Electrical and Chemical Characterisation of Ultrathin Transistor Gate Dielectric Layers

A thesis for the Degree of Doctor of Philosophy

Presented to Dublin City University

by

Robert O'Connor, B.Sc.
School of Physical Sciences
Dublin City University

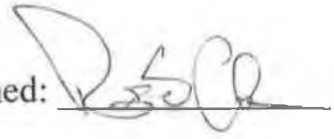
Research Supervisor:

Dr. Greg Hughes

January 2005

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent of which the work has been cited and acknowledged

Signed:  I.D Number: 97482697

Date: 11/4/05

Acknowledgements

Many thanks to everyone who helped me both directly and indirectly during the last 3 years; to Intel Ireland and Enterprise Ireland for funding. Thanks to IMEC for providing most of the samples used in this work, and to Robin Degraeve and many others at IMEC for passing on their expertise in the field, without which the project would have failed. Within DCU there are countless people to acknowledge, particular thanks to Pat Wogan for help with measurement setups and occasional automotive assistance! Away from DCU, thanks to friends and family for their support, both financial and otherwise, and lastly and most importantly, thanks to my project supervisor, Dr. Greg Hughes for listening and helping over the years and even travelling all the way to New York to make measurements!

Abstract

This work examines the suitability of both ultrathin Silicon Oxynitride (SiON) and Hafnium Silicate (HfSiON) layers to be used as high-permittivity gate dielectrics to serve as a replacement to silicon dioxide (SiO_2) in future Metal-Oxide-Semiconductor (MOS) technologies. The reason SiO_2 needs to be replaced is the extremely high levels of leakage current displayed in ultrathin layers required for sub-90nm CMOS. The main part of the thesis consists of an electrical characterisation section, where the layers are evaluated in terms of their electrical reliability when fabricated into MOS devices, so as to determine their maximum operating voltage and performance during expected device lifetime. Techniques used include constant voltage stress, constant current stress, ramped voltage stress, charge pumping to determine interface state densities and stress induced leakage current measurements. Conventional methods of determining reliability are also evaluated. Results show that these methods cannot be blindly applied in ultrathin regime, and that finding an alternative dielectric material is a major challenge.

The second part of the thesis consists of a chemical characterisation section, where blanket layers of the materials are examined using a range of surface analysis techniques. X-ray photoelectron spectroscopy (XPS), synchrotron based photoemission and secondary ion mass spectroscopy (SIMS) are used to determine the chemical composition and chemical depth profile information of the layers and to probe the electronic structure of the valence bands and allow the valence band offsets to be determined.

Publications Resulting From This Work:

Robert O'Connor, Robin Degraeve, Ben Kaczer, Anabela Veloso, Greg Hughes, Guido Groeseneken.
Weibull slope and voltage acceleration of ultrathin oxynitrides.
Microelectronic Engineering. 72 (2004) pp 61-65.

Robert O'Connor, Greg Hughes, Robin Degraeve, Ben Kaczer.
Temperature accelerated breakdown in ultra-thin SiON dielectrics
Semicond. Sci. Tech. 19 (2004) pp. 1254-1258.

Robert O'Connor, Greg Hughes, Robin Degraeve, Ben Kaczer.
Progressive breakdown in ultra-thin SiON dielectrics and its effect on transistor performance.
Microelectronics Reliability 45 (2005) pp 869-874.

Robert O'Connor, Greg Hughes, Robin Degraeve, Ben Kaczer, Thomas Kaureauf.
Reliability of HfSiON gate dielectrics.
Semicond. Sci. Tech. 20 (2005) pp. 68-71.

Robert O'Connor, Greg Hughes, Robin Degraeve, Ben Kaczer, Thomas Kaureauf.
Charge trapping in MOSFETS with HfSiON gate dielectrics.
In Press, Accepted to Microelectronic Engineering (2004).

Robert O'Connor, Stephen McDonnell, Greg Hughes.
Low voltage stress-induced leakage current in 1.4-2.1nm SiON and HfSiON gate dielectric layers.
Submitted to Semiconductor Science and Technology.

Robert O'Connor, Greg Hughes.
A synchrotron radiation study of high-k gate dielectrics
In preparation.

A. Veloso, F. N. Cubaynes, A. Rothschil1, S. Merten1, R. Degraeve, R. O'Connor, C. Olsen, L. Date, M. Schaekers, C. Dachs, M. Jurczak.
Ultrathin oxynitride gate dielectrics by pulsed-rf DPN for 65nm general-purpose CMOS applications.
Proc. of the 33rd European Solid-State Device Research Conference ESSDERC'2003, pp. 239-242 (September 2003).

B. Kaczer, R. Degraeve, R. O'Connor, Ph. Roussel, G. Groeseneken.
Implications of progressive wear-out for lifetime extrapolation of ultra-thin (EOT~1nm) SiON films.

Presented at IEDM 2004.

Chapter 1: Introduction

It is a well-known fact that the digital revolution of the last 40 years has been based on the development of increasingly faster and smaller microprocessors, memory units and other microelectronic devices that make data transfer faster. It is also common knowledge that these advanced devices have been made possible due to intensive research into the fundamental building block of these devices, namely the Metal Oxide Semiconductor Field Effect Transistor (MOSFET). The advancement of MOSFET technology has steadily followed Moore's Law [1] (as shown in figure 1.1) for over 30 years, and the number of transistors per microprocessor has now exceeded 10^8.

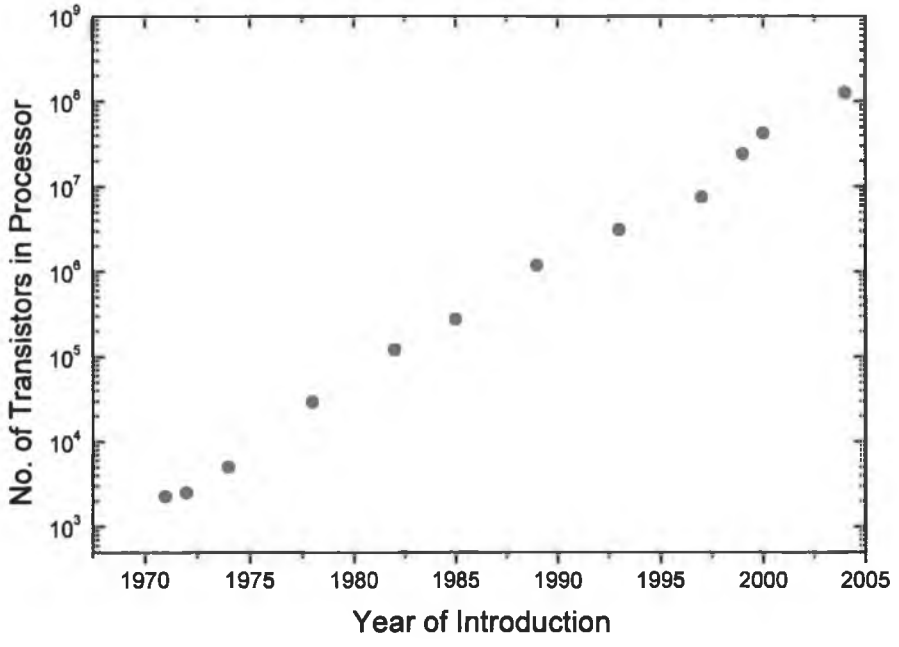


Figure 1.1 The number of MOS transistors on Intel processor versus year of introduction. Moore's Law states that the number should double every 18-24 months. Data from INTEL.

1.1 The MOS Transistor

In the microprocessor, the MOSFET essentially works as a switch for controlling digital logic, allowing Boolean calculations. The MOSFET is fabricated as shown in figure 1.2 by implanting a p-type (or n-type) silicon substrate with n-type (p-type) wells which act as source and drain. A layer of insulating material is used to cover the channel between the source and the drain. This insulator allows a conductive channel to be formed at the substrate-insulator interface when a sufficient gate bias is applied, thereby enabling conduction from source to drain.

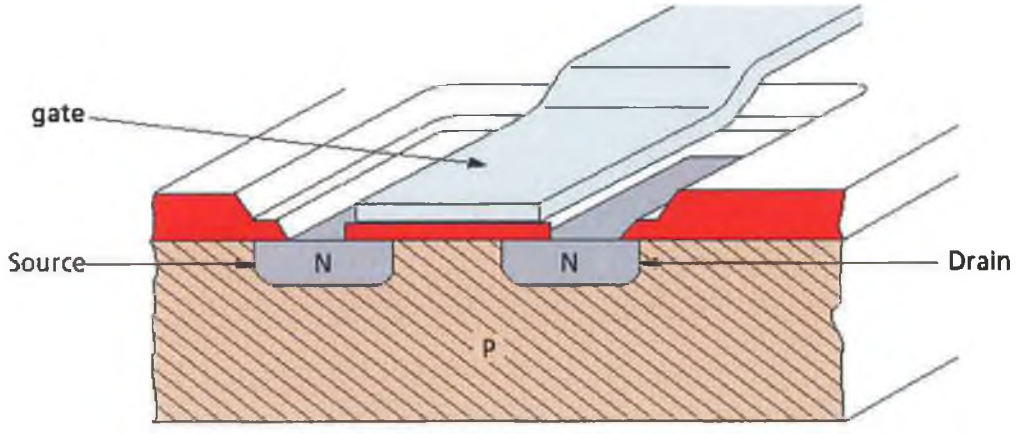


Figure 1.2 Schematic cross section of an n-MOSFET.

The gate voltage controls the switching action by controlling the carrier density in the channel. Above the 'threshold voltage', current flows and the switch is on, below this voltage almost zero current flows and the switch is off. A drain current-gate voltage (I_dV_g) curve for such a device is shown in figure 1.3, where a fixed voltage of 0.05V is applied to the drain and the gate voltage is increased.

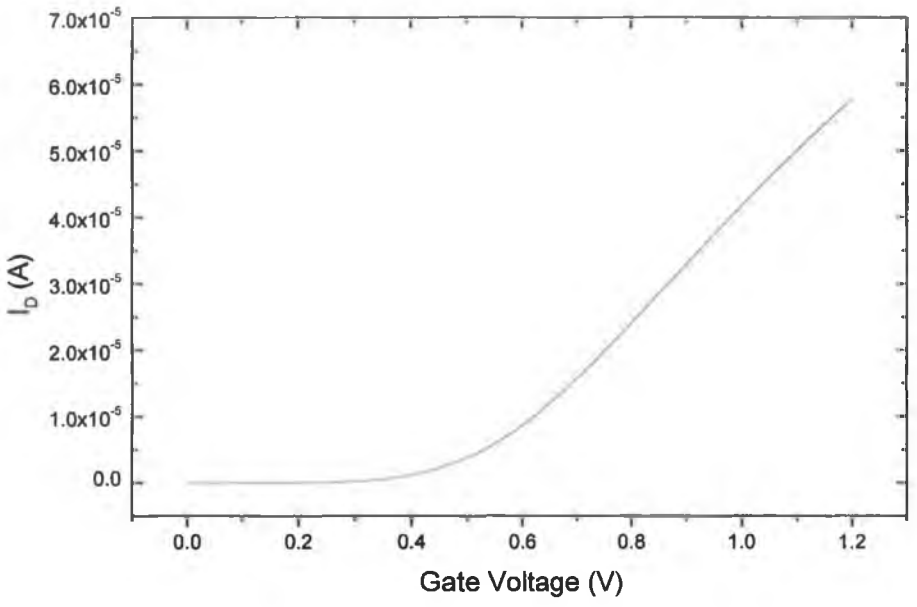


Figure 1.3 Sample I_dV_g curve for a MOSFET

The threshold behaviour can be clearly seen above a gate voltage of 0.5V. Another plot commonly used to analyse a MOSFET is shown in figure 1.4. The plot shows the drain current (I_d) versus the drain voltage (V_g) characteristics for a range of gate voltages and illustrates the how the I-V curve for the channel is modified by the changing gate voltage.

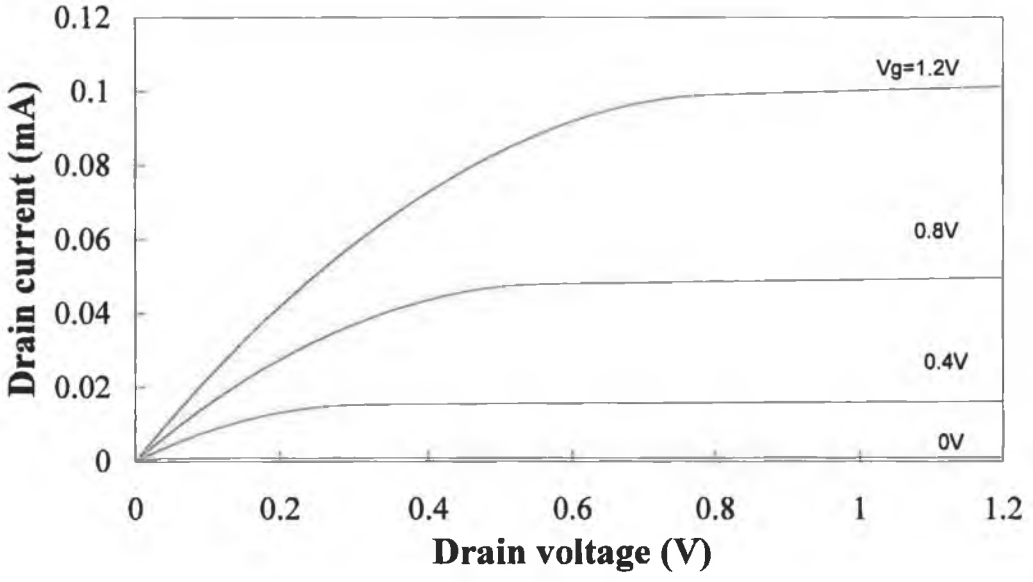


Figure 1.4 Sample I_dV_d curves for a MOSFET, showing the effect of increasing gate voltage.

1.2 Scaling Limits and the End of SiO_2

The gate dielectric plays a critical role in MOSFET technology. From the beginning of integrated circuit fabrication from silicon, the material used as the gate dielectric has been SiO_2. It is the native oxide of Si and has excellent material properties including large bandgap (~8.8eV), high resistivity (>10^{12} Ω.m), and it forms a high quality interface with the Si substrate with very low interface state densities (<$5x10^{10}cm^{-2}$)[2]. All of these properties make it the ideal choice as dielectric layer, and indeed it is because of these properties that the Si/SiO_2 system has been so successful in comparison to other possible semiconductor candidates, such as germanium.

Over the years, SiO_2 has continued to be the dielectric of choice during aggressive scaling that has led to a reduction in channel length (the distance from source to drain in the silicon substrate) from 6μm for the 8080 Processor in 1974 to 0.18μm for the Pentium 4 in 2000. Already channel lengths as low as 65nm are being currently tested for integration into microprocessors[3]. One of the problems associated with this approach is that we cannot assume that device properties will be maintained for as long as the down-scaling continues. Over the years, scaling laws have often appeared to be a limiting factor. Decreasing all device dimensions increases current density and power loss, but a reduction in supply voltage to counteract this limits the speed, so there are many trade offs to be considered.

As dimensions are further reduced, a problem with the SiO_2 layer becomes evident. Electrons can tunnel quantum mechanically through barrier layers, which have a thickness in the nanometer range. The tunnel current is exponentially dependent on oxide thickness and in ultra-thin layers it can result in the degradation of the insulating properties of the dielectric, within the expected lifetime, which can lead to device failure. The tunnel current vs. oxide thickness plot [4] in figure 1.5 illustrates the extraordinary current densities present in ultrathin layers. The horizontal lines, indicating the maximum allowable leakage currents for portable and desktop applications, show that an alternative dielectric with lower leakage current is required below 1.2nm thickness.

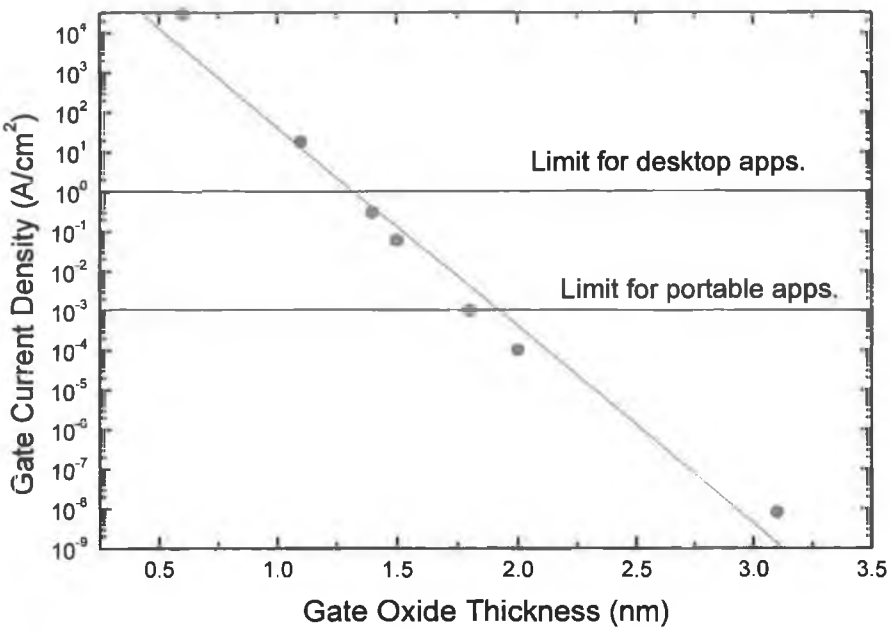


Figure 1.5 Oxide leakage current @1.5V for 35nm channel length nMOS devices. From ref. 2.

Indeed below 1.2nm the idea of a layer of SiO_2 begins to lose meaning as a large fraction of the layer is made of the structural transition region from Si to SiO_2 leaving perhaps 2-3 monolayers of 'bulk' SiO_2. At these ultrathin dimensions, thickness uniformity also becomes an issue due to the fact that a non-uniform layer leads to conduction hot-spots[5].

The basic necessity for an alternative material is that it can give the same capacitance as an ultrathin layer of SiO_2, but can be fabricated thicker to stem the leakage current problem. The basic equation dictating the capacitance of a dielectric layer is given by:

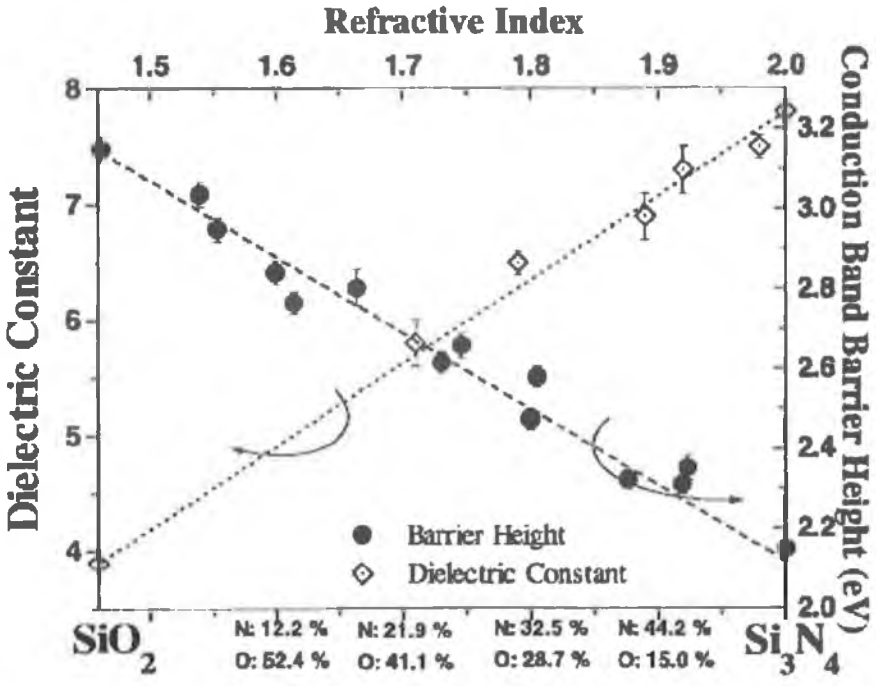
$$C_{ox} = \frac{\varepsilon_r \varepsilon_0 A}{d} \qquad \text{Eq. 1.1}$$

where ε_r (sometimes denoted by 'k') is the relative permittivity of the material, ε_0 is the permittivity of free space, A is the area of the layer, and d is the thickness. It is clear that using a material with a higher relative permittivity (dielectric constant) than SiO_2 (k=3.82) would allow the thickness of the layer to be increased while maintaining the same capacitance. This increase in thickness

has the effect of reducing the leakage current due to a reduction in the quantum mechanical tunneling current.

The addition of nitrogen to SiO_2 to make a SiO_xN_y dielectric layer was found to increase the dielectric constant slightly[6]. As the nitrogen concentration is increased in the layer, the dielectric constant rises linearly, as seen in figure 1.6. As well as raising the dielectric constant, SiON gate dielectrics have been shown to suppress the migration of boron from p+ poly-Si gates into the channel, which is a problem with SiO_2 that causes threshold voltage shifts [7]. The nitrogen incorporation also leads to a better quality Si/SiO_2 interface. One reason which has been proposed to explain this improvement is that the introduction of a trivalent atom accommodates more flexible local bonding arrangements than is possible with the tetravalent silicon and divalent oxygen at the interface. Note also from the figure that nitrogen incorporation has the negative effect of lowering the barrier height to tunnelling carriers. The obvious question that arises from studying figure 1.6 is 'Why not use pure Si_3N_4 as the dielectric, as you could have a layer twice as thick with the same capacitance?'

The reason is that at higher nitrogen concentrations, transconductance (the conductivity of the channel in a transistor) is lowered, the threshold voltage is shifted and, at high concentrations near the interface, the device mobility is adversely affected [6]. While SiO_xN_y layers are clearly convenient gate dielectrics, involving only a modest change in processing steps, it is accepted that their use in advanced processors will be limited and that a move completely away from silicon oxide based dielectrics will eventually be needed as aggressive dimensional scaling continues.


Refractive Index
1.5
1.6
1.7
1.8
1.9
2.0
Dielectric Constant
Conduction Band Barrier Height (eV)
8
7
6
5
4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
Barrier Height
Dielectric Constant
SiO2
N: 12.2 %
O: 52.4 %
N: 21.9 %
O: 41.1 %
N: 32.5 %
O: 28.7 %
N: 44.2 %
O: 15.0 %
Si3N4

1.3 Charge transport mechanisms in gate dielectrics

When a ramp voltage is applied across an oxide, it becomes clear that the oxide is not a perfect insulator, as a current is detected. Figure 1.7 below shows the I-V characteristics for 3 different oxides prepared under different conditions, with a thickness of ~1.5nm. The sudden increase in current at a voltage of about 3.5V results from a breakdown of the insulating properties of the layers. From the point where this breakdown occurs we can then calculate the breakdown field ($E=V/t_{ox}$) to be 2.33×10^9 V/m.

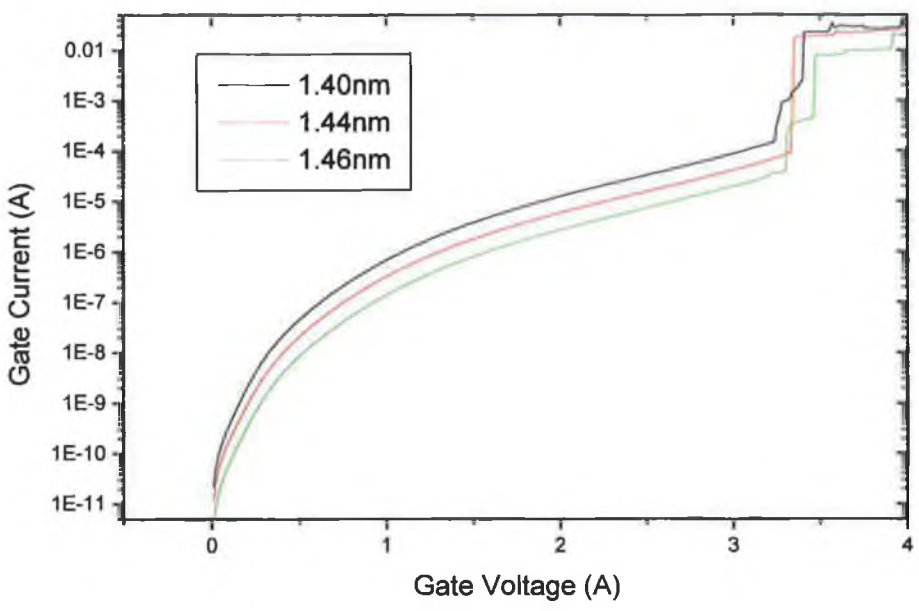


Figure 1.7 Tunnel current in 1μm x 1μm capacitor with active area encompassing source and drain, n+ poly-si gate and active area, p-type substrate. Thicknesses quoted are optically measured.

The predominant conduction process in dielectric layers of this thickness is quantum mechanical electron tunnelling through the oxide[9]. During I-V measurements on thin oxides where the gate voltage is lower than 3.7V, the electrons no longer travel through a triangular energy barrier into the oxide conduction band as happens for thicker oxides. The electrons tunnel directly from anode to cathode into the conduction band of the substrate. This process is

shown in figure 1.8. The easiest way to differentiate between the two is that direct tunnelling shows no temperature dependence, as the same barrier is encountered by tunnelling carriers independent of thermal energy, whereas with FN tunnelling the barrier becomes smaller as thermal energy is increased. The two processes are described by different equations, but the important factor is that the tunnel current is exponentially dependent on oxide thickness. In thicker oxides, Fowler-Nordheim[10] tunnelling dominates, where the electrons travel through the SiO_2 conduction band into that of the silicon substrate.

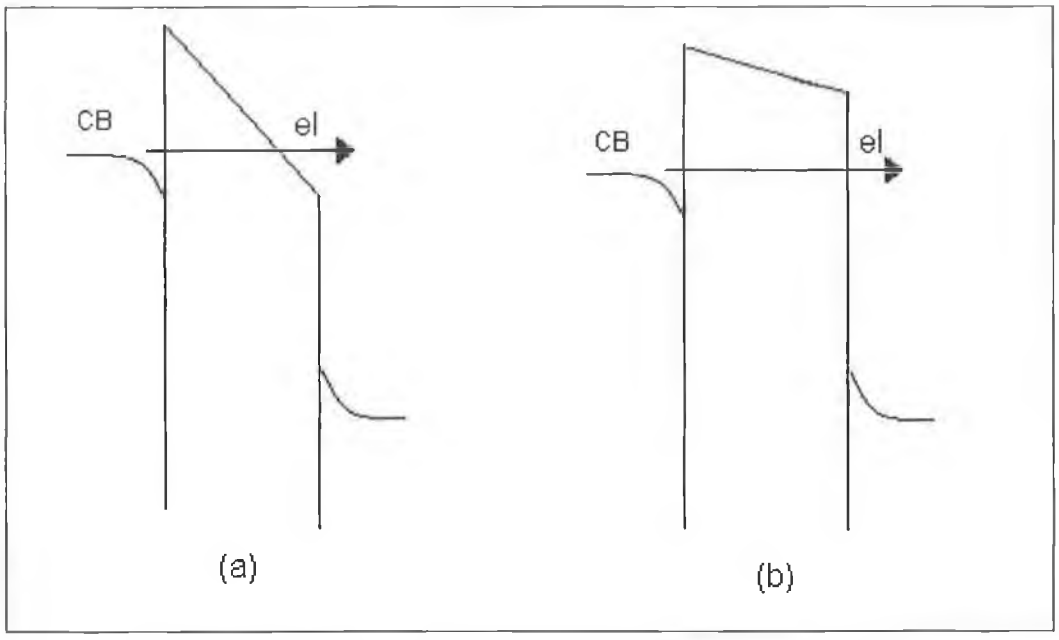


Figure 1.8 Tunnelling mechanisms though SiO_2. 'CB' represents the start of the silicon conduction band

Fowler-Nordheim Tunnelling is described by the equation[13]:

$$J_{FN} = A.E_{ox}^2 \exp\left[-\frac{B}{E_{ox}}\right] \qquad \text{Eq 1.2}$$

where $A = \frac{q^3}{16\pi^2 \hbar \Phi_B}$ and $B = \frac{4}{3}\frac{(2m^*)^{1/2}}{q\hbar}\Phi_B^{3/2}$.

q is the electron charge, Φ_B is the barrier height, and m^* is the effective electron mass .

Direct tunnelling is more difficult to describe and involves using the WKB approximation. One model describes the current as[12]:

$$J_{DT} = AB\exp\left(\frac{\left(-4\pi(2q.m^*)^{1/2})T_{ox}B^{1/2}\right)}{h}\right) - A\Phi_B\exp\left(\frac{\left(-4\pi(2q.m^*)^{1/2}T_{ox}\Phi_B^{1/2}\right.}{h}\right)$$

(Eq 1.3)

where $A = \Phi_B - {}^{V_{ox}}\!/_{2}$ and $B = \left({}^{q^2}\!/_{2\pi hT_{ox}^2}\right)$, V_{ox} is the oxide voltage, and T_{ox} is the oxide thickness. Clearly direct tunnelling is a lot more difficult to describe mathematically. Tunnel current mechanisms are discussed in more detail in Chapter 3.

1.4 The Si/insulator Interface

The ideal device will have a periodic crystal of Si terminated at the interface where a layer of amorphous SiO_2 begins. This however is only an ideal case, and in Si/SiO_2 systems is not possible. Termination is never perfect due to bond orientation and strain between the two layers as shown in Figure 1.9. Hydrogen is used to passivate dangling Si bonds at the interface as the uppermost layer of Si cannot be co-ordinated to four oxygen atoms as required for stoichiometric SiO_2[11]. Imperfections at the Si/SiO_2 surface can also degrade the quality of the interface.

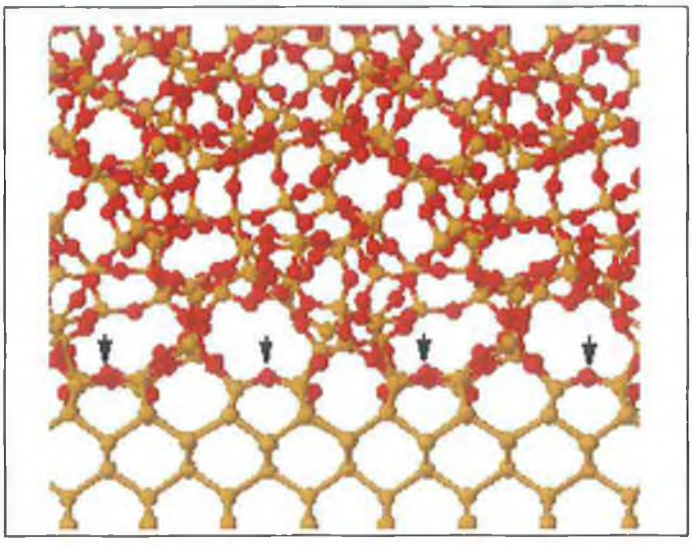


Figure 1.9 Schematic diagram of the Si/SiO_2 interface showing the transition from amorphous SiO_2 to crystalline Si. This transition cannot be made without introducing structural imperfections. Yellow-silicon, Red-oxygen

A more realistic description of the interface is crystalline Si merging into some interfacial oxide SiO_x where x is less than 2 and then into bulk SiO_2. The resulting interface trapped charge (from mid-bandgap interface states) contributes to threshold voltage shifts and accelerates oxide breakdown, and can also degrade transconductance (and thus channel mobility) due to scattering[12]. As mentioned earlier, as the gate dielectric is scaled down toward 1nm, the interface becomes a more significant fraction of the total layer, and the understanding of interface properties becomes all the more critical. These interfacial problems become more pronounced as SiO_2 is replaced by new materials, which may behave quite differently on a Si substrate.

1.5 Summary and Outline

It is clear that dielectric scaling poses a significant problem for the semiconductor industry. Chapter 2 outlines the techniques used for the study of the chemical composition of blanket dielectric layers, including X-ray Photoelectron Spectroscopy (XPS), synchrotron radiation based photoemission studies, and Secondary Ion Mass Spectroscopy (SIMS). XPS provides information on the chemical composition while the synchrotron based

photoemission studies probe the electronic structure of the valence bands and allow the valence band offsets to be determined. SIMS provides some useful depth profile information. This work in Chapters 3,4, and 5 focuses on time dependent dielectric breakdown (TDDB), and the external parameters that influence reliability estimates based on TDDB measurements. Chapter 3 outlines the techniques used to characterise devices, how reliability estimates are arrived at, the causes of breakdown and the effect of oxide breakdown on a device. Chapter 4 deals with work carried out on ultra-thin SiON dielectric layers for use in the 65nm process node, their reliability, the temperature dependence of time-to-breakdown, the progressive nature of the breakdown and the effects of oxide degradation on MOSFET parameters and the implications for a real circuit. Chapter 5 deals with a similar procedure for HfSiON layers, and outlines the differences in characterisation methods and reliability between the two materials. The final part of the thesis (Chapter 6) deals with the chemical characterisation of blanket layers of the SiON and HfSiON dielectrics. These layers are the same thickness as those used in the device structures. Chapter 7 then concludes the work, summarising the findings from the electrical and chemical characterisation of the dielectric layers, and includes a brief discussion on the options available for CMOS dielectric layers for future generations.

1.6 References

[1] G.E. Moore, Cramming more components onto integrated circuits, Electronics 38 (8) 1965.

[2] M.L. Green, E.P. Gusev, R. Degraeve, E.L. Garfunkel, Ultrathin (<4nm) SiO_2 and Si-O-N gate dielectric layers for silicon microelectronics: Understanding the processing, structure, and physical and electrical limits, J. Appl. Phys. 90 (5) (2001) pp. 2057-2121.

[3] A. Veloso, F. N. Cubaynes, A. Rothschild, S. Mertens, R. Degraeve, R. O'Connor, C. Olsen, L. Date, M. Schaekers, C. Dachs, M. Jurczak, Ultra-thin Oxynitride Gate Dielectrics by Pulsed-RF DPN for 65 nm General Purpose CMOS Applications, *Proc. of the 33rd ESSDERC'2003*, pp. 239-242.

