{n_{BD}}.$$

The final breakdown distribution is give by F_{HBD}, regardless of the first breakdown characteristics. A relationship between the final hard breakdown and the total breakdown event are established by the prevalence, given by

$$Ln(-Ln(1 - F_{FHBD})) = Ln(-Ln(1 - F_{BD})) + Ln(\alpha_{HBD}).$$

According to this theory, the same type of defect paths triggered the hard breakdown and soft breakdown, and the hard breakdown event would not be a consequence of lateral propagation of the soft breakdown spot to nearby regions.

Another statistically independent successive breakdown model is also established based on this theory. The generation of breakdown paths is uniform and uncorrelated. This approach assumes that several statistically independent soft breakdowns occur

before the leakage current in the gate dielectric is large enough to cause circuit failure and those breakdown filaments are stable with subsequent stress time. It has been shown that the prevalence method and the successive breakdown approach are statistically compatible and can be related [117].

Different from the statistically independent breakdown mode, a different mode called progressive breakdown has been observed where the breakdown path formed by the initial breakdown continue wear out as the stress continues [118-121]. This new breakdown behavior has been confirmed by many reports [122-124]. In order to precisely project the lifetime of a device, the progressive breakdown should be studied separately: the initial breakdown and the evolution of the breakdown path.

5.1.4 Impact of Soft Breakdown

At first, it is reported that there is no significant degradation after the soft breakdown on device characteristics, i.e. the device does not fail after soft breakdown under constant current stress and constant voltage stress [116].

It is also found that hard breakdown occurs more likely for short channel devices in overlap regions and long channel device survives after soft breakdown [125]. Thus narrow devices are more sensitive to soft breakdown.

The observation that some circuits remain functional even after a gate oxide breakdown has raised much interest in studying the physical mechanisms of post breakdown conduction in ultrathin gate oxides[118, 126-129]. Reported by many reports, the first gate oxide breakdown dose not necessarily imply the failure of the product, thus the circuit lifetime projection can be extended[126].

5.1.5 Dielectric Breakdown of HfO_2

As a result of the scaling of gate dielectrics, materials other than SiO_2 are been incorporated into the device structures. HfO_2 is among the promising candidates for replacing SiO_2. Thus, detailed study of the HfO_2 becomes important in order to develop lifetime extrapolation models.

In order to increase the stability, an interfacial layer of SiO_2 is always inserted between the silicon substrate and the HfO_2 bulk oxide [130-135]. Thus a question is raised that if there is a breakdown event, where is the exact location of this breakdown event [136-139]. In the following experiments, we used the CVS tests to study the breakdown characteristics of HfO_2 and compared them with SiO_2.

The devices used in this work are HfO_2/Hf silicate hybrid stack pMOSFETs as shown in Fig. 5.2. The gate dielectrics were fabricated by metal organic chemical vapor deposition (MOCVD), post deposition annealing (PDA) with NH_3 at 700 ^oC for 60 s. The equivalent oxide thickness (EOT) is 2.2 nm. Between the HfO_2 bulk oxide and silicon substrate, there is an interfacial layer of SiO_2 with a thickness of 1.0 nm.

Compared with a control SiO_2 sample with an oxide thickness of 1.6 nm, there are two distinct features.

First, the temperature acceleration factors are quite similar, as indicated by Fig. 5.5. The activation energy for the HfO_2 devices is 0.76 eV, and for SiO_2 devices is 0.83 eV. However, this does not guarantee that the same breakdown path is formed, as it is find that the activation energy will change with different gate voltages [34, 46, 140-142].

However, when comparing the voltage acceleration factors, there is a clear difference between these two kinds of devices. When plotted in a Logarithm-Logarithm scale, as in Fig. 5.6, the SiO_2 device shows a slope of 44.2, which is very consistent with other reports [44, 57]. For the HfO_2 devices, a slope of 18.9 is obtained. Note that the number 44 is universal to SiO_2 with different oxide thickness, stress voltages, and temperatures, according to the power law distribution. Thus, if the breakdown happens at the interfacial layer, a similar slope should still be observed. This may indicate a different breakdown pattern than the SiO_2 devices.

One of the possible reasons is that when the HfO_2 devices are stressed with subjection injection, the breakdown is formed at the HfO_2 bulk oxide inside of the SiO_2 interfacial layer, as shown in Fig. 5.7. This observation is supported by some reports [71, 136], while contradictory to other reports [14, 72]. More work is still needed in this field.

We also performed Substrate Hot Electron Injection (SHEI) experiment with HfO_2 devices. Similar to the SiO_2 devices, SHEI significantly shortens the T_{bd} of devices [23]. Which suggest the breakdown mechanisms of HfO_2 are similar to those of SiO_2 devices. As shown by Fig.5.8 and Fig. 5.9, the SHEI substantially increased the gate current, and consequently reduced the lifetime of the devices.

5.2 Progressive Breakdown of SiO_2

5.2.1 Introduction

As briefly discussed in chapter 3, the lifetime extrapolated from the accelerated tests largely rely on the accurate physical model established for the degradations. It is still unclear if the physical mechanism of the post first breakdown is the same with the first breakdown. Thus, the understanding of the physical mechanism of the soft breakdown becomes critical. Since some MOSFET digital circuits could remain functional after gate oxide breakdown, the separate consideration of soft breakdown and hard breakdown events is necessary to set up an adequate application-specific reliability assessment methodology.

We studied the two kinds of soft breakdown phenomena, the stable soft breakdown and the unstable soft breakdown. It is found that the hardness of the first breakdown will influence the evolution and progression of the subsequent breakdown event. For different breakdown phenomenon, the acceleration factors are different, which indicate different physical mechanisms.

5.2.2 Devices and Experimental Setup

The devices used in this work are nMOSFETs with a gate oxide thickness ranging from 1.6 nm to 2.2 nm. The device channel length and width are 0.25 μm and 10 μm respectively. All devices are stress with constant voltages in the accumulation mode. The first breakdown criterion was set to be a 100 % increase in the gate current I_g. According to the percolation theory, the first breakdown happens when a low resistance filament is formed by the defects generated in the gate oxide. This filament

causes a sudden surge of I_g through the oxide. The power dissipation through the filament can locally produce a large amount of heat, causing permanent structural damage along the path of the filament.

In order to limit the power dissipation through the filament during the first breakdown, discrete ohm resistors were inserted between the gate of the device and the power supply. Since the gate oxide ohm resistance is much larger than the ohm resistors used in the study, which were 10 $\mathrm{k}\Omega$, 30 $\mathrm{k}\Omega$ and 500 $\mathrm{k}\Omega$, the reduced voltage by the ohm resistor was relative small. Thus, we use the same gate voltage for different test and consider them the same. However, after the first breakdown, the resistance of the gate oxide reduced largely by the filament. Thus, the different resistor can determine the seriousness of the first breakdown. With a larger ohm resistor, the magnitude of the leakage current was much smaller, the power dissipated was reduced, and the breakdown was relatively softer than those with a smaller value ohm resistor were.

5.2.3 Results and Discussion

The stable and unstable soft breakdown phenomenon

Fig. 5.10 and Fig. 5.11 show the typical behavior of stable and unstable soft breakdown. Various phases during the evolution of hard breakdown are identified. After the wear-out period, where the percolation path is first formed and caused the first breakdown, the digital phases was defined as the region in time when the gate current increased from 1 μA to 10 μA and analog phase was the region when the current increased from 10 μA to 200 μA. Different from the analog phase, which

indicates a monopoly increases of current; the digital phase is usually accompanied by noisy fluctuations. This is also confirmed by other researchers [143]. The slope of the analog phase is defined as the current degradation rate. The residual time is the time between the wear out and the final breakdown. A temperature and voltage dependence of both the digital and analog phases were found and will be studied later.

During our tests, both of two degradation modes happen. However, the stable soft breakdown mode is preferable if the voltage polarity used in the second stress were reversed from the first breakdown. In order to specifically study the stable soft breakdown, this test condition is used to generate the stable soft breakdown events.

With a stable filament, the gate current continuously increased with very small noise until final breakdown. Even though the digital and analog phases can still be defined, the length of the analog phase for different stress conditions were almost independent of the stress voltage and temperature, which suggest that it is more likely due to the behavior of the hard breakdown current due to the series resistance of the test structures.

Analysis of the unstable soft breakdown

Noticeably, there are two distinct phases in the evolution of an unstable soft breakdown, the digital phase and the analog phase, as shown in Fig. 5.12. The digital phases features a relative current level accompanied a significant noise level; while the analog phase shows a continuous increasing noisy current. In our study, for better clearance, we define the digital phase to be the time for the current increases from 1

μA to 10 μA, and the analog phase to be that from 10 μA to 200 μA. This is observed to be consistent with most of the test data.

Based on our analysis, it appears that the digital phase and the analog phase have different physical mechanisms and they are not correlated with each other. Fig. 5.13 shows the time length of both phases. We observe a largely scattered relationship between them.

The correlation between the first breakdown and the final breakdown

This is a simple methodology to detect a nanometer-precise position of each breakdown spot along the channel length direction [144, 145]. The relative breakdown position is calculated from $I_d/(I_d + I_s)$ at negative V_g and source and drain are grounded with the substrate. Depending on whether the majority of the electrons injected through the breakdown path are collected by the source or the drain, $I_d/(I_d + I_s)$ will be close to 0 or close to 1 respectively. If $I_d/(I_d + I_s)$ is close to 0.5, this means the breakdown position is around the middle of the channel.

Using this technique, the breakdown locations of the first breakdown and final breakdown were studied. Fig. 5.14 and Fig. 5.15 show the ratio obtained after the first breakdown and the final breakdown. The ratios of the unstable filaments are generally correlated. However, for the stable filaments, the ratios are uncorrelated, which clearly indicated the independence of the first breakdown and the final breakdown.

The voltage and temperature acceleration factors

The accelerating factors for the post breakdown were studied. It is found that the voltage and temperature acceleration factors used to predict the residual time will depend on the type of breakdown filament formed after the first breakdown.

Fig. 5.16 and Fig. 5.17 show the voltage and temperature acceleration factors for the evolution of unstable filaments. It was observed that the temperature dependence was not Arrhenius and exhibited a much better fit when plotted to $T_{bd} \propto \exp(AT)$ with T as the temperature and A as a constant. Note that the temperature acceleration is voltage independent. The exact mechanism by which the unstable breakdown filament continues to grow and result in a hard breakdown event is still unknown

Fig. 5.18 compares the voltage dependence of the residual time to hard breakdown for devices stressed with stable filaments, unstable filaments and that of the first breakdown. Clearer, the voltage dependence of the devices where the breakdown filament is stable is identical to the dependence observed for the firs bd. This confirms that the stable breakdown filaments share a common origin with the first breakdown and are physically different that those unstable filaments.

The temperature dependence of those different filaments further proves that the stable filament mode has the same origin as the first breakdown as shown in Fig. 5.19. The temperature dependence of the residual time for the stable filaments is almost identical to the first breakdown and quite different from unstable filaments.

The influence of SHEI

The Substrate Hot Electron Injection (SHEI) experiment has been extensively used in the gate dielectric integration tests [23]. By using SHEI, the current level through the gate dielectric can be controlled separate from the electric field.

We conducted SHEI on both of the stable soft breakdown and the unstable soft breakdown. Fig. 5.20 and Fig. 5.21 show the influence of SHEI to the analog and digital phases. Clearly, for the unstable soft breakdown, both phases are shortened by

approximately the same amount, indicating that the additional tunneling current affected both phases to a similar degree. However, for the stable soft breakdown, the additional tunneling current did not affect the analog phase. It because that this analog phase is not related to the initial breakdown filament as assumed for the unstable soft breakdown. This analog phase maybe related to the behavior of the hard breakdown current due to the series resistance of the test structure and other factors. The reduction of the digital phase is the SHEI acceleration of the intrinsic oxide defect generation and wearout.

The influence of first breakdown hardness

As mentioned early, various resistors are used for the first breakdown. It is observed that different test set will influence the degree of the first breakdown hardness.

Fig. 5.22 shows the effect of different resistor values on the conductance of the post soft breakdown. The filament leakage current shown in the plot was measured with V_g of 1.5 V. All the devices were stressed at 150 ^{o}C with 3.9 V applied to the gate until the first breakdown was detected. Clearly, the magnitude of the leakage current increases as the resistance decreases. Thus the resistor can roughly control the first breakdown hardness.