[4] M. L. Green, T. W. Sorsch, G. L. Timp, D. A. Muller, B. E. Weir, P. J. Silverman, S. V. Moccio and Y. O. Kim, Understanding the limits of ultrathin SiO_2 and Si-O-N gate dielectrics for sub-50 nm CMOS, Microeletron. Eng. 48 (1-4) (1999) pp. 25-30.

[5] J.H. Stathis, Reliability limits for the gate insulator in CMOS technology, IBM J. Research and Dev 46 (2/3) (2002) pp. 265-286.

[6] X. Guo, T.P. Ma, Tunneling leakage current in oxynitride: dependence on oxygen/nitrogen content, IEEE Elec. Dev. Letters, 19 (6) (1998) pp. 207-209.

[7] C.T. Wu, A. Mieckowski, R. S. Ridley Sr., G. Dolny, T. Grebs, J. Linn and J. Ruzyllo, Effect of nitridation on the reliability of thick gate oxides, Microelectron. Reliab. 43 (1) (2003) pp. 43-47.

[8] G.D. Wilk, R.M. Wallace, J.M. Anthony, High-k gate dielectrics: Current status and materials properties consideration, J. Appl. Phys. 89 (10) (2001), pp. 5243-5275.

[9] M. Hirose, Electron tunnelling through ultrathin SiO_2, Materials Sci, and Eng, B41 (1996) pp. 35-38.

[10]R. H. Fowler, L. Nordheim, Electron emission in intense electric fields, Proc. R. Soc. London, Ser. A 119, 173 (1928).

[11] E.H. Poindexter, MOS interface states: overview and physicochemical perspective, Semicond. Sci. Tech. 4 (1989) pp.961-969.

[12] Y.B. Park, D.K. Schroder, Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM, IEEE Trans. Electron. Devices, 45 (6) pp. 1361-1368 (1998).

Chapter 2: Chemical Characterisation Techniques for Dielectric Layer Analysis

2.1 Introduction

It is the electrical characteristics of a device that are the ultimate factor in evaluating the effectiveness of a particular dielectric layer. However, the chemical composition, elemental depth profile, and interface quality of the layer impact on the electrical properties that it exhibits. Accordingly, chemical characterisation techniques are used to obtain information about the valence and conduction band offsets, which have a direct influence on the electrical characteristics of the structures.

The SiO_2/Si interface has been studied thoroughly in the past 30 years, and is relatively well understood. However, as technology has evolved and the requirement for a new dielectric material has become evident, many new material systems are being studied. The incorporation of nitrogen into the SiO_2, which raises the dielectric constant slightly was undertaken and, more recently, the use of metal oxides such as HfO_2 and TiO_2 which have significantly higher dielectric constants is being actively investigated. More complex systems have also come to be studied, like gate stacks with two or three interfaces within the layered structure[1].

This chapter outlines the techniques use to characterise both SiON and HfSiON blanket dielectric layers deposited on silicon substrates. The SiON layers were grown in the same process as the device wafers that are electrically characterised in Chapter 4 of this thesis, and therefore directly represent the same oxide layers. This allows for the possibility of attributing particular electrical characteristics to different processing conditions. The hafnium silicate blanket layers were grown for test purposes and were not integrated into device structures. Nevertheless, some interesting chemical analysis can be carried out by comparing results from different processing conditions.

Techniques used for the chemical analysis were X-ray photoelectron spectroscopy (XPS) (and angle resolved XPS), Soft XPS using a synchrotron radiation source, X-ray absorption spectroscopy (XAS) and Secondary Ion Mass Spectroscopy (SIMS). The underlying principles and strengths of each technique are briefly outlined in this chapter.

2.2 X-ray Photoelectron Spectroscopy

2.2.1 Photoemission

The basic photoemission process is shown in figure 2.1[2]. When light of suitable photon energy is incident on a solid, electrons in the material can absorb the photons as described by the photoelectric effect. In this work, the x-ray photon source was a conventional Mg Kα line producing 1253.6eV photons. Because of the energy of the incident photon, the electrons that have binding energy less than 1250eV can be excited from below the Fermi level to above the vacuum level of the solid[3].

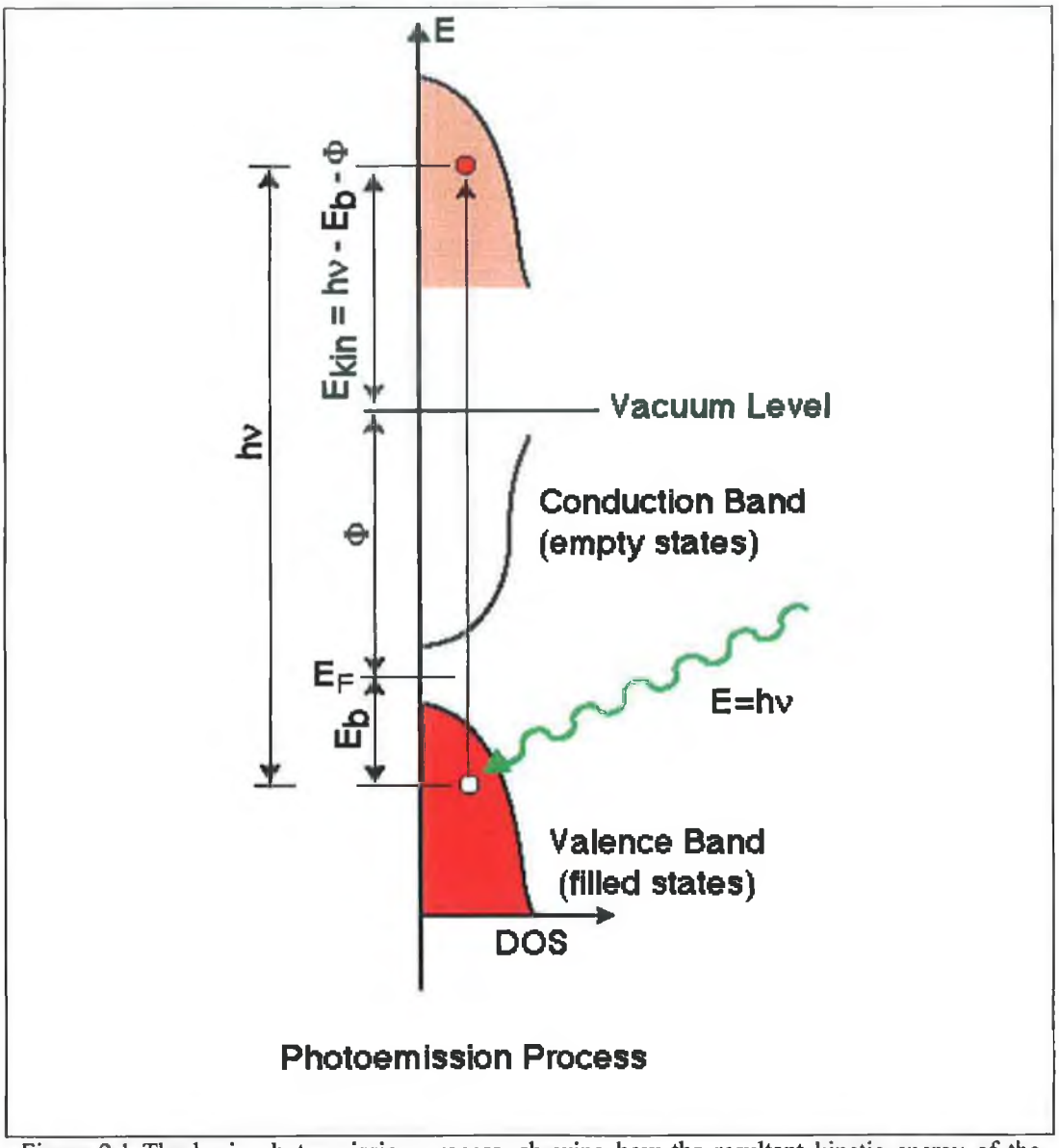


Figure 2.1 The basic photoemission process, showing how the resultant kinetic energy of the photoelectron is deduced.

The kinetic energy of such an electron is described in equation 1.

$$E_{Kin} = h\nu - E_B - \phi \qquad \text{Eq. 2.1}$$

That is, the kinetic energy (E_k) of an emitted electron is dependent on the incident photon energy ($h\nu$), the binding energy of the electron in the parent atom (E_B) and the work function (ϕ) of the solid. Because the binding energy of each core level electron is characteristic of the atom from which it came, analysis of the range of kinetic energies of the emitted electrons allows elemental

analysis. Valence band states can also be probed. They are however broader and more complicated as valence band electrons are involved in bonding with other atoms. A typical survey XPS spectrum of an ultrathin HfSiON layer is shown in figure 2.2.

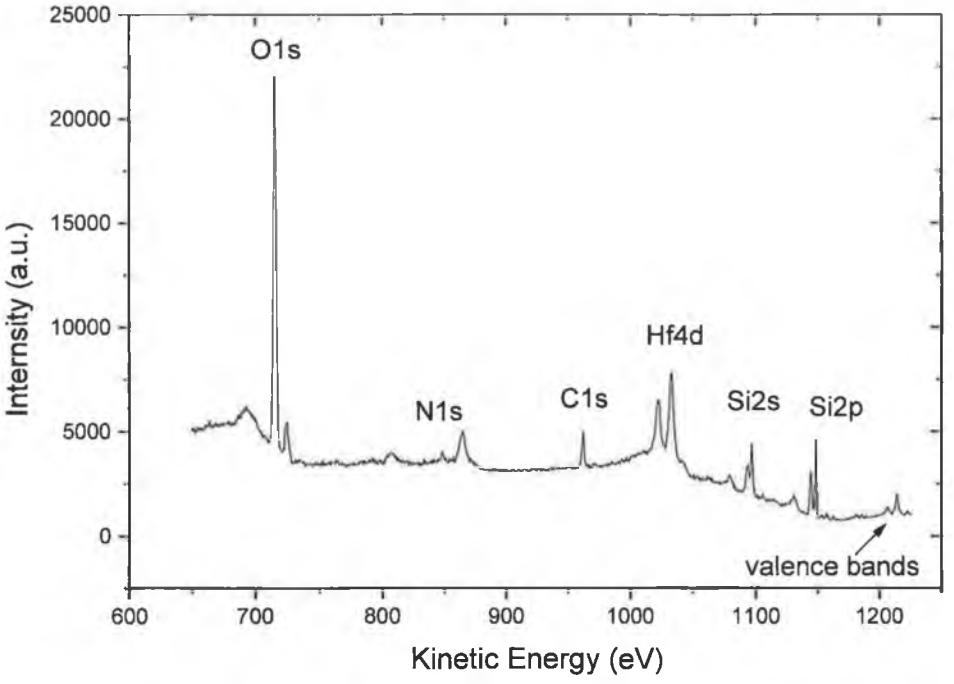


Figure 2.2 A typical wide scan for a HfSiON sample showing peaks associated with all of the expected elements.

The peaks associated with elements silicon, hafnium, oxygen, nitrogen, and carbon (which contaminates the sample surface) can be identified, as can the valence band levels at higher kinetic energy. The entire spectrum sits on a background of secondary electrons, which have been scattered during the photoemission process.

This wide energy scan is primarily used to determine the individual elements present in the sample. Further higher-resolution scans are carried out in the region of each elemental peak. In this case, Si2p, Hf4d, C1s, N1s, and O1s would be investigated. Shown in figure 2.3 is a Si2p spectrum from this HfSiON sample. There are 2 peaks, one at 1150.25eV and one at 1146.6eV. These kinetic energies can be converted to binding energies of 102.75eV and 106.4eV

respectively. The lower binding energy peak is attributed to silicon atoms in the underlying substrate, while the higher binding peak is related to silicon atoms in an SiO_2 environment in the dielectric overlayer. The difference in binding energy reflects the difference in the chemical environment of these silicon atoms and is conventionally referred as to the chemical shift. The fact that the silicon substrate signal can be observed illustrates that the sampling depth of the technique is sufficient to 'see' through to the silicon substrate. The sampling depth is limited by the mean free path of an electron in the material in question. That is, the average distance an electron will travel in the material without undergoing a collision. Obviously, electrons will be detected from deeper down in a material with a high mean free path.

The dependence of electron mean free path on electron kinetic energy is illustrated in Figure 2.4[4]

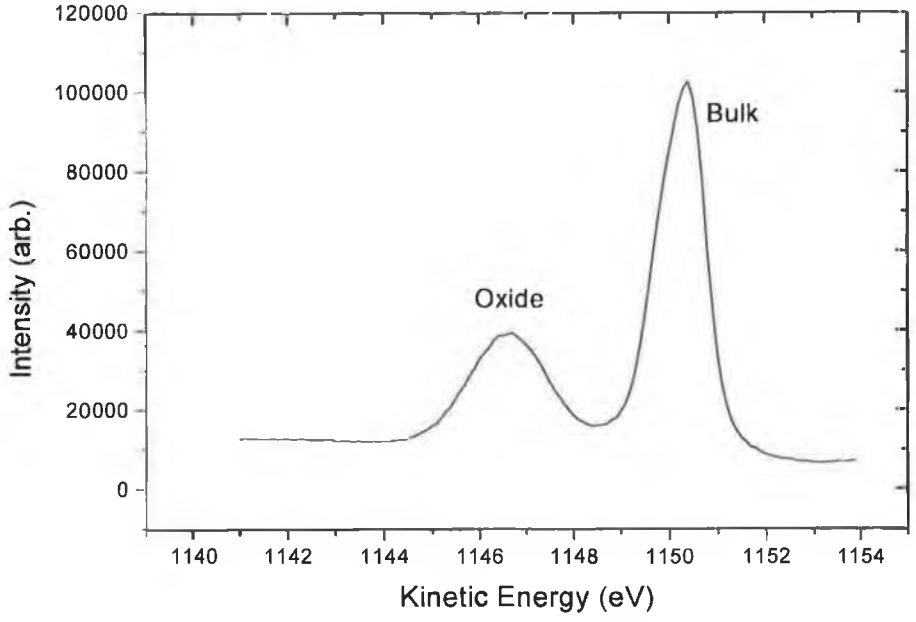


Figure 2.3 A typical Si2p spectrum. The signal from the bulk is larger than the oxide signal, indicating that the sampling depth is much greater than the oxide thickness

The photoelectrons detected in our experiments, where measured $E_K < 1000eV$, will thus originate from less than 10nm into the surface, due to exponential drop-off in signal with depth. Because the dielectric layers under consideration are no

more than ~5nm physical thickness, both the overlayer and the substrate can be simultaneously probed.

Because the samples under consideration are insulators, they are slow to replace electrons that leave during the photoemission process, leaving the layer positively charged. This charging causes the spectrum to move to higher binding energy. Because of this effect, all spectral peaks are referenced to the known binding energy of the surface C1s peak (285eV).

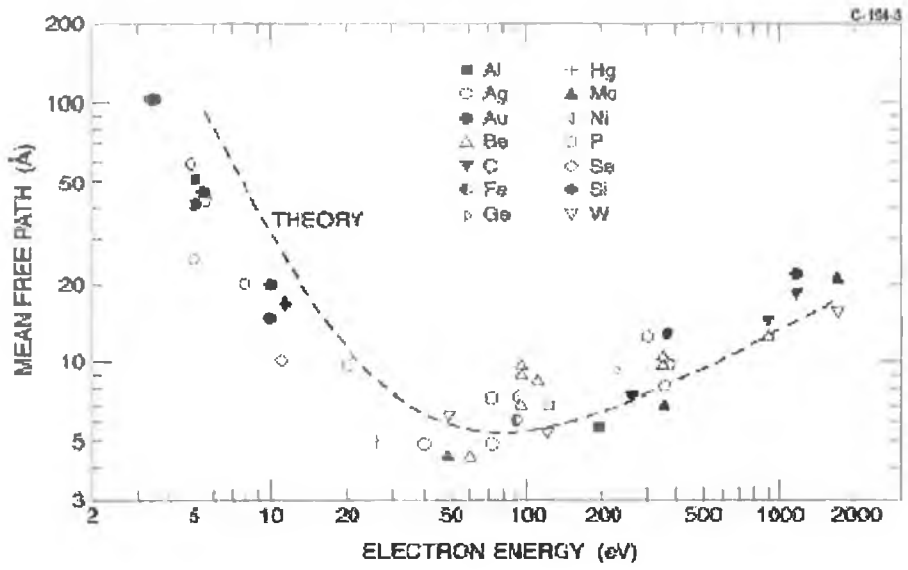


Figure 2.4 The universal electron mean free path curve.

2.2.2 Experimental Setup

XPS is performed under vacuum, so as photoelectrons can enter the electron analyser without being scattered and the sample surface remains of constant chemical composition for the duration of the experiment. The vacuum system used consists of two rotary backed oil diffusion pumps. The system is shown in figure 2.5a. Samples are transferred from air through the fast entry lock on the right through a gate valve into an ultrahigh vacuum (UHV) preparation chamber. The sample is then transferred through another gate valve into the analysis chamber containing the x-ray source and the analyser. The hemispherical electron energy analyser is schematically shown in figure 2.5b. The excited electrons are focused through an electrostatic lens, and then pass through the

hemispherical analyser. Only electrons of energy E_0 will pass through the hemispheres undeflected. E_0 is swept through the entire energy range under consideration to acquire a spectrum.

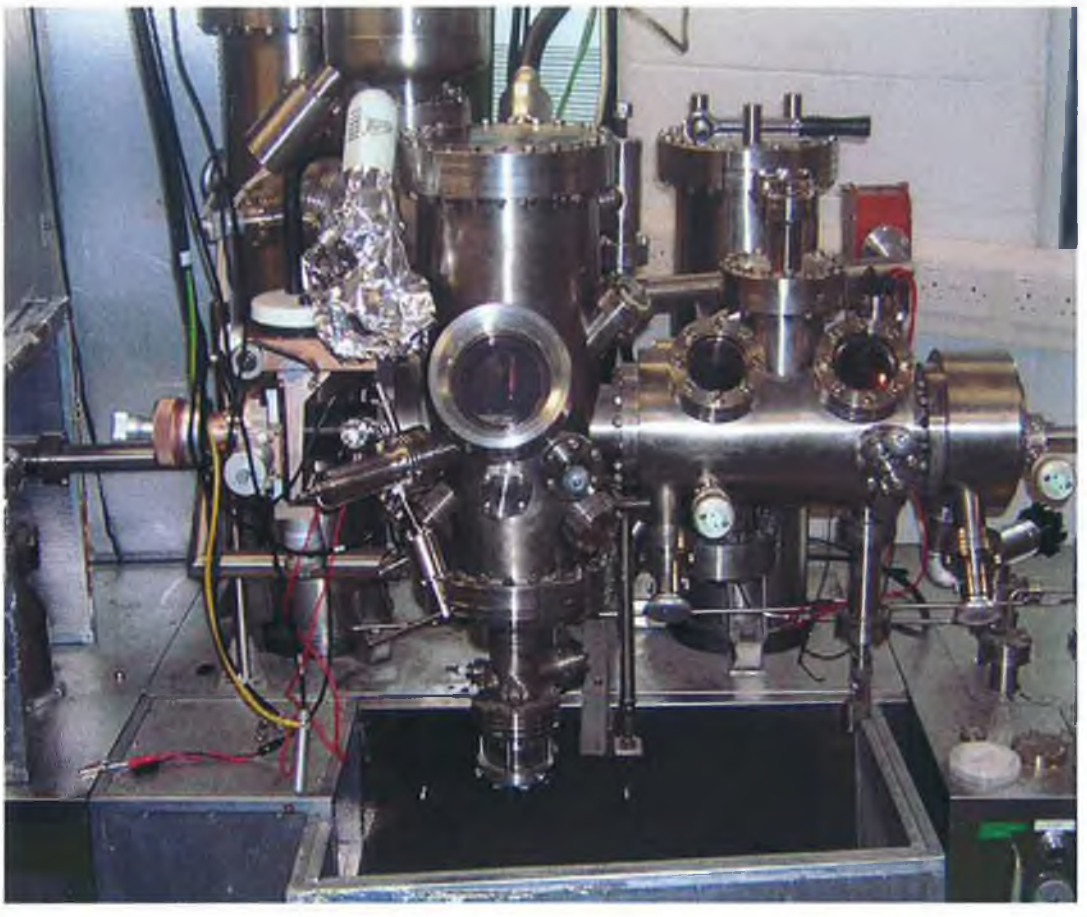


Figure 2.5a XPS system. Samples are inserted through the fast entry lock on the right, through the preparation, and into the main analysis chamber where measurements are performed under ultrahigh vacuum

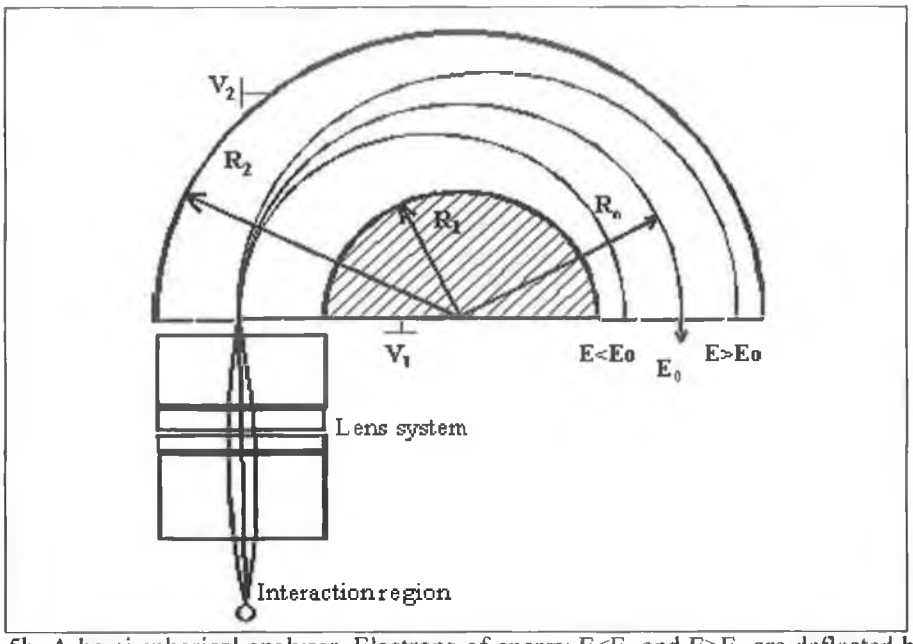


Figure 2.5b. A hemi-spherical analyser. Electrons of energy $E<E_0$ and $E>E_0$ are deflected by the potential applied to the analyser walls.

The photoelectron current is amplified by an electron multiplier know as a channeltron. The multiplier amplifies the signal by accelerating the electrons through a potential difference of 3.5KeV, causing the electrons to collide with the multiplier surface, releasing many more electrons in an electron cascade. The signal is then fed to a computer, which displays the number of counts as a function of electron kinetic energy.

2.2.3 Experimental Data Analysis

Before making quantitative comparisons of elemental composition, the 'photo-ionisation cross-section[5] for the elements in question must be taken into account. This is basically a measure of the probability of a particular atomic level being photo-ionised by an incident photon. Each photoemission peak is scaled to account for this before quantitative analysis can be carried out. The scaling factors are known as Relative Sensitivity Factors (RSF) and are used to normalise the peak heights in order to extract elemental concentrations from the spectra.

For the purpose of our study, each peak was divided by the appropriate elemental RSF, and the normalised peak heights were determined. The background was

removed by simple linear subtraction. The peak heights for each element could then be compared in order to determine elemental concentrations. An example is shown below in table 2.1

Element	Counts	RSF	Normalised Number of Counts
Silicon	3000	.25	12000
Oxygen	26400	.66	40000
Nitrogen	1680	.42	4000
Total	31080		56000

Table 2.1: Sample elemental peak heights and normalisation in XPS

In this simplistic case, the silicon concentration would be 12000/56000= 21.5%, 71.4% for oxygen, and 8.1% for nitrogen.

The wide energy scan in figure 2.2 shows a carbon 1s peak at just above 950eV kinetic energy. This peak is attributed to hydrocarbon contamination on the surface of the sample[6], which forms when the sample is exposed to air. Generally this surface layer is only a couple of monolayers deep and represents <10% of the total signal. However care must be taken, as carbon can continue to adsorb to the surface when in the analysis chamber. The additional carbon comes from the oil diffusion pumps, which help maintain the vacuum[7]. The effect of the carbon on measurements can be better understood by using angle resolved XPS.

2.3 Angle resolved XPS (ARXPS)

Angle resolved XPS was used in an attempt to study the depth profile of the layers. A more commonly used method to determine depth profile information is to perform an XPS scan followed by Ar^+ bombardment and then another XPS scan, in a sequential manner. However, Ar bombardment is found to damage the layer, introducing sub-oxides to the top surface of the layer[8] and contributing to

intermixing across the interface. Also, because the layers under consideration are so thin, control of the uniformity of the bombarding process becomes difficult.

ARXPS involves a modification of conventional XPS, wherein the sample is rotated about it's own axis normal to the x-ray beam, and XPS spectra are taken at a number of angles with respect to the surface normal. This is shown schematically in figure 2.6a. When the sample is rotated to an angle θ with respect to the detector, the path length for an electron to escape the layer at depth 'd' into the material becomes $x = \frac{d}{\cos\theta}$. This is illustrated in figure 2.6b.

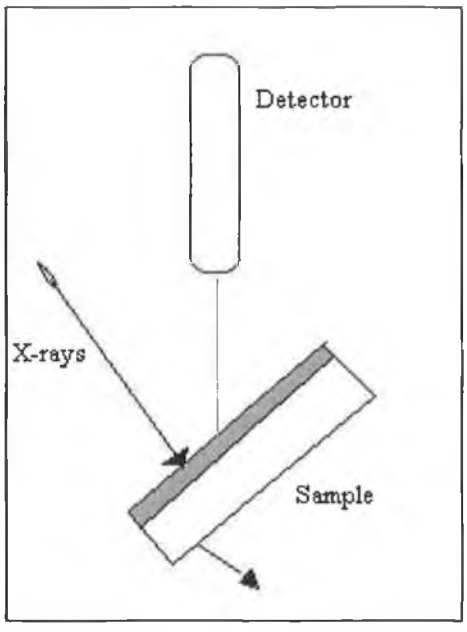


Figure 2.6a Angle Resolved XPS. The sample is rotated about its own axis, thus varying the sample depth

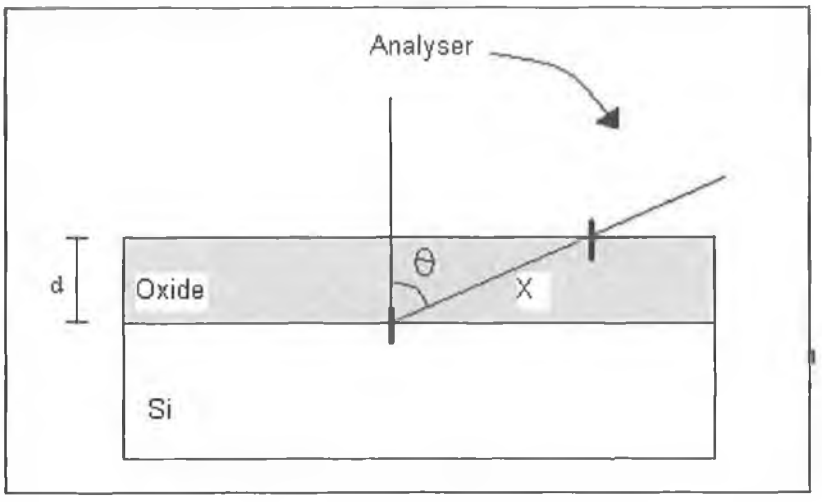


Figure 2.6b Path length difference in ARXPS.

Because the XPS signal falls off exponentially with depth, a scan taken at an angle $>0^{\circ}$ will have a greater sensitivity to the upper part of the layer, and at extreme angles, virtually all of the signal comes from the overlayer. Assuming a sampling depth of 3.5nm for example, and an overlayer of thickness 2nm, 55% of detected electrons will originate from the overlayer. Moving to an angle of 30° means that 66% of detected electrons come from the overlayer and at 60°, the number rises to 75%. So, in this manner a chemical compositional depth profile of the layer can be determined.

If an element is distributed uniformly in the material, its associated peak height should ideally remain constant with angle. However, surface hydrocarbon contamination attenuates the signal with a cos θ dependence and so the signal from a uniformly distributed element will in fact have a cos θ dependence because of the presence of this layer. If the peak drops off faster than this, it is likely that the element is concentrated deeper in the layer, as the off normal emission angle measurement is more sensitive to the surface. If the peak intensity drops off slower than cos θ, the element is primarily located at the surface. In the example shown in figure 2.7, for a SiON sample, the oxide silicon signal falls off slightly faster than cos θ, indicating that it has a slightly higher concentration in the lower part of the layer, whereas the nitrogen drops significantly slower than cos θ, indicating that it is heavily surface localised.

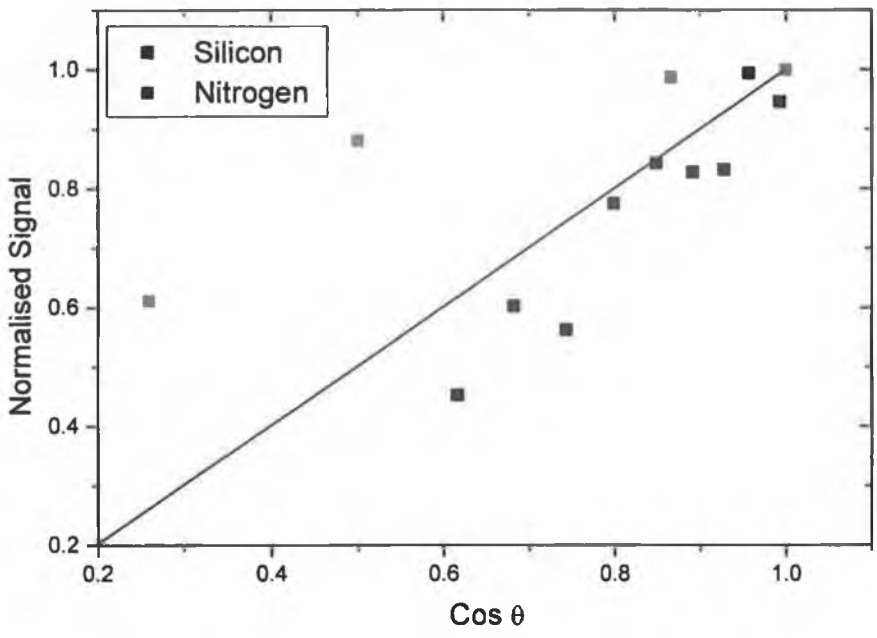


Figure 2.7 Normalised Si and N peaks show the change of signal with angle. The variation from cos θ behaviour indicates that the elements are not uniformly distributed through the layer.

De-convoluting the data so as to obtain a quantitative depth profile is an extremely complicated process and was not examined in great detail. Results are presented in Chapter 6 using an ARXPS software module by Tougaard[9], to model the layers based on ARXPS intensities.

2.4 Soft XPS Using Synchrotron Radiation

Soft XPS measurements were taken on the U4A beamline at the National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory, New York. In the synchrotron, electrons travel under UHV at high velocity in a circular path, 'steered' by a magnetic field. While moving at velocities near the speed of light, the electrons emit electromagnetic radiation tangential to their direction of travel. This broadband electromagnetic radiation is made available to individual beamlines tangential to the storage ring. A synchrotron is shown schematically in figure 2.8[10].

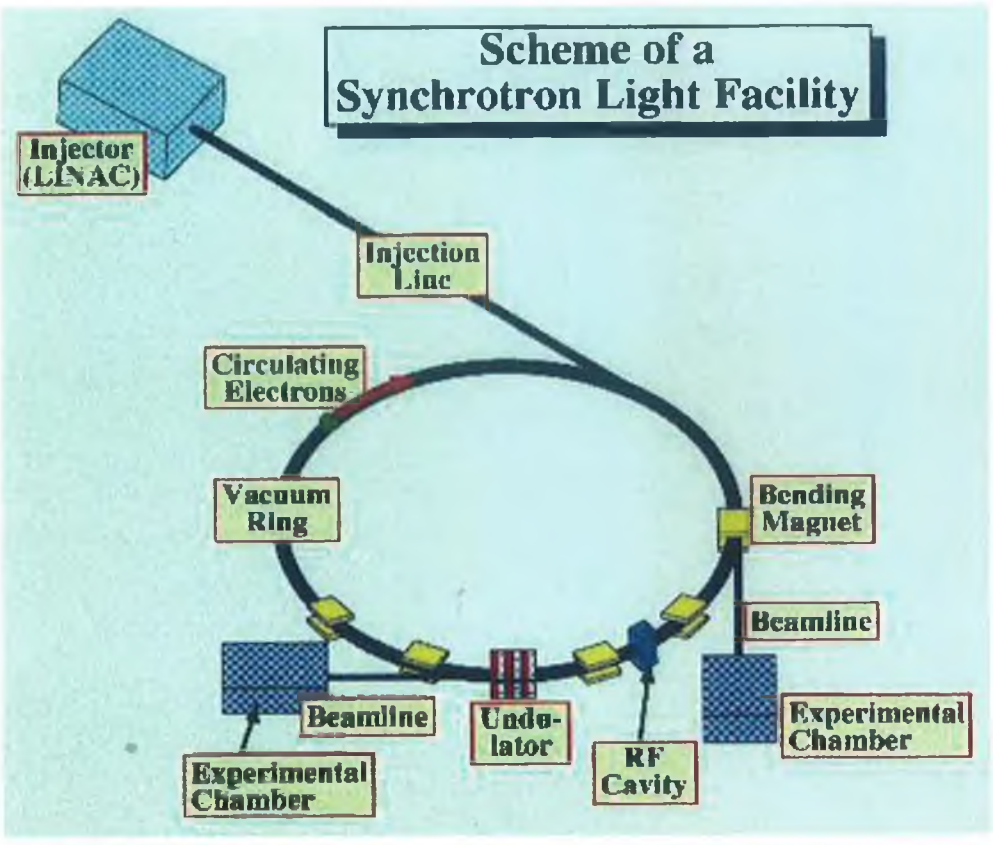


Figure 2.8 Schematic diagram of a synchrotron facility, showing how the light is produced

The main advantage of using synchrotron light is the broad spectral range of photon energies available which when coupled with a monochromator, enables very high resolution spectra to be acquired.