Fig. 5.23 shows the effect variety the gate series resistance value on the residual time. The residual time is the time until the final hard breakdown occurs after the first soft breakdown. Four distributions are shown with the stress conditions for the first breakdown shown in the inset. The second stress condition used for obtaining hard breakdown was the same for all of the distributions. The distribution appear bimodal,

especially the distribution obtained with the lowest resistor value. The bimodal residual time distributions have been reported and the concept of mixing of stable and unstable breakdown modes is used to explain this [121]. As indicated, the devices in the experienced the hardest first breakdown are those with 10 kΩ gate series resistor, and they exhibited the largest post-breakdown leakage currents. Furthermore, this distribution also appears to be more bimodal than the other distributions for devices that experienced softer first breakdown. Fig. 5.23 also shows a trend that devices having harder first breakdown generally exhibit shorter residual time.

5.2.4 Conclusion

The phenomenon of progressive breakdown of ultrathin SiO2 is studied. Two kinds of breakdown filaments are observed and studied. It is found that the stable breakdown filament is due to a separate breakdown filament independent from the initial breakdown. However, the unstable filament is the continuing wearout of the initial breakdown filament. The voltage stress acceleration factors and the temperature dependence of the unstable filament are found to be different from the initial breakdown. The influence of the initial breakdown hardness to the unstable filaments is also tested. It is found that with a lower current compliance in the circuit, the generated breakdown is more likely to be soft breakdown.

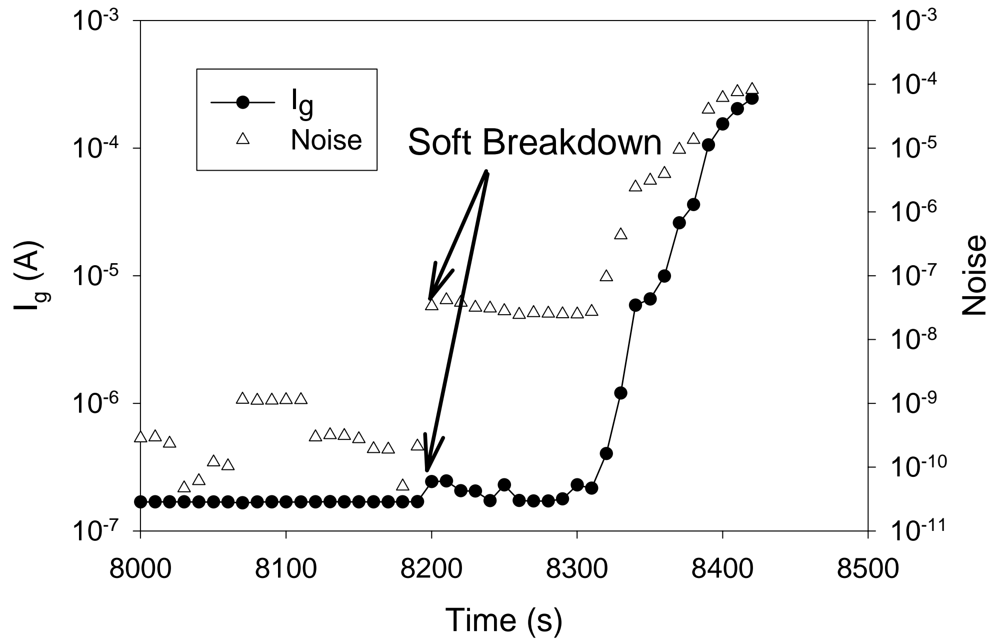


Fig. 5.1. The soft breakdown of gate dielectrics.

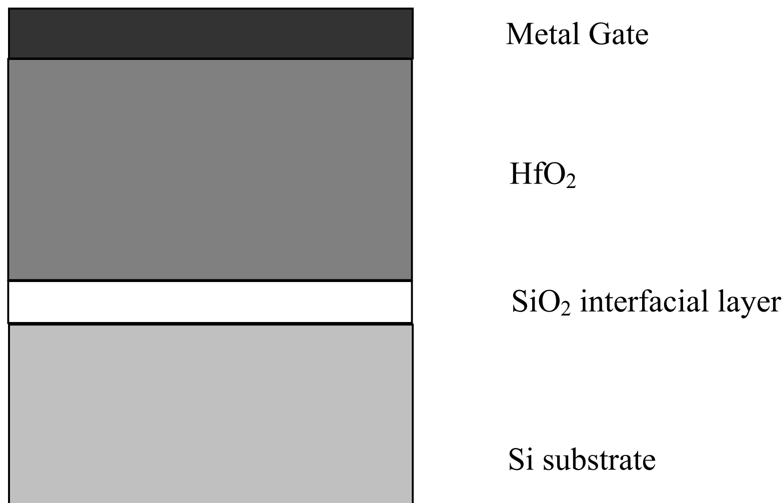


Fig. 5.2. The schematic structure of the HfO_2 device. (Drawn not to scale.)

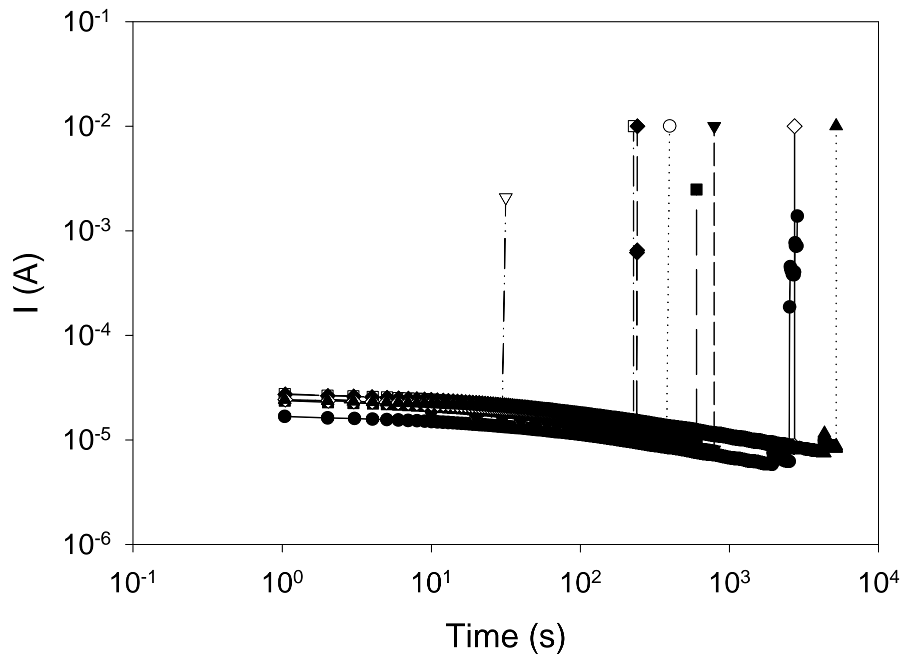


Fig. 5.3. Constant voltage stress of the HfO_2 devices.

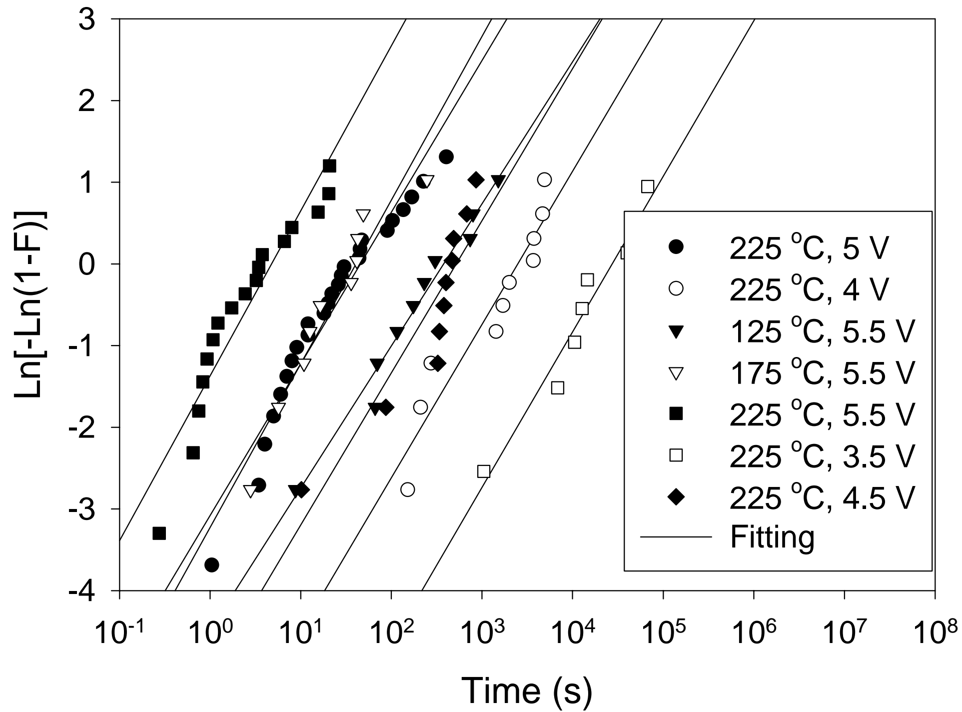


Fig. 5.4. Lifetime distribution of HfO_2 devices.

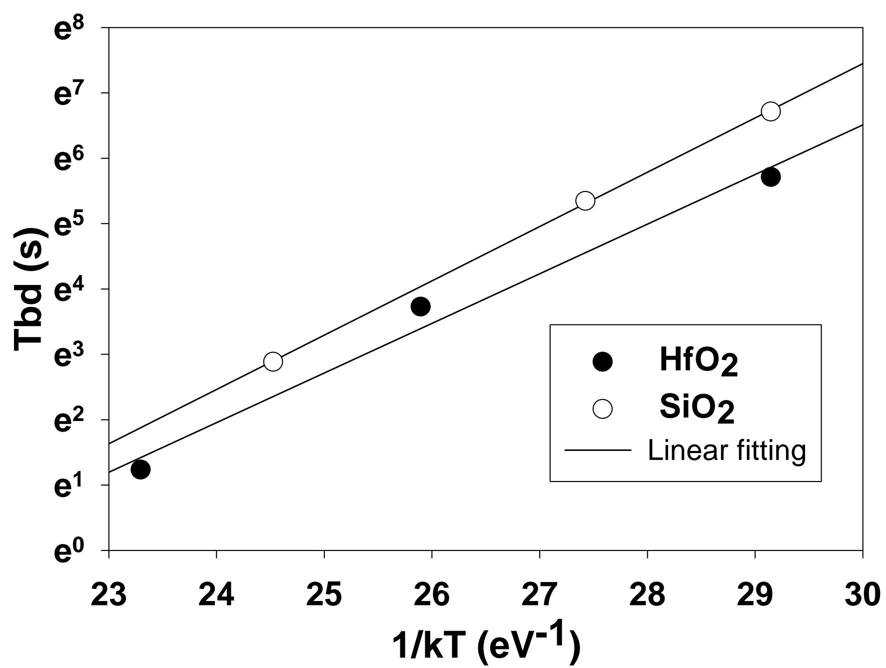


Fig. 5.5. Comparison of the temperature dependence of HfO_2 and SiO_2 devices.

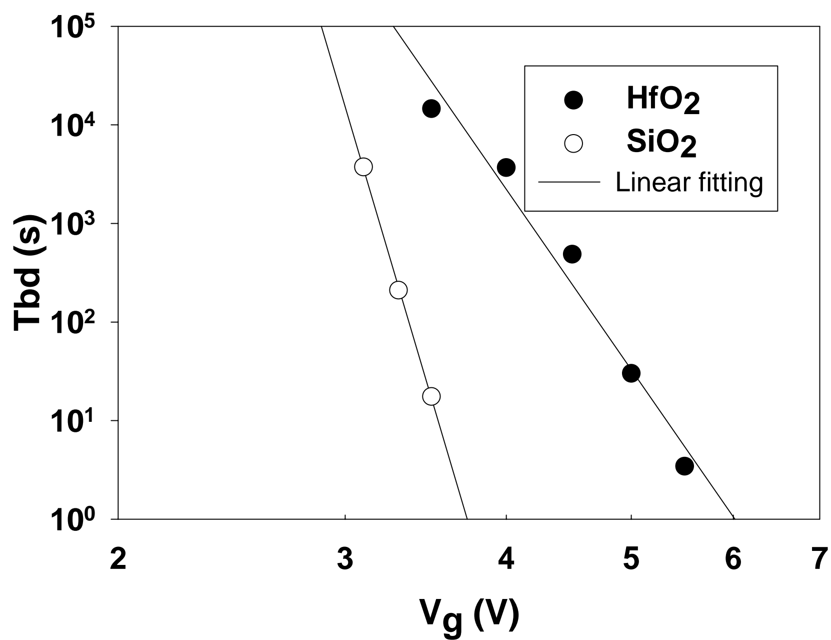


Fig. 5.6. Comparison of the voltage acceleration of HfO_2 and SiO_2 devices.