The I14A beamline at the NSLS is part of the VUV ring at the facility. The beamline structure is shown in figure 2.9. The light passes through a front end mirror, an entrance slit, a 3 toroidal grating monochromator, an exit slit and then into the analysis chamber where it is incident on the sample. The monochromator is used to select the photon energy. The 3 gratings have 288, 822, and 2400 lines/mm giving a range of photon energies. Typical scans were carried out with a step size of 0.05eV, a dwell time of 0.2s, in the photon energy range 20-180eV with an overall resolution <200meV on the Si 2p peak. This photon energy range is ideal for the study of the Si 2p core level, which has a binding energy of approximately 100eV, and also for the investigation of valence band features.

The vacuum in the chamber is maintained by a rotary-backed turbo-molecular pump and ion pumps. All samples were loaded simultaneously to avoid having to break vacuum, and to ensure identical experimental conditions for each sample.

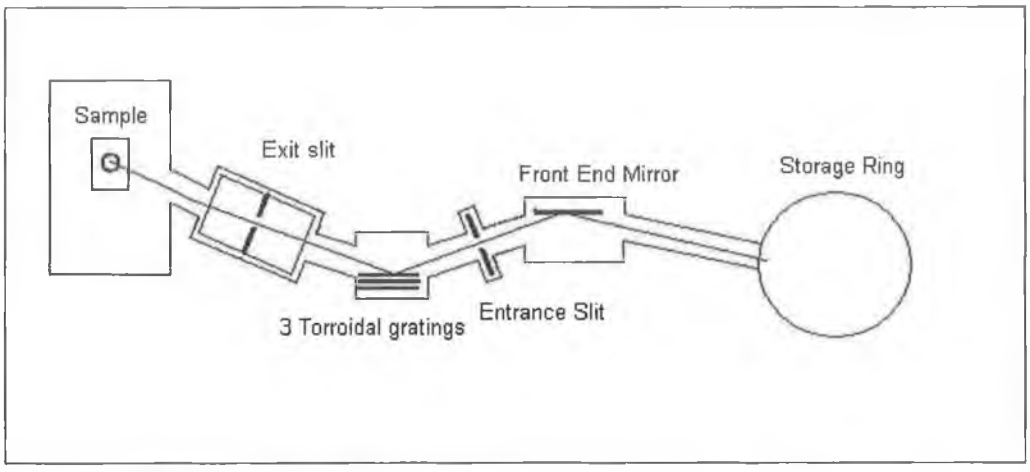


Figure 2.9 Schematic Diagram of the U4A beamline at the NSLS, showing the path followed by the light from the storage ring to the sample

Some considerations for synchrotron data analysis were:

Spin Orbit Stripping:

The finite resolution of the technique means that sometimes close-lying peaks cannot be resolved. In particular the $Si2p_{3/2}$ and $Si2p_{1/2}$ components are not well resolved. The two components in the Si2p peak arise from coupling between the magnetic fields of spin and angular momentum in the 'p' orbital. For this reason the silicon $2p_{1/2}$ component was removed from the peak before analysis was performed. The software used also removes the background[11]. An example of a Si2p peak before and after stripping is shown in figure 2.10. The peak is much narrower when the $Si2p_{1/2}$ component is removed.

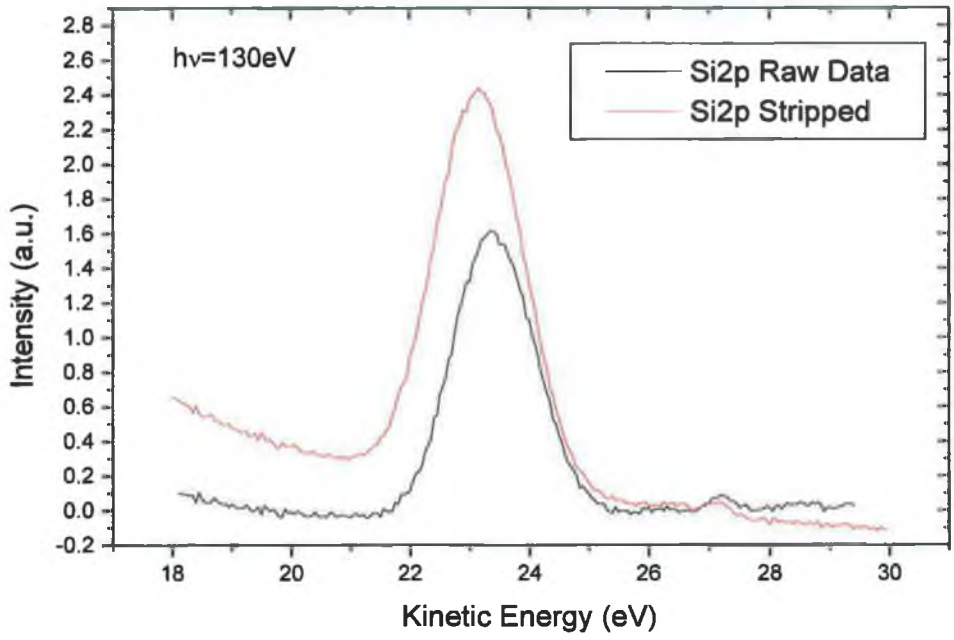


Figure 2.10 Si2p peak before an after spin orbit stripping. The true single peak width can only be measured after stripping.

Fitting:

Where necessary, peaks were fitted using a standard fitting program. Fitting peaks assists in determining the number of chemical environments present, and resolving the energy separation between them. Shown in figure 2.11 is a Hf 4f doublet peak at a photon energy of 100eV. In this case, the software calculates the best fit with an energy separation of 1.65eV between the peaks, and a branching (height) ratio of 0.73. The effect of varying preparation methods and fitting parameters was studied, to investigate peak broadening or shifting as a function of sample preparation procedures.

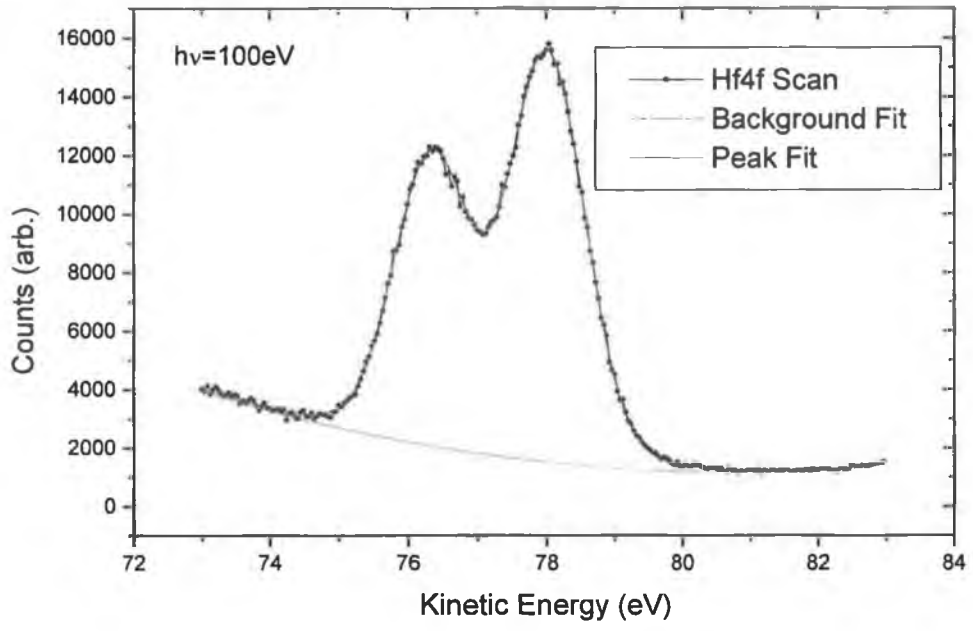


Figure 2.11 Hf4f peak fitted with a combined Gaussian-Lorentzian. The background is fitted with a polynomial.

2.5 X-ray Absorption Spectroscopy (XAS)

Conventional Soft XPS can be used to probe valence band electrons and thereby give information about the valence band offset. This represents the barrier height to tunnelling carriers in the valence band of the layer in question, as shown in figure 2.12. X-ray Absorption Spectroscopy (XAS) can in turn be used to calculate the conduction band offset, i.e. the barrier to tunnelling carriers in the conduction band. In XAS, monochromated synchrotron radiation light is used to scan through the absorption edge of a particular element causing the creation of core holes. The process results in the emission of x-rays due to de-exciting electrons filling the core holes. The intensity of this x-ray emission, which is measured in an x-ray spectrometer, is proportional to the partial density of states in the conduction band of that particular element. By comparing the XAS spectra for oxygen, for example, from SiO_2 and HfSiON, the conduction band offset can be measured.

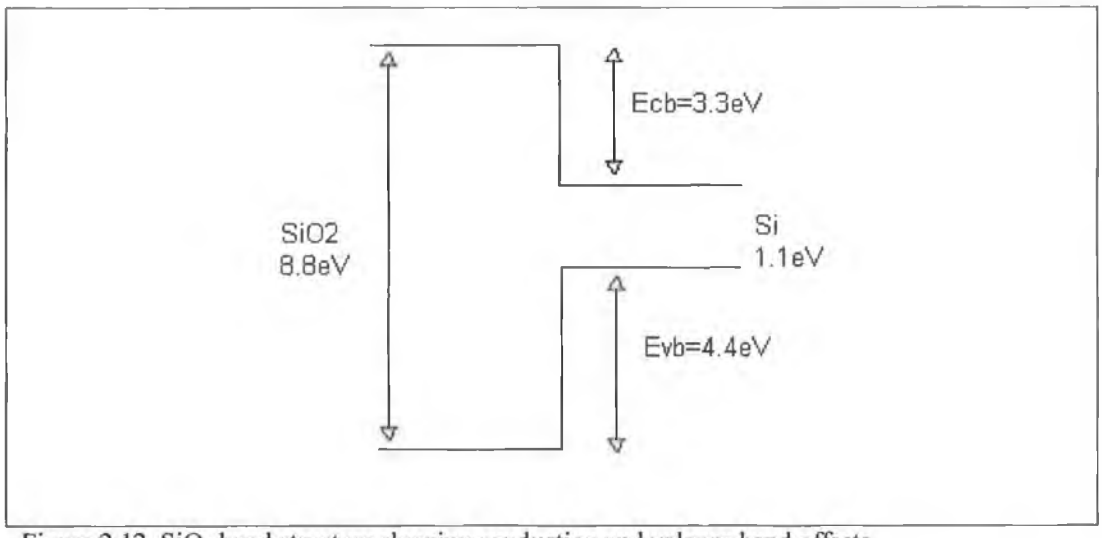


Figure 2.12. SiO_2 band structure showing conduction and valence band offsets

Previous results have shown that the incorporation of nitrogen alters the band structure of the Si/SiO_2 system[12]. In chapter 6, results are shown for HfSiON both with and without nitrogen, to examine whether similar effects are seen.

2.6 Secondary Ion Mass Spectroscopy (SIMS)

Secondary Ion Mass Spectroscopy (SIMS) is a valuable tool used to provide depth profile information about solid layers. During a SIMS measurement, the sample surface is slowly sputtered away with an ion beam. SIMS is thus a destructive technique. While the sputtering takes place, the fragments of the layer that are displaced are continually analysed in a mass spectrometer, so a depth profile can be constructed. The incident beam used in this study was a neutral Ar beam generated in a charge neutralisation chamber in order to avoid sample charging. The SIMS process is shown schematically in figure 2.13[13]. The displaced material consists of small concentrations of positive and negative ions of the elements present in the sample. These ions are analysed by a quadrupole mass spectrometer, which can be set up to sequentially accept ions of a certain mass/charge ratio. As the voltages on the rods of the spectrometer are varied, ions of varying charge to mass ratio are detected. An example mass/counts bar chart is shown in figure 2.14.

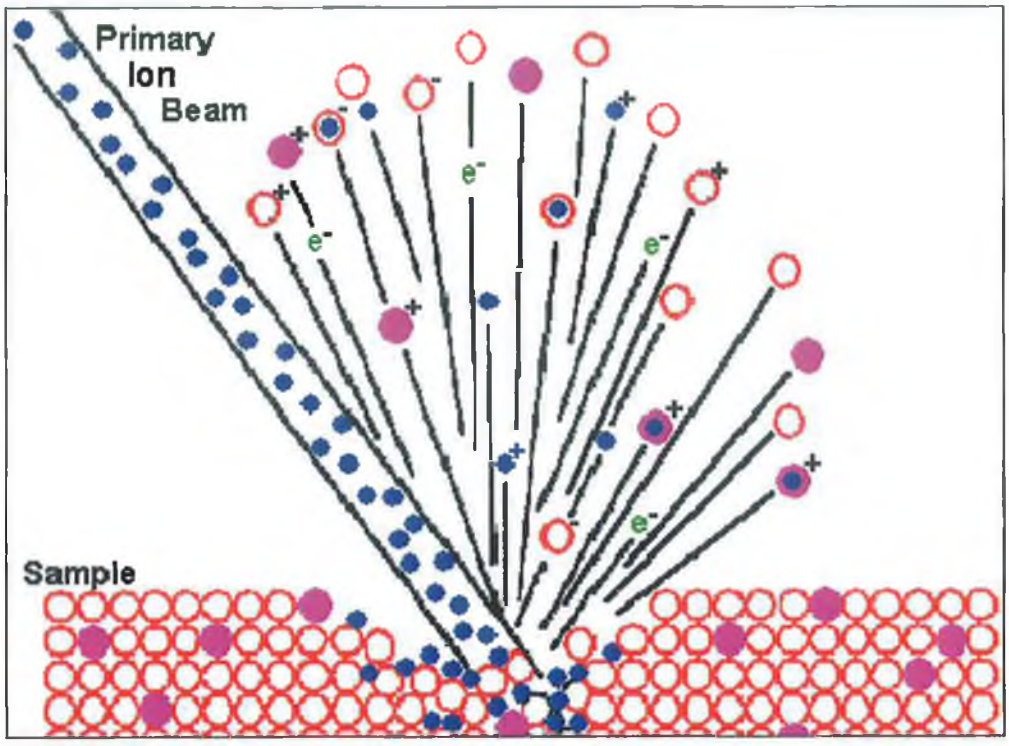


Figure 2.13 Schematic diagram of a SIMS experiment, showing how the primary beam interacts with the sold releasing secondary ions.

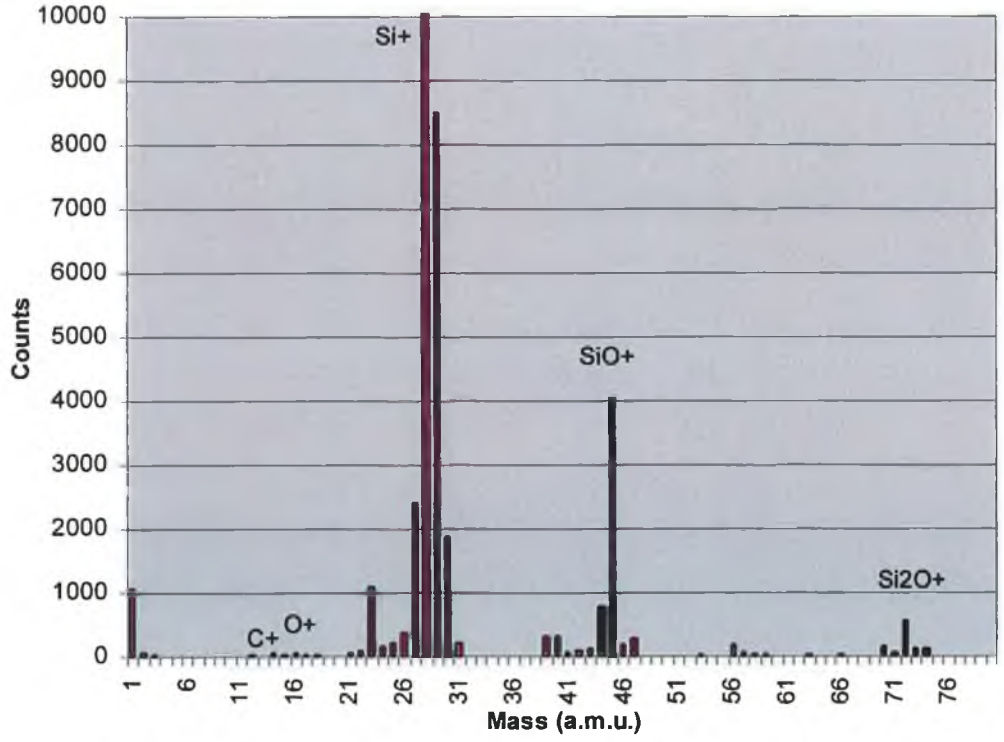


Figure 2.14 A typical SIMS spectrum; showing peaks associated with compounds of Si, N, C, and O atoms.

Silicon is seen at atomic mass 28 as expected. This peak represents Si^+, or singly ionised silicon atoms. There is also a very small contribution from Si at 14, which represents Si^{2+}, which overlaps with N^+. The overlap occurs because the technique cannot differentiate between two species with the same charge/mass ratio. The contribution of Si at 14 is quite small in comparison with the N, and so is ignored. There is also a peak attributed to Si at 56 a.m.u. due to Si_2^+ i.e. two silicon atoms which are bonded together and pass through the analyser as a single mass fragment.

For this work, oxygen, silicon, nitrogen, and carbon (and hafnium for the HfSiON layers) were measured as a function of sample depth, by setting the ion beam at a very low sputter rate, and allowing the measurement to continue for several hours, as the layer is etched away. SIMS results are compared with depth profile information from ARXPS data.

2.7 Summary

The techniques outlined in this chapter are powerful tools in obtaining vital chemical information about dielectric layers. XPS and ARXPS yield quantitative elemental concentration information and the angular aspect makes depth profiling of the layers possible in a non-destructive manner. This information can be cross-compared with SIMS results to give confidence in results. In this way, the effect of varying sample preparation on the final layer can be studied. The synchrotron based techniques can give information which can be related to the electronic properties of the layers, with conduction band offset being determined from X-ray absorption measurements, and valence band offset being measured from Soft-XPS. The effect on preparation conditions on these parameters can also be studied.

2.8 References

[1]Z. Xu, B.Kaczer, J. Johnson, D. Wouters, G. Groeseneken, Charge trapping in metal-ferroelectric-insulator-semiconductor structure with $SrBi_2Ta_2O_9/Al_2O_3$ /SiO_2 stack, J. Appl. Phys. 96 (3) (2004) pp. 1614-1619

[2]http://www.zaiko.kyushu-u.ac.jp/~watanabe/
SemiconSS/SCElectronic/PhotoEmission1.jpg

[3] M.Cardona, L.Ley (Eds.). Photoemission in Solids, Pats 1 and 2, Springer, Berlin 1978.

[4]M.P. Seah, W.A. Dench, Quantitative electron spectroscopy of surfaces: a standard data base for electron inelastic mean free paths in solids. Surface Interface Analysis, 1, (1979) pp. 2–11.

[5] J. Schofield, Quantitative XPS analysis of the chemical composition of S1(Ag-O-Cs) photocathodes and its relation to the photosensitivity, J. Electron Spectrosc. Relat. Phenom. 8 (1976) pp. 129

[6] K. Persy and N. Gurker, XPS—Investigation of contamination layers, Journal of Electron Spectroscopy and Related Phenomena, 13, (2) (1978) pp. 91-95.

[7] L. Holland, Sources of surface contamination in vacuum evaporation systems, Vacuum, 13, (5) (1963), pp. 173-180.

[8] E. Paparazzo et al, Evidence of SiO_x suboxides at Ar ion etched silica surfaces, J. Vac. Sci. Technol. A, 8 (1990) pp. 2231-2235

[9] QUASES-ARXPS, Version 1.1, S. Tougaard and T.S. Lassen, Denmark.

[10] Trieste Synchrotron Website: http://www.elettra.trieste.it/visitors.

[11] J. Roche, PhD Thesis, Page 71, Dublin City University (2002)

[12] C. McGuinness, G. Hughes, J. Roche, D. Fu, J.E. Downes and K.E. Smith, Electronic structure of thin film silicon oxynitrides measured using soft x-ray emission and absorption, Journal of Applied Physics 94 (2003) pp. 3919-3922

[13] Charles Evans and Associates Website: http://www.cea.com/cai/simstheo/ionsput.htm

Chapter 3: Electrical Characterisation Methods for Gate Dielectrics

3.1 I_GV_G Tunnel Current Measurements

For the transistor gate dielectric thicknesses investigated in this thesis, the application of a gate voltage lower than the barrier height encountered by tunnelling electrons results in a small current flow. This is due to quantum mechanical electron tunnelling through the insulator potential barrier. A current-voltage (I-V) measurement can be used to determine the breakdown field of the layers under investigation as shown in figure 3.1. For these ultra-thin SiON samples, the breakdown voltage is approximately 24MV/cm. As tunnel current is exponentially dependent on barrier thickness, a reduction in thickness of just 0.2nm increases the current by a factor of 5.

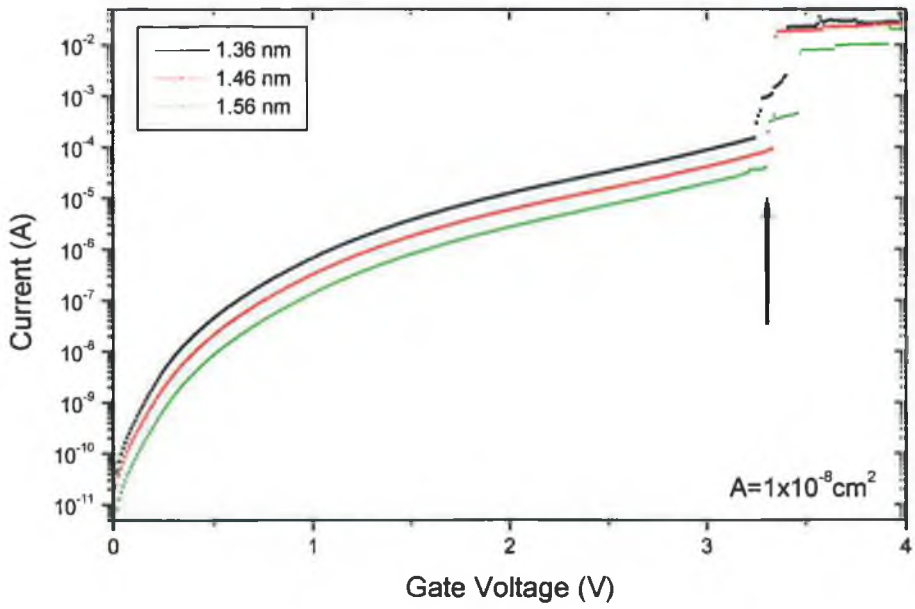


Figure 3.1 Tunnel current characteristics for 1.36-1.56nm SiON samples. Breakdown voltage is approx 3.3V for these samples.

The significant increase in leakage current by reducing the gate oxide thickness is shown in figure 3.2, where the leakage current density is shown for a range of thicknesses.

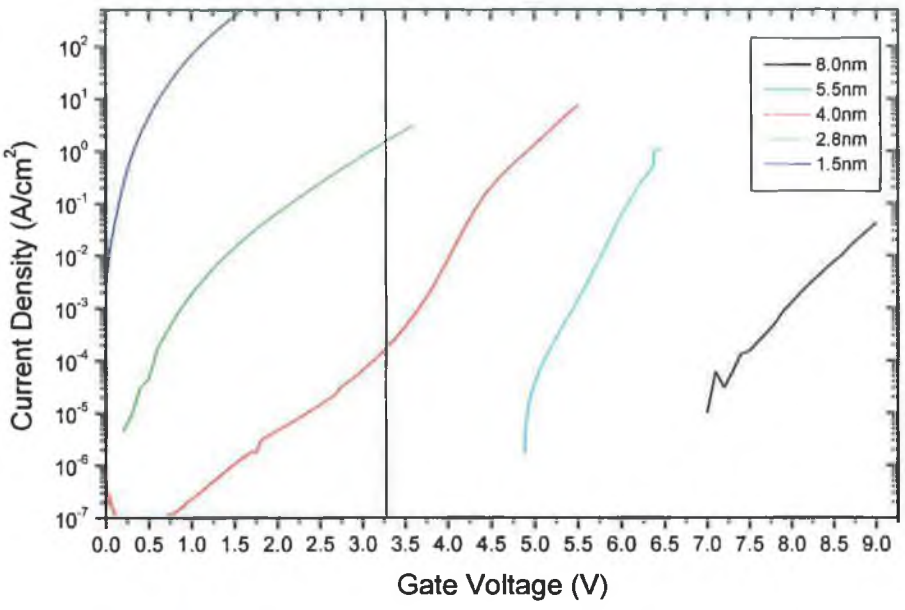


Figure 3.2. Tunnel current characteristics for various oxide thicknesses illustrate the exponential rise in current for decreasing thickness. The transition from Fowler-Nordheim to direct tunnelling can also be seen

There are two possible current transport mechanisms, as discussed in Chapter 1, depending on the shape of the barrier encountered by tunnelling electrons. Fowler-Nordheim tunnelling takes place where the applied voltage across the oxide is greater than the barrier height (fig 3.3a), and the electrons tunnel through a triangular barrier into the SiO_2 conduction band prior to entering the silicon conduction band. However, when the oxide voltage is less than the barrier height, the electrons encounter a trapezoidal barrier (fig 3.3b), and tunnel directly into the conduction band of the Si. Depending on the substrate Si orientation, doping and other factors, the barrier height is generally accepted to be in the range 2.5eV-3.3eV for SiO_2[1] and nitrided SiO_2[2]. Therefore, depending on the oxide thickness the predominant tunnelling mechanism can vary in that in thick layers, the probability of direct tunnelling is greatly reduced. This is evident from figure

3.2, where the change in gradient can be clearly seen for the 4.0nm oxide. Direct tunnelling in the thinner layers causes a large increase in the current density and in some transport models is described using the Fowler-Nordheim equation and some multiplicative factor[3].

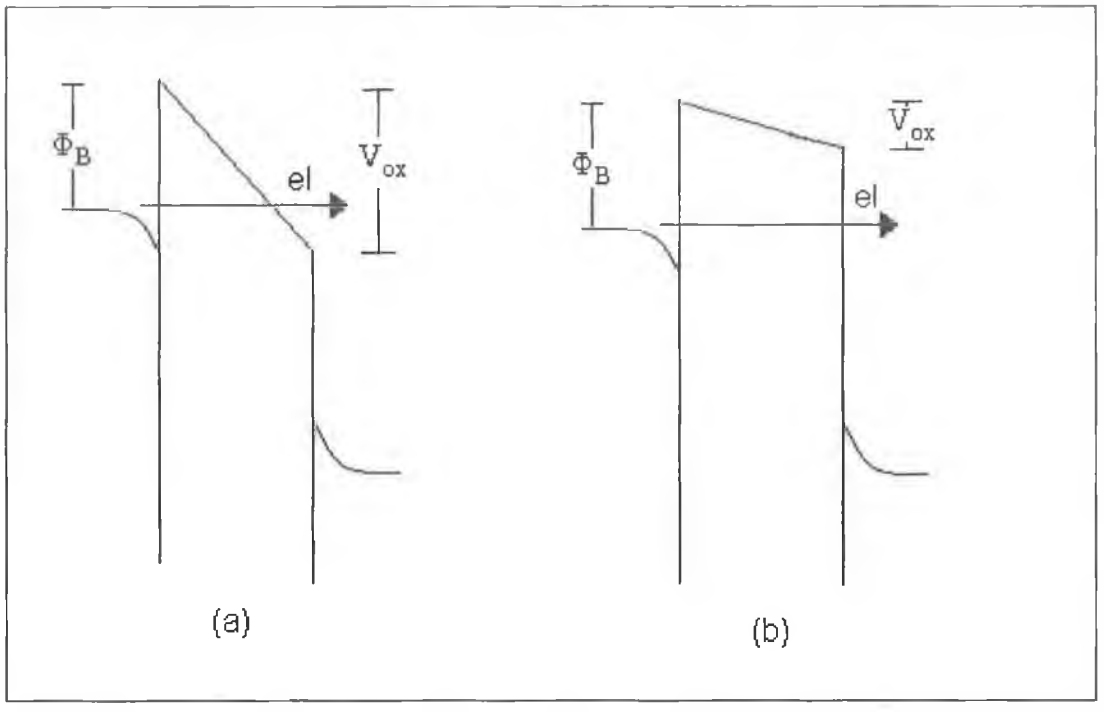


Figure 3.3 (a) Fowler-Nordheim tunnelling through a triangular barrier, into the oxide conduction band. (b) Direct tunnelling through a trapezoidal barrier.

Though ultrathin oxides generally operate in the direct tunnelling regime, and Fowler-Nordheim tunnelling is no longer an issue, as high-k materials are incorporated with physical thicknesses of greater 4nm, Fowler-Nordheim tunnelling could again be a factor in I-V measurements and can yield some interesting information about the layers.

If the Fowler-Nordheim current profile is examined and $\ln(J/E_{ox}^2)$ vs. 1/F (fig 3.4a) is plotted, where F is the oxide field and J is the current density, a straight line plot (from the FN-tunnelling equation) is expected. However in 3-6nm oxides, a slight oscillatory component on the straight line (fig 3.4b) is present, which arises from the quantum reflection of electrons with the correct wavelength at the SiO_2 interface[4].

Information about the SiO_x transition region width at the interface can be gained from the shape and width of the oscillations[5]. This is because the SiO_x region

reduces the distance through which the electrons have to tunnel, as there are available states in this region, and the shape and position of the oscillations reflect this. Using the interference model proposed by Mao et al[4], we have observed similar differences in applied voltage and observed voltage at the current oscillation extrema, corresponding to a transition width of ~.35nm, approximately 10% of the total for the 4nm oxides used.

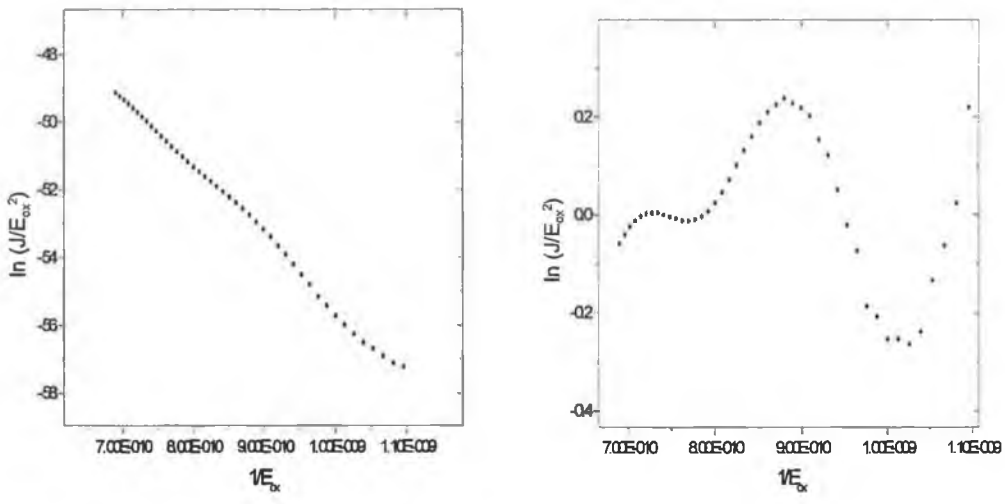


Figure 3.4 A Fowler Nordheim plot, with the tunnel current oscillations extracted from the straight line fit. The oscillations are extracted by subtracting the linear fit from the FN-equation, from the actual experimental data

Direct tunnelling proves more difficult to describe accurately as shown in figure 3.5. Experimental data for a 2.8nm oxide is shown alongside the model for direct tunnelling outlined in chapter 1. Tweaking of parameters, such as barrier height (3.5V in the example), effective mass ($0.45m^*_0$ in the example), and oxide thickness is required to acquire the close fit in figure 3.5 and even still, like many approximations[6], it has significant shortcomings in the low-field range. The model also begins to deviate from experimental data at the onset of Fowler-Nordheim tunnelling at around 3.3V as expected.

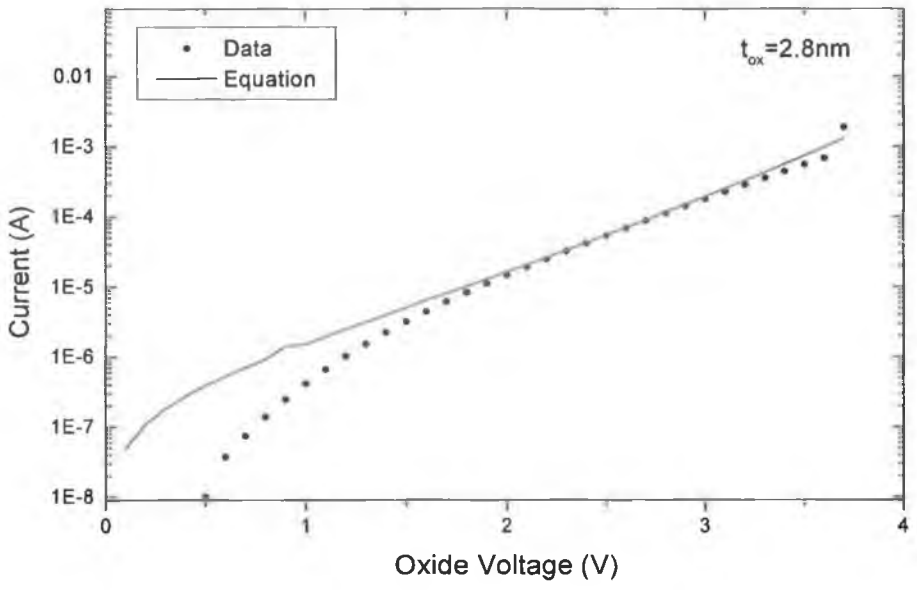


Figure 3.5 Experimental direct tunnelling data fitted with a model based on the WKB approximation for a capacitor with active area, n+ gate, p-type substrate. The model is quite inaccurate in the low field range.