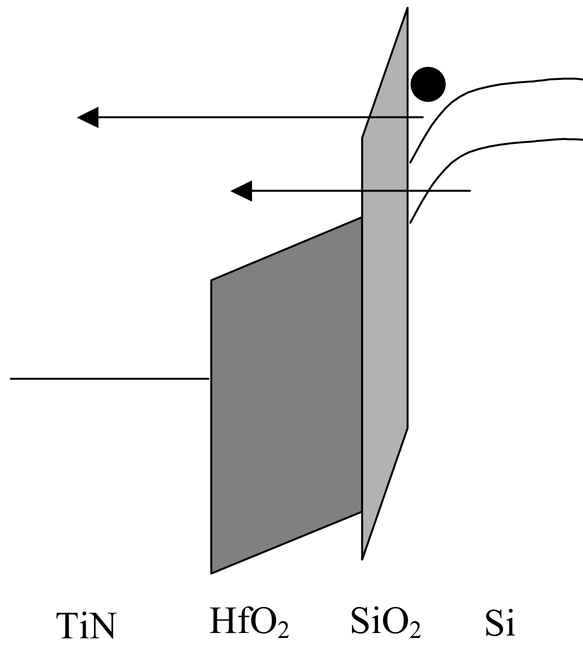


Fig. 5.7. The substrate injection stress of the HfO_2 device.

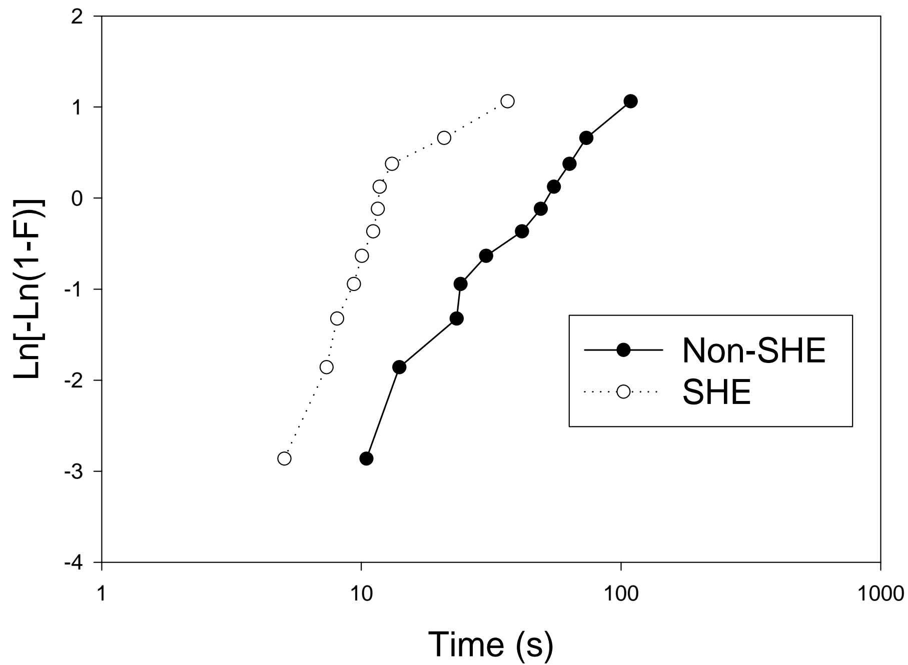


Fig. 5.8. The influence of substrate hot electron injection to the device lifetime distribution.

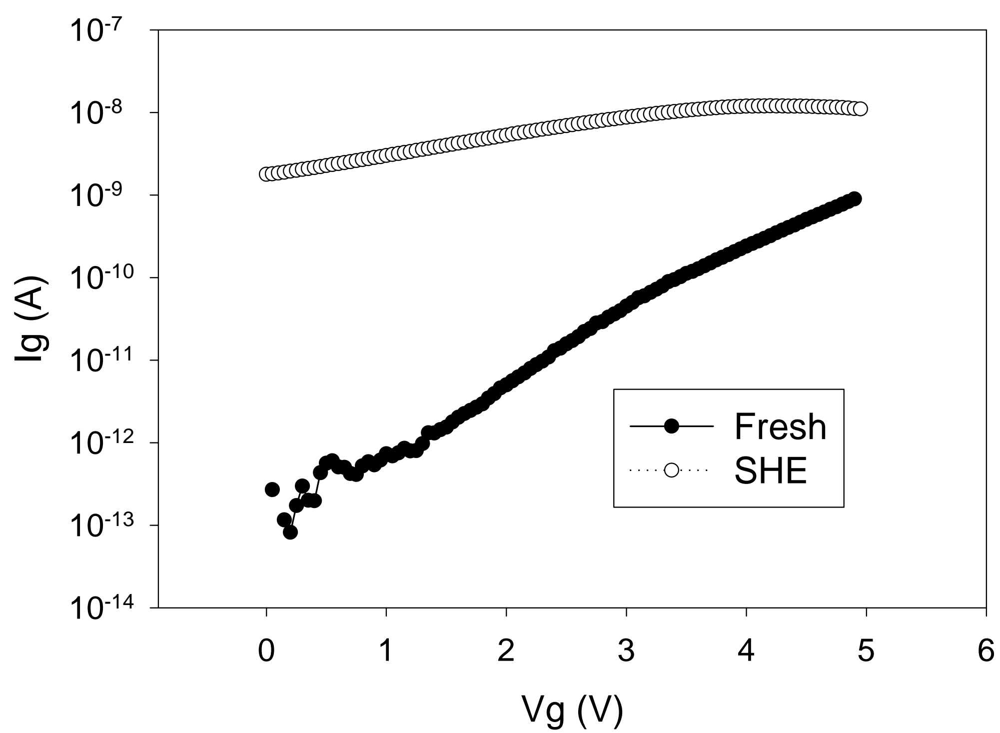
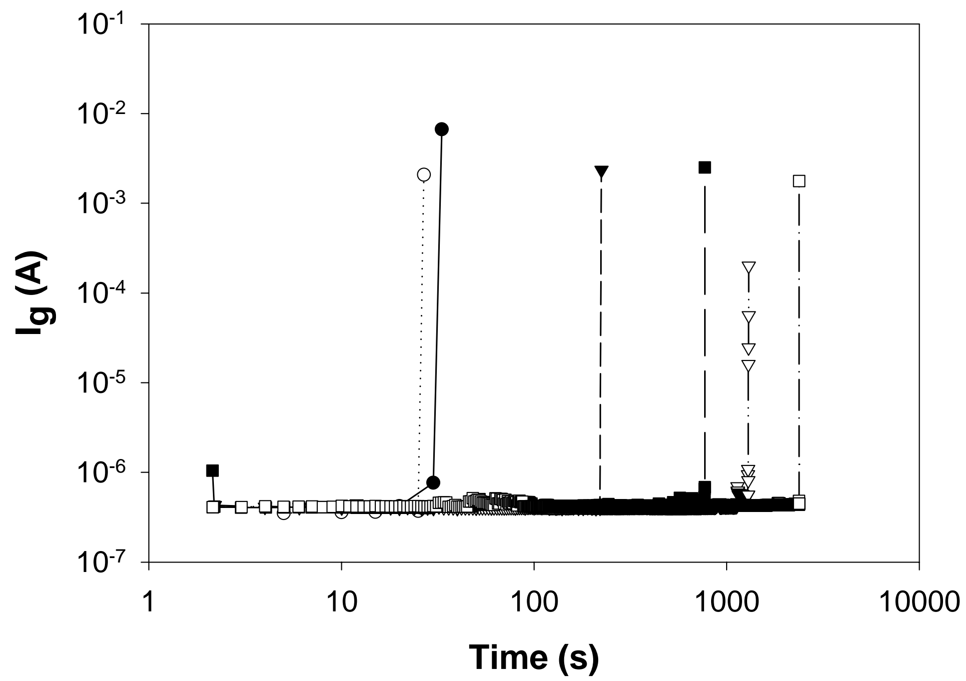


Fig. 5.9. The substrate hot electron injection of the HfO_2 device.



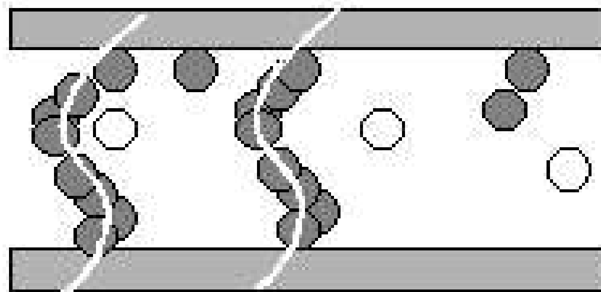
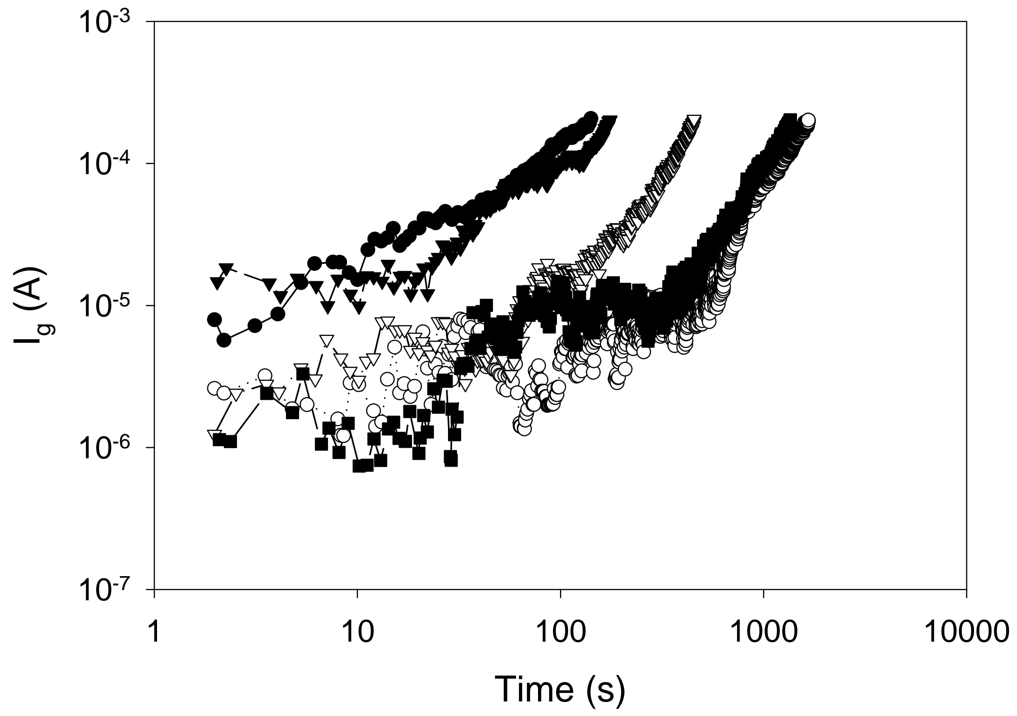


Fig. 5.10. Stable soft breakdown of the gate oxide. The upper one shows the gate current time evolution and the lower one shows the forming of the breakdown path.



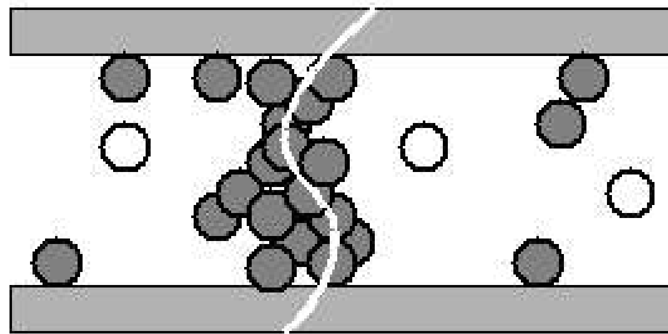


Fig. 5.11. Unstable soft breakdown. The upper one shows the gate current time evolution and the lower one shows the forming of the breakdown path

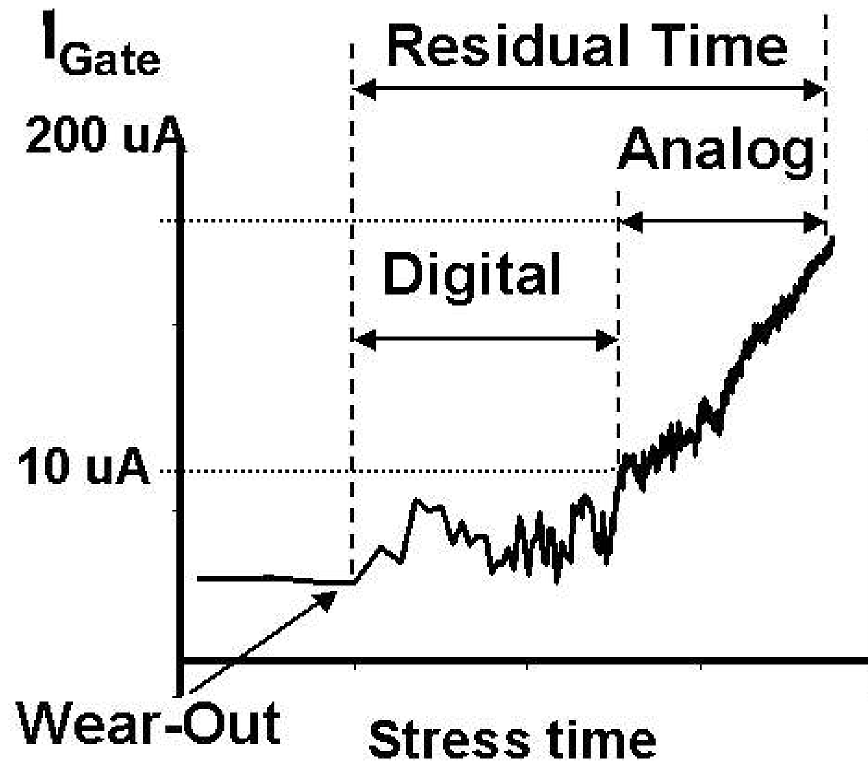


Fig. 5.12. The digital and analog phases.

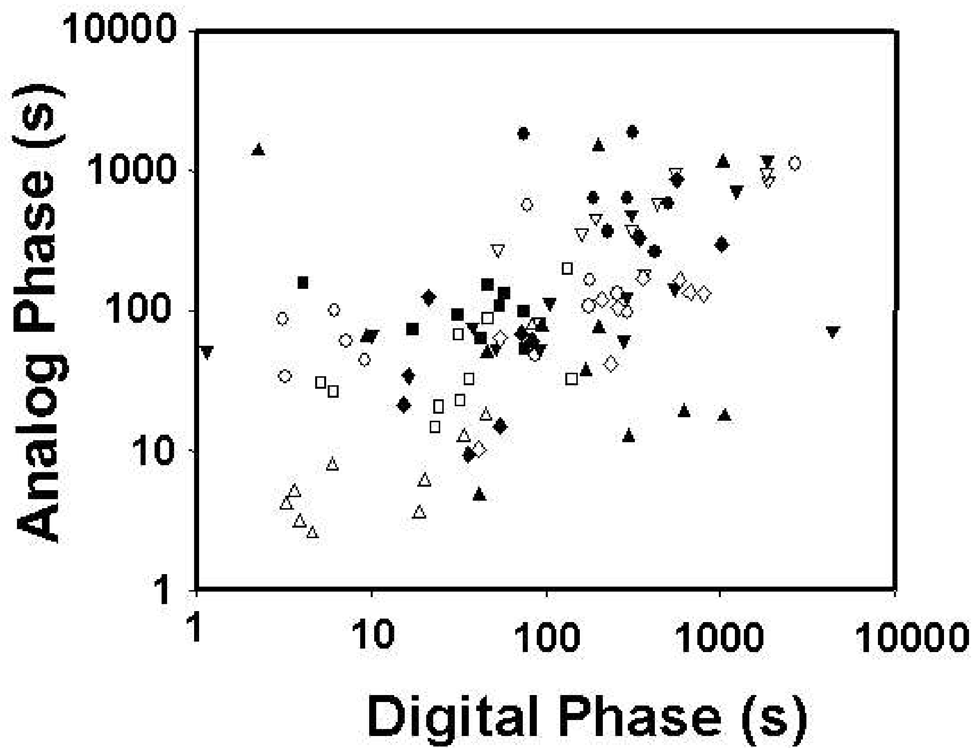


Fig. 5.13. The scattered plot showing the length of the analog and digital phase of the unstable soft breakdown are uncorrelated.

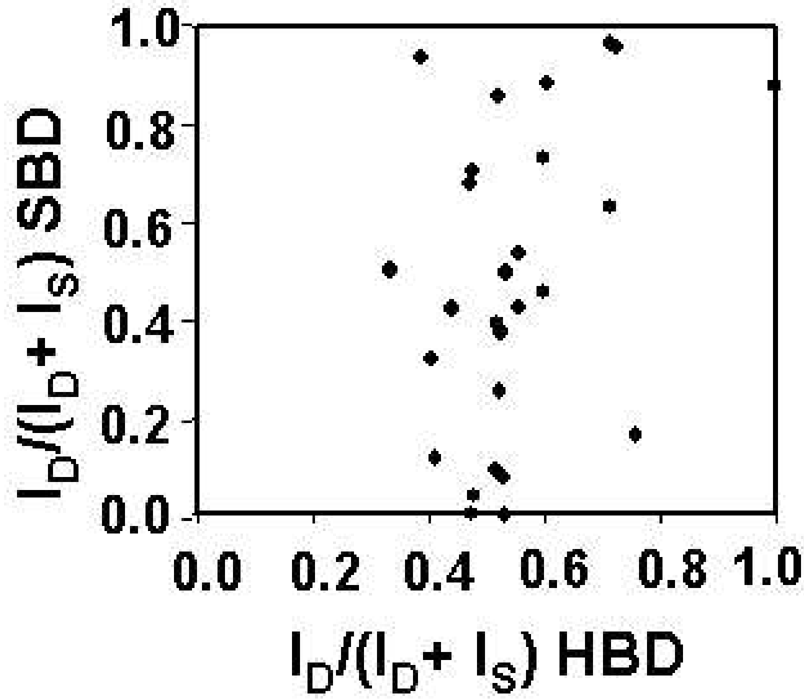


Fig. 5.14. The correlation between the first breakdown and final breakdown of stable soft breakdown.

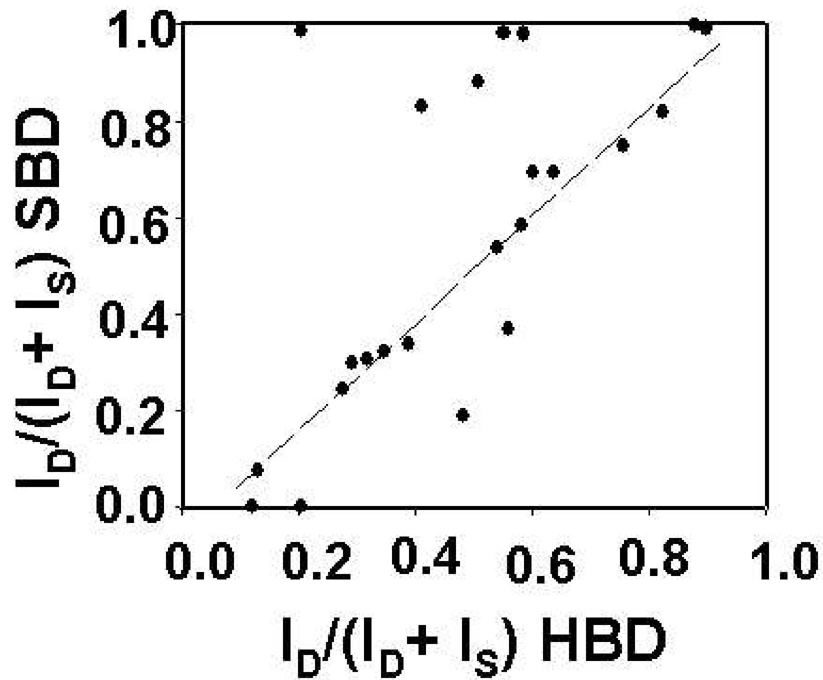


Fig. 5.15. The correlation between the first breakdown and final breakdown of unstable soft breakdown.

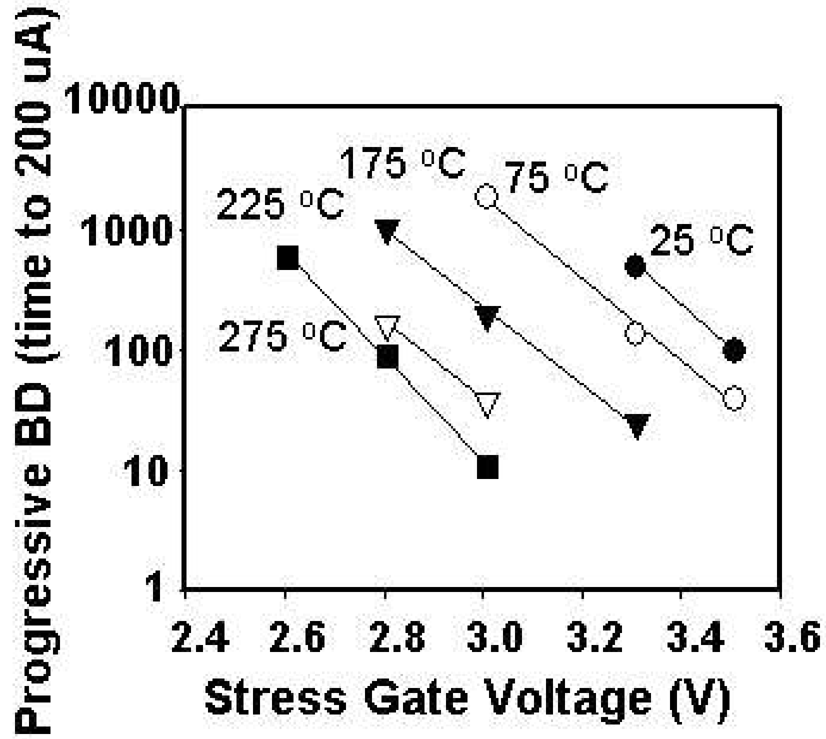


Fig. 5.16. The voltage acceleration of the unstable soft breakdown, which shows a temperature independent acceleration factor.

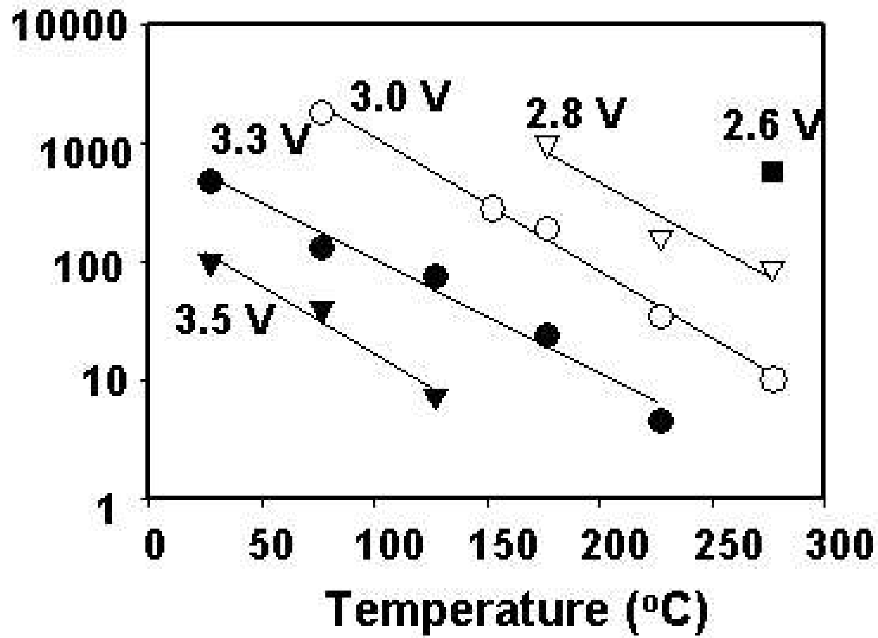


Fig. 5.17. The temperature dependence of the unstable soft breakdown, which shows a un-Arrhenius relationship.