3.2 Time Dependent Dielectric Breakdown

The primary function of the gate oxide is to act as an insulating layer for the lifetime of the device in which it is incorporated. However, when an oxide is stressed using either constant voltage, or constant current stress at high field, it is found to gradually lose its insulating properties over time, until eventually the layer suffers electrical breakdown. The main focus of the electrical measurement based part of this thesis is on the reliability of the layers under consideration, the degradation of the layers during their lifetime, and the operation of MOSFETs as oxide parameters degrade. This section deals with the characterisation methods used in the study.

Dielectric breakdown has long been attributed to the build-up of some sort of defect (now assumed to be electron traps) over time where, upon reaching a local critical defect density, oxide breakdown occurs. Early models consisted of dividing the oxide into cells, and as the electrons traps built up at random positions, they filled up these cells[7]. Breakdown was triggered when any one cell had exceeded the critical number of traps, which form a conductive path from

anode to cathode. However, problems with this method included the fact that the model could trigger breakdown once the critical number of traps was reached in a cell, even if the traps didn't reach from anode to cathode, but were condensed in the middle of the cell. Also, a local critical density could be reached but the traps may be distributed in neighbouring cells, and breakdown would not be triggered.

An advancement of this model using percolation theory was proposed by Degraeve et al[8], whereby the traps form at random in the oxide, but they can only form a conductive chain with other traps if they are within a certain percolation distance of each other. When a conductive chain occurs from anode to cathode, breakdown occurs. The process is illustrated schematically in figure 3.6

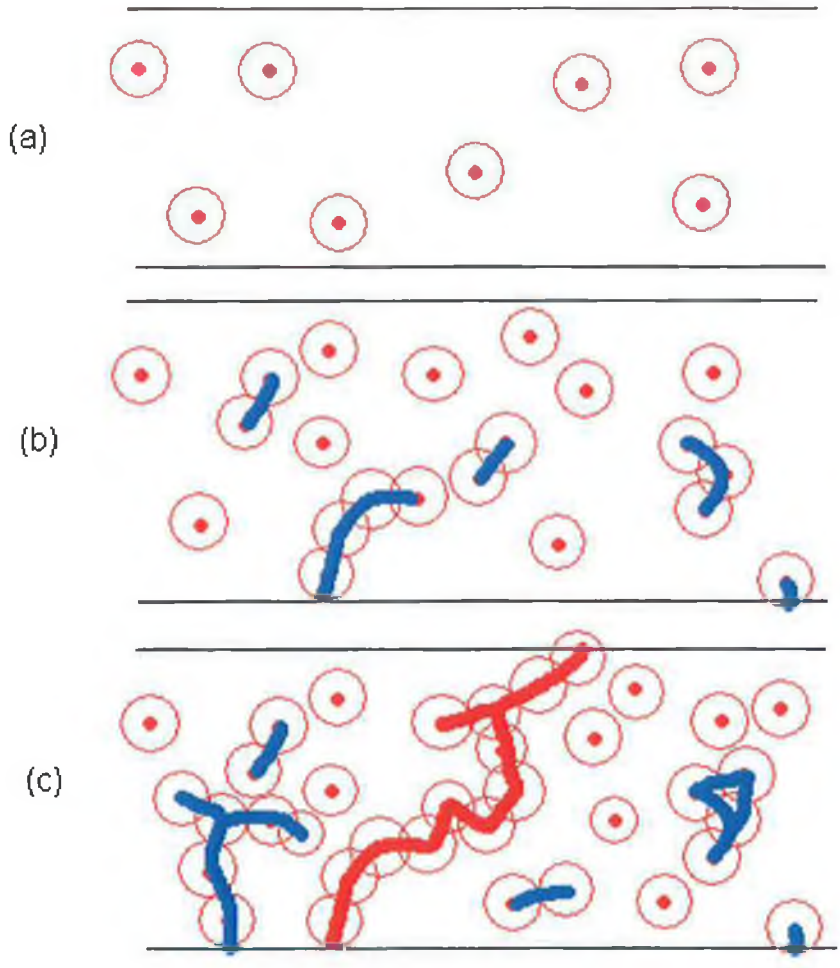


Figure 3.6 Buildup of electron traps (a) leading to the formation of localized conduction paths within the oxide (b), and ultimately the formation of a conduction path from anode to cathode causing oxide breakdown (c).

From this figure, a number of effects of the random nature of trap generation on breakdown are clear. Firstly, because of the random location of trap generation, the time-to-breakdown will be statistically distributed; it becomes clear as well that the trap generation rate is a critical factor in determining time to breakdown[9]. Also, the effect of decreasing dielectric thickness on the time-to-breakdown parameter (t_{bd}) becomes obvious. As fewer traps are required to form a breakdown path, the overall density at which breakdown occurs will generally be lower, and breakdown will occur much quicker for the same trap generation rate[10]. The next section outlines the statistics involved in describing oxide breakdown.

3.2.1 Weibull Statistics

If a number of identical oxide samples are stressed under the same conditions, a wide variety of t_{bd} values will occur. The reason for this, as mentioned, is the buildup of traps at random locations, leading to the possibility of a conduction path forming at both low and high trap densities, depending on the particular trap location profile for each oxide. The distribution is described by Weibull statistics [9]. This distribution is used a lot in engineering applications where 'weakest link' processes occur. The function is:

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \qquad \text{Eq 3.1}$$

where, β is the Weibull Slope of the distribution (shape factor) and η is the 63% (scale factor) which is the time at which 63% of a certain sample set will have broken, independent of the number of samples tested.
Furthermore, by taking the natural log of both sides it can be shown that,

$$\ln\left[-\ln\left(1 - F(t)\right)\right] = \beta.\ln(t) - \beta.\ln(\eta) \qquad \text{Eq 3.2}$$

Therefore, a plot of **ln[-ln(1-F(t))] vs. ln(t)** gives a straight line with a slope β. This is the usual way in which a set of breakdown data is presented, and a sample Weibull plot is shown below in fig 3.7, for a 1.5nm oxide. It has a Weibull slope of 2.2. The slope is calculated from a "maximum likelihood fit", which ensures a more accurate extraction than a least squares fit. The reason for this is that the "maximum likelihood" method gives more weight to time intervals in which breakdown occurs more often i.e. from 80-300 seconds in figure 3.7. The slope is then better fitted to the times at which breakdown has occurred more often. In thicker oxides, the distribution is made of 2 parts, an intrinsic breakdown slope and a flatter extrinsic one, at the low end of the t_{bd} scale, attributed to process induced defects and oxides which are already broken prior to stress[11]. In today's low defect density, small area samples, the two modes are no longer distinguishable. This work focuses on intrinsic breakdown, which is due to electron flux and trap generation rather than processes related effects.

The Weibull slope and 63% values are key factors in investigating how the oxide layers will behave at lower voltages. A low Weibull slope is equivalent to a very wide distribution of breakdown times, whereas a higher slope means all the breakdowns occur in a narrower time window. The significance of the Weibull slope in reliability projections will be made clear in the next section.

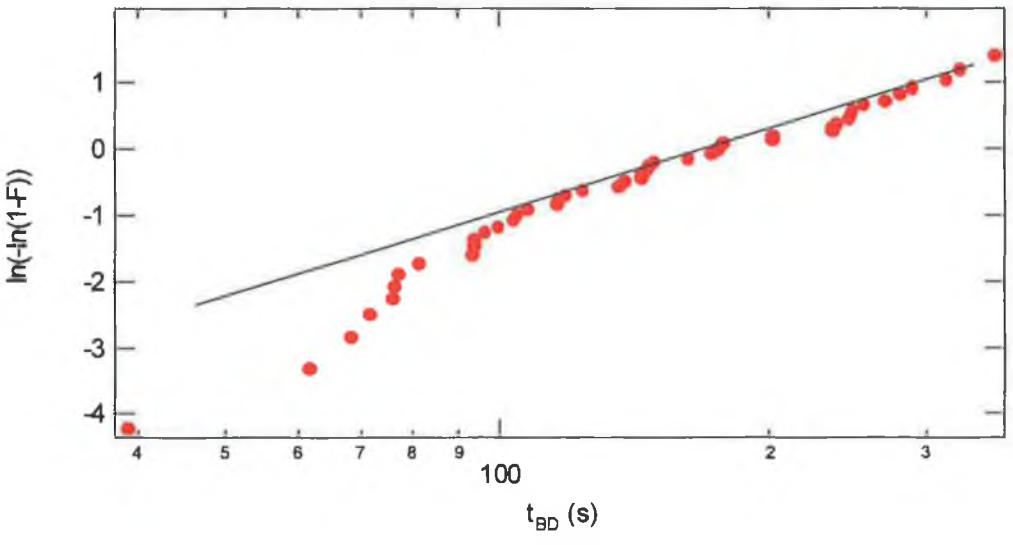


Figure 3.7 Sample Weibull plot for a 1.5nm oxide with an area of $1x10^{-8}cm^2$. Weibull plots are fitted with the maximum likelihood method to extract the Weibull slope.

As mentioned, the trap density at breakdown becomes lower as oxide thickness is reduced. If we look at t_{bd} statistics for decreasing thickness, we see that the Weibull slope also reduces for reducing thickness[8]. The reason for this is that as thickness is reduced, the number of traps required to form a breakdown path is reduced. Thus there is a wide statistical variation in the time it will take to form a path. In a thicker oxide, where 6-7 traps are required to form a path, the statistical spread is far less and the distribution is much steeper.

Figure 3.8 shows this relationship for our data, and for that of Degraeve et al[8], and the effect as predicted by the percolation model. The reason for the high Weibull slope observed at 1.5nm is discussed in chapter 4.

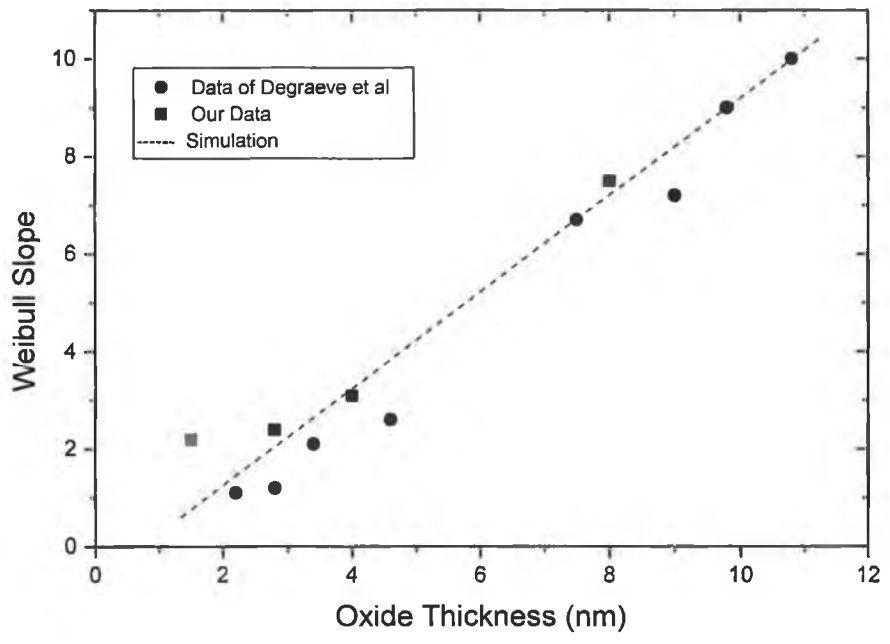


Figure 3.8 Variation of Weibull slope with oxide thickness for oxide from 1.1 to 11nm. The percolation model provides a good fit to the data

The area of the test structure under consideration is also important in determining the time to breakdown. The distributions of devices of different areas are related by[12]:

$$\frac{\eta_1}{\eta_2} = \left(\frac{A_2}{A_1}\right)^{1/\beta} \qquad \text{Eq 3.3}$$

where η is the 63% breakdown time, A is the device area, and β is the Weibull slope. So, as area is decreased, the time to breakdown increases as expected. Thus for a thin oxide where β is low, e.g. 2.2, a reduction in area of 1 order of magnitude results in a t_{bd}(63%) almost 3 times higher.

3.2.2 Reliability

When fabricating an oxide layer for use in a device, it is generally expected that the layer will maintain its insulating properties and perform as expected for the lifetime of the device which is normally taken as 10 years. However, as the t_{bd} is statistically distributed, and many measurements are needed to accurately observe the distribution, it is impractical to stress oxides at operating voltage because the modal t_{bd} will be ~10 years, and in some individual oxides, 100's of years! The solution is to use accelerated lifetime tests, whereby an elevated stress voltage is used such that the modal t_{bd} occurs within a measurable time frame, generally less than 10000s. Measurements are made at multiple stress voltages, and it is then possible to make an extrapolation to the operational voltage and examine reliability. A data set for such a reliability examination is shown in figure 3.8b

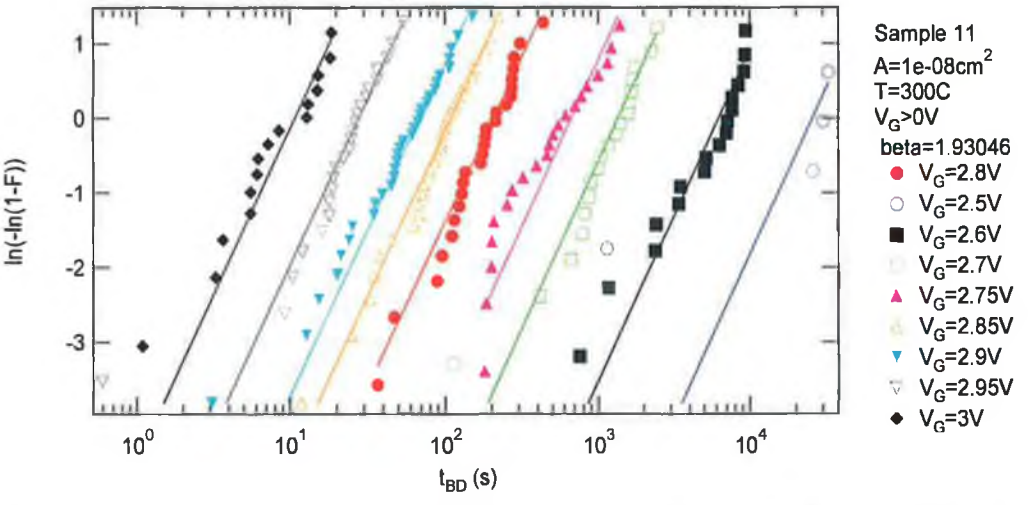


Figure 3.8b Weibull plots at many stress voltages used to make a reliability extrapolation to the low field range. Weibull slope is seen to be independent of stress voltage. Devices: Capacitors with active area, p+ gate and active, n-type substrate. t_{ox}=1.11nm EOT. Breakdown criteria=large current jump

In this case stress voltages from 2.5V-3V are used for the stress. Obviously, using many measurements means better accuracy in the Weibull slope, and more points on the extrapolation line to the low field range. In this case we have used ~200 samples and so the error in the Weibull slope $(\frac{1}{\sqrt{n}}) = 0.07$, so the Weibull slope is 1.93 ± 0.13. To make the reliability estimate, we take the $t_{bd}(63\%)$ values, corresponding to y=0 on the Weibull plot, and plot $t_{bd}(63\%)$ vs. gate voltage as shown in figure 3.9, and extrapolate back to operating voltage.

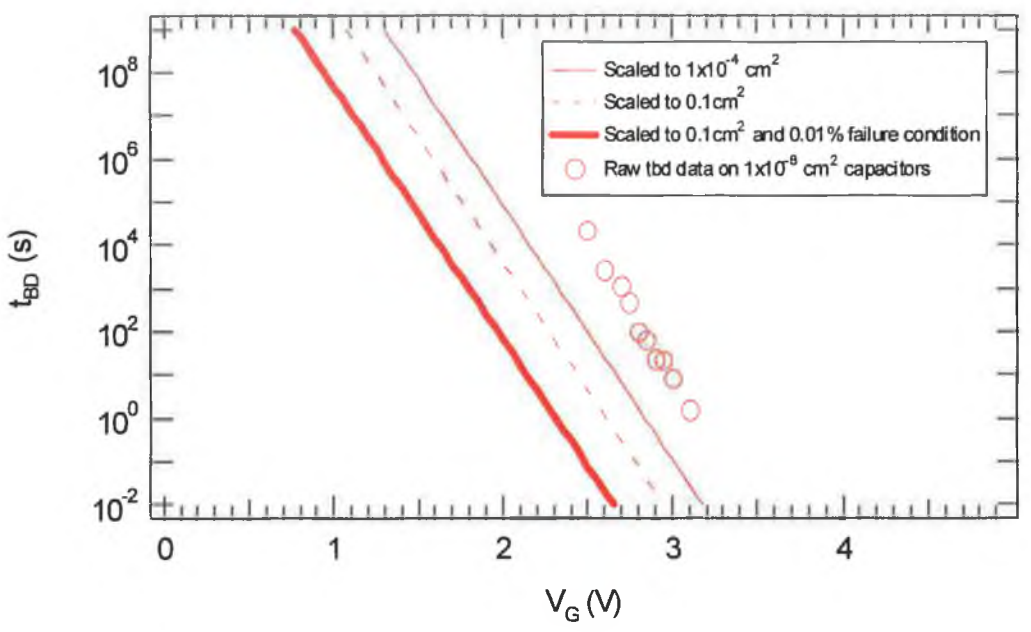


Figure 3.9 Reliability extrapolation using $t_{bd}63\%$ data from figure 3.8 including area and percentile scaling. Maximum operating voltage for any given lifetime requirements can be deduced from the plot.

The raw data (open circles) acquired on devices of area $1x10^{-8}cm^2$ is first scaled to an area of $1x10^{-4}cm^2$ using the area scaling equation outlined earlier in this chapter. This allows data collected at a number of areas to be normalised to one reference area, however in this example only one area is used. The next step in making a reliability estimate is to scale to a typical chip area ($0.1cm^2$) making the assumption that individual oxide breakdowns will stop the chip functioning. Since the area of a chip is so much larger than the area of individual gate dielectrics, the reliability estimate is essentially a worst-case scenario. Furthermore, our data points represent the $t_{bd}(63\%)$ data, which is obviously an unacceptably high failure rate, and thus the data is further scaled to a 0.01%

failure rate condition, giving the final reliability extrapolation (bold line) and allowing a maximum operating voltage for a required lifetime to be deduced. From the above diagram, an operating voltage of 0.8V would give a projected lifetime well in excess of $3x10^8$s (10 years)

The Weibull slope affects the area scaling as mentioned previously; a high slope equates to t_{bd} distributions at different areas being closer together, and so the extrapolation lines (scaled to $0.1cm^2$ and $1x10^{-4}cm^2$) are drawn closer to the raw data (acquired at $1x10^{-8}cm^2$). This obviously improves the maximum operating voltage for a desired lifetime. The Weibull slope also affects the percentile scaling (bold line). The reason for this is illustrated in figure 3.10. A distribution with a high β has a $t_{bd}(63\%)$ that is closer to the low percentile values than a distribution with a high $t_{bd}(63\%)$.

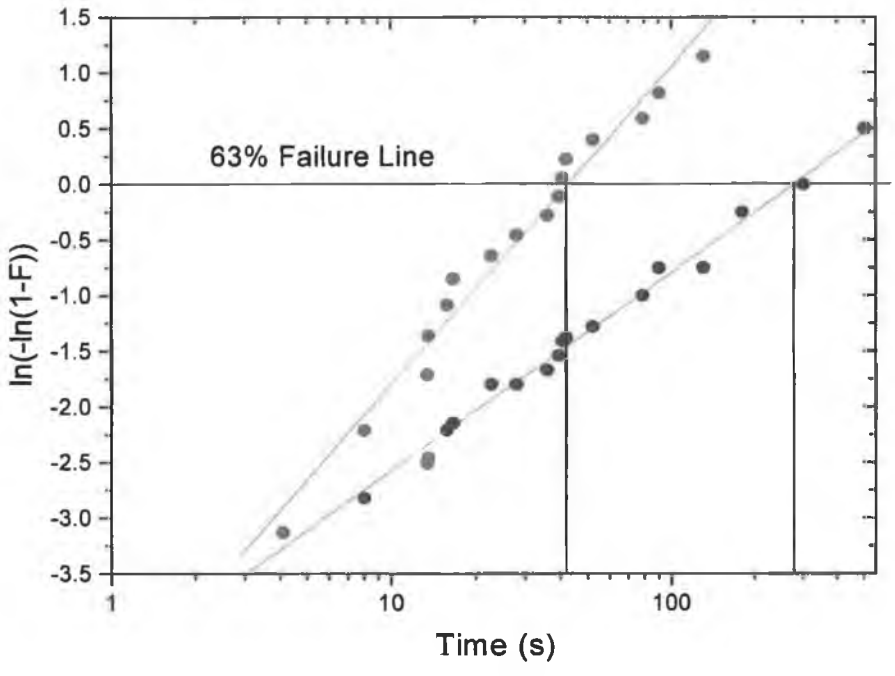


Figure 3.10 The effect of Weibull slope on percentile scaling. A low slope stems from a wide statistical spread and means a greater gap between the 63% time and the low percentile breakdowns.

For example in figure 3.10, low percentile values are found at ~5 seconds. The $t_{bd}(63\%)$ for the distribution with the high slope is ~40s and for the low slope ~280s. So much less scaling is required in going from $t_{bd}(63\%)$ to low $t_{bd}(0.01\%)$ in the sample with the high slope. In figure 3.9, this would mean the bold line would be drawn closer to the raw data, thus improving the reliability outlook. It is clear then that for area and percentile scaling, a high β is desirable to improve the reliability prediction. The inherent properties of the oxide, which determine β, will be discussed in detail later in the thesis.

On examining figure 3.9, the slope of the extrapolation is obviously another important factor in determining reliability. The slope is known as the voltage acceleration[13] (γ) and is again an inherent property of the oxide. There are however a number of issues with making such an extrapolation. Firstly, there is nothing to suggest that the oxide suffers that same degradation at low field where it is impractical to measure, as it does in the measurement range. This however is difficult to confirm due to the time such measurements take, and the argument is based on theory rather than practice. The other more relevant issue is the extrapolation method used. In figure 3.9 we have used a linear extrapolation to the low field ($t_{BD}=t_0.\exp(-\gamma.V_G)$) to gauge reliability. However Wu et al[14], have suggested that the correct extrapolation obeys a power law, where the voltage acceleration (γ) is gate voltage dependent (i.e. $\gamma=(\delta(\ln(t_{bd}))/\delta V)$. There are a number of reports in the literature discussing the correct model and the physical origin of the field acceleration model obeyed by the oxide[15,16]. These will be discussed in the next section.

The difference between the two models in determining t_{bd} at low field is shown in figure 3.11. When extrapolated using the linear extrapolation model, a t_{bd} of $\sim 1\times10^{14}$s is expected. Using the power law model on the same data yields a t_{bd} of $>1\times10^{20}$s in this case. Again however, the timescale for making measurements to support either model is impracticable, and both are supported by theory rather than experiment. For this work, all reliability estimates are made using the linear extrapolation model, as it represents the worst case.

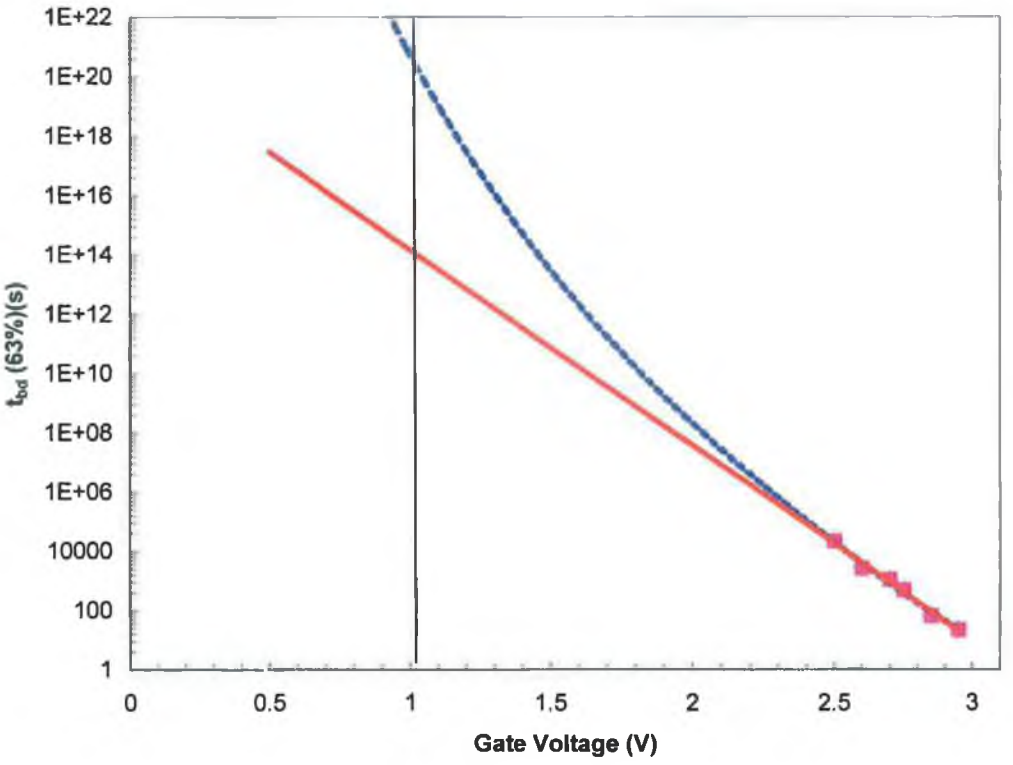


Figure 3.11 Using the power law model (blue) for reliability extrapolations leads to extremely favourable reliability outlooks when compared with the linear extrapolation model

3.2.3 Breakdown Mechanisms

It is known that the degradation and breakdown of gate oxides is due to charge traps forming within the layer during its lifetime. However, the trapping process; that is, the physical origin of the traps, is not well understood. This section outlines the types of charge traps that form in the oxide, and the most commonly referred to mechanisms in the literature as to how they are created.

1. ***Fixed Oxide Charge***[17]

During a constant voltage stress on an nMOSFET using substrate injection in thick SiO_2 layers (>4nm), the current is found to initially increase, indicating the presence of positive charge trapped in the oxide. After a time, the increase stops

and the current begins to decrease, as negative charge trapping dominates, though this effect is only slight for layers of this thickness. In ultra-thin SiON, a very small increase in the current is detected, again indicating the presence of positive trapped charge. No decrease is observed before breakdown for ultra-thin layers. These observations suggest that positive trapped charge resides at the interface, and because there is little bulk in ultrathin layers, the negative charge buildup, which may form in the bulk, is negligible.

Extending the discussion to high k materials, hafnium silicate shows only a decrease in gate current, for thicker layers, indicating that negative trapped charge in the bulk is dominant, if it follows the same trends as SiON.

In thinner HfSiON layers, the rapidly increasing leakage current during voltage stress (Stress induced leakage current[18] is discussed in the next section) makes it impossible to judge whether positive or negative charge trapping is dominant.

2. Interface Traps

Another widely studied mechanism is the presence of traps at the insulator/Si interface[19]. It is generally accepted that interface traps are created through the same mechanism as bulk traps, but due to the strained layer of SiO_x at the interface (which originates from the structural mismatch between Si and the insulator) where the density of localised states is increased, the trap generation rate is different from that in the bulk[20]. The interface trap densities are measured by charge pumping in this work, a technique that is outlined in section 3.4. The interface trap generation rates for both SiON and HfSiON are discussed in Chapters 4 & 5 respectively.

3. Neutral Electron Traps

During the stressing of an oxide, neutral electron traps form throughout the layer[21]. There is extensive information in the literature to suggest that it is these traps that build up, form a percolation path and result in the breakdown of the oxide[8]. However, because the traps are neutral, they require a filling step to make them electrically visible. Neutral trap densities cannot be measured for ultrathin

oxides, as the full traps have a tendency to empty before they can be measured. There are 3 models that suggest how these traps are formed:

The *anode hole injection*[22] model assumes that some electrons that arrive at the anode have sufficient energy to release an energetic hole back into the oxide. These energetic holes then create oxide traps as they tunnel. However, this model fails to explain the temperature dependence of oxide degradation.

The *hydrogen release*[23] model views the holes injected at the anode as not taking place in the trap generation process, but that the hole injection mechanism happens alongside the release of hydrogen (used to passivate the interface) into the oxide at the anode that generates traps. Thus, the measured hole fluence is simply a monitor of the trap generation process as both are dependent on the energy released at the anode by tunnelling electrons.

Finally, the *electric field* [21] model assumes that the energy released by tunnelling electrons at the anode is not responsible for trap generation, but that simply the electric field gives sufficient energy to the oxide to stimulate the process that leads to trap generation.

3.3 Stress Induced Leakage Current

It was noted in the 80's by Maserijian and Zamani[24], that the I-V curve of an oxide that had undergone stress differed from that of a fresh oxide, in that the stressed I-V displayed a higher current in the low field range, a so called stress induced leakage current (SILC). The effect is shown in figure 3.11, where a device is stressed for a set time, and then an I-V measurement is performed, and so on until breakdown. The effect is dominant at low field as at the high field, the I-V curves converge again. The predominant mechanism responsible for SILC is believed to be trap assisted tunnelling through neutral electron traps formed during stress [25]. Direct tunnelling at high field swamps this effect, and this is the reason why the I-V curves are similar at high field.

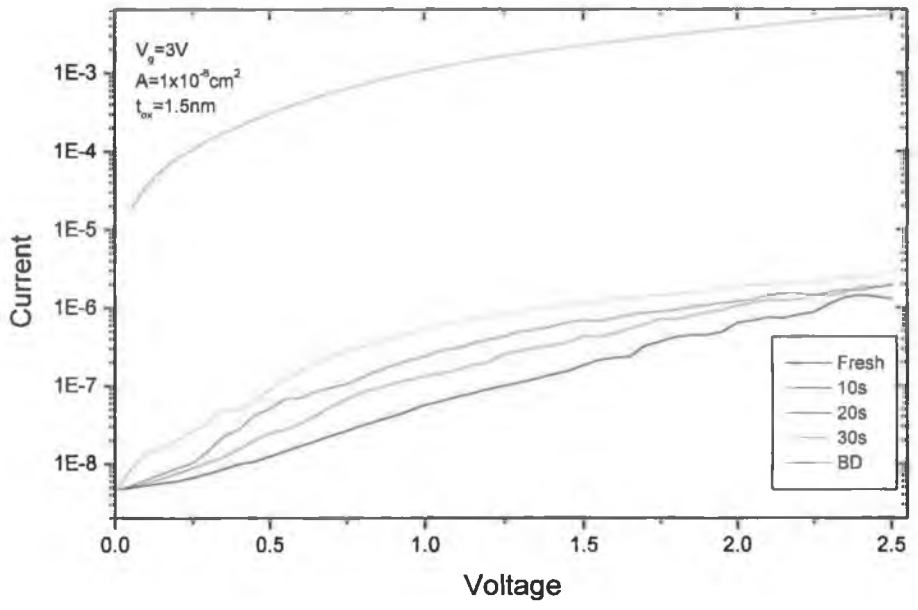


Figure 3.12 The buildup of SILC during a constant voltage stress. The increase in SILC can be directly correlated to the number of neutral electron traps in the layer.

A more conventional way to measure the SILC is to stress the layer at high field, then stop the stress and measure the SILC at low field, such that no further degradation is taking place during the sense measurement. The time interval between the stress and sense measurements is important, as SILC has a transient component associated with the discharging of traps near to the interface. This is less important in thinner oxides, where the steady state trap assisted tunnelling component dominates the current flow.

An example of a SILC measurement for an ultrathin layer is shown in figure 3.13 taken from[26]. The slight rise in SILC as the stress continues is attributed to positive charge trapping, and the formation of traps near to the interfaces that assist the tunnelling process. However, measurements at 2V and 2.25V show a sudden increase in SILC. This corresponds to the formation of a 'well placed' trap, in the middle of the oxide, the location of which is ideal for the trap assisted tunnelling process. The formation of such a trap in a thin oxide is often an indication that breakdown will soon follow.

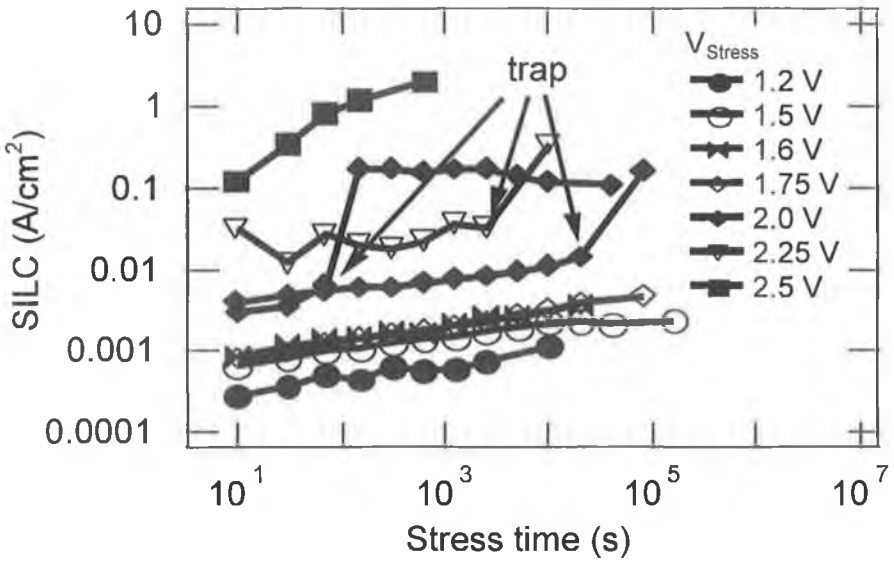


Figure 3.13 Increase in SILC with stress time at a number of stress voltage during CVS. The formation of a "well placed" trap causes a sharp rise in the SILC[26].