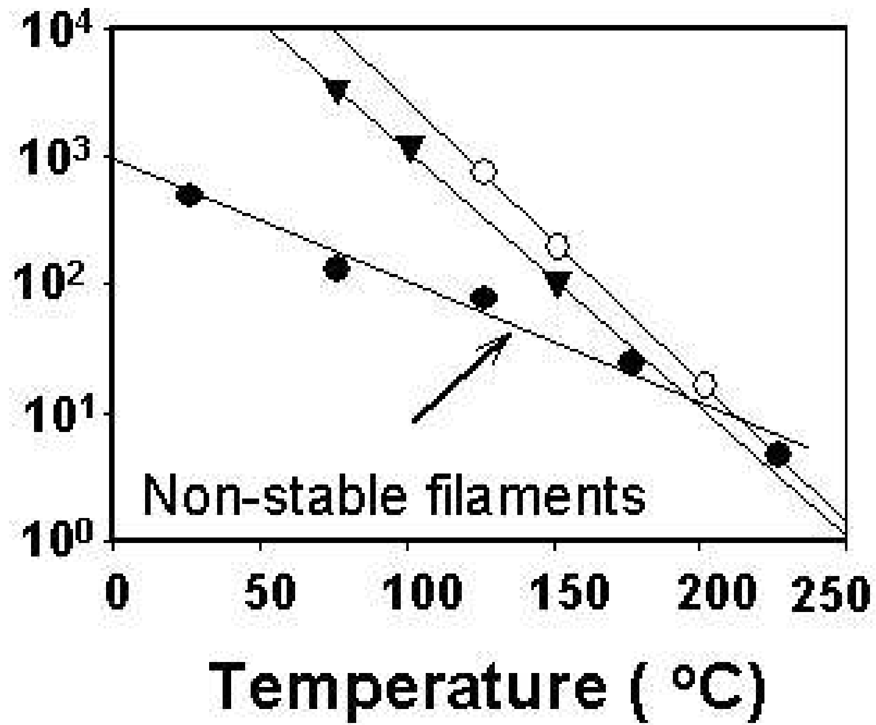


Fig. 5.18. The temperature dependence of the unstable, stable soft breakdown and the first breakdown.

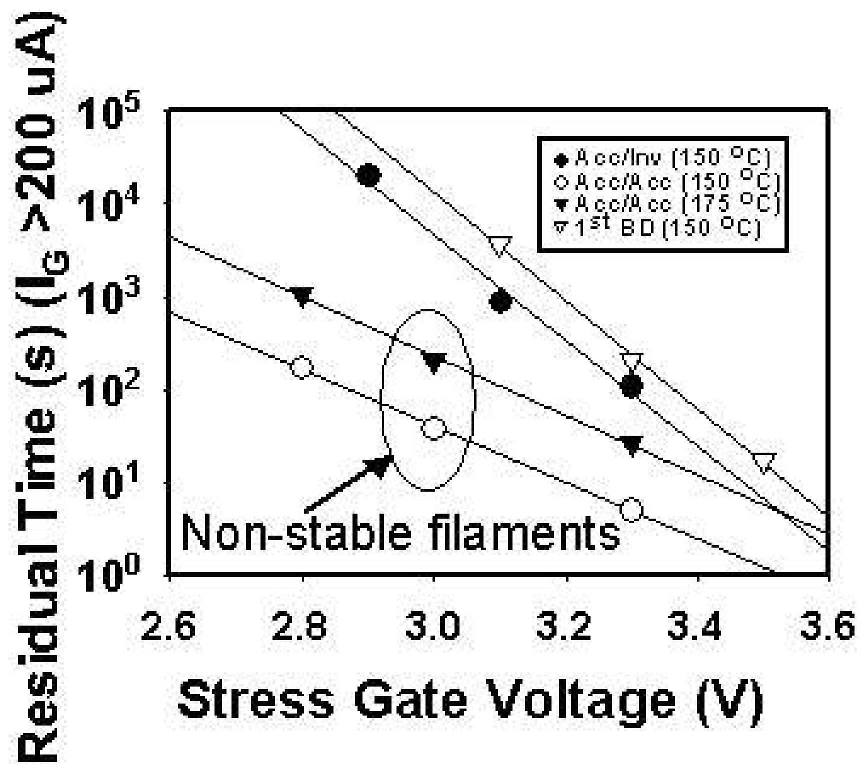


Fig. 5.19. The voltage acceleration of the residual time for unstable, stable soft breakdown and the first breakdown.

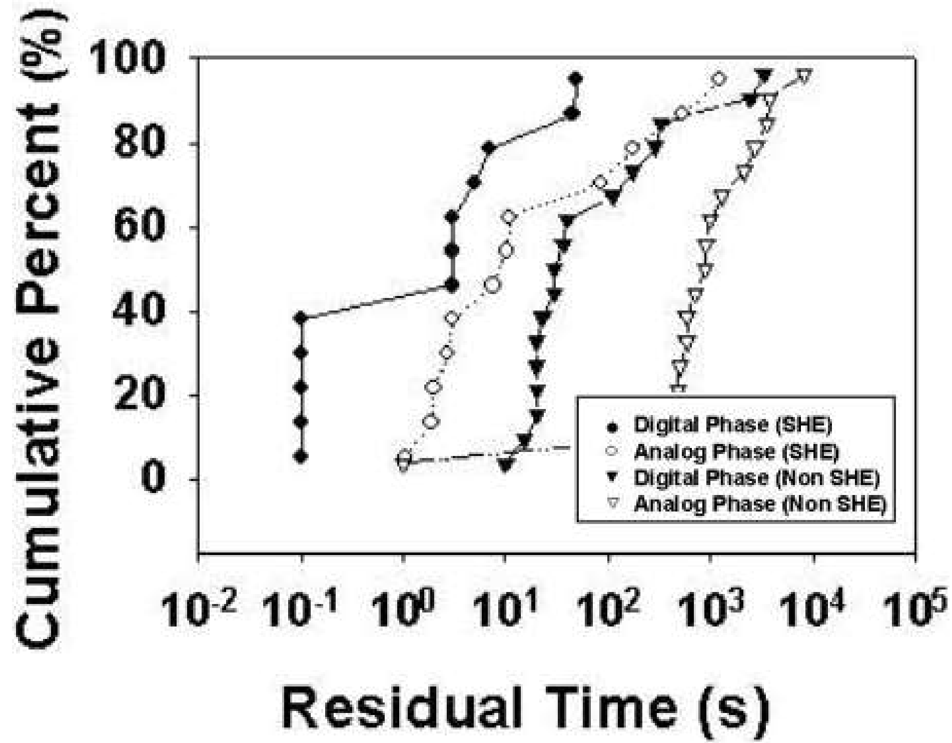


Fig. 5.20. SHEI test for the unstable breakdown filament.

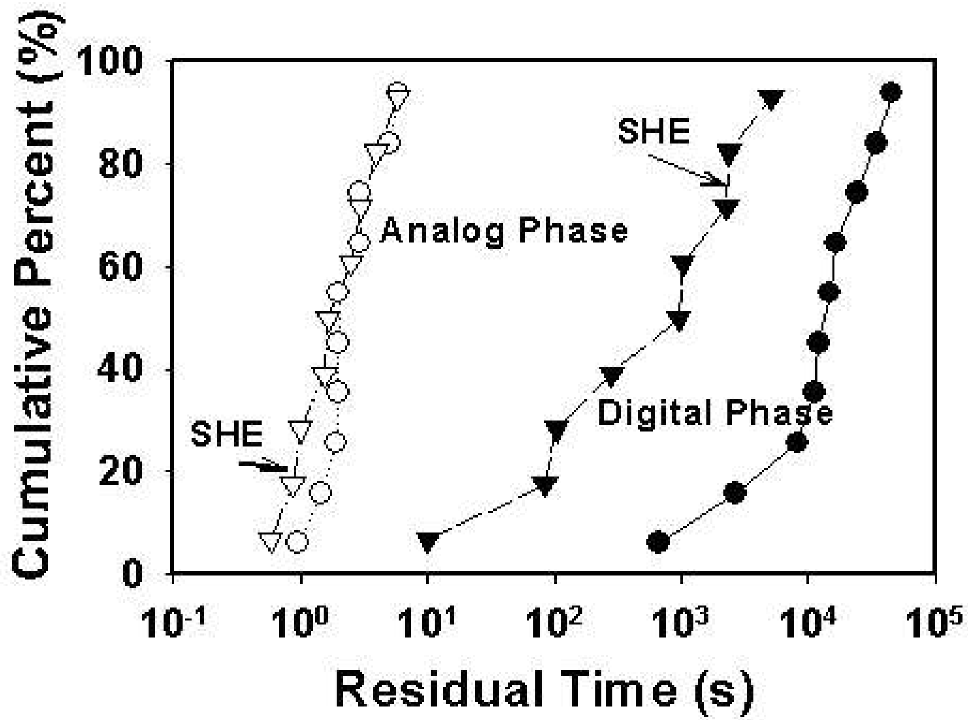


Fig. 5.21. SHEI test for the stable breakdown filament.

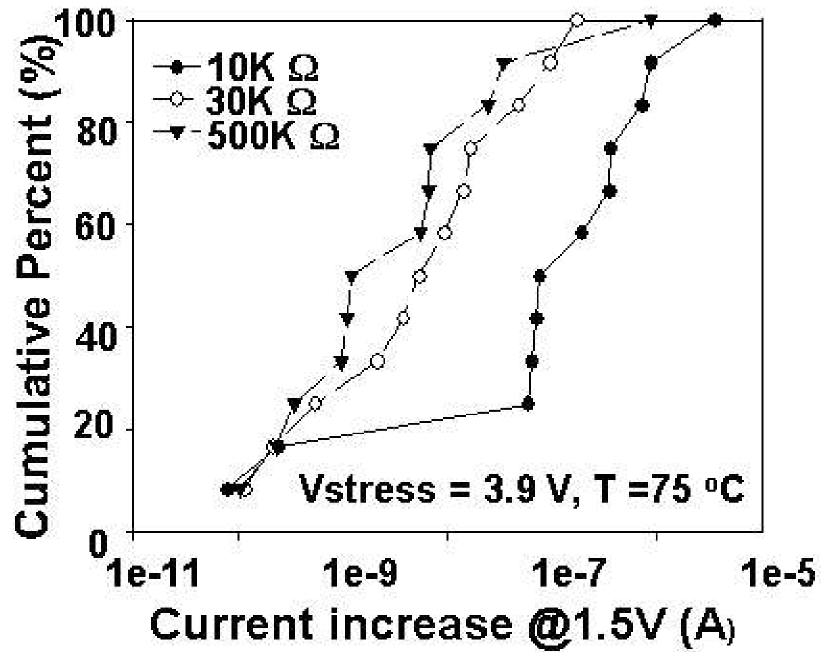


Fig. 5.22. The influence of the current compliance to the consequent stress leakage current.

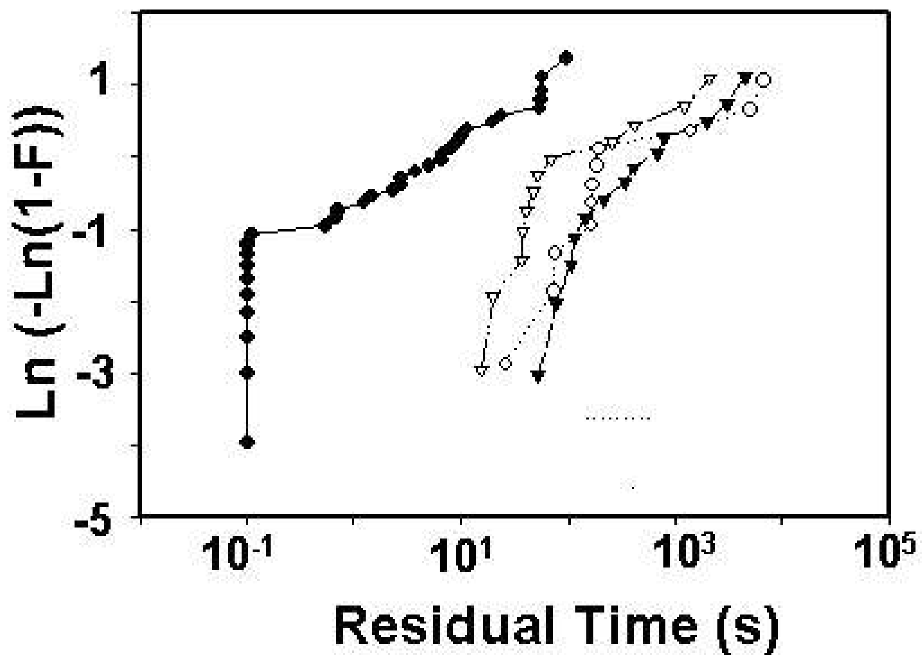


Fig. 5.23. The influence of first breakdown hardness to the residual time.

Chapter 6

Conclusion

In this work, extensive studies are conducted on the reliability assessment of advanced gate dielectrics. New test methodologies and physical mechanism interpretations are established with ultrathin SiO_2 and HfO_2 devices. These work concentrate in several parts.

First, the dynamic NBTI phenomenon was studied with ultrathin SiO_2 devices. We found with a dynamic stress condition, the device lifetime can be largely extended due to the reduced NBTI degradation. This reduction is contributed to the annealing of fixed oxide charges during the stress off period. A mathematical model is also established to explain this phenomenon. Using the measured parameters, the model coincide the experimental data well.

Second, the bulk trap contribution to the HfO_2 was tested, especially to the NBTI degradation. Compared with SiO_2 devices, the NBTI degradation has a similar trend. However, it is found that they have different frequency response than the SiO_2 devices. This difference is later found due to the traps inside the gate dielectrics. Detailed studies show that the NBTI degradation at dc stress and dynamic stress conditions have different temperature acceleration factors due to the bulk traps. The

disappearance of this difference by insetting a detrapping period further proves this observation.

Third, the progressive breakdown of ultrathin SiO_2 is studied. By a two-step test methodology, the degradation of the progressive breakdown path was monitored in terms of the activation energy, the voltage acceleration factor. Two kinds of breakdown filaments, the stable one and the unstable one, were studied. By comparison, the stable filament is found to be a breakdown filament independent of the original breakdown filament. However, the unstable filament is the continuing degradation of the original filament.

At last, the research of the HfO_2 devices breakdown is conducted. By comparing with the breakdown of SiO_2 devices, the breakdown caused by the substrate injection is attributed to the HfO_2 layer. More work is needed.

Appendices

Publications:

"The Contribution of HfO_2 Bulk Oxide Traps to Dynamic NBTI in pMOSFETs," B. Zhu, J. S. Suehle, E. Vogel, J. B. Bernstein, IEEE International Reliability Physics Symposium, 2005.

*"Detailed St*udy and Projection of Hard Breakdown Evolution in Ultra-Thin Gate Oxides," J. S. Suehle, B. Zhu, Y. Chen, and J. B. Bernstein, Microelectronics Reliability 45, 2005.

"Mechanism of Dynamic NBTI in Deep Sub-Micron p-MOSFETs," B. Zhu, J. S. Suehle, Y. Chen, and J. B. Bernstein, IEEE Integrated Reliability Workshop Final Report, 2004.

"Mechanism for Reduced NBTI Effect Under Pulsed Bias Stress Conditions," B. Zhu, J. S. Suehle, Y. Chen, and J. B. Bernstein, IEEE International Reliability Physics Symposium, 2004.

"Acceleration Factors and Mechanistic Study of Progressive Breakdown in Small Area Ultra-thin Gate Oxides," J. S. Suehle, B. Zhu, Y. Chen, and J. B. Bernstein, IEEE International Reliability Physics Symposium, 2004.

"Negative Bias Temperature Instability of Deep Sub-micron p-MOSFETs under Pulsed Bias Stress," B. Zhu, J. S. Suehle, Y. Chen, and J. B. Bernstein, IEEE Integrated Reliability Workshop Final Report, 2002.

Bibliography

[1] S. Wolf, *Silicon Processing for the VLSI Era, Vol. 3: The Submicron MOSFET*, 1 ed: Lattice Press, 1994.

[2] A. Kingon, J. Maria, and S. Streiffer, "Alternative dielectrics to silicon dioxide for memory and logic devices," *NATURE*, vol. 406, pp. 1032-1038, 2000.

[3] "International Technology Roadmap for Semiconductors," 2004, Ed.: Semiconductor Industry Association.

[4] J. Park, B. Park, M. Cho, C. Hwang, K. Oh, and D. Yang, "Chemical vapor deposition of HfO2 thin films using a novel carbon-free precursor," *JOURNAL OF THE ELECTROCHEMICAL SOCIETY*, vol. 149, pp. G89-G94, 2002.

[5] K. Forsgren, A. Harsta, J. Aarik, A. Aidla, J. Westlinder, and J. Olsson, "Deposition of HfO2 thin films in HfI4-based processes," *JOURNAL OF THE ELECTROCHEMICAL SOCIETY*, vol. 149, pp. F139-F144, 2002.

[6] D. Gilmer, R. Hegde, R. Cotton, J. Smith, L. Dip, R. Garcia, V. Dhandapani, D. Triyoso, D. Roan, A. Franke, R. Rai, L. Prabhu, C. Hobbs, J. Grant, L. La, S. Samavedam, B. Taylor, H. Tseng, and P. Tobin, "Compatibility of silicon gates with hafnium-based gate dielectrics," *MICROELECTRONIC ENGINEERING*, vol. 69, pp. 138-144, 2003.

[7] Y. Yang, W. Zhu, T. Ma, and S. Stemmer, "High-temperature phase stability of hafnium aluminate films for alternative gate dielectrics," *JOURNAL OF APPLIED PHYSICS*, vol. 95, pp. 3772-3777, 2004.

[8] T. Moon, M. Ham, M. Kim, I. Yun, and J. Myoung, "Growth and characterization of MOMBE grown HfO2," *APPLIED SURFACE SCIENCE*, vol. 240, pp. 105-111, 2005.

[9] J. Hong, T. Moon, and J. Myoung, "Microstructure and characteristics of the HfO2 dielectric layers grown by metalorganic molecular beam epitaxy," *MICROELECTRONIC ENGINEERING*, vol. 75, pp. 263-268, 2004.

[10] P. E. H. Nicollian, W.R.; Hu, J.C., "Experimental evidence for voltage driven breakdown models in ultrathin gate oxides," presented at IEEE International Reliability Physics Symposium, 2000.

[11] T. Tomita, H. Utsunomiya, T. Sakura, Y. Kamakura, and K. Taniguchi, "A new soft breakdown model for thin thermal SiO2 films under constant current stress," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 46, pp. 159-164, 1999.

[12] M. Alam, B. Weir, and P. Silverman, "A study of soft and hard breakdown - Part I: Analysis of statistical percolation conductance," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 49, pp. 232-238, 2002.

[13] A. Kerber, E. Cartier, P. Roussel, L. Pantisano, T. Kauerauf, G. Groeseneken, H. Maes, and U. Schwalke, "Charge trapping and dielectric reliability of SiO2-Al2O3 gate stacks with TiN electrodes," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 50, pp. 1261-1269, 2003.

[14] R. Degrave, T. Kauerauf, A. Kerber, E. Cartier, B. Govoreanu, P. Roussel, L. Pantisano, P. Blomme, B. Kaczer, and G. Groeseneken, "Stress polarity

dependence of degradation and breakdown of SiO/sub 2//high-k stacks," presented at IEEE International Reliability Physics Symposium Proceedings, 2003.
[15] Richard S. Muller, Theodore I. Kamins, and M. Chan, *Device Electronics for Integrated Circuits*, 3 ed: John Wiley & Sons, 2003.
[16] S. M. Sze, *Physics of Semiconductor Devices*. New York: Wiley, 1981.
[17] A. RUSU and C. BULUCEA, "DEEP-DEPLETION BREAKDOWN VOLTAGE OF SILICON-DIOXIDE SILICON MOS CAPACITORS," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 26, pp. 201-205, 1979.
[18] E. N. Wu, E.; Aitken, J.; Abadeer, W.; Han, L.K.; Lo, S., "Structural dependence of dielectric breakdown in ultra-thin gate oxides and its relationship to soft breakdown modes and device failure," presented at International Electron Devices Meeting, 1998.
[19] E. Wu, J. Stathis, and L. Han, "Ultra-thin oxide reliability for ULSI applications," *SEMICONDUCTOR SCIENCE AND TECHNOLOGY*, vol. 15, pp. 425-435, 2000.
[20] A. Chou, K. Lai, K. Kumar, P. Chowdhury, and J. Lee, "Modeling of stress-induced leakage current in ultrathin oxides with the trap-assisted tunneling mechanism," *APPLIED PHYSICS LETTERS*, vol. 70, pp. 3407-3409, 1997.
[21] M. Houng, Y. Wang, and W. Chang, "Current transport mechanism in trapped oxides: A generalized trap-assisted tunneling model," *JOURNAL OF APPLIED PHYSICS*, vol. 86, pp. 1488-1491, 1999.
[22] M. Houssa, M. Tuominen, M. Naili, V. Afanas'ev, A. Stesmans, S. Haukka, and M. Heyns, "Trap-assisted tunneling in high permittivity gate dielectric stacks," *JOURNAL OF APPLIED PHYSICS*, vol. 87, pp. 8615-8620, 2000.
[23] E. Vogel, J. Suehle, M. Edelstein, B. Wang, Y. Chen, and J. Bernstein, "Reliability of ultrathin silicon dioxide under combined substrate hot-electron and constant voltage tunneling stress," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 47, pp. 1183-1191, 2000.
[24] D. Bauza and G. Ghibaudo, "Analytical study of the contribution of fast and slow oxide traps to the charge pumping current in MOS structures," *SOLID-STATE ELECTRONICS*, vol. 39, pp. 563-570, 1996.
[25] D. Bauza, "Rigorous analysis of two-level charge pumping: Application to the extraction of interface trap concentration versus energy profiles in metal-oxide-semiconductor transistors," *JOURNAL OF APPLIED PHYSICS*, vol. 94, pp. 3239-3248, 2003.
[26] R. PAULSEN and M. WHITE, "THEORY AND APPLICATION OF CHARGE-PUMPING FOR THE CHARACTERIZATION OF SI-SIO2 INTERFACE AND NEAR-INTERFACE OXIDE TRAPS," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 41, pp. 1213-1216, 1994.
[27] R. PAULSEN, R. SIERGIEJ, M. FRENCH, and M. WHITE, "OBSERVATION OF NEAR-INTERFACE OXIDE TRAPS WITH THE CHARGE-PUMPING TECHNIQUE," *IEEE ELECTRON DEVICE LETTERS*, vol. 13, pp. 627-629, 1992.
[28] T. Hori, *Gate Dielectrics and MOS ULSI*. New York: Springer-Verlag, 1997.

[29] J. BREWS, "SUB-THRESHOLD BEHAVIOR OF UNIFORMLY AND NONUNIFORMLY DOPED LONG-CHANNEL MOSFET," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 26, pp. 1282-1291, 1979.

[30] M. Ohring, *Reliability & Failure of Electronic Materials & Devices*, 1 ed: Academic Press, 1998.

[31] A. K. Sharma, *Semiconductor Memories: Technology, Testing, and Reliability*, 1 ed: Wiley-IEEE Press, 2002.

[32] G. Klutke, P. Kiessler, and M. Wortman, "A critical look at the bathtub curve," *IEEE TRANSACTIONS ON RELIABILITY*, vol. 52, pp. 125-129, 2003.

[33] J. Suehle, "Ultrathin gate oxide reliability: Physical models, statistics, and characterization," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 49, pp. 958-971, 2002.

[34] K. SCHUEGRAF and C. HU, "RELIABILITY OF THIN SIO2," *SEMICONDUCTOR SCIENCE AND TECHNOLOGY*, vol. 9, pp. 989-1004, 1994.

[35] K. SCHUEGRAF and C. HU, "HOLE INJECTION SIO2 BREAKDOWN MODEL FOR VERY-LOW VOLTAGE LIFETIME EXTRAPOLATION," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 41, pp. 761-767, 1994.

[36] M. FISCHETTI, "MODEL FOR THE GENERATION OF POSITIVE CHARGE AT THE SI-SIO2 INTERFACE BASED ON HOT-HOLE INJECTION FROM THE ANODE," *PHYSICAL REVIEW B*, vol. 31, pp. 2099-2113, 1985.

[37] J. Kim, J. Sanchez, T. DeMassa, M. Quddus, D. Smith, F. Shaapur, K. Weiss, and C. Liu, "Surface plasmons and breakdown in thin silicon dioxide films on silicon," *JOURNAL OF APPLIED PHYSICS*, vol. 84, pp. 1430-1438, 1998.

[38] D. DiMaria, E. Cartier, and D. Buchanan, "Anode hole injection and trapping in silicon dioxide," *JOURNAL OF APPLIED PHYSICS*, vol. 80, pp. 304-317, 1996.

[39] M. d. W. Rasras, D.; Groeseneken, G.; Kaczer, B.; Degraeve, R.; Maes, H.E., "Photo-carrier generation as the origin of Fowler-Nordheim-induced substrate hole current in thin oxides," presented at International Electron Devices Meeting, 1999.

[40] D. DIMARIA and J. STASIAK, "TRAP CREATION IN SILICON DIOXIDE PRODUCED BY HOT-ELECTRONS," *JOURNAL OF APPLIED PHYSICS*, vol. 65, pp. 2342-2356, 1989.

[41] J. H. S. E. Cartier, and D. A. Buchanan, "Passivation and Depassivation of Silicon Dangling Bonds at the Si/SiO2 Interface by Atomic Hydrogen," *Appl. Phys. Lett.*, vol. 63, pp. 1510-1512, 1993.

[42] R. STAHLBUSH and E. CARTIER, "INTERFACE DEFECT FORMATION IN MOSFETS BY ATOMIC-HYDROGEN EXPOSURE," *IEEE TRANSACTIONS ON NUCLEAR SCIENCE*, vol. 41, pp. 1844-1853, 1994.