The level of SILC is a good measure of the neutral electron trap density in the oxide[27], and can be used for degradation and breakdown predictions. SILC is a problem in memory devices, as stored charge can leak out, causing information to be lost. Also, in MOSFETS, the SILC causes a higher consumption of power, which can be problematic for portable devices.

3.4 Detection of Interface Traps by Charge Pumping

Traps in the strained Si/dielectric interfacial region are of particular interest as it is these traps that are closest to the critical channel region of the device. For this reason, a reliable method of determining the interface state density is important in evaluating new dielectric layers. Charge pumping is employed as the primary method of interface state detection, as it is particularly effective for ultra-thin layers. This is because it remains accurate even when the gate leakage current is high, and its analysis is simpler than that of Deep Level Transient Spectroscopy (DLTS) and more reliable than methods based on 1/f noise analysis[28].

For this measurement, the source and drain are grounded, and the substrate current is measured, while a pulsed bias is applied to the gate. There are two methods by which the pulse is varied to execute the measurements as shown in

figure 3.14. Either the amplitude is fixed and the base level is swept so that the base level starts in accumulation and finishes in inversion, or the base level is kept constant and the amplitude is increased so that at maximum amplitude the device is pulsed from inversion to accumulation. The resulting charge pumping current profiles are shown in the bottom of figure 3.14

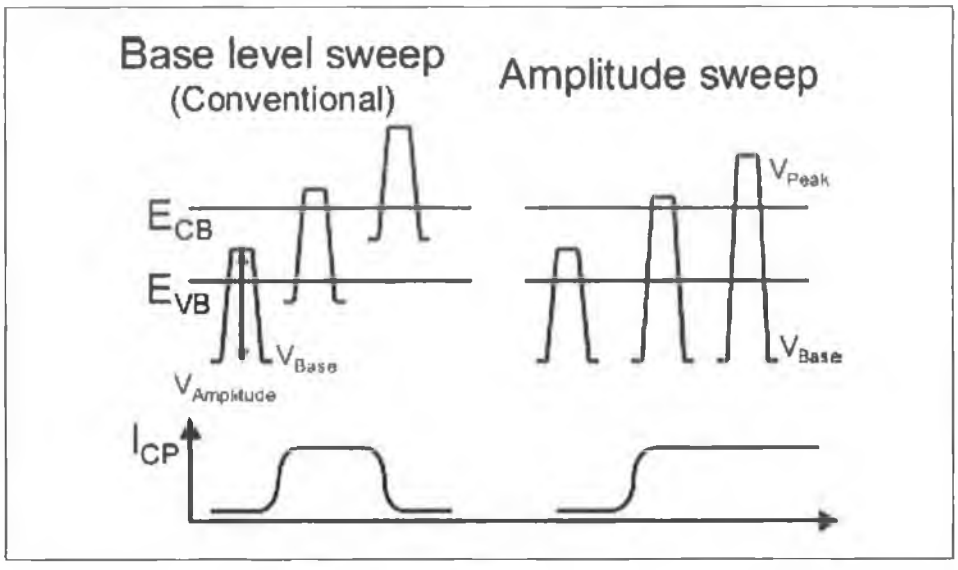


Figure 3.14 The base level sweep and amplitude sweep charge pumping methodologies. Both can be used to determine the interface trap density.

The origin of the charge pumping (CP) current is as follows. With the channel being swept from accumulation to inversion, electrons from the source and drain fill the existing interface traps. When the device is swept back to accumulation, electrons rush back into the source and drain, with the exception of those that are trapped. These electrons recombine with the majority holes, resulting in a recombination current being detected at the substrate. Clearly the magnitude of the recombination current depends on the number of interface traps.

The trap density is described by:

$$N_{it} = \frac{I_{cp}}{qAf} \qquad \text{Eq 3.4}$$

In equation 3.4, I_{cp} is the charge pumping current, q is the electron charge, A the device area, and f the pulse frequency. Most measurements in this work are high frequency (~1-3MHz). As only traps very close to the interface can respond to such a high frequency, we get a true measure of the interface trap density, without measuring the bulk trap density.

In ultrathin oxides, the charge pumping current is superimposed on the leakage current and sometimes evaluation of the charge pumping current proves difficult. However, this is easily rectified by making a low frequency measurement where there is no CP component, and subtracting it from the high frequency measurement. Charge Pumping is used in Chapters 4 and 5, in a stress and sense set-up, to characterise the interface quality by measuring the increase in the number of interface traps during electrical stressing.

3.5 Device Degradation

The constant degradation of oxide insulating properties during stress obviously has an effect on MOSFET parameters. Device performance is usually monitored using the I_DV_D and I_DV_g curves. The significance of these measurements is outlined here. A more detailed analysis is found in Chapter 4 for our SiON samples and Chapter 5 for HfSiON.

The I_DV_D curves measure the drain current while sweeping the drain voltage from 0V to operating drain voltage. This is carried out a number of gate voltages, and from these curves we can extract the saturation drain current, which is the drain current in operating conditions. A simulated[29] set of I_dV_d curves for an nFET is shown in figure 3.15 for a channel length L= 90nm and a channel width W=0.1μm device, with t_{ox}=1.5nm.

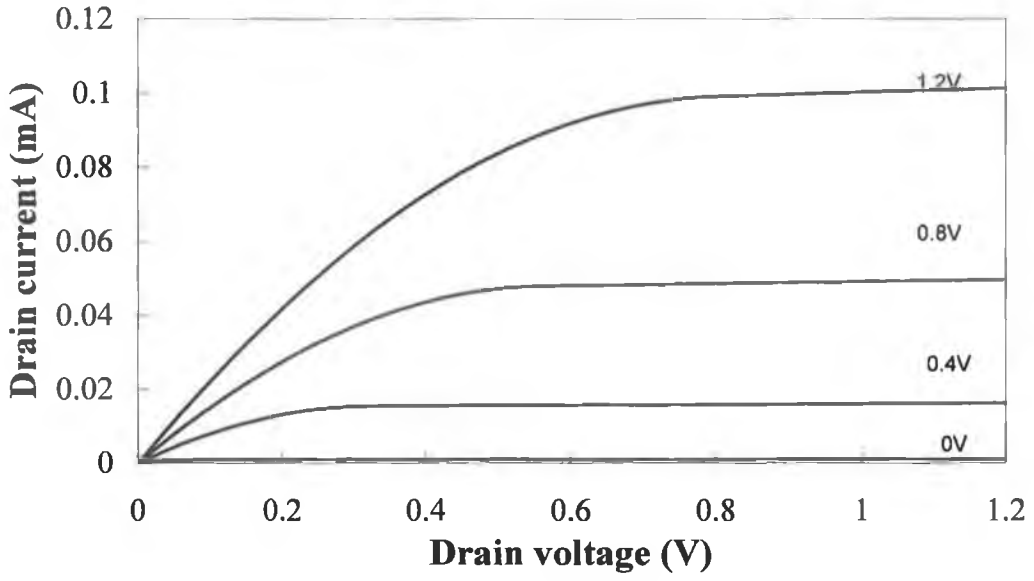


Figure 3.15 Simulated I_dV_d curves for a MOSFET at a number of gate voltages.

The linear and saturation regions are clearly visible. The saturation region is due to the channel being pinched off. The simulation yields an on-current (V_g=1.2V and V_D =1.2V) of 101μA and an off-current (V_g=0V and V_D =1.2V) of 1μA.

Show in figure 3.16, is a simulated I_dV_g plot for the same device, which is acquired by measuring the drain current while sweeping the gate voltage. In this case, the threshold gate voltage is 0.4V. From the I_dV_g the transconductance profile of the channel can be extracted, which is a good indication of channel mobility.

Transconductance (g_m) is the rate of change of drain current with respect to gate voltage for constant drain voltage. It is described by equation 3.5.

$$g_m = \frac{W}{L}\mu C_i V_d \qquad \text{Eq 3.5}$$

where W and L are the width and length of the channel respectively, μ is the channel mobility, C_i is the oxide capacitance and V_D is the drain voltage.

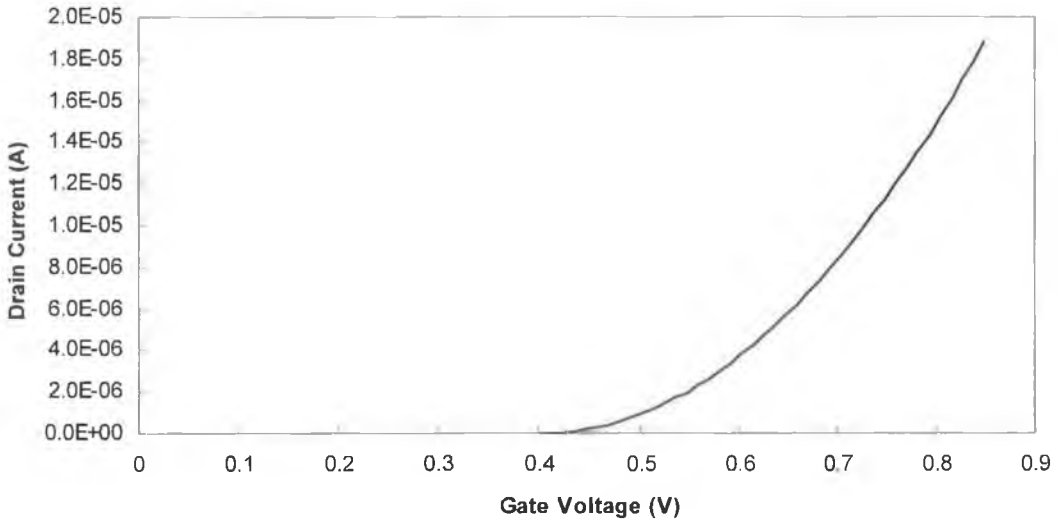


Figure 3.16 Simulated I_dV_g curve for a MOSFET. In this case the threshold voltage is 0.4V. This I-V curve can be used to monitor the effect of oxide degradation on device performance

Upon the application of a gate bias, oxide degradation begins to affect the I_dV_d and I_dV_g profiles of the device. If there is a buildup of negative fixed charge, it will act against the gate voltage so as to raise the threshold voltage, lowering the on-state current and reducing the maximum transconductance. Positive fixed charge will act with the gate voltage and lower the threshold and increase the off-state current, as carriers are still attracted to the channel by the oxide charge. Interface charge can lead to Coulombic scattering of carriers in the channel, causing a further reduction in drive current[30]. Finally, SILC provides a mechanism for charge carriers in the channel that would normally contribute to the drive current being lost to the gate. As the stress continues, this loss of carriers becomes more severe, lowering the drive current even further. In logic applications, the difference between on and off currents needs to be as high as possible, to ensure accurate signal transfer. This is threatened by fixed oxide charge raising the off current and SILC and interface trapped charge lowering the on-current. So, in addition to conventional oxide reliability, the reliability in current levels, and the stability of the threshold voltage also must be considered with ultra-thin, or high-k dielectric layers.

Figure 3.17 shows these effects in a real device after a constant voltage stress. The reduction and change in position of maximum transconductance are evident, as is the noise in the curve, thought to be due to Coulomb scattering from

interface traps. The drop off in transconductance above 0.8V becomes more pronounced during stress. The drop off in the fresh sample is attributed to the gate leakage current[31], and mobility degradation associated with a high carrier concentration near to the interface. Any further drop off after stress is attributed to SILC. In the I_dV_g curves, only a very slight shift in threshold voltage is evident in this particular example.

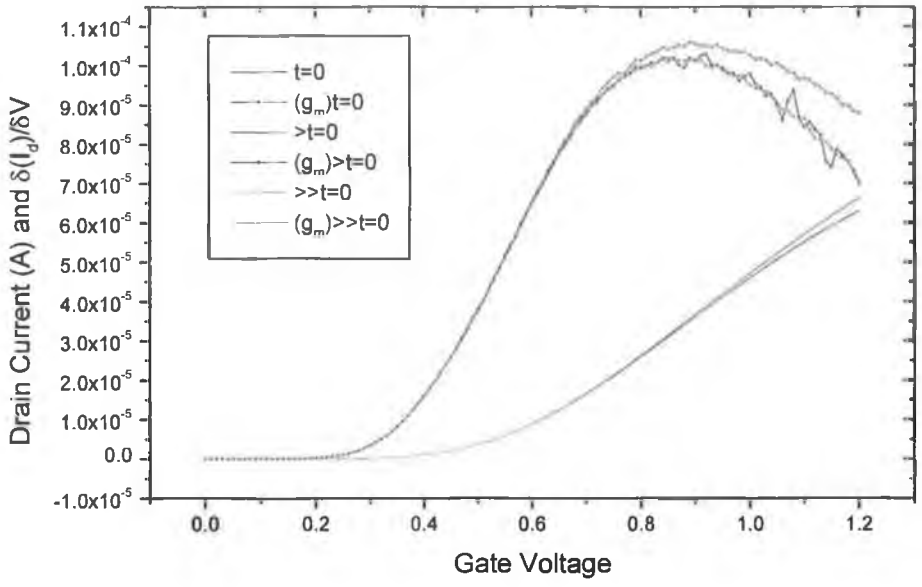


Figure 3.17 I_dV_g and transconductance curves for a device during constant voltage stress. The effect on both is evident.

3.6 Summary

In this chapter, electrical and statistical methods of measuring and describing oxide breakdown have been outlined. The following important conclusions on gate oxide integrity have been drawn: The problem of gate leakage becomes a larger problem as oxide thickness is reduced. Origins of gate leakage current in gate dielectrics and thickness dependence have been outlined. Information from Fowler-Nordheim I-V curves can be used to extract the width of the SiO_x transition region. The problem of time-dependent dielectric breakdown, its statistical nature, and its oxide field dependence has been discussed, and we have

shown that it is a threat to MOSFET long-term reliability. Breakdown mechanisms and trap generation have been discussed and the literature suggests that it is a percolation path of neutral electron traps from anode to cathode that causes breakdown. The origin of these neutral traps is still speculative. Three models exist in the literature that describe the possible origin of the traps, the anode hole injection model, the hydrogen release model, and the electric field or 'thermochemical' model. Each has experimental evidence to support it. Stress induced leakage current (SILC) is a consequence of the buildup of neutral traps, which increases the leakage current at low field and causes excessive power consumption.

3.7 References

[1] M. Hirose, Electron tunnelling through ultrathin SiO_2, Materials Science and Engineering, B41 (1996) pp. 35-38.

[2] X. Guo, T.P. Ma, Tunneling leakage current in oxynitride: dependence on oxygen/nitrogen content, IEEE Elec. Dev. Letters, Vol.19, No. 6, June 1998, pp. 207-209.

[3] K. F. Schuegraf, C. Hu, Reliability of thin SiO_2, Semicond. Sci. Technol., 9 (1994) pp.989-1004.

[4] L. Mao, C Tan, M Xu, Measurements of the widths of transition regions at Si-SiO_2 interface in metal-oxide-semiconductor structures from quantum oscillations in Fowler-Nordheim tunnelling current, Solid State Communications, 119 (2001) pp67-71.

[5] J.C. Poler, K. K. McKay, E. A. Irene, Characterization of the Si/SiO_2 interface morphology from quantum oscillations in Fowler-Nordheim tunnelling currents, J. Vac. Sci. Technol. B 12(1), Jan/Feb 1994.

[6] Khairurrijal, W. Mizubayashi, S. Miyazaki, M. Hirose, Analytic model of direct tunnel current through ultrathin gate oxides, Journal of Applied Physics, Vol. 87, No. 6, pp. 3000–3005.

[7] J. Suñé, I. Placencia, N. Barniol, E. Farrés, F. Martín, X. Aymerich, On the breakdown statistics of very thin SiO_2 films, Thin Solid Films, 185 (1990) pp. 347-362.

[8] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, H.E. Maes, A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides, IEDM Tech. Dig., pp. 863-866, 1995.

[9] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, H.E. Maes, New insights in the relation between electron trap generation and the statistical properties of oxide breakdown, IEEE Trans. Elec. Cev. 45(4) (1998) pp. 904-911.

[10] R. Degraeve, P. Roussel, G. Groeseneken, , H.E. Maes, A new analytic model for the description of the intrinsic oxide breakdown statistics of ultra-thin oxides, Microelecron Reliab. 36 (1996) pp1639-1642 .

[11] R Degraeve, PhD Thesis, pg. 20, KU Leuven, Belgium, 1998.

[12] T. Nigam, R. Degraeve, G. Groeseneken, M.M. Heyns, H.E. Maes, Constant current charge-to-breakdown: still a valid tool to study the reliability of MOS structures? Proc. IRPS 1998 pp. 62-69.

[13] J.W. McPherson, D.A. Baglee, Acceleration factors for thin gate oxide stressing, Proc. IRPS 1985 pp. 1-5.

[14] E. Wu, J. Aitken, E. Nowak, A. Vayshenker, P. Varekamp, G. Hueckel, J. McKenna, D. Harmon, L.-K. Han, C. Montrose, R. Dufresne, and R.-P. Vollertsen, "Voltage-Dependent Voltage Acceleration of Oxide Breakdown for Ultra-Thin Oxides," *IEDM Tech. Digest*, pp. 54–57 (2000).

[15] J. Sune, M. Nafria, E. Miranda, X. Oriols, R. Rodriguez, X. Aymerich, Failure Physics of ultra0thin SiO_2 gate oxides near their scaling limit, Semicond. Sci. Technol. 15 (2000) pp. 445-454.

[16] E. Cartier, Characterization of the hot electron induced degradation of thin SiO_2 gate oxides, Microelectron. Reliab. Vol. 38 No.2 (1998) pp. 201-211.

[17] D.J. DiMaria, D. Arnold, E. Cartier, Impact Ionization and positive charge formation in silicon dioxide films on silicon, Appl. Phys. Lett., Vol. 60, no. 17, pp. 2118-2120 (1992).

[18] D.J. DiMaria, E. Cartier, Mechanis for stress-induced leakage current in thin silicon dioxide films, J. Appl. Phys., 78 (6) 1995, pp. 3883-3894.

[19] D.J. DiMaria, D. Arnold, E. Cartier, Impact Ionization, trap creation, degradation, and breakdown in sililcon dioxide films on silicon, J. Appl. Phys. 1993; 73(7) pp. 3367-3384.

[20] G. Bersuker, Y. Jeon, H. Huff, Degradation of thin oxides during electrical stress, Microeletron. Reliab. 41 (2001) pp. 1923-1931.

[21] D. J. Dumin, J. R. Maddux, R. S. Scott, R. Subramoniam, A model relating wearout to breakdown in thin oxides, IEEE Trans. Electron Devices 41, (1994) pp. 1570-1580.

[22] K. F. Scuegraf, Hu C. Metal oxide semiconductor field effect transistor substrate current during Fowler-Nordheim tunnelling and silicon dioxide reliability, J. Appl. Phys. 1994; 76(6) pp. 3695-3700.

[23] E. Cartier, J.H. Stathis, Atomic hydrogen induced degradation of the Si/SiO_2 structure, Microelectron. Eng. 28, (1995) pp. 3-10.

[24] J. Maserijian, N. Zamani, Behaviour of the Si/SiO_2 interface observed by Fowler-Nordheim tunnelling, J. Appl. Phys. 53 (1982) pp. 559-567.

[25] B. Ricco, G, Gozzi, M Lanzoni, Modelling and simulation of stress-induced leakage current in ultrathin SiO_2 films, IEEE Trans. Elec. Dev, 42 (7) (1998) pp. 1554-1560.

[26] T. Kauerauf, R.O' Connor, M. Aoulaiche, R. Degraeve, M. Houssa, G. Groeseneken, H. Maes, Similarities in reliability evaluation between HfSiON and SiON, To be published.

[27] J. DeBlauwe, R. Degraeve, R. Bellens, J. V. Houdt, Ph. Roussel, G. Groeseneken, H.E. Maes, study of dc stress-induced leakage current (SILC) and it's dependence on oxide nitridation, Proc. ESSDERC 1996, pp. 361-364.

[28] G. Groeseneken, H.E. Maes, N. Beltran, R.F. De Keersmaecker, A reliable approach to charge-pumping measurements in MOS transistors, IEEE Trans. Electron Devices Vol. ED-31 No. 1 (1984) pp. 42-53.

[29] Simulator from 'Principles of Semiconductor Devices' section 7.7.1 by Bart Van Zeghbroeck available at http://ece-www.colorado.edu/~bart/book/book/.

[30] Y.B. Park and D.K. Schroder, "Degradation of Thin Tunnel Gate Oxide Under Constant Fowler-Nordheim Current Stress for a Flash EEPROM", *IEEE Trans. Electron Dev.* 45, 1361-1368, June 1998.

[31] P.M. Zeitzoff, C.D. Young, G.A. Brown, K. Yudong, Correcting effective mobility measurements for the presence of significant gate leakage current, IEEE Electron Device Letters, Vol. 24 (2003) pp. 275-277.

Chapter 4: Electrical Characterisation of Ultra-thin SiON layers

4.1 Introduction

This chapter contains a reliability study of ultrathin SiON layers, prepared for use in the 65nm technology node, using the measurement techniques and applying the trap generation theory outlined in Chapter 3. Section 4.2 describes how the samples were fabricated, and the difference between the method used to grow the oxides in this work and more conventional methods. The reliability of the layers is then examined in terms of Weibull slopes, voltage acceleration and maximum operating voltage, and some interesting results are explained by the progressive wear out of the breakdown spot. The study is extended to examine time dependent dielectric breakdown at high temperature using the same measurements. The possibility that devices may continue to function in the post breakdown phase while the gate current remains modest is then investigated. These results would indicate that this strongly depends on device aspect ratio. Finally, the implications of oxide breakdown for a real circuit are briefly outlined.

4.2 Sample Fabrication

There is a need to optimise the properties of SiON due to the problems which have been encountered with implementing the proposed new high-k dielectric materials. Problems such as low mobility and high charge trapping with these materials have extended the life of SiO_2 based layers well beyond what would have been predicted even a few years ago. The samples from IMEC used in this work were grown in three steps[1].

First, a base layer of SiO_2 was grown using in-situ steam generation for pmos devices and rapid thermal oxidation for nmos devices, leading to base oxide

thicknesses of 0.8nm-1.2nm. The nitridation step consisted of exposing the oxides to a high density, pulsed RF N_2 plasma. Plasma based nitridation has taken over in recent years from the widely used high temperature thermal nitridation technique, because it can be better controlled leading to a more predictable nitrogen profile[2]. Using a pulsed plasma, the ions should strike the surface with a lower energy than for a continuous wave plasma, possibly allowing for the incorporation of more nitrogen in the layer, while minimising any damage to the layers. Various frequencies, plasma effective powers, and nitridation times were used to produce a range of different samples. Finally, a post nitridation anneal in O_2 was carried out to stabilize the layers, different anneal times being used for nmos and pmos layers. Resulting layers in nmos devices had an EOT of ~1.4nm and a physical thickness of ~1.5nm, while layers in pmos had EOT~1.1nm and physical thickness of 1.6nm. Thus the dielectric constant was approximately the same as that of SiO_2 (3.9) in the nmos layers, and about 5.6 in pmos. These layers were then fabricated into transistors and capacitors of various aspect ratios and capped with polysilicon gates for electrical testing.

Table 4.1 summarises the process conditions and the resulting equivalent oxide thicknesses (EOT) for all samples used in the reliability study. All samples are capacitors with active area encompassing source and drain. The structure is shown schematically in figure 4.0.

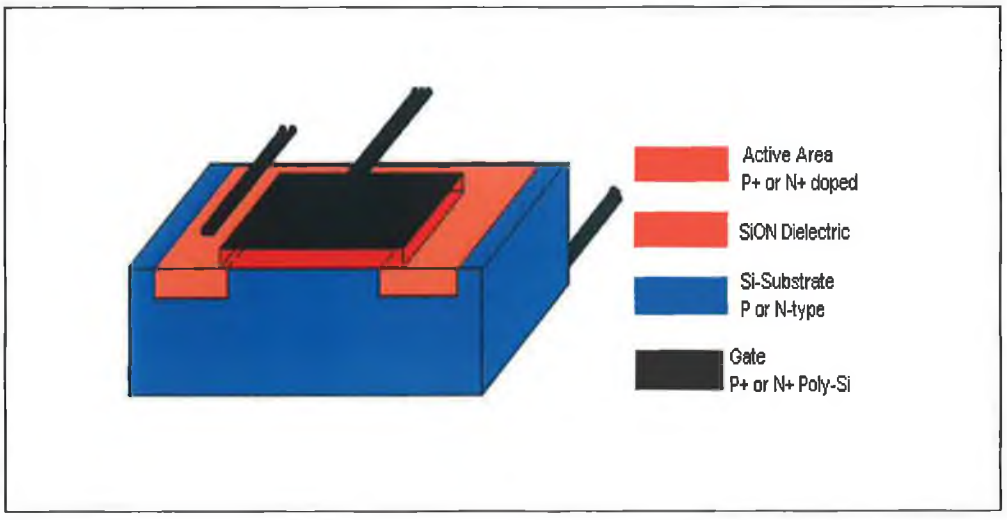


Figure 4.0: Schematic diagram of the test structures used in the reliability study.

SAMPLE	OXIDATION	ANNEAL	EOT (nm)	Type
1	RTO + DPN 15 sec, 10mTorr, 20% duty cycle, 10kHz, 100% N_2, 1125W	15 sec, 1000°C, O_2, 0.5 Torr	1.44	nmos w/ p-substrate
2	RTO + DPN 15 sec, 10mTorr, 20% duty cycle, 10kHz, 100% N_2, 1750W	15 sec, 1000°C, O_2, 0.5 Torr	1.44	nmos w/ p-substrate
3	RTO + DPN 15 sec, 10mTorr, 20% duty cycle, 10kHz, 100% N_2, 2375W	15 sec, 1000°C, O_2, 0.5 Torr	1.46	nmos w/ p-substrate
4	0.7 nm ISSG + DPN 30 sec, 5mTorr, 40%N_2, 100W	1 sec, 1000°C, O_2, 0.5 Torr	1.10	pmos w/ n-substrate
5	0.8 nm ISSG + DPN 30 sec, 5mTorr, 40%N_2, 100W	1 sec, 1000°C, O_2, 0.5 Torr	1.13	pmos w/ n-substrate
6	0.9 nm ISSG + DPN 30 sec, 5mTorr, 40%N_2, 100W	1 sec, 1000°C, O_2, 0.5 Torr	1.11	pmos w/ n-substrate
7	1.0 nm ISSG + DPN 30 sec, 5mTorr, 40%N_2, 100W	1 sec, 1000°C, O_2, 0.5 Torr	1.12	pmos w/ n-substrate

Table 4.1 SiON layer preparation conditions

4.3 Reliability

The reliability of the gate-dielectric layer is known to be of utmost importance in the fabrication of ULSI devices and has been a growing concern as the layers become thinner, as already outlined. This section investigates the reliability of our ultra-thin SiON layers nitrided using decoupled plasma nitridation, to ascertain the lifetime of devices that have been fabricated using them, and to compare the reliability with that of conventional SiO_2 layers. As the layers are so thin, leakage currents of $10A/cm^2$ are common at 1V. Because of this high leakage current, only the smallest devices can be used for reliability forecasts due to series resistance effects[3]. In larger area devices, an appreciable part of the voltage will fall across the contact pad, leading to a lower oxide voltage, and consequently a reliability prediction that would be an over-estimation. The effect is shown in figure 4.1, where the I_gV_g curves for a typical sample are shown for

various areas. The larger devices have a lower current density at high voltage, in the Fowler-Nordheim regime.

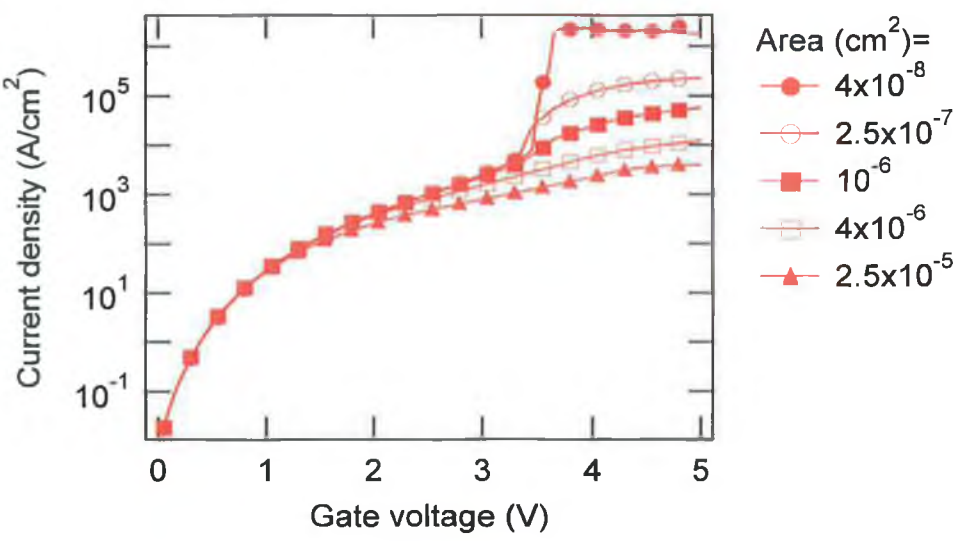


Figure 4.1: Current density vs. gate voltage. For large devices the current is significantly lower at high voltage, because not all of the voltage drops across the oxide. This leads to over-estimated lifetimes.

For this reason most of the samples used in this work are 1μm x 1μm (i.e. 1 x10^{-8} cm^2). A Constant Voltage Stress (CVS) reliability study was first performed for both nmos and pmos capacitor structures using both positive and negative gate voltage. This involves applying a constant gate voltage and monitoring the gate current until such time as oxide breakdown occurs. The initial current was recorded and only results for samples within 5% of the average current were used, to ensure all measurements were made on uniform thickness layers (a variation in the initial current is an indication of different electron tunnelling distances). Results for pmos-gate injection were discarded, as in this case, electrons are injected from the valence band, which represents a bigger barrier to tunnelling carriers, and reliability is not an issue in the samples under investigation[4]. CVS measurements were carried out on all wafers in order to determine the effect of varying nitrogen concentrations and process conditions on device reliability. However, this proved difficult, as the layers with the highest nitrogen concentration were also the thickest, so any improvement in reliability was likely due to thickness effects. Figure 4.2 shows a sample current vs. time (I-t) trace for a CVS measurement where breakdown occurs at ~432 seconds. In

these thin layers, the current simply rises progressively and extraction of the t_{bd} proves very difficult.

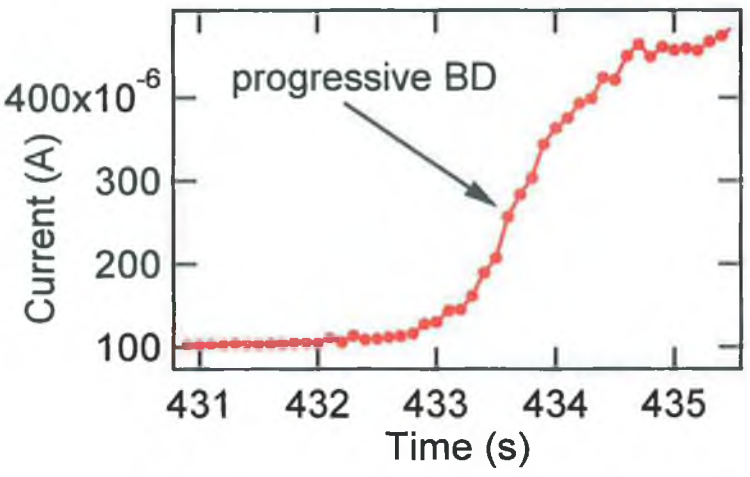


Figure 4.2: Sample current-time trace during a constant voltage stress. The current is monitored until oxide breakdown occurs. Then t_{bd} is recorded. Note the progressive rise in current after the formation of the breakdown path at ~432s.

In these cases, a gate current noise-analysis algorithm[5] was used to extract the t_{bd}. This algorithm accurately determines the time-to-soft breakdown, which for this work is treated as 'real' breakdown, without the device having suffered the thermal runaway effects that lead to hard breakdown.

The severity of the breakdown is related to the alignment and positioning of traps within the oxide as schematically illustrated in Figure 4.3. If the breakdown occurs because of a single, well-placed trap (within the percolation distance of the anode and cathode) there may only be a slight increase in gate current and noise, whereas if the breakdown arises from a chain of two traps the increase in current will be much greater, because the tunnelling distance for electrons for each step is significantly reduced, and thus the tunnelling probability is higher than if there was only one trap. This leaves a question mark over which should be treated as the real breakdown, and how reliability extrapolations should be made.

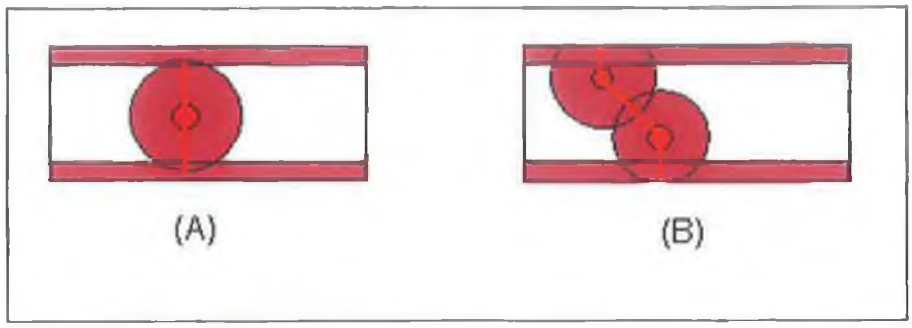


Figure 4.3: A conduction path made up of a single well placed trap within the percolation distance of both anode and cathode (A) will lead to only a slight increase in gate current, and though it may be detected as a breakdown, devices may continue to function. (B) shows a two-trap percolation path, which will lead to a larger increase in current.

We will demonstrate later in this chapter that several parameters change based on reliability criteria and that it is necessary to examine the functionality of a device after a so-called 'soft' breakdown to determine whether such soft-breakdowns should be used in estimating reliability.

The t_{bd} Weibull distributions were fitted with a maximum likelihood method, to ensure accurate extraction of the Weibull slope (β) and the 63%-value (η). Each oxide was stressed at a number of stress voltages and β was determined for the oxide by using all t_{BD} distributions. Log (η) was plotted vs. V_g (figure 4.4). The slope of this line is defined as the acceleration factor (γ).

The data points in figure 4.4 represent real t_{bd} data. By simple linear extrapolation and applying the area and percentage scaling rules discussed in Chapter 3, the maximum operating voltage at the specification condition of 10 years (~$3x10^8$ seconds) lifetime, 0.1 cm^2 and 0.01% failures was determined.

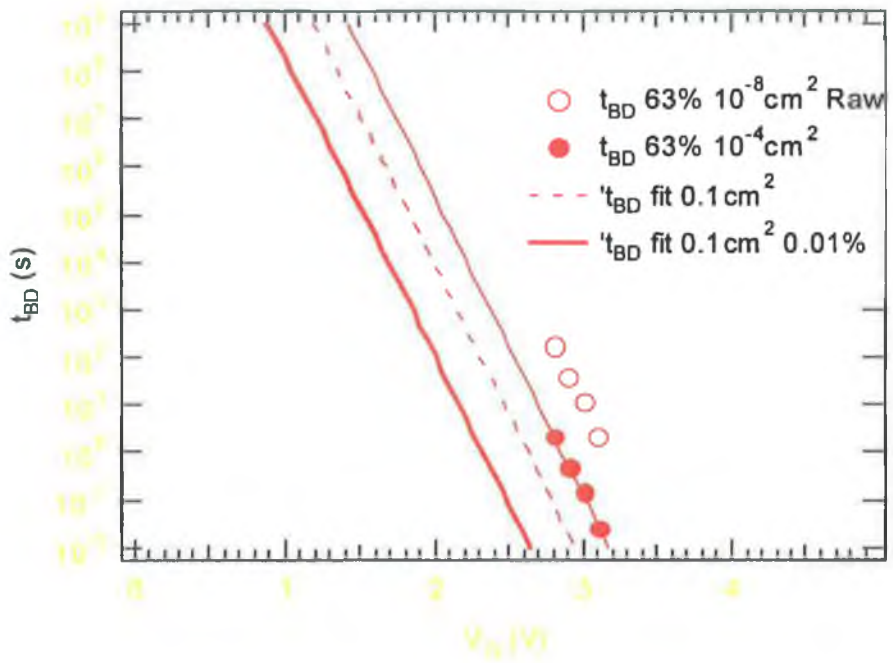


Figure 4.4: Reliability extrapolation to determine maximum operating voltage. Open circles represent raw data. Closed circles are the data scaled to $1x10^{-4}cm^2$

Figure 4.5 shows a comparison between the voltage acceleration factors (γ), measured on the SiON layers and those observed for conventional SiO_2. The trend for SiO_2 seems to support voltage dependent voltage acceleration. The fact that γ may be dependent on gate voltage was demonstrated in the work of Alam et al [6]. It is clear that γ is generally lower than expected for SiON dielectrics. On examining figure 4.4, it could then be concluded that lower γ (lower slope of extrapolation) would lead to a lower operating voltage if the device were to be operational for 10 years. We observed, however, maximum operating voltages of 0.8 V or higher as shown in figure 4.6.

Also, it can be seen from figure 4.6, that SiON with EOT 1.1 nm in pmos devices can in some cases be equally reliable as 1.45 nm thick samples fabricated in nmos devices.

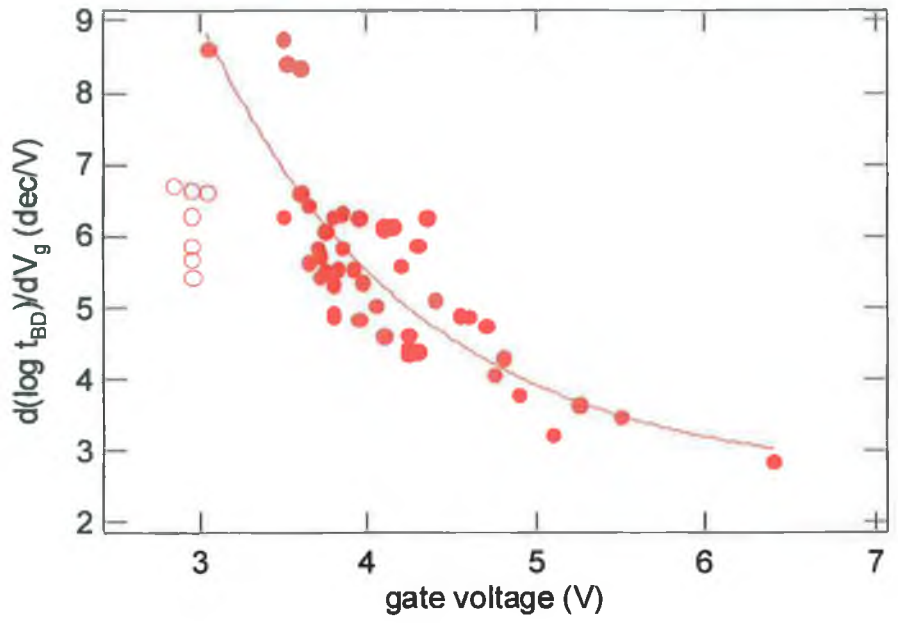


Figure 4.5 Comparison of γ for SiO_2 (closed circles) and SiON (open circles) shows much lower voltage acceleration rates for SiON, which would indicate lower reliability, but the high β values lead to similar reliability to SiO_2 at operating voltage. Solid line is a fit to the SiO_2 data.

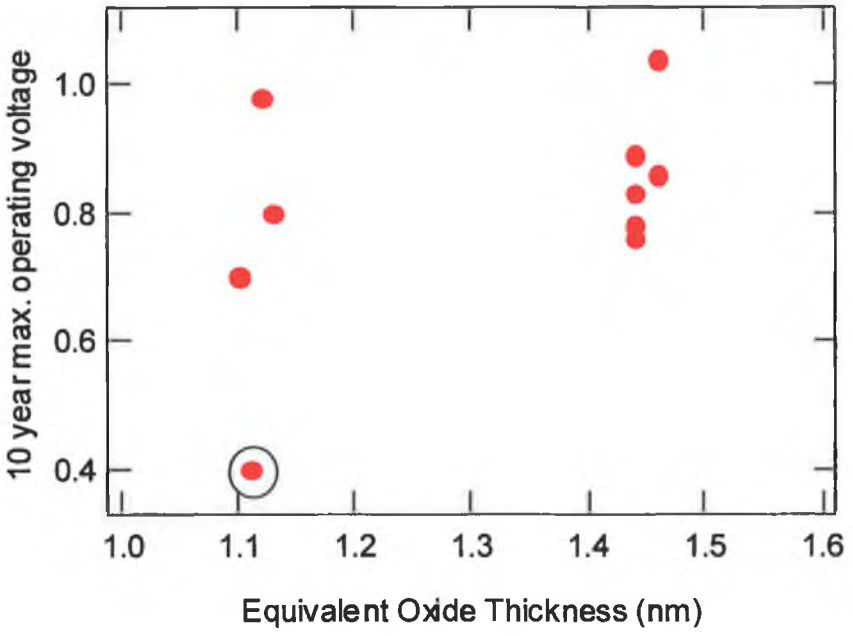


Figure 4.6 The same maximum operating voltage is sufficient for several SiON recipes (req=0.8V). The oxides with EOT of ~1.4nm are nMOS devices and those of ~1.1nm are pMOS. Equal reliability can be attained for both. The circled point corresponds to a sample with low β and low γ.

The reason for this high reliability is the very high values for β, which were all in the range 1.5 to 2.4 for SiON layers, while for SiO_2 layers in this thickness

range, values in the 1 to 1.5 range are expected[7]. The reason for this expectation is simply that for very thin oxides, breakdown occurs at a much lower trap density, because a percolation path of only a few traps will cause breakdown. The statistical spread of trap density on creating a path of only a few traps is much larger than for a path of many traps in a thick layer, and so thin layers have a lower Weibull slope[8]. The high β values recorded for the SiON layers reduces the area dependence of the t_{bd} (explained in Chapter 3), which has the effect of drawing the dashed and bold lines in figure 4.4, closer to the raw data extrapolation. It also results in a reduction in the gap between the 63% and the low percentile t_{bd}, which causes the 0.01% fit to be drawn closer to the raw data. With the reliability at specification fits in figure 4.4 being drawn closer to the raw data (i.e. being shifted to the right on the plot), this means that they will intercept the 10-year line at a higher operating voltage. Moreover, β valves for the SiON dielectrics seem to relate to γ as shown in fig. 4.7, an effect that has not been noted before in the literature. The decrease of γ, which degrades the reliability, is compensated by a high β, which improves the reliability.

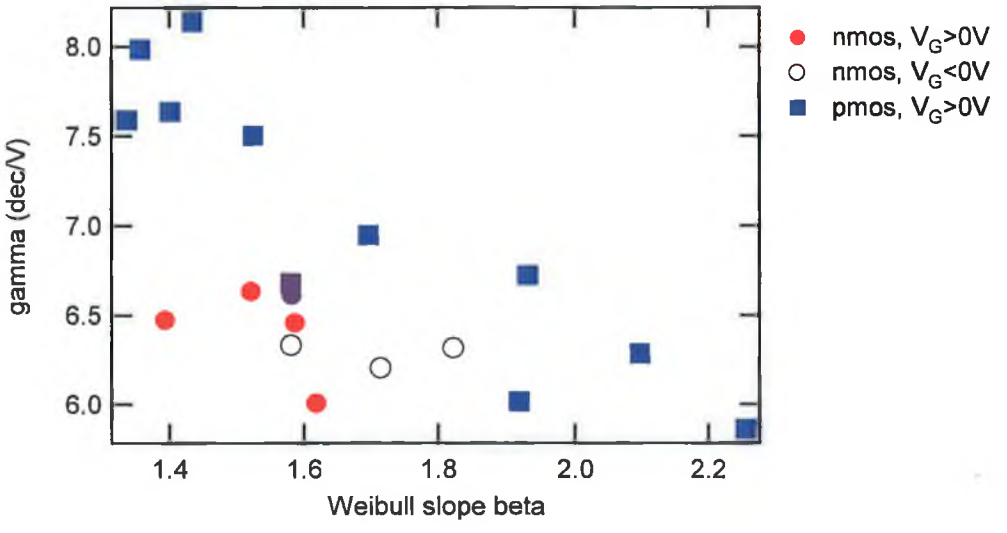


Figure 4.7. γ vs. β shows that there is a correspondence between the quantities. Low γ, which determines reliability, is compensated by high β, which improves it.

4.4 Further examination of Weibull Slopes, Trap Generation and Breakdown

In analysing the reason for the unexpectedly high Weibull slopes measured, some more information is required about the origin of the slope. It has already been outlined that the Weibull slope is strongly dependent on the number of traps that constitute the breakdown path. Taking a percolation distance of 0.9nm as predicted by the percolation model and assuming that the interface acts as a sheet of traps available to terminate a breakdown path[9], one well-placed trap in the centre of a 1.5nm oxide is enough to cause breakdown.In terms of statistics, the critical trap density to form such a path is Weibull distributed with a slope (β_{ot}) of 1. If a two-trap conduction path made of two not so well placed traps causes breakdown, β_{ot} would be 2. However, time-to-breakdown experiments do not give the Weibull slope of the critical trap density. The trap density is related to the stress time by a power law, $D_{ot}= C.t^m$ where C is dependent on stress voltage. From this, it follows that the distribution of the critical trap density is related to the distribution of the time-to-breakdown by the trap generation rate or, more simply, $\beta_{tbd}=m*\beta_{ot}$.This theory can now be applied to examine the origin of the high β. For conventional ultra-thin (1.5-2.5nm) SiO_2, m is between 0.5 and 0.7, and β_{ot} is 2 (as two traps break the oxide), resulting in β=1 to 1.4. The very high β in the SiON samples can be explained in two ways:

(1) Three traps are needed to break the oxide, i.e. β~2 and m~0.7 so β_{ot}~3. Assuming a percolation distance of 0.9nm, the effective 'radius of conduction' of a trap is 0.45nm. It is obvious then, that the oxides used in this study are just thin enough to allow a two-trap conduction path to break the oxide, which theoretically can happen from 1.35nm-1.8nm. So, for three-trap breakdown, the percolation distance would no longer be 0.9nm, for the model to be correct, it would have to be changed to 0.6nm.

(2) m is higher than in SiO_2 i.e. $\beta \sim 2$ and $\beta_{ot} \sim 2$ so $m \sim 1$.

This would mean the percolation distance remains unchanged from that of SiO_2, but that SiON instead has a higher trap generation rate, and that the incorporation of the nitrogen promotes trap generation by some mechanism.

In order to discriminate between these two possibilities, the pre-breakdown events in 47 current-voltage traces for one selected oxide recipe with a t_{bd} Weibull slope of 2.2 were analysed. The pre-breakdown current steps can be interpreted as the formation of non-breaking conduction paths. If the current level (ΔI) used as the breakdown trigger is reduced below 6μA, the measured Weibull slope changes[10].

The measured relation between β and ΔI is shown in figure 4.8. β drops in 1 single step from 2 to about 0.8 as the current step is reduced to lower values. This suggests a change from 2 to 1 trap conduction paths. At low current triggers, paths of a single trap are causing the current increase ($\beta_{ot}=1$). From this observation, it can be concluded that the trap generation rate is near the expected value of $m \cong 0.7$. So, if the trap generation rate is as expected, why does the Weibull slope grow larger than the expected value of 1.6 when the current step is increased?

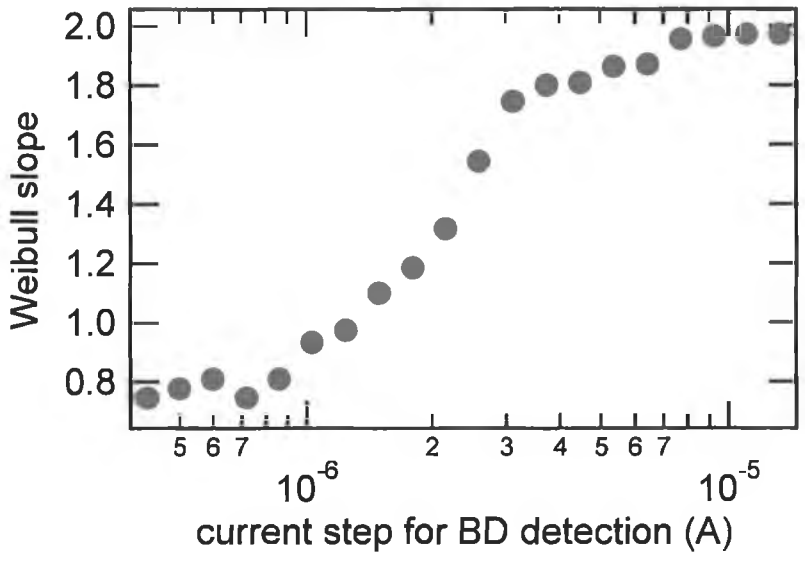


Figure 4.8: Measured Weibull slope for increasing breakdown current trigger. In this case the Weibull slope rises above the expected value of 1.6 for two trap breakdown paths.

We tried to further examine the trap generation rate by measuring the increase of the interface states during stress by charge pumping. This was done by applying a CVS to the oxide, and then interrupting the stress with a 3MHz charge pumping measurement, to measure the number of interface traps. Apart from pure interface states, bulk traps very near to the interface can also be sensed in very thin SiO_2. The interface trap density provides an estimate for the bulk trap density.[11]

Our measurements (figure 4.9), however, show a very low interface state generation rate, not at all corresponding with the m=0.8 deduced from the breakdown analysis.

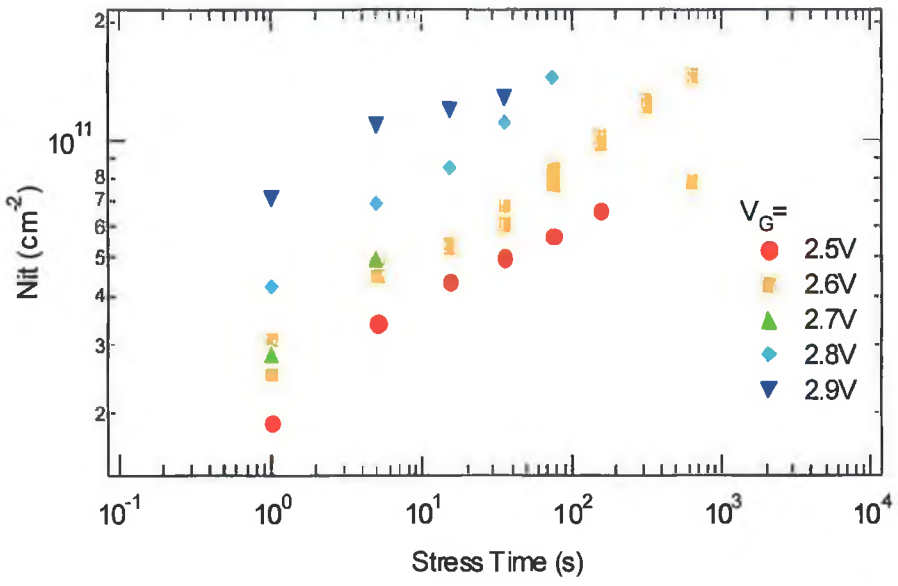


Figure 4.9 Number of interface traps as a function of stress shows a very low generation rate. The rate is much lower (~0.2) than the expected value of close to 1 from our statistical analysis.

The behaviour thus cannot be verified to be due to an increased trap generation rate. Also, bulk and interface trap generation rates seem to vary greatly. The true explanation for the high Weibull slopes, low voltage accelerations, and correlation between the two, was found to lie in the progressive wearout (the growth of a conduction path after it's formation at the instant of breakdown) of the breakdown path, and the ability of the layers to resist such wearout[12].

If we consider the onset of progressive breakdown after the formation of a conduction path, we find that if the breakdown trigger uses a small change in the

gate current (ΔI), there will be a significant time delay between the triggering of a breakdown event and a substantial rise in gate current as given by the differential equation which can be used to describe progressive wearout:

$$\frac{dI(t)}{dt} = KI(t)^b \exp\left\{a\left[V_g - R_s I(t)\right]\right\} \qquad \text{Eq 4.1}$$

where K is a constant, b is an adjustable wearout exponent, a is the voltage acceleration of the leakage path wearout, and R_s is a series resistance.
After this stage, when the current begins to grow appreciably, its growth is limited by a series resistance and continues to grow logarithmically. The three different phases of breakdown are schematically illustrated in figure 4.10. The effect of stress voltage on the progressive wearout is discussed later in the chapter.

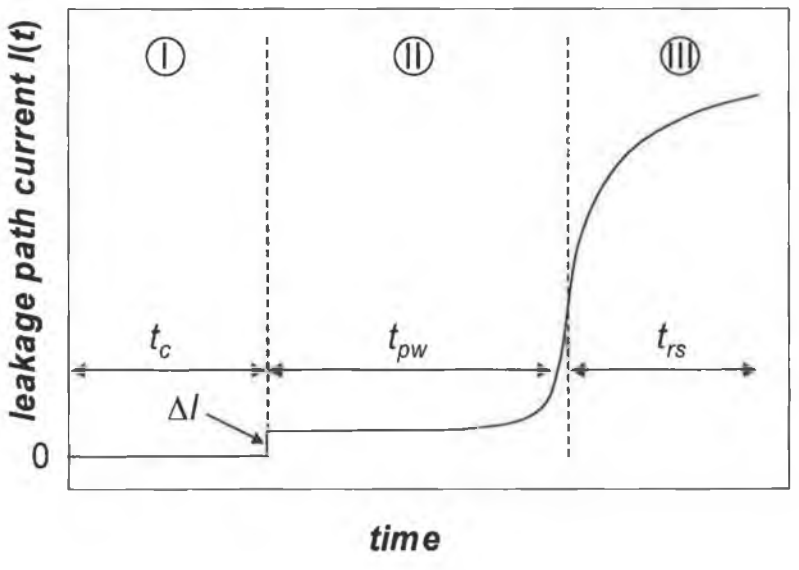


Figure 4.10: The breakdown process. In part I, the percolation model describes the buildup of traps, and the current through the path is 0 as it has not yet formed. After the path is created (t_c) in part II, it begins to grow slowly and allow more current to flow through it. The actual formation of a path may only result in a very small current jump (ΔI). After a certain time (t_{pw}), the current shows a significant increase, and it eventually saturates, due to series resistance[12].

This simple description can be used to explain the unusual Weibull slope and voltage acceleration values. For very small current steps, the Weibull slopes are constant around 0.8 (from figure 4.8). Also, at these very low current steps, the

voltage acceleration is much closer to the expected value of ~10. So, at low current step, the breakdown behaviour is as expected in line with the predictions of the percolation model.

However, when we begin to trigger on a higher current step, the actual formation of the breakdown path has gone unnoticed, and our trigger occurs some time in stage II of figure 4.10, where the path is already wearing out. In this regime, percolation theory no longer applies and the growth of the breakdown path dominates measurements.

This is made very clear in figure 4.11 where the transition from the regime where percolation theory applies, to that where it does not is immediately apparent. It is in this second stage where high Weibull slopes and low voltage acceleration will be measured. The low voltage acceleration is actually a measure of the voltage dependence of the wear-out of the conduction path. Depending on the 'resistance' against the onset of progressive breakdown, both the Weibull slope and the acceleration factor vary. In other words, both the Weibull slope and the acceleration factor in this second phase are a measure of how robust the oxide is against progressive breakdown.

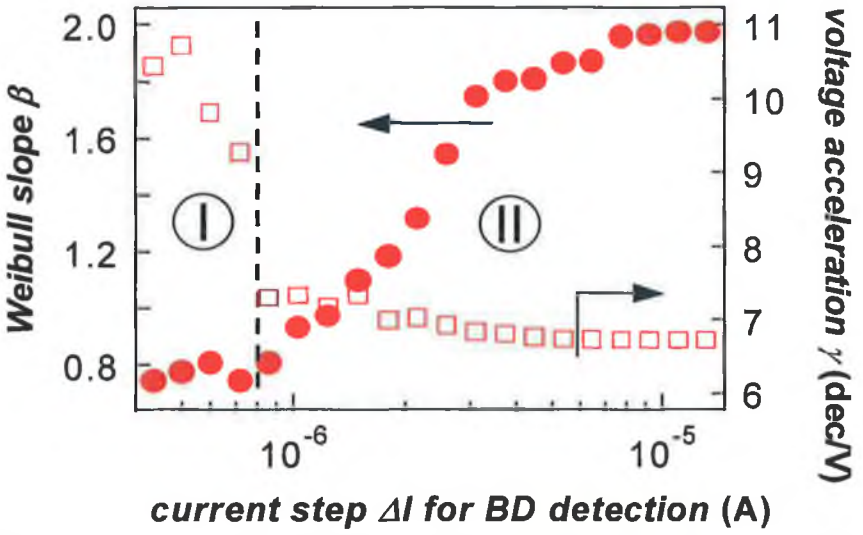


Figure 4.11: Measured β and γ vs. current step. At low current step in region I, expected values are found. Only when the current step is increased, such that measurements are made in the wearout stage, do we measure high β and low γ. The relationship between the two quantities is also only visible at high current step.

Thus in ultrathin oxides, using the intuitive idea that an oxide breakdown occurs when there is a significant rise in the gate current level is no longer valid. The

progressive wear-out of the conduction path after breakdown needs to be understood before a correct reliability extrapolation can be made.

4.5 Elevated Temperature Effects

4.5.1 Temperature Accelerated Breakdown

Another important factor in assessing reliability is the effect of elevated temperature on oxide performance. Gate oxides have to perform at temperatures above 100°C when in microprocessor devices[13], and so it is necessary to study their behaviour at and above this temperature. In the high temperature study undertaken in this work, the devices underwent constant voltage stress, at temperatures ranging from 20°C to 200°C, and the times to breakdown were recorded. In this case, breakdown was defined in the conventional manner as a significant rise in the gate current, indicating the formation of an electrically conducting path from anode to cathode. However, in the light of the above discussion, the results again show that this definition of breakdown is incomplete. As with the room temperature study, the time-to-breakdown (t_{bd}) Weibull distributions were fitted with a maximum likelihood slope method, to ensure accurate extraction of the Weibull slope (β) and the 63%-value (η). Each oxide was stressed at a number of stress voltages and β was determined for the oxide by using all t_{bd} distributions. Oxides fabricated into pmos and nmos structures were both tested, and in all cases, a positive gate voltage was used. The voltage acceleration factor γ, was subsequently determined from a plot of Log (η) vs. V_g.

Figure 4.12 shows a plot of Log (η) vs. V_g. for one oxide recipe grown in a pmos device, for temperatures ranging from 20° to 200°C (293K-473K). From this it is clear that elevated temperature greatly reduces the t_{bd} for a given stress voltage. Note that the voltage acceleration, determined from the slope of this graph, is essentially temperature independent. The voltage acceleration for all samples is shown in figure 4.13, and shows quite a wide scatter indicating that voltage

acceleration does not depend largely on temperature. The effect is almost identical for nmos and pmos devices. The constant voltage acceleration with increasing temperature agrees with some previous works[14], which state that voltage acceleration is temperature independent in the studied temperature range.

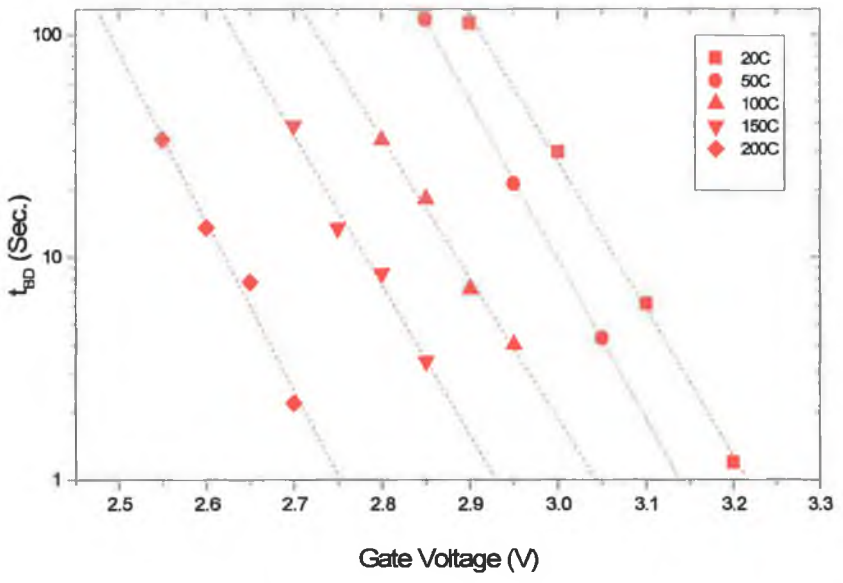


Figure 4.12. The 63% time-to-breakdown plotted vs stress voltage at different temperatures. A parallel shift of the t_{BD}-V_G trace is observed with change of T. The voltage acceleration appears to be temperature independent, this is further illustrated in figure 4.13

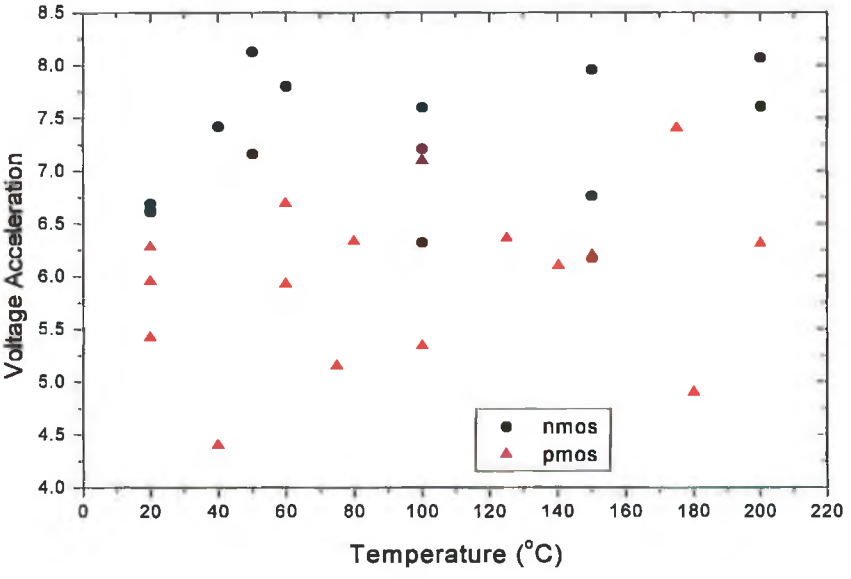


Figure 4.13 Voltage acceleration vs. Temperature for all oxides investigated. The plot shows a wide scatter in voltage acceleration for both nmos and pmos structures. This was confirmed for a number of breakdown trigger current levels.

Figure 4.14 shows the effect of increased temperature on the t_{bd} for a particular oxide at a set voltage of 2.72 V. The data are well described by a power law. The t_{bd} is two orders of magnitude lower at 400K than at 300K. If we extrapolate this back to the low field operating voltage range using a worst case linear extrapolation, keeping the operating temperature 50K lower could mean many years of additional device lifetime, so clearly controlling operating temperature and understanding temperature effects should be of concern to industry.

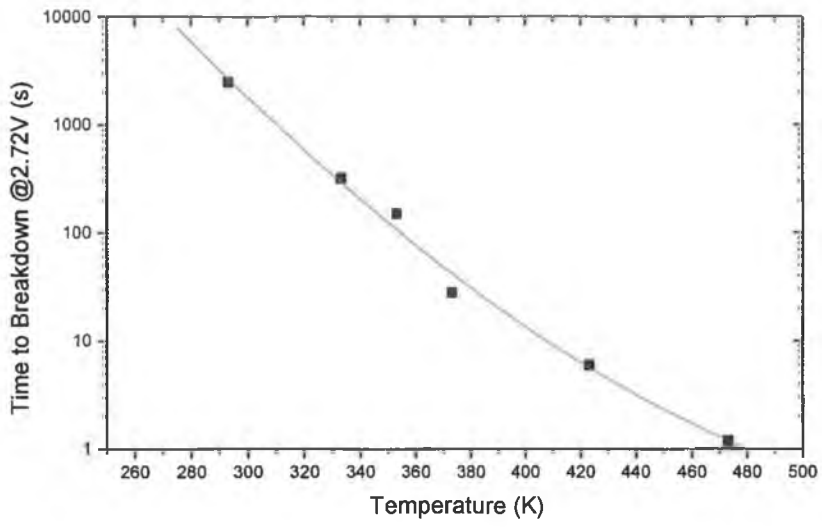


Figure 4.14. The time to breakdown at 2.72V for a variety of stress temperatures. A reduction of the stress temperature greatly increases the time to breakdown, indicating that the trap generation rate has a strong thermal dependence.

The power law behaviour in figure 4.14 equates to temperature dependent temperature acceleration. We will use our findings to prove that this behaviour supports the voltage dependent voltage acceleration model proposed by Wu et al[15]. The Weibull slope is affected by increased temperature as shown in figure 4.15 but the effect is only visible for pmos structures. Also visible from figure 4.15 is a clear difference in Weibull slope between pmos and nmos structures across the temperature range studied. Since making a reliability projection to the low field range involves the Weibull slope one would assume that this difference between pmos and nmos would result in a difference in reliability. However, we must be sure that this is a real effect before making such a statement.

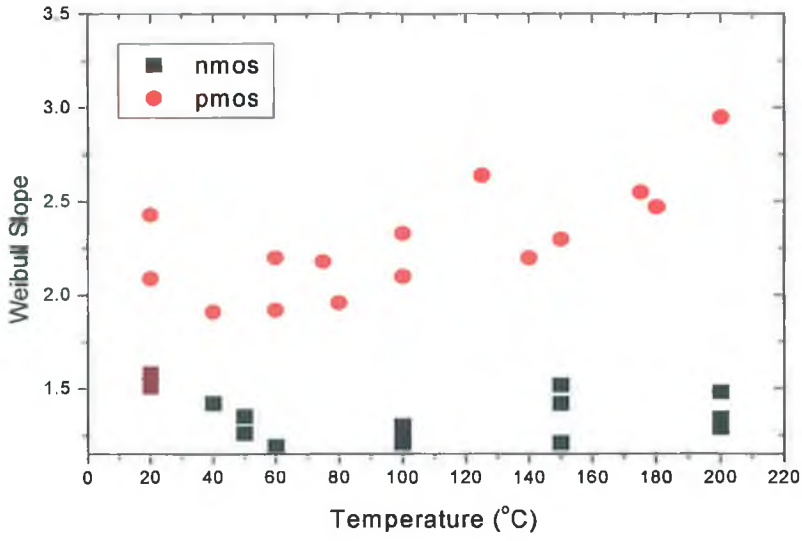


Figure 4.15: Weibull slope vs. temperature for pmos and nmos devices. Pmos layers show increased Weibull slope at high temperature.

As outlined in the previous section, the Weibull slope is strongly dependent on the breakdown trigger used in measurements, and so the difference in Weibull slope between pmos and nmos structures is found to disappear when the breakdown trigger is lowered to a current level where the percolation model applies. Figure 4.16 shows the effect of elevated temperature on Weibull slope for a range of trigger currents for the pmos structures. As the breakdown trigger is lowered, the Weibull slope in general drops, but no matter what the breakdown trigger, the slope is still found to increase for increasing temperature.