[43] J. STATHIS and E. CARTIER, "ATOMIC-HYDROGEN REACTIONS WITH P(B) CENTERS AT THE (100) SI/SIO2 INTERFACE," *PHYSICAL REVIEW LETTERS*, vol. 72, pp. 2745-2748, 1994.

[44] J. R. Wu, E.; MacDonald, B.; Li, E.; Tao, J.; Tracy, B.; Fang, P., "Anode hole injection versus hydrogen release: the mechanism for gate oxide breakdown," presented at IEEE International Reliability Physics Symposium, 2000.
[45] J. McPherson and R. Khamankar, "Molecular model for intrinsic time-dependent dielectric breakdown in SiO2 dielectrics and the reliability implications for hyper-thin gate oxide," *SEMICONDUCTOR SCIENCE AND TECHNOLOGY*, vol. 15, pp. 462-470, 2000.
[46] J. W. McPherson and D. A. Baglee, "Acceleration factors for thin gate oxide stressing," presented at IEEE International Reliability Physics Symposium, 1985.
[47] "JESD 35-A: Procedure for the Wafer Level Testing of Thin Dielectrics," in *JEDEC Solid State Technology Association*, E. I. A. J. E. D. E. C. 14.2, Ed., 2001.
[48] J. Stathis, "Percolation models for gate oxide breakdown," *JOURNAL OF APPLIED PHYSICS*, vol. 86, pp. 5757-5766, 1999.
[49] E. Anolick and G. Nelson, "Low field time dependent dielectric integrity," presented at IEEE International Reliability Physics Symposium, 1979.
[50] D. Crook, "Method of determining reliability screens for time dependent dielectric breakdown," presented at IEEE International Reliability Physics Symposium, 1979.
[51] I. C. Chen, S. Holland, and C. HU, "A quantitative physical model for time-dependent breakdown in SiO2," presented at IEEE International Reliability Physics Symposium, 1985.
[52] J. LEE, I. CHEN, and C. HU, "MODELING AND CHARACTERIZATION OF GATE OXIDE RELIABILITY," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 35, pp. 2268-2278, 1988.
[53] C. Hu; and Q. Lu, "A unified gate oxide reliability model," presented at IEEE International Reliability Physics Symposium Proceedings, 1999.
[54] K. Okada and K. Yoneda, "A consistent model for time dependent dielectric breakdown in ultrathin silicon dioxides," presented at International Electron Devices Meeting, 1999.
[55] K. P. Cheung, "A physics-based, unified gate-oxide breakdown model," presented at International Electron Devices Meeting.
[56] J. M. McKenna, E. Y. Wu, and S.-H. Lo, "Tunneling current characteristics and oxide breakdown in P+ poly gate PFET capacitors," presented at IEEE International Reliability Physics Symposium, 2000.
[57] E. Y. Wu, J. Aitken, E. Nowak, A. Vayshenker, P. Varekamp, G. Hueckel, J. McKenna, D. Harmon, L.-K. Han, C. Montrose, and R. Dufresne, "Voltage-dependent voltage-acceleration of oxide breakdown for ultra-thin oxides," presented at International Electron Devices Meeting, 2000.
[58] J. H. Stathis, "Physical and predictive models of ultra thin oxide reliability in CMOS devices and circuits," presented at IEEE International Reliability Physics Symposium.
[59] E. Wu, J. Sune, and W. Lai, "On the Weibull shape factor of intrinsic breakdown of dielectric films and its accurate experimental determination -

Part II: Experimental results and the effects of stress conditions," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 49, pp. 2141-2150, 2002.

[60] J. Suñé and E. Y. Wu, "Hydrogen-Release Mechanisms in the Breakdown of Thin SiO2 Films," *Physical Review Letters*, vol. 92, pp. 087601.

[61] Y. Chin and B. Chiou, "Effects of underlayer dielectric on the thermal characteristics and electromigration resistance of copper interconnect," *JAPANESE JOURNAL OF APPLIED PHYSICS PART 1-REGULAR PAPERS SHORT NOTES & REVIEW PAPERS*, vol. 42, pp. 7502-7509, 2003.

[62] J. Lloyd, E. Liniger, and S. Chen, "Time dependent dielectric breakdown in a low-k interlevel dielectric," *MICROELECTRONICS RELIABILITY*, vol. 44, pp. 1861-1865, 2004.

[63] N. Matsukawa and K. Kanebako, "A new model for TDDB lifetime distribution of SiO2," *MICROELECTRONIC ENGINEERING*, vol. 48, pp. 121-124, 1999.

[64] P. Nukala and S. Simunovic, "Scaling of fracture strength in disordered quasi-brittle materials," *EUROPEAN PHYSICAL JOURNAL B*, vol. 37, pp. 91-100, 2004.

[65] 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," presented at International Electron Devices Meeting, 1995.

[66] J. Sune, "New physics-based analytic approach to the thin-oxide breakdown statistics," *IEEE Electron Device Letters*, vol. 22, pp. 296-298, 2001.

[67] G. Paulzen, "Q(bd)(-) dependencies of ultrathin gate oxides on large area capacitors.," *MICROELECTRONIC ENGINEERING*, vol. 36, pp. 321-324, 1997.

[68] 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," presented at International Electron Devices Meeting, 1995.

[69] B. E. Weir, P. J. Silverman, M. A. Alam, F. Baumann, D. Monroe, A. Ghetti, J. D. Bude, G. L. Timp, A. Hamad, T. M. Oberdick, N. X. Zhao, Y. Ma, M. M. Brown, D. Hwang, T. W. Sorsch, and J. Madic, "Gate oxides in 50 nm devices: thickness uniformity improves projected reliability," presented at International Electron Devices Meeting, 1999.

[70] E. Y. Wu, W. W. Abadeer, L.-K. Han, S.-H. Lo, and G. R. Hueckel, "Challenges for accurate reliability projections in the ultra-thin oxide regime," presented at IEEE International Reliability Physics Symposium Proceedings, 1999.

[71] W. Y. Loh, B. J. Cho, M. S. Joo, M. F. Li, D. S. H. Chan, S. Mathew, and D.-L. Kwong, "Analysis of charge trapping and breakdown mechanism in high-k dielectrics with metal gate electrode using carrier separation," presented at IEEE International Electron Devices Meeting, 2003.

[72] R. Degraeve, A. Kerber, P. Roussell, E. Cartier, T. Kauerauf, L. Pantisano, and G. Groeseneken, "Effect of bulk trap density on HfO/sub 2/ reliability and yield," presented at IEEE International Electron Devices Meeting, 2003.

[73] F. P. Kamarinos G, "How will physics be involved in silicon microelectronics," *JOURNAL OF PHYSICS D-APPLIED PHYSICS*, vol. 29, pp. 487-500, 1996.
[74] K. H. Hess, A.; McMahon, W.; Cheng, K.; Lee, J.; Lyding, J, "The physics of determining chip reliability," *IEEE Circuits and Devices Magazine*, vol. 17, pp. 33-38, 2001.
[75] P. L. L. Levin, R., "Crossroads for mixed-signal chips," in *IEEE Spectrum*, vol. 39, 2002, pp. 38-43.
[76] D. Schroder and J. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *JOURNAL OF APPLIED PHYSICS*, vol. 94, pp. 1-18, 2003.
[77] P. V. G. a. D. M. Brown, "DENSITY OF SiO2[Single Bond]Si INTERFACE STATES," *Applied Physics Letters*, vol. 8, pp. 31-33, 1966.
[78] M. Makabe, T. Kubota, and T. Kitano, "Bias-temperature degradation of pMOSFETs: mechanism and suppression," presented at IEEE International Reliability Physics Symposium, 2000.
[79] M. S. B. E. Deal, A. S. Grove, and E. H. Snow, "Characteristics of the surface-state charge (Qss) of thermally oxidized silicon," *J. Electrochem. Soc.*, vol. 114, pp. 266, 1967.
[80] S. R. Hofstein, "Stabilization of MOS devices," *Solid State Electron.*, vol. 10, pp. 657-70, 1967.
[81] D. LU, G. RUGGLES, and J. WORTMAN, "EFFECTS OF PROCESSING CONDITIONS ON NEGATIVE BIAS TEMPERATURE INSTABILITY IN METAL-OXIDE-SEMICONDUCTOR STRUCTURES," *APPLIED PHYSICS LETTERS*, vol. 52, pp. 1344-1346, 1988.
[82] S. K. H. a. R.F.DeKeersmaecker, "Hole trapping and interface state generation during bias-temperature stress of SiO2 layers," *Appl.Phys.Lett.*, vol. 47, pp. 381-383, 1985.
[83] D. J. Breed, "Non-ionic room temperature Instabilities in MOS devices," *Solid-State Electron.*, vol. 17, pp. 1229, 1974.
[84] D. J. Breed, "A new model for the negative voltage instability in MOS devices," *Applied Physics Letters*, vol. 26, pp. 116-118, 1975.
[85] C. BLAT, E. NICOLLIAN, and E. POINDEXTER, "MECHANISM OF NEGATIVE-BIAS-TEMPERATURE INSTABILITY," *JOURNAL OF APPLIED PHYSICS*, vol. 69, pp. 1712-1720, 1991.
[86] S. C. JEPPSON KO, "NEGATIVE BIAS STRESS OF MOS DEVICES AT HIGH ELECTRIC-FIELDS AND DEGRADATION OF MNOS DEVICES," *JOURNAL OF APPLIED PHYSICS*, vol. 48, pp. 2004-2014, 1977.
[87] S. M. OGAWA S, SHIONO N, "INTERFACE-TRAP GENERATION AT ULTRATHIN SIO2 (4-6NM)-SI INTERFACES DURING NEGATIVE-BIAS TEMPERATURE AGING," *JOURNAL OF APPLIED PHYSICS*, vol. 77, pp. 1137-1148, 1995.
[88] G. GERARDI, E. POINDEXTER, P. CAPLAN, M. HARMATZ, W. BUCHWALD, and N. JOHNSON, "GENERATION OF PB CENTERS BY HIGH ELECTRIC-FIELDS - THERMOCHEMICAL EFFECTS," *JOURNAL OF THE ELECTROCHEMICAL SOCIETY*, vol. 136, pp. 2609-2614, 1989.

[89] M. Krishnan, V. Kol'dyaev, E. Morifuji, K. Miyamoto, T. Brozek, and X. Li, "Series resistance degradation due to NBTI in PMOSFET," *MICROELECTRONICS RELIABILITY*, vol. 42, pp. 1433-1438, 2002.
[90] S. Zafar, Y. Lee, and J. Stathis, "Evaluation of NBTI in HfO2 gate-dielectric stacks with tungsten gates," *IEEE ELECTRON DEVICE LETTERS*, vol. 25, pp. 153-155, 2004.
[91] S. F. H. O. Yuan Chen; Jonathon Zhou; Tedja, A.S., "Stress-induced MOSFET mismatch for analog circuits," presented at IEEE International Integrated Reliability Workshop Final Report, 2001.
[92] V. K. Reddy, A.T.; Marshall, A.; Rodriguez, J.; Natarajan, S.; Rost, T.; Krishnan, S., "Impact of negative bias temperature instability on digital circuit reliability," presented at IEEE International Reliability Physics Symposium Proceedings, 2002.
[93] F. E. L. Pagaduan, J.K.J. Vedagarbha, V. Lui, K. Hart, M.J. Gitlin, D. Takaso, T. Kamiyama, S. Nakayama, K., "The effects of plasma-induced damage on transistor degradation and the relationship to field programmable gate array performance," presented at IEEE International Reliability Physics Symposium.
[94] N. S. Yung-Huei Lee; Mielke, B.; Stadler, S.; Nachman, R.; Hu, S., "Effect of pMOST bias-temperature instability on circuit reliability performance," presented at IEEE International Electron Devices Meeting, 2003.
[95] G. Chen, M. Li, C. Ang, J. Zheng, and D. Kwong, "Dynamic NBTI of p-MOS transistors and its impact on MOSFET scaling," *IEEE ELECTRON DEVICE LETTERS*, vol. 23, pp. 734-736, 2002.
[96] W. Abadeer and W. Ellis, "Behavior of NBTI under AC dynamic circuit conditions," presented at IEEE International Reliability Physics Symposium Proceedings.
[97] S. Rangan, N. Mielke, and E. C. C. Yeh, "Universal recovery behavior of negative bias temperature instability," presented at IEEE International Electron Devices Meeting, 2003.
[98] B. S. Doyle, B. J. Fishbein, and K. R. Mistry, "NBTI-enhanced hot carrier damage in p-channel MOSFETs," presented at International Electron Devices Meeting.
[99] M. A. Alam, "A critical examination of the mechanics of dynamic NBTI for PMOSFETs," presented at IEEE International Electron Devices Meeting, 2003.
[100] B. Zhu, J. S. Suehle, and J. B. Bernstein, "Mechanism for reduced NBTI effect under pulsed bias stress conditions," presented at IEEE International Reliability Physics Symposium Proceedings, 2004.
[101] J. Zhang and W. Eccleston, "Positive bias temperature instability in MOSFET's," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 45, pp. 116-124, 1998.
[102] V. Huard and M. Denais, "Hole trapping effect on methodology for DC and AC negative bias temperature instability measurements in PMOS transistors," presented at IEEE International Reliability Physics Symposium Proceedings, 2004.