A plausible explanation is that as we move toward higher temperature, the breakdown is formed with almost no increase in current and so our breakdown is *always* in phase II of figure 4.10, meaning that we never measure t_{bd} in the percolation model regime at high temperature. It is difficult to confirm this, as at very low current step triggers, current noise becomes a problem.

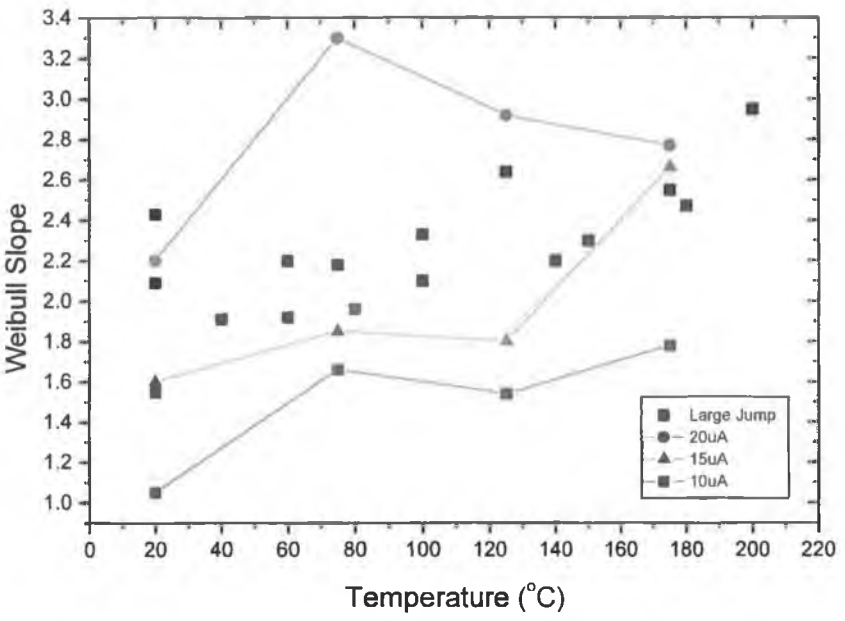


Figure 4.16. Weibull slope vs. Temperature for pmos layers showing different breakdown triggers. As the breakdown trigger is made lower, the increase in Weibull slope becomes less severe.

This increase of Weibull slope at high temperature, whether a real effect or not, makes reliability extrapolations even more complicated. The effect of elevated temperature operation on t_{bd} would indicate a much lower projected maximum operating voltage if the device was to be operational for 10 years as the extrapolations for reliability estimates lower time-to-breakdown. The increasing Weibull slopes for increasing temperature in pmos structures also improves the reliability projection by reducing area dependence of the t_{bd}, which has the effect of drawing the dashed and bold lines in figure 4.4, closer to the raw data extrapolation. It also results in a reduction in the gap between the 63% and the low percentile t_{bd}, which causes the 0.01% fit to be drawn closer to the raw data. The result of such uncertain reliability projections makes them essentially meaningless, as we will calculate good reliability for the pmos structures, even though they have lower t_{bd} than the nmos samples during high-field measurements. The results of the study demonstrate this. Figure 4.17 shows the maximum operating voltage for all nmos and pmos samples. While the nmos samples suffer degraded reliability at elevated temperature as expected, pmos samples gain reliability which is due to favorable scaling that comes with a high

β. Note that the extraction of maximum operating voltage carries an error of ~0.2V, but even with such an error bar, the trend is visible.

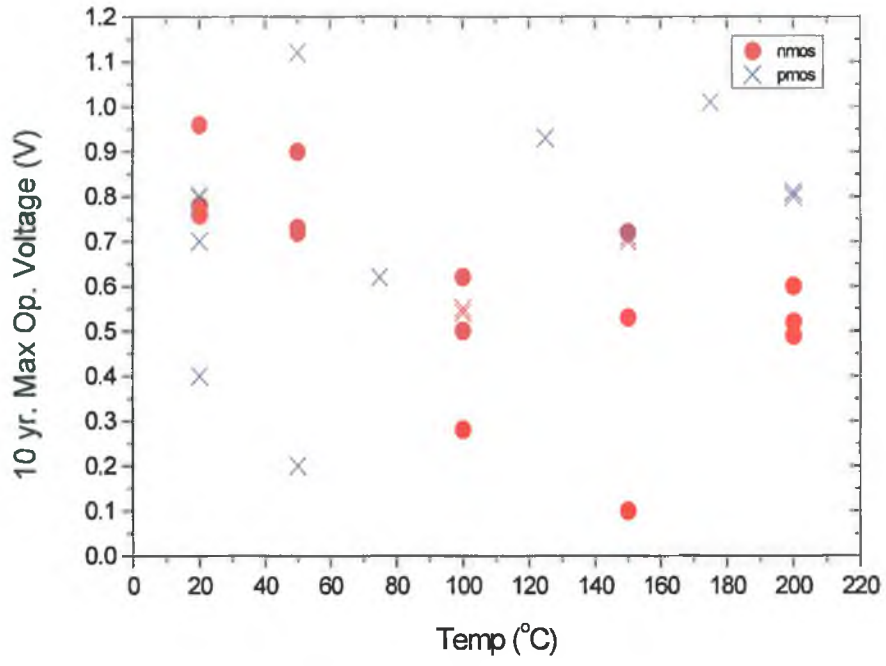


Figure 4.17 The10 year max operating voltage vs. temperature for all oxide recipes. Pmos structures show an overall rise in reliability at higher temperatures, where nmos lose reliability. This discrepancy is most likely due to inaccurately detecting breakdown events in the pmos structures.

4.5.2 Interdependence of Temperature and Voltage During Stress

The trap generation process is becoming more widely attributed to processes that include the breaking of Si-O bonds[16], and in electrical stress measurements this is attributed to the electric field across the oxide, either directly or by a multi-step process. What is often overlooked is the fact that temperature can also assist in the bond breaking process and even though the bond energy of a Si-O bond is around 8eV, the decreased breakdown times at elevated temperature are an indication of the thermal contribution.

To date, little work has been carried out on the relationship between the stress voltage and the temperature, and how the two are interrelated.

To observe the relationship between the two, we took a fixed breakdown time of 10 seconds, at a number of different stress temperatures, and plotted the voltage required to achieve this breakdown time, which was acquired from the fit.
A sample plot is shown in figure 4.18 and a good linear fit is obtained between the two quantities. Similar linearity is found for all samples investigated.

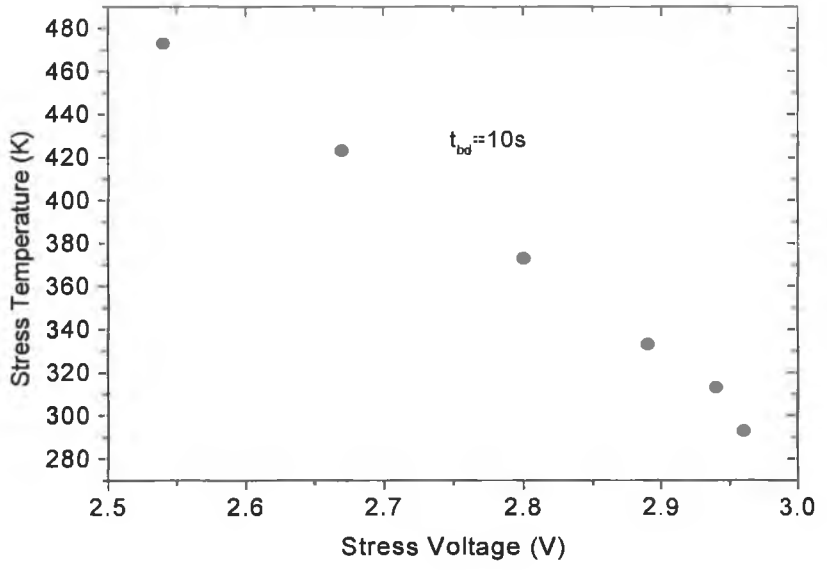


Figure 4.18: Stress voltage required to attain a fixed breakdown time of 10sec at a variety of different stress temperatures. The linear relationship suggests the two quantities can be used interchangeably in a constant voltage stress.

The slope of the fit is approximately 400 for all samples. The significance of the linearity is that increasing the stress temperature by 400K is directly equivalent to raising the stress voltage by 1V, or because the voltages used are much smaller than this, an increase of 0.1V is the same as a 40K temperature increase. So the same t_{bd} can be achieved at say 400K and 2.72V as for 300K and 2.97V. This correlation would suggest that traps are generated by the same means for electrical stress as for thermal stress, and further strengthens the thermo-chemical model[17].
Knowing the constant by which the two are related could then allow the lifetime of devices that are to be operated at elevated temperatures to be tested at room temperature. For example, if these devices were required to operate at a temperature of 600K, one could simply stress the devices at a modest 400K,

while increasing the stress voltage by 0.5V, thereby simulating the higher temperature.

On examining data for conventional SiO_2 dielectric layers, the temperature/stress voltage relationship is found to be ≈200K/V[18]. This implies that the SiON layers in this study have about double the thermal strength of conventional layers and should thereby be more suited to higher temperature operation.

The fact that a linear relationship exists allows a normalisation of data collected at a number of stress temperatures to one reference temperature, in the same way that data for different area devices is often normalised to a reference area. This is illustrated in figure 4.19 where a set of data such as that in figure 4.12 is normalised to 100°C (373K).This approach is only useful in a temperature range where the change in γ with temperature is small.

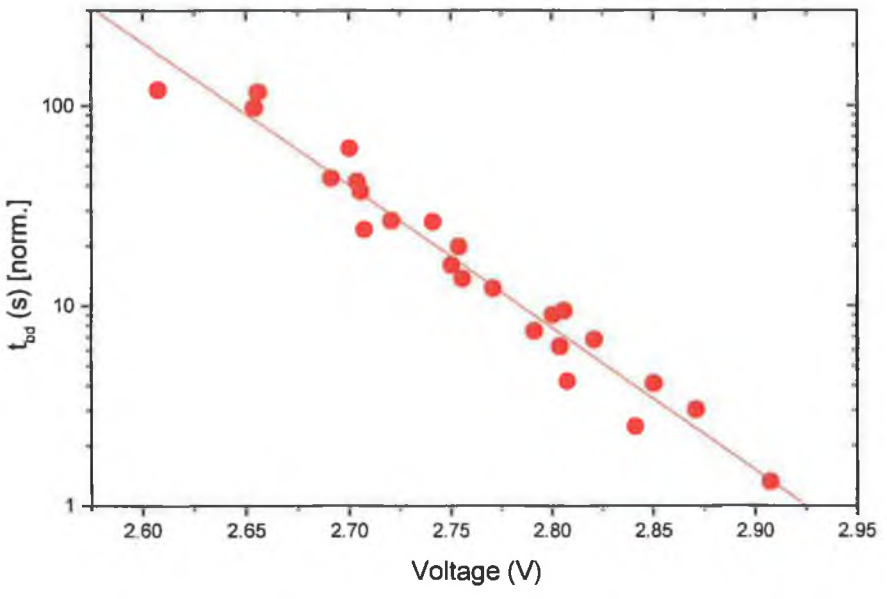


Figure 4.19: The t_{bd} data for temperatures from 20-200°C (293-473K) normalised to 100°C using a V/T constant of 400°C/V

The 'equivalent voltage' can be calculated for any voltage/temperature (where $\delta\gamma/\delta t$ is small) combination by using equation 4.2 below, where V_{act} is the actual stress voltage used, T_{act} is the actual temperature used, and T_{norm} is the temperature we wish to normalise to, with α the V/T constant.

$$V_{norm} = V_{act} + \left(\frac{T_{act} - T_{norm}}{\alpha} \right) \qquad \text{Eq 4.2}$$

This method of characterising the thermal behaviour of oxide layers could prove useful to industry as downscaling continues and temperature effects continue to be an issue.

We can use the knowledge that stress voltage and stress temperature are directly equivalent and apply it to the data shown in figure 4.14. If we convert the temperature data to equivalent voltage, we see that the temperature dependent temperature acceleration of figure 4.14 is directly equivalent to voltage dependent voltage acceleration. This supports the model proposed by Wu et al[15], by allowing the low voltage range to be probed within an acceptable timescale, by increasing the temperature of the stress.

4.6 Progressive Breakdown and Device Functionality

4.6.1 Progressive Wearout

The majority of literature on the topic of dielectric breakdown is based on the behaviour of the devices in capacitor test structures. In the past, this has been an effective way of studying them, as a hard breakdown event suffered in a thick-dielectric, large area capacitor essentially destroys the layer's insulating properties. This breakdown results in the current subsequently rising by 10's of milliamps in some cases. It is clear that only a transistor of very large area could continue to function after having undergone such a gate oxide failure, since when a gate voltage is applied, most carriers would be lost to the gate and so the formation of a conductive channel between source and drain would be unlikely. However, more recently, as this work and that of others has shown[19], experimental measurements are made at stress conditions where breakdown shows a more progressive nature, and the current at breakdown rises more slowly. This gives rise to the idea that if the current increases slowly after

breakdown at stress fields of typically 20MV/cm then it would be expected to rise much more slowly at typical operating fields of (6MV/cm) and devices could possibly continue to function, despite having suffered an oxide breakdown.

We have carried out breakdown measurements and analysed the post breakdown characteristics of capacitor devices and continued the analysis to observe the effects of breakdown on real FET devices. This work demonstrates the ability of some FETs to function as switches even after they have undergone breakdown.

For these experiments the capacitors underwent constant voltage stress, and the times to breakdown were recorded as normal. Then the stress was allowed to continue after oxide breakdown until such a time as the current had risen by 100μA from that at the moment before breakdown, and the time for the current to grow to this level ($t_{100\mu A}$) was recorded. This is illustrated in figure 4.20.

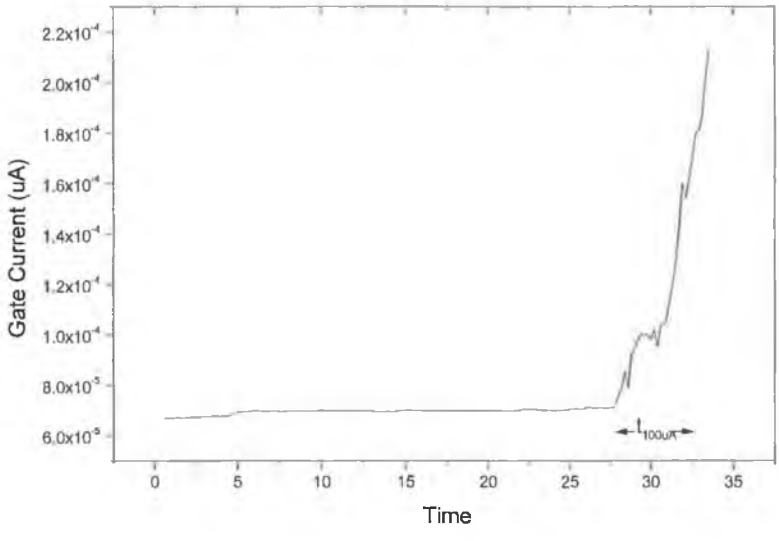


Figure 4.20. I-t trace for a 1.44nm oxide at 2.8V stress voltage. t_{100uA} is defined as the time taken for the gate current to grow by 100μA from the value recorded at the instant before breakdown.

It is found that the time for the current to reach this level is also Weibull distributed with a Weibull slope that is generally higher than the corresponding t_{bd} slope, This agrees with previous results suggesting high Weibull slopes are associated with progressive wearout. This lower statistical spread suggests that current increase following the initial breakdown is no longer solely determined by trap generation at random locations within the oxide and that the more

predictable growth of the existing breakdown path is the primary method by which the current increases.

4.6.2 Factors Affecting Progressive Breakdown

The phenomenon of progressive breakdown is voltage dependent with an average voltage acceleration of about 5.5 decades / 1V for all samples investigated, independent of thickness and nitrogen concentration. That is, for a reduction in stress voltage of 1V, the t_{100uA} will be increased by 5.5 orders of magnitude. This value is comparable with low voltage acceleration values attained in the earlier reliability study proving once again, that the values obtained were a measure of progressive wear-out voltage acceleration and not that associated with trap generation.

Figure 4.21 shows a sample of stress voltage against $t_{100\mu A}$ for one oxide recipe at different temperatures. It has already been reported that temperature causes acceleration of oxide breakdown[14], and from figure 4.21 it is clear that it also causes an acceleration of the growth of the breakdown path. For the oxide shown in figure 4.21 at 150°C and 2.6V, the t_{100uA} is of the order of 4 seconds. The t_{bd} for this dielectric recipe under the same conditions is 6.7 seconds.

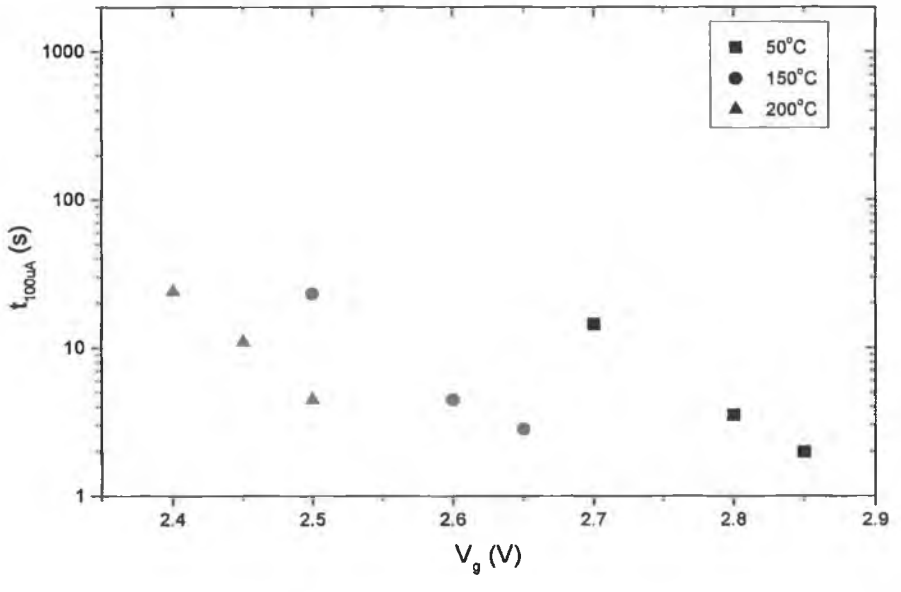


Figure 4.21 Time for the current path to allow 100μA leakage vs. Gate Voltage for one oxide recipe at different temperatures. Temperature has a large effect on growth of the conduction path after breakdown.

If using this data, we then make an extrapolation back to the low field range, and use the area and percentile scaling rules to give an industrial reliability specification, taking an operating voltage of 0.8V, we get a t_{bd} of $1.3x10^8$s (over 4 years), and a t_{100uA} of $1x10^8$s (nearly 3.2 years). From this analysis it is clear that if MOSFETs could continue to function after a gate dielectric breakdown with the gate current remaining modest, then some of this post breakdown time could be added to a reliability projection for the device, indicating that gate oxide failure isn't as much of a threat to reliability as it might initially appear.

It can be concluded that the progressive breakdown time is dependent on temperature, and the effect of elevated stress temperature is very similar in the pre-breakdown and post-breakdown regimes.

Shown in figure 4.22 is the response of all oxide recipes to varied stress temperature. The stress temperature affects the $t_{100\mu A}$ by reducing it by about one order of magnitude for every 100°C increase, to a first approximation. An order of magnitude at operating voltage represents an increase in lifetime of perhaps a few years, and thus keeping the temperature low can further slow the progressive breakdown.

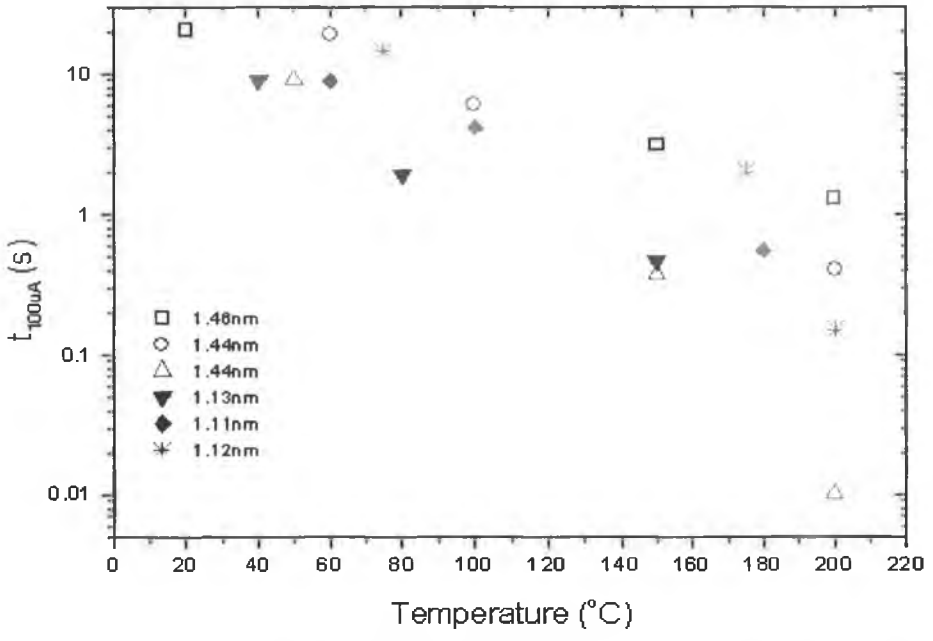


Figure 4.22 Effect of elevated temperature on $t_{100\mu A}$. An increase of approximately 100°C in stress results in an order of magnitude reduction in $t_{100\mu A}$. Open symbols are for nmos devices, closed symbols for pmos. Thicknesses are EOT.

We have shown that $t_{100\mu A}$ is related to stress voltage and temperature, two factors which also affect to t_{bd}. It follows then that t_{bd} and $t_{100\mu A}$ should be related. The $t_{100\mu A}$ was found to correlate with the t_{bd} in that an increase in t_{bd} of 1 decade, corresponds to an increase in $t_{100\mu A}$ of approximately 0.7 decades, at all voltages and temperatures investigated.

Figure 4.23 shows the correlation between t_{bd} and $t_{100\mu A}$. This plot includes all data for a number of wafers at various stress voltages and temperatures, and shows that in general a long t_{bd} leads to a longer $t_{100\mu A}$. These results indicate that a device subjected to stress conditions which cause a short t_{bd}, will suffer an accelerated post-breakdown current rise, and vice versa. In summary, the $t_{100\mu A}$ is accelerated by stress voltage and stress temperature, in a similar manner to t_{bd} and can be correlated to t_{bd}.

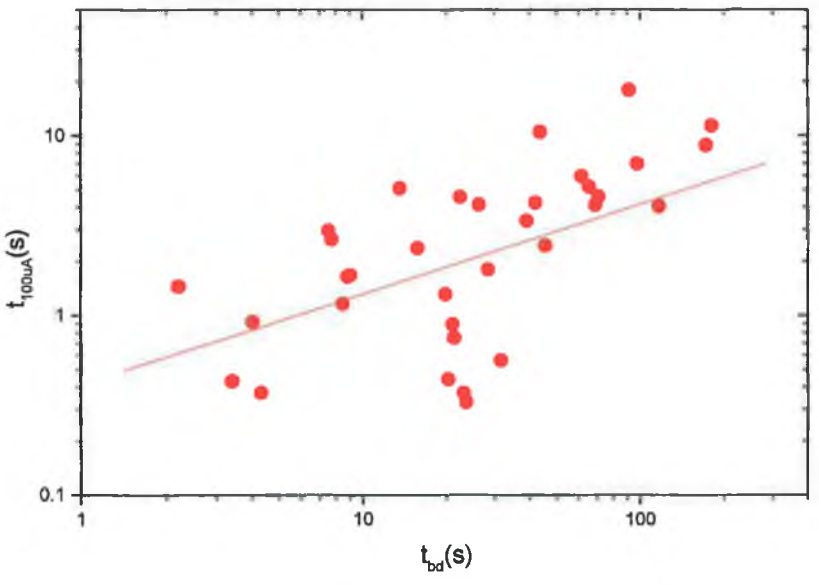


Figure 4.23. The $t_{100\mu A}$ is dependent on t_{bd}. Data was collected on a number of wafers at various stress temperatures and voltage combinations that lead to a range of t_{bd} values.

4.6.3 FET Operation in the Progressive Phase

To evaluate the functionality of n-MOSFETS after gate oxide breakdown, devices of various channel lengths and widths were studied. The devices were fabricated on the same wafers as the capacitors discussed in the previous section.

The devices had channel lengths between 0.11μm and 0.25μm and widths from 0.4μm to 10μm. The study consisted of acquiring a set of characteristic I_dV_d curves at a variety of gate voltages and examining the ratio of on-state and off-state currents as a function of dielectric degradation.

The on-state current is the drain current (I_d) with a drain-source voltage (V_{ds}) equal to 4 times the threshold voltage (~0.3V for these devices) and the same voltage applied to the gate. The off-state current is (I_d) with (V_{ds}) equal to 4 times the threshold voltage and gate voltage = 0. Typical profiles of I_d as a function of V_{ds} for on-state and off-state gate voltages are illustrated in fig 4.24. The ratio of I_{on}/I_{off} should be at least 15-20 per micron channel length to ensure accurate signal transfer in an operational device.

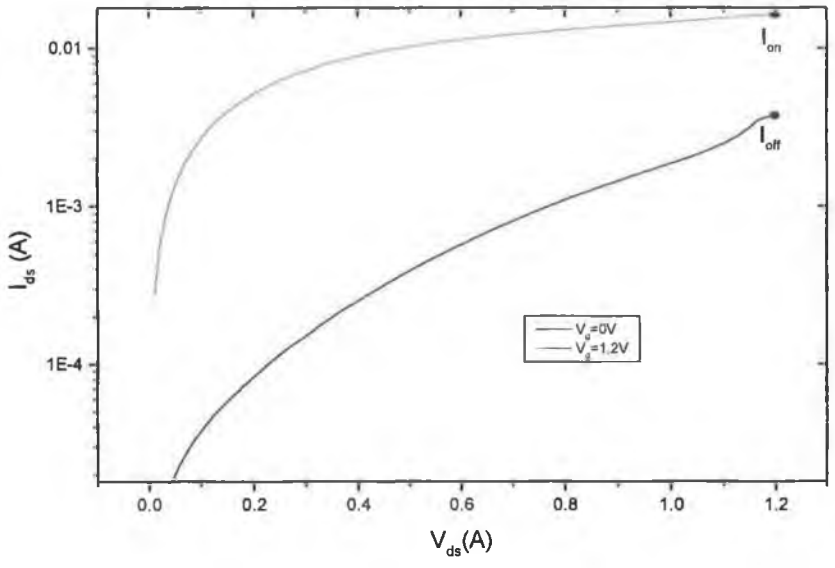


Figure 4.24 I_{on} is defined as I_{ds} with V_{ds} and V_g set at 4 times the threshold voltage. I_{off} is Ids with V_{ds} set at 4 times the threshold voltage and Vg set at zero.

After acquiring the parameters I_{on} and I_{off} from the curves mentioned above, the gate oxide was stressed at high field. The stress was interrupted and the same parameters were noted again. The parameters were re-measured at the instant of oxide breakdown and repeatedly after breakdown. It was found that in general, performance at $t_{100\mu A}$ was very similar to performance at breakdown for large channel width devices, so it was decided to continue the stress until multiple

breakdowns had occurred. The I_{on}/I_{off} ratio for a L=0.25μm W=10μm device is shown in figure 4.25, alongside the I-t trace for the stress at 2.9V.

The first point to note is that during the initial stress, both I_{on} and I_{off} are seen to rise steadily by a small amount. This is attributed to positive charge trapping at the SiON/Si interface, attracting carriers to the channel. The effect is more pronounced for short channel devices. The current ratio continues to rise very slowly until a breakdown event occurs when both I_{on} and I_{off} in some cases are altered. The usual result is a lowering of the I_{on}/I_{off} ratio, degrading transistor performance. We see an initial rise in the ratio (I_{on}/I_{off}), and then it stabilises, even though gate oxide breakdowns have occurred (at ~ 80s and 600s). The reason for this lies in the localised nature of the breakdown path, and is illustrated in figure 4.26.

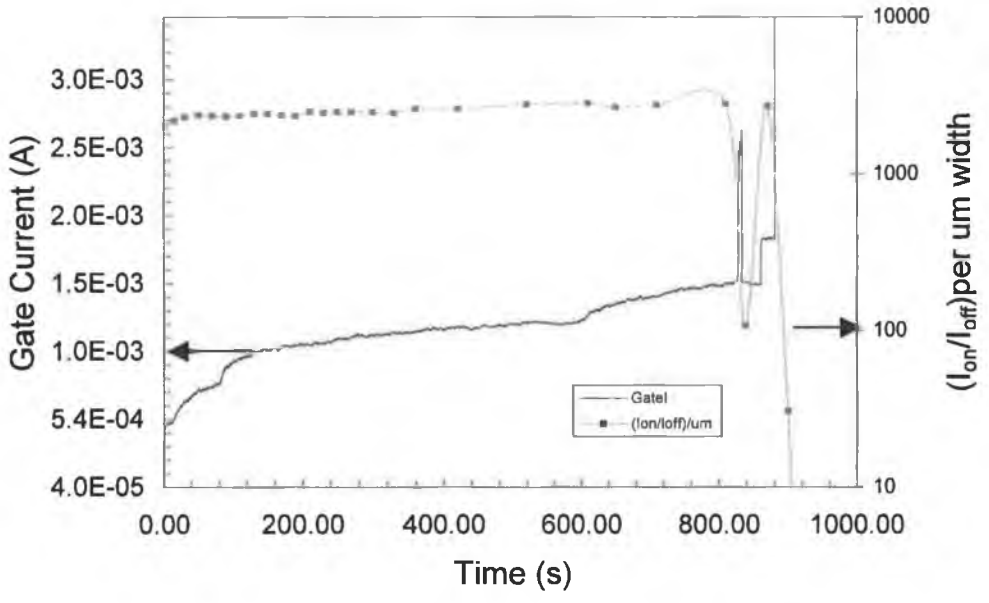


Figure 4.25 Variation of I_{on}/I_{off} with gate current during a constant voltage stress. Soft breakdown events do not affect the switching capabilities of FETs with a large channel width.

Though the gate current rises after a breakdown event (figure 4.25 at~ 80s and 600s), sometimes by up to an order of magnitude, the rise is a localised effect, and only inhibits transistor performance in the region of the breakdown path. Conduction from source to drain in the on-state is still possible in the shaded areas in the diagram, as an inversion layer can be maintained. Eventually hard breakdown occurs, and the inversion channel cannot be maintained. This leads to

a dramatic reduction in the I_{on}/I_{off} ratio, and renders the transistor dysfunctional as a digital switch. As channel width is reduced, a single oxide breakdown can have grave consequences for transistor operation. This is shown in figure 4.27 where I_{on}/I_{off} after a single oxide breakdown, is compared with I_{on}/I_{off} for a fresh sample, at various channel widths.

As the channel width is reduced, the breakdown path affects a higher fraction of the channel area until the channel width is in the sub-micron regime, when a single breakdown becomes fatal for the device.

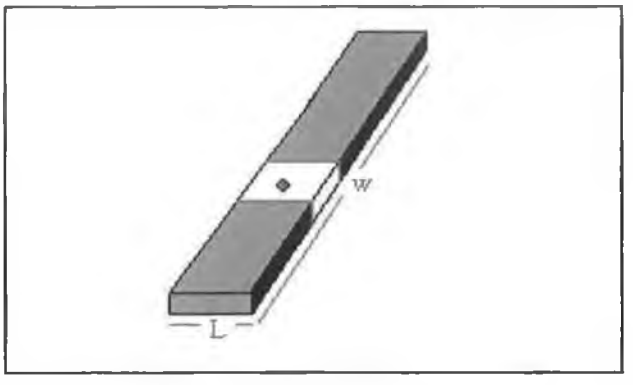


Figure 4.26 Large aspect ratio devices are essentially unaffected by a gate oxide breakdown that occurs in the channel, as the breakdown is localized event, and an inversion channel can be maintained across the majority of the device.

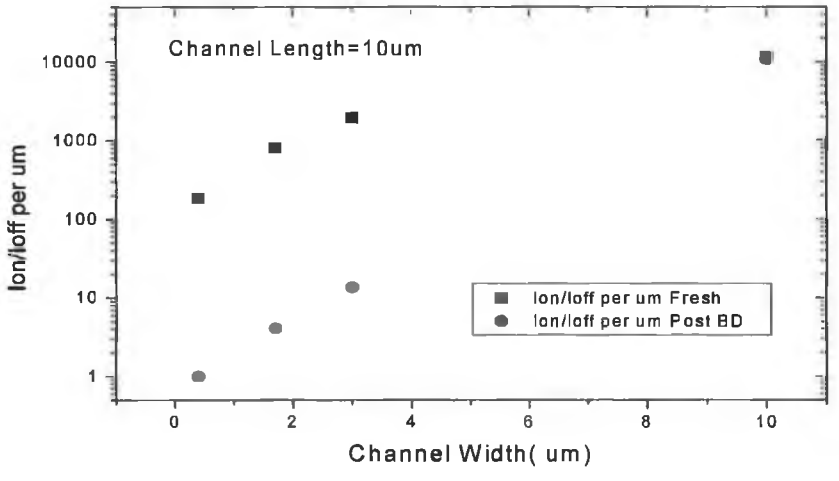


Figure 4.27 I_{on}/I_{off} at various channel widths before and after breakdown (where breakdown is defined as a large current jump). At low widths, the breakdown path affects a large fraction of the area, thus an inversion layer cannot be maintained

The fact that a single breakdown can have catastrophic effects at widths of the order of 1μm, suggests that the area affected by a breakdown is much larger than

once assumed which is in agreement with the results of Cester et al[20]. The spatial location of the breakdown path is very important in determining how the device will behave. Obviously degradation in the I_{on}/I_{off} ratio can be caused by a rise in I_{off} or a drop in I_{on}. The determining factor that causes the degradation is based on the location of the breakdown path between the source and drain.