[103] Y. Kim, G. Gebara, M. Freiler, J. Barnett, D. Riley, J. Chen, K. Torres, J. Lim, B. Foran, F. Shaapur, A. Agarwal, P. Lysaght, G. A. Brown, C. Young, S. Borthakur, H.-J. Li, B. Nguyen, P. Zeitzoff, G. Bersuker, D. Derro, R. Bergmann, R. W. Murto, A. Hou, H. R. Huff, E. Shero, C. Pomarede, M. Givens, M. Mazanez, and C. Werkhoven, "Conventional n-channel MOSFET devices using single layer HfO2 and ZrO2 as high-k gate dielectrics with polysilicon gate electrode," presented at International Electron Devices Meeting.

[104] C. Hobbs, H. Tseng, K. Reid, B. Taylor, L. Dip, L. Hebert, R. Garcia, R. Hegde, J. Grant, D. Gilmer, A. Franke, V. Dhandapani, M. Azrak, L. Prabhu, R. Rai, S. Bagchi, J. Conner, S. Backer, F. Dumbuya, B. Nguyen, and P. Tobin, "80 nm poly-Si gate CMOS with HfO2 gate dielectric," presented at International Electron Devices Meeting, 2001.

[105] C. D. Young, G. Bersuker, G. A. Brown, C. Lim, P. Lysaght, P. Zeitzoff, R. W. Murto, and H. R. Huff, "Charge trapping in MOCVD hafnium-based gate field dielectric stack structures and its impact on device performance," presented at IEEE International Integrated Reliability Workshop Final Report, 2004.

[106] G. Chen, K. Y. Chuah, M. F. Li, D. S. H. Chan, C. H. Ang, J. Z. Zheng, Y. Jin, and D. L. Kwong, "Dynamic NBTI of PMOS transistors and its impact on device lifetime," presented at IEEE International Reliability Physics Symposium Proceedings, 2003.

[107] C. Weintraub, E. Vogel, J. Hauser, N. Yang, V. Misra, J. Wortman, J. Ganem, and P. Masson, "Study of low-frequency charge pumping on thin stacked dielectrics," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 48, pp. 2754-2762, 2001.

[108] M. Depas, T. Nigam, and M. Heyns, "Soft breakdown of ultra-thin gate oxide layers," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 43, pp. 1499-1504, 1996.

[109] E. Miranda, J. Sune, R. Rodriguez, M. Nafria, and X. Aymerich, "Soft breakdown fluctuation events in ultrathin SiO2 layers," *APPLIED PHYSICS LETTERS*, vol. 73, pp. 490-492, 1998.

[110] K. Okada, "Soft breakdown of oxide-nitride-oxide stacked gate dielectrics used in metal-oxide-nitride-oxide-silicon-based flash memories," *APPLIED PHYSICS LETTERS*, vol. 83, pp. 5542-5544, 2003.

[111] M. Radhakrishnan, K. Pey, C. Tung, and W. Lin, "Physical analysis of hard and soft breakdown failures in ultrathin gate oxides," *MICROELECTRONICS RELIABILITY*, vol. 42, pp. 565-571, 2002.

[112] F. Crupi, R. Degraeve, G. Groeseneken, T. Nigam, and H. Maes, "On the properties of the gate and substrate current after soft breakdown in ultrathin oxide layers," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 45, pp. 2329-2334, 1998.

[113] E. Snyder and J. Suehle, "Fast breakdown detection in ultra-thin dielectrics," *SOLID STATE TECHNOLOGY*, pp. S4-S8, 2000.

[114] G. B. Alers, B. E. Weir, M. A. Alam, G. L. Timp, and T. Sorch, "Trap assisted tunneling as a mechanism of degradation and noise in 2-5 nm oxides," presented at IEEE International Reliability Physics Symposium, 1998.

[115] S. Bruyere, E. Vincent, and G. Ghibaudo, "Quasi-breakdown in ultra-thin SiO2 films: occurrence characterization and reliability assessment methodology," presented at IEEE International Reliability Physics Symposium, 2000.

[116] T. Pompl, H. Wurzer, M. Kerber, R. C. W. Wilkins, and I. Eisele, "Influence of soft breakdown on NMOSFET device characteristics," presented at IEEE International Reliability Physics Symposium, 1999.

[117] E. Y. Wu and J. Sune, "Successive breakdown events and their relation with soft and hard breakdown modes," *Electron Device Letters, IEEE*, vol. 24, pp. 692-694, 2003.

[118] B. P. S. Linder, J.H.; Frank, D.J.; Lombardo, S.; Vayshenker, A., "Growth and scaling of oxide conduction after breakdown," presented at IEEE International Reliability Physics Symposium Proceedings, 2003.

[119] F. Monsieur, E. Vincent, G. Pananakakis, and G. Ghibaudo, "Wear-out, breakdown occurrence and failure detection in 18-25 angstrom ultrathin oxides," *MICROELECTRONICS RELIABILITY*, vol. 41, pp. 1035-1039, 2001.

[120] J. Sune, G. Mura, and E. Miranda, "Are soft breakdown and hard breakdown of ultrathin gate oxides actually different failure mechanisms?" *IEEE ELECTRON DEVICE LETTERS*, vol. 21, pp. 167-169, 2000.

[121] E. Wu, J. Sune, B. Linder, J. Stathis, and W. Lai, "Critical assessment of soft breakdown stability time and the implementation of new post-breakdown methodology for ultra-thin gate oxides [MOSFET]," presented at IEEE International Electron Devices Meeting, 2003.

[122] B. Linder and J. Stathis, "Statistics of progressive breakdown in ultra\-thin oxides," *MICROELECTRONIC ENGINEERING*, vol. 72, pp. 24-28, 2004.

[123] J. Stathis, B. Linder, R. Rodriguez, and S. Lombardo, "Reliability of ultra-thin oxides in CMOS circuits," *MICROELECTRONICS RELIABILITY*, vol. 43, pp. 1353-1360, 2003.

[124] E. Miranda, A. Cester, and A. Paccagnella, "Stochastic modeling of progressive breakdown in ultrathin SiO2 films," *APPLIED PHYSICS LETTERS*, vol. 83, pp. 5014-5016, 2003.

[125] F. Crupi, B. Kaczer, R. Degraeve, A. De Keersgieter, and G. Groeseneken, "Location and hardness of the oxide breakdown in short channel n- and p-MOSFETs," presented at IEEE International Reliability Physics Symposium, 2002.

[126] B. Kaczer, R. Degraeve, G. Groeseneken, M. Rasras, S. Kubicek, E. Vandamme, and G. Badenes, "Impact of MOSFET oxide breakdown on digital circuit operation and reliability," presented at International Electron Devices Meeting, 2000.

[127] R. Rodriguez, J. H. Stathis, and B. P. Linder, "Modeling and experimental verification of the effect of gate oxide breakdown on CMOS inverters," presented at IEEE International Reliability Physics Symposium Proceedings, 2003.

[128] F. V. Monsieur, E.; Roy, D.; Bruyere, S.; Vildeuil, J.C.; Pananakakis, G.; Ghibaudo, G., "A thorough investigation of progressive breakdown in ultra-thin oxides. Physical understanding and application for industrial reliability assessment," presented at IEEE International Reliability Physics Symposium Proceedings, 2002.
[129] F. Monsieur, E. Vincent, G. Ribes, V. Huard, S. Bruyere, D. Roy, G. Pananakakis, and G. Ghibaudo, " Evidence for defect-generation-driven wear-out of breakdown conduction path in ultra thin oxides," presented at IEEE International Reliability Physics Symposium Proceedings, 2003.
[130] G. Wilk, R. Wallace, and J. Anthony, "High-kappa gate dielectrics: Current status and materials properties considerations," *JOURNAL OF APPLIED PHYSICS*, vol. 89, pp. 5243-5275, 2001.
[131] A. Chatterjee, R. A. Chapman, K. Joyner, M. Otobe, S. Hattangady, M. Bevan, G. A. Brown, H. Yang, Q. He, D. Rogers, S. J. Fang, R. Kraft, A. L. P. Rotondaro, M. Terry, K. Brennan, S.-W. Aur, J. C. Hu, H.-L. Tsai, P. Jones, G. Wilk, M. Aoki, M. Rodder, and I.-C. Chen, "CMOS metal replacement gate transistors using tantalum pentoxide gate insulator," presented at International Electron Devices Meeting, 1998.
[132] Kizilyalli IC, Huang RYS, and R. PK, "MOS transistors with stacked SiO2-Ta2O5-SiO2 gate dielectrics for giga-scale integration of CMOS technologies," *IEEE ELECTRON DEVICE LETTERS*, vol. 19, pp. 423-425, 1998.
[133] Roy PK and K. IC, "Stacked high-epsilon gate dielectric for gigascale integration of metal-oxide-semiconductor technologies," *APPLIED PHYSICS LETTERS*, vol. 72, pp. 2835-2837, 1998.
[134] Lu Q, Park D, Kalnitsky A, Chang C, Cheng CC, Tay SP, King TJ, H. CM, and, "Leakage current comparison between ultra-thin Ta2O5 films and conventional gate dielectrics," *IEEE ELECTRON DEVICE LETTERS*, vol. 19, pp. 341-342, 1998.
[135] Park D, King Y, Lu Q, King TJ, Hu CM, Kalnitsky A, Tay SP, and C. CC, "Transistor characteristics with Ta2O5 gate dielectric," *IEEE ELECTRON DEVICE LETTERS*, vol. 19, pp. 441-443, 1998.
[136] W. Loh, B. Cho, M. Joo, M. Li, D. Chan, S. Mathew, and D. Kwong, "Charge trapping and breakdown mechanism in HfAlO/TaN gate stack analyzed using carrier separation," *IEEE TRANSACTIONS ON DEVICE AND MATERIALS RELIABILITY*, vol. 4, pp. 696-703, 2004.
[137] S. Huang and T. Wu, "Thickness dependence of structural and electrical characteristics of ZrO2 thin films as grown on Si by chemical-vapor deposition," *JOURNAL OF VACUUM SCIENCE & TECHNOLOGY B*, vol. 22, pp. 1940-1948, 2004.
[138] S. Lee, C. Choi, A. Kamath, R. Clark, and D. Kwong, "Characterization and reliability of dual high-k gate dielectric stack (poly-Si-HfO2-SiO2) prepared by in situ RTCVD process for system-on-chip applications," *IEEE ELECTRON DEVICE LETTERS*, vol. 24, pp. 105-107, 2003.

[139] S. Lee and D. Kwong, "TDDB and polarity-dependent reliability of high-quality, ultrathin CVD HfO2 gate stack with TaN gate electrode," *IEEE ELECTRON DEVICE LETTERS*, vol. 25, pp. 13-15, 2004.

[140] J. S. Suehle, E. M. Vogel, B. Wang, and J. B. Bernstein, "Temperature dependence of soft breakdown and wear-out in sub-3 nm SiO2 films," presented at IEEE International Reliability Physics Symposium, 2000.

[141] J. S. Suehle, P. Chaparala, C. Messick, W. M. Miller, and K. C. Boyko, "Field and temperature acceleration of time-dependent dielectric breakdown in intrinsic thin SiO2," presented at IEEE International Reliability Physics Symposium, 1994.

[142] K. C. Boyko and D. L. Gerlach, "Time dependent dielectric breakdown at 210 Å oxides," presented at International Reliability Physics Symposium, 1989.

[143] T. U. Sakura, H.; Kamakura, Y.; Taniguchi, K., "A detailed study of soft- and pre-soft-breakdowns in small geometry MOS structures," presented at IEEE International Electron Devices Meeting, 1998.

[144] R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, "Relation between breakdown mode and breakdown location in short channel NMOSFETs and its impact on reliability specifications," presented at IEEE International Reliability Physics Symposium, 2001.

[145] R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, "Relation between breakdown mode and location in short-channel nMOSFETs and its impact on reliability specifications," *Device and Materials Reliability, IEEE Transactions on*, vol. 1, pp. 163-169, 2001.