It has been shown [21] that the location can be determined from the equation,

$$x = \frac{I_S}{I_D + I_S} \qquad \text{Eq 4.3}$$

where $0 \leq x \geq 1$, with x indicative of, but not equal to, the position of the breakdown between source and drain. x = 0 corresponds to a breakdown in the source, x = 1 corresponds to a breakdown in the drain, and 0<x>1 corresponds to a channel breakdown.

The 1st case, a breakdown in the source, has no affect on the off-state current, but on-state current is significantly lower as carriers exiting the source, attracted by the drain voltage, are also attracted by the gate voltage through the near-ohmic contact with the gate, thus lowering the measured drain current. An example of a source breakdown is shown in figure 4.28.

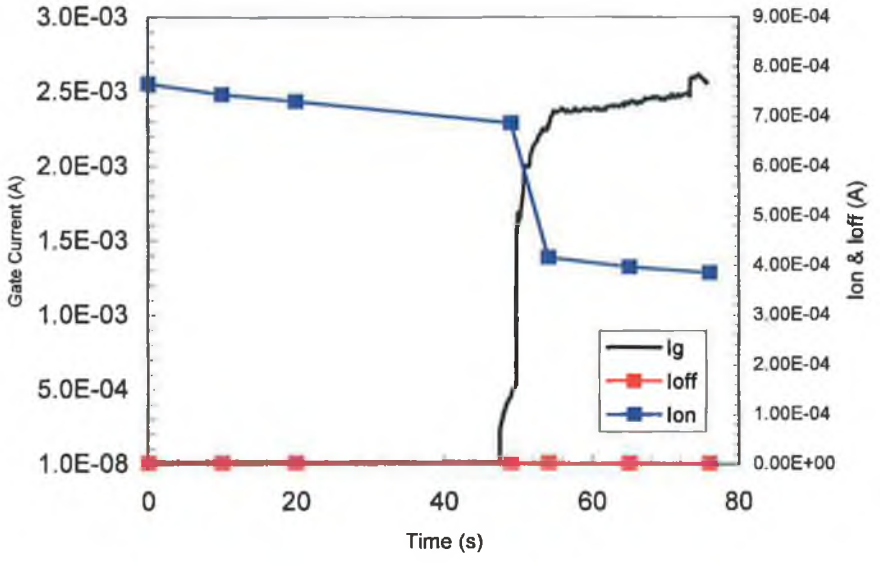


Figure 4.28 The effect of a gate-source breakdown on I_{on} & I_{off} for a W=0.1μm x L=0.4μm device

The second case, illustrated in figure 4.29 is a gate to drain breakdown. In this case, the off-state current will significantly increase, as the drain voltage attracts carriers from the gate, which can flow through the breakdown path and be detected as a drain current. The current is much larger as the flow through the nearly Ohmic breakdown path, greatly outweighs the contribution from the drain-substrate reverse bias pn-junction. In the on-state, the application of a voltage to the gate results in zero potential difference between the gate and drain, so the gate-drain current stops. Thus the transistor behaves as if there were no breakdown and I_{on} should not be affected.

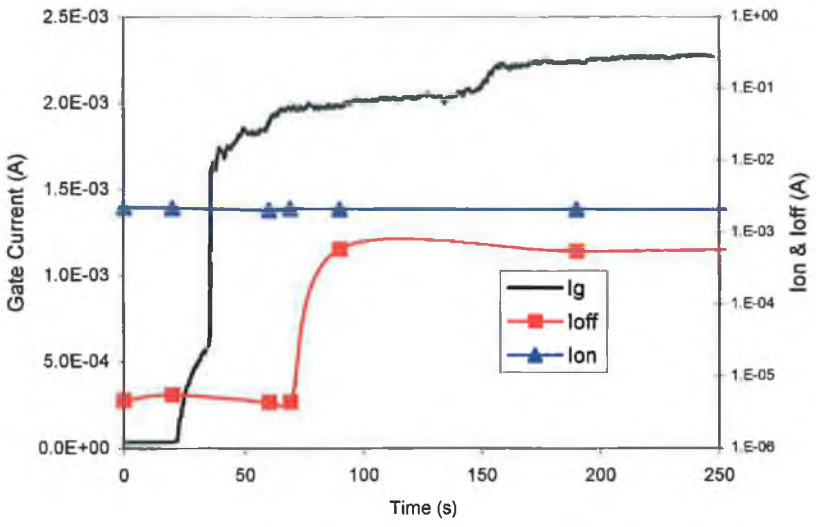


Figure 4.29 The effect of a gate –drain breakdown on I_{on}&I_{off}

The final case is a gate-to-channel breakdown. The off-current in this case is dependent on breakdown location. The closer the breakdown path is to the drain, the larger the off-current. This is because of parasitic bipolar behaviour between the gate, substrate and drain. If the breakdown is close to the drain, electrons have to travel only a small distance through the substrate and have a higher probability of being detected, than if the breakdown path was far from the drain. This effect is shown in figure 4.30, where the ratio of I_{off} before and after breakdown is plotted against breakdown location. Breakdowns in or near the source cause almost no increase in I_{off} as expected, but show a decrease in I_{on}. Channel breakdowns show both parameters are affected, and a drain breakdown shows an increase in I_{off}.

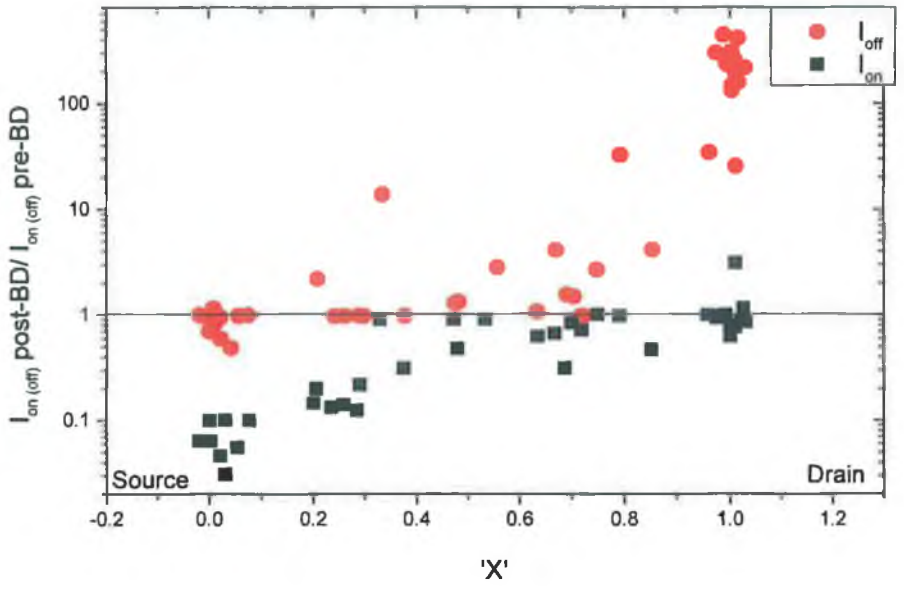


Figure 4.30. The ratio of I_{off} after and before breakdown, and also I_{on} after and before breakdown , as a function of relative breakdown position. Location x = 0 corresponds to a source breakdown, x = 1 corresponds to a drain breakdown.

For a channel breakdown, the on-current reduction is dependent on channel width, as discussed earlier. If the channel width is sufficient, an inversion layer can be maintained away from the breakdown path and the device can continue to function, with only a slight reduction in I_{on}. For small width devices, I_{on} is reduced significantly. Thus, the aspect ratio, and breakdown location all affect the post-breakdown behaviour when a gate-channel breakdown occurs.

Generally, we could not find any evidence that device operation was significantly different at $t_{100\mu A}$ than at t_{bd} for any of the areas studied. In devices with wide channels, multiple breakdowns are required to cause any degradation in device performance. In narrow channel devices, a single breakdown is enough to cause catastrophic failure of the device before it gets to $t_{100\mu A}$ and any further progression of the breakdown is found to make very little difference to I_{on} and I_{off}.

4.7 Impact of breakdown on circuit operation

If we take the rather simple case of a CMOS NAND gate shown in figure 4.31[22], and examine the impact of a gate oxide breakdown at various locations in a particular FET, we see that different locations will have different effects on the logic operation.

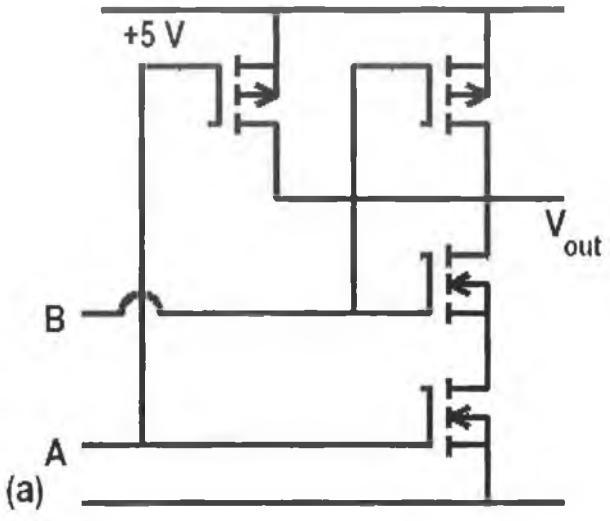


Figure 4.31. A CMOS NAND gate, constituting 4 MOSFETS. A gate oxide breakdown in any of the FETs could cause erroneous outputs from the device.

In this circuit the upper two FETs are normally on p-FETs, and the lower two are normally off n-FETs. Thus, only if a voltage is applied to inputs A and B will the circuit be grounded giving a zero output at V_{out}. However, in the case of a gate oxide breakdown in the source of the lowermost FET, input A will be constantly connected to ground, meaning that V_{out} will be +5V no matter what combination of inputs are used. In the case of a drain breakdown, if both A and B inputs are connected, A is connected directly to V_{out}, but as shown earlier, an inversion channel can be maintained with a drain breakdown, and so the device will be grounded and function correctly giving a 0 output. However if only input B is connected, the increased off-state current associated with a drain breakdown may cause the device to act as if it is grounded, though this is quite speculative. It is more difficult to predict what will happen if a channel breakdown occurs, as the

outcome will vary depending on the breakdown location along the length of the device. It is clear then, that an oxide breakdown in this simple device can have affect the logic operation, and the truth table for the device. Breakdowns in the other FETs will have different consequences affecting the output in different ways. Though the manifestation of these device level problems are impossible to predict given the relative complexity of a modern logic circuit to the one we have studied, it remains clear that gate oxide breakdown can affect logic functions, and there is due cause in researching new devices and materials to ensure that oxide breakdown doesn't become the limiting factor in the progress of the semiconductor industry.

4.8 Summary

This chapter outlined reliability concerns for ultra-thin SiON layers. There was very little difference in performance for sample prepared in various manners, except for the expected thickness effects. Initial reliability estimates using Weibull slopes, voltage acceleration, and extrapolation were found to give unexpectedly high Weibull slopes and low acceleration factors. Firstly, the possibility of an unexpectedly high trap generation rate was investigated, but this was ruled out by charge pumping measurements, and by monitoring the Weibull slope as a function of breakdown trigger current. The values obtained were subsequently found to be due to the time interval between the formation of a conduction path, and a significant increase in gate current. The breakdown triggers used were within this time interval and thus the measured Weibull slopes were a reflection of the growth of the conduction path, and not trap generation within the oxide before breakdown.

Increased temperature was found to accelerate the breakdown significantly and the measured increase in Weibull slope for increasing temperature was interpreted once again as being due to using a breakdown current trigger that was in the post breakdown phase. Temperature and voltage were linked using a simple relationship, whereby increasing the temperature of a stress by 40K was

found to be directly equivalent, in terms of breakdown time, to increasing the voltage by 0.1V.

The growth of gate current after a gate oxide breakdown was studied in terms of the rising of the leakage current to 100μA, and the factors that accelerate the growth of a breakdown path were investigated. It was found that the same factors that accelerate breakdown are responsible, i.e. temperature, and gate voltage.

The pre- and post-breakdown operation of a variety of transistors with different areas was investigated. For large area devices, progressive breakdown is not an issue as multiple breakdown paths are required to degrade the functionality of the device. In small area devices, simply the formation of breakdown path is enough to destroy the device, and the growth of the path causes no further degradation. Furthermore, we have outlined what effect breakdown path location has on I_{on} and I_{off}, with breakdowns at or near the source destroying the device by reducing I_{on}, and breakdowns at or near the drain causing an increase in I_{off}.

4.9 References

[1]A. Veloso, F. N. Cubaynes, A. Rothschild, S. Mertens, R. Degraeve, R. O'Connor, C. Olsen, L. Date, M. Schaekers, C. Dachs, M. Jurczak, Ultra-thin Oxynitride Gate Dielectrics by Pulsed-RF DPN for 65 nm General Purpose CMOS Applications, *Proc. of the 33rd ESSDERC'2003*, pp. 239-242

[2]Y. He, M. Xu and C. Tan, Effects of plasma nitridation on ultra-thin gate oxide electrical and reliability characteristics, Solid-State Electronics, In Press.

[3] F. Pio, L. Ravazzi, C. Riva, Influence of series resistance in oxide parameter extraction from accelerated tests data, Proc. ESREF' 92, pp. 105-108, 1992.

[4] C. Salm J. H. Klootwijk, Y. Ponomarev, P. W. M. Boos, D. J. Gravesteijn, and P. H. Woerlee, Gate Current and Oxide Reliability in p+ Poly MOS Capacitors with Poly-Si and Poly-Ge0.3Si0.7 Gate Material, IEEE Elec. Dev. Lett. Vol. 19 No. 7 July 1998. pp. 213-215.

[5] Ph. Roussel, R. Degraeve, G.V. Van den Bosch, B. Kaczer ,G. Groeseneken, Accurate and robust noise-based trigger algorithm for soft breakdown detection in ultrathin gate dielectrics, IEEE Trans. Devic and Materials Reliability 1 (2) (2001) pp. 120-127.

[6] M.A. Alam, J. Bude, A. Ghetti, Field acceleration for oxide breakdown – Can an accurate anode hole injection model resolve the E vs. 1/E controversy?, *Proc. IRPS*, pp.. 21-26, 2000.

[7]R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, H.E. Maes, New insights in the relation between electron trap generation and the statistical properties of oxide breakdown, IEEE Trans. Elec. Dev. 45(4) (1998) pp. 904-911.

[8]J.H. Stathis, Quantitative model of the thickness dependence of breakdown in ultra-thin oxides, Microelectron. Eng. 36 (1997) pp. 325-328.

[9] R. Degraeve, B. Kaczer, G. Groeseneken, Degradation and breakdown in thin oxide layers: mechanisms, models and reliability prediction, Microelectron. Reliab. 39 (1999) pp. 1445-1460.

[10] R. Degraeve, B. Kaczer, F. Schuler, M. Lorenzini, D. Wellekens, P. Hendrickx, J. Van Houdt, L. Haspeslagh, G. Tempel and G. Groeseneken , Statistical model for SILC and pre-breakdown current jumps in ultra-thin oxide layers, IEDM Techn. Dig. (2001) pp. 121-124.

[11] D. J. DiMaria, J. H. Stathis, Explanation for the oxide thickness dependence of breakdown characteristics of metal-oxide-semiconductor structures, Appl. Phys. Lett. 70 (20) (1997) pp. 2708-2710.

[12] B. Kaczer, R. Degraeve, R. O'Connor, Ph. Roussel, G. Groeseneken, Implications of progressive wear-out for lifetime extrapolation of ultra-thin (EOT~1nm) SiON films, Accepted to IEDM 2004.

[13] B. Kaczer et al Proc. ESSDERC 1999, pp. 356-359.

[14] E. Wu, J. Suñé, W. Lai, E. Nowak, J. McKenna, A. Vayshenker and D. Harmon, Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin oxides, Mircoelectron. Eng. 59 (2001) pp. 25-31.

[15] E. Wu, J. Aitken, E. Nowak, A. Vayshenker, P. Varekamp, G. Hueckel, J. McKenna, D. Harmon, L.-K. Han, C. Montrose, R. Dufresne, and R. -P. Vollertsen, "Voltage-Dependent Voltage Acceleration of Oxide Breakdown for Ultra-Thin Oxides," *IEDM Tech. Digest*, pp. 54–57 (2000).

[16] G. Bersuker, Y. Jeon and H. R. Huff, Degradation of thin oxides during electrical stress, Microelectron. Reliab. 41 (2001) pp. 1923-1931.

[17] D.J. Dumin, J.R. Maddux, R.S. Scott, R. Subramoniam, A model relating wearout to breakdown in thin oxides, IEEE Trans. Electron Dev, 41 (9) (1994) pp. 1570-1580

[18] R. Degraeve, Private Correspondence

[19] F. Monsieur, Degradation mechanisms during the ultra-thin oxide breakdown, Proc. INFOS 2003, pg. IT1.

[20] A. Cester, S. Cimino, A. Paccagnella, G. Ghiaini, and G. Guegan, Collapse of MOSFET Drain Current After Soft Breakdown and its Dependence on the Transistor Aspect Ratio W/L, Proc. IRPS 2003, pp 189-195

[21] B. Kaczer, R. Degraeve, M. Rasras, A. De Keersgieter, K. Van de Mieroop and G. Groeseneken, Analysis and modeling of a digital CMOS circuit operation and reliability after gate oxide breakdown: a case study, Microelec. Rel. 42, Iss. 4-5, (2002) pp. 555-564.

[22] Brian Lawless, Fundamental Analogue Electronics, Prentice Hall, 1997.

Chapter 5: Electrical Characterisation of Hafnium Silicate Dielectric Layers

5.1 Introduction

The search for a replacement dielectric for SiO_2 is an active area of research. It has been stated that the only problem with SiO_2 is its relatively low-k value, and the only benefit of high-k materials is the fact that the dielectric constant is high. While this statement is not strictly correct, it illustrates the problem with high-k materials. Many materials have been considered thus far with significant research into Al_2O_3, Ta_2O_5, TiO_2, HfO_2, and ZrO_2 to name but a few. However, no clear candidate has emerged, because, in replacing SiO_2, new materials must meet a number of requirements, which are not easy to fulfill.

The alternative material must be thermodynamically stable on the Si substrate, forming a high quality interface with a low interface state density (ideally less than $5x10^{10}/cm^2$). One of the major problems with high-k materials is the formation of a metal-oxide/Si_xO_y transition region at the substrate interface, which lowers the dielectric constant. However, due to the excellent properties of the SiO_2/Si interface, gate stacks of the form metal-oxide/SiO_2/Si have been investigated, as good channel mobility can be maintained with such an interface[1]. The material must also have a sufficiently high barrier to tunnelling electrons, as leakage current depends exponentially on this barrier height. The barrier heights for selected materials taken from Robertson[2] are shown in figure 5.1. Clearly, the most suitable material from this point of view is Al_2O_3, but it has a dielectric constant of ~8, only twice that of SiO_2. So it would be at best, a short-term solution.

Materials with extremely high k values have also been extensively studied, for example TiO_2 (k= ~80) and HfO_2 (k = ~30). However TiO_2 has displayed very high leakage currents, poor mobility and proves difficult to grow uniformly[3], and HfO_2 has shown similar problems[4].

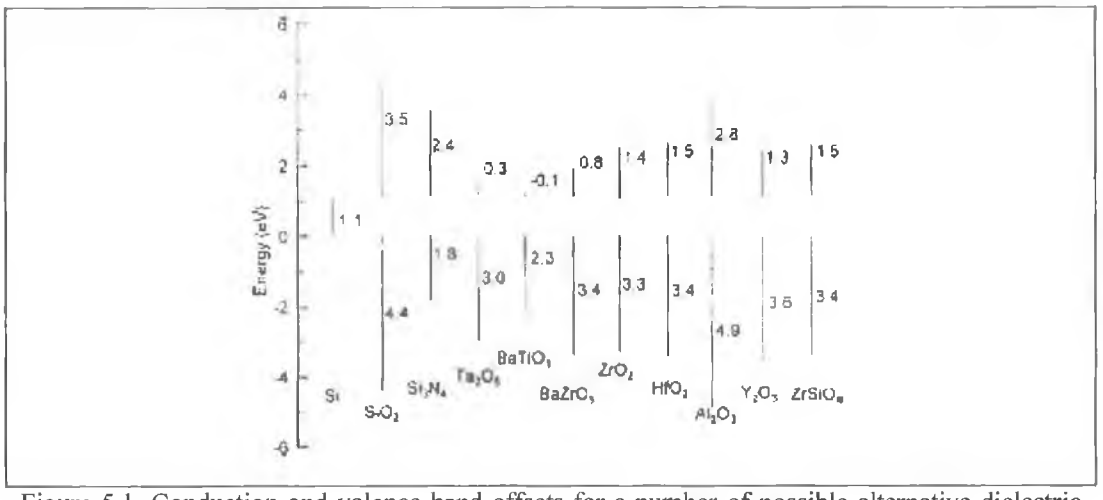


Figure 5.1. Conduction and valence band offsets for a number of possible alternative dielectric materials, from Robertson[1].

Other issues with materials with such high-k values include high interface trap densities and excessive threshold voltage shifts in devices. The problems with implementing metal oxides as alternative dielectrics mean that a short-term alternative may be necessary. To this end, silicates are currently the subject of a lot of research[5]. These materials are a mixture of metal oxide with SiO_2, which take advantage of the properties of the SiO_2 while benefiting from the high dielectric constant of the metal oxide. Resulting layers usually have a k value of about 10-15 and show good stability and low defect densities [4]. This chapter examines the suitability of HfSiON layers with EOT from 1.4-2.1nm as a short-term replacement dielectric.

5.2 Sample Preparation

The devices used for these experiments were both square and rectangular gated transistors with polysilicon gates with areas from $10^{-4}cm^2$ to 10^{-8} cm^2. Polysilicon gates have thus far proved incompatible with many high-k layers, because of the thermodynamic instability they show with silicon[6], forming interfacial layers, however silicates are reported to show better compatibility with polysilicon gates. The devices were grown on test wafers used for evaluating sub-90nm processes at IMEC.

The wafers were prepared by carrying out an IMEC clean which leaves a layer of SiO_2. The HfSiO layers were then grown using MOCVD in a standard 8-inch wafer MOCVD reactor with showerhead at 600°C and 7.25 Torr. Substrate resistivity was 20±4 Ωcm. Hafnium tetra-diethylamide (TDEAH) and tri(dimethylamino)silane (TDMAS) were used as precursors, resulting in HfSiO-layer physical thicknesses from 1.0nm to 3.0nm. Then, a post deposition anneal was carried out in a NH_3 atmosphere at 800°C for 1min. The processing sequence used, has led to thicknesses of 1.38-2.14 nm Equivalent Oxide Thickness (EOT). CV measurements show a dielectric constant of 11.3 for the layers.

All samples were n-MOSFET structures with p-type substrates and contacts made with source, drain, gate and substrate. Device sizes used in experiments are detailed throughout the chapter.

5.3 Reliability

An initial Current-Voltage (I-V) sweep was performed on all thicknesses investigated (1.38nm, 1.53nm, 1.67nm, 2.14nm) to examine the leakage current through the layers, compared with that in conventional SiO_2 and SiON layers. Figure 5.2 shows the I-V curves for a 1.6nm HfSiON sample, compared with SiON layers of similar EOT discussed in chapter 4, and for thicker layers of SiO_2. The benefit of using HfSiON in terms of leakage current is evident, with many orders of magnitude in leakage current reduction at 1V.

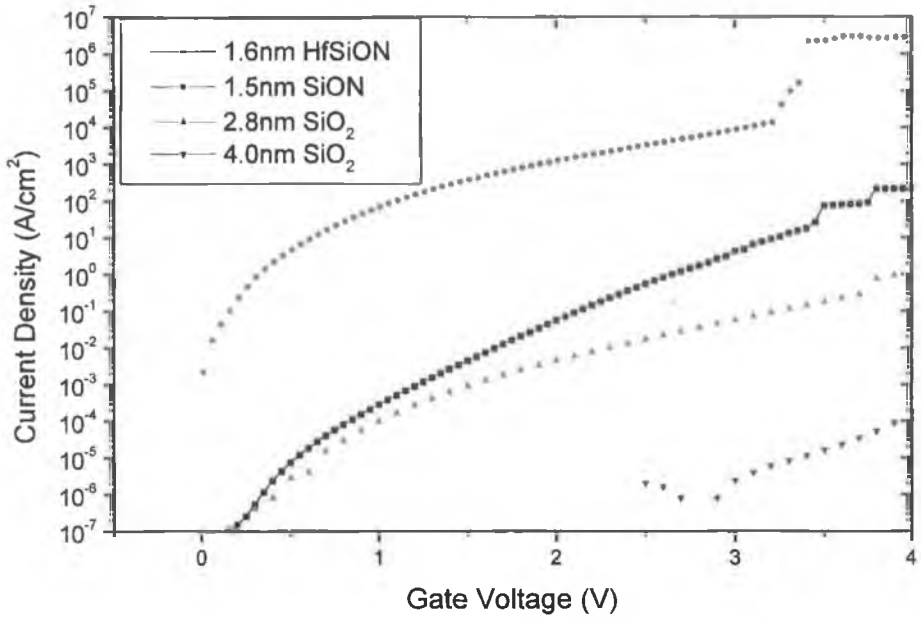


Figure 5.2. I-V curves for gate dielectric layers of varying thickness. HfSiON shows an improvement in leakage current over SiON of similar EOT.

The predominant current tunnelling mechansim was investigated by performing I-V sweeps at high-temperature. If direct tunnelling dominates, no temperature dependence should be evident. However if some trap-assisted method is contributing to conduction (for example Poole-Frenkel conduction due to the thermal excitation of traps) then temperature dependence should be observed. The results in figure 5.3 show that for the thinnest layers, there is no temperature dependence and direct tunnelling dominates. For thicker layers, the leakage current is higher at high temperatures showing that traps assisted tunnelling is present in the conduction mechanism even before any stress has been applied to the layer.

Instead of making the measurement at V_g =1V, the change in threshold voltage (V_t) during stress was taken into account, and results are taken at a gate voltage such that V_g-V_t=1V, so the substrate is always 1V in accumulation above V_t.

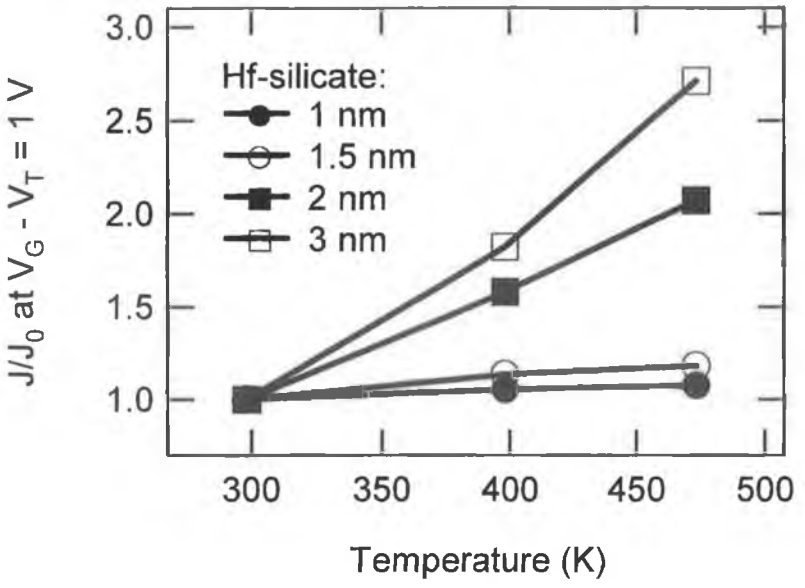


Figure 5.3.Temperature dependence of leakage current in HfSiON layers. Thin layers show no temperature dependence (characteristic of direct tunnelling) and thick layers show some temperature dependence indicating that traps in the layer assist the tunnelling process.

Reliability was evaluated using accelerated lifetime constant voltage stress measurements similar to those discussed in chapter 4. A sample I-t trace at +3.3V for a gate dielectric thickness of 2.13 nm, with channel width and length of 10μm is shown in figure 5.4. The fall in the measured current following the application of the stress voltage is attributed to the buildup of negative charge in the dielectric layer. In this particular example, breakdown occurs at a lower current than the initial current, making current level breakdown detection difficult. In thinner dielectric layers, this initial decrease in stress current is much less pronounced as charge trapping has a much less significant effect. The occurrence of features such as that at ~400s is not well understood but may indicate the presence of a switching process that is related to trapping and de-trapping of charge within the oxide.

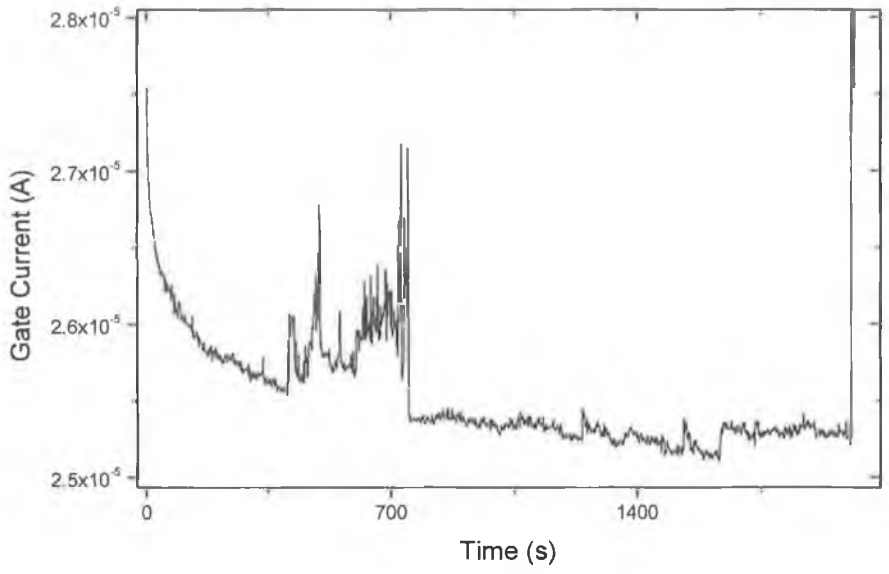


Figure 5.3: An I-t trace for a 2.13nm HfSiON layer. Negative charge-trapping causes an initial decrease in gate current.

A reliability estimate for a 1.67nm oxide is show in figure 5.4 by plotting V_g vs. t_{bd}. The data points in the figure represent real t_{bd} data. Using the same analysis as for SiON, by simple linear extrapolation and applying the area and percentage scaling rules, the maximum operating voltage at the specification condition of 10 years ($3x10^8$s) lifetime, fit to a $0.1cm^2$ area (dashed line) and a 0.01%failure condition (bold line) was determined. In thin layers (<1.5nm), no clear breakdown occurs and breakdown detection again proves difficult, with the current rise being progressive in nature. Care must be taken in extracting the correct Weibull slopes and voltage acceleration values. From figure 5.4 the 10-year maximum operating voltage can be extracted to be approximately 0.85V. Within error bars, this is almost the same as for SiON of similar EOT.

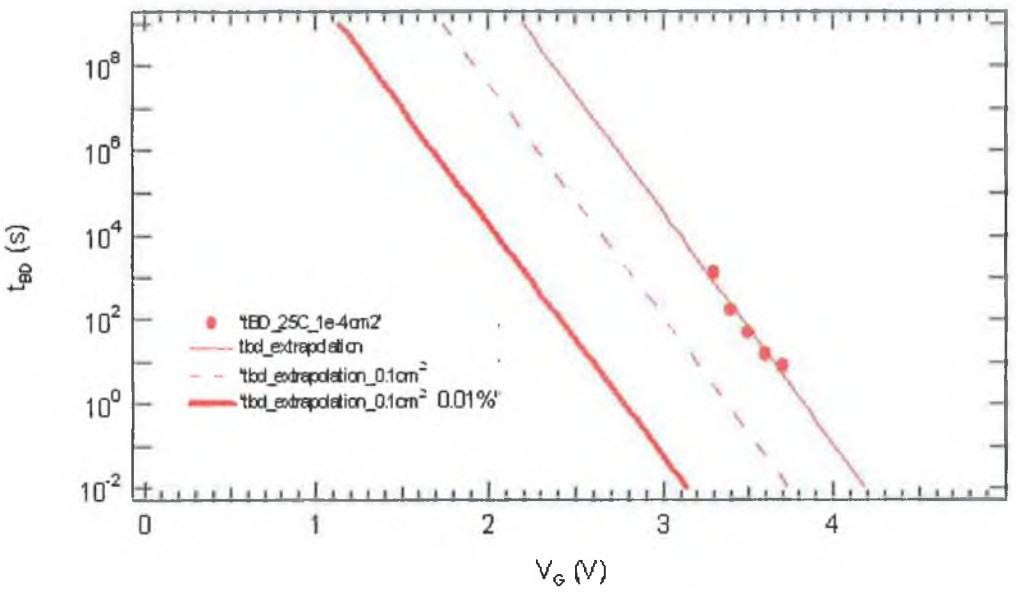


Figure 5.4: Example of measured t_{BD} data for a 2.13nm HfSiON layer and extrapolation. The slope of this fit is γ. The dashed and bold lines are calculated for an area of 0.1 cm^2 and for 0.1cm^2 and 0.01% failure respectively. Breakdown is taken as a large current jump, such as that in figure 5.3 at ~2000 sec.

5.3.1 Weibull Slopes

For all samples, the Weibull slopes were quite low (~0.9-1.2). Following the percolation model[7], the thinnest oxide, which has a physical thickness of about 3.6nm, should have a Weibull slope of about 3 as a chain of ~4 traps is needed to cause breakdown. The measured Weibull slope had a value of ~1.12. A possible explanation for this discrepancy is now proposed. Considering the percolation model, and knowing that the oxide physical thickness is ~3.6 nm the following conclusions can be drawn. The trap density D_{ot} at which breakdown occurs is Weibull distributed with a slope β_{ot} equal to the number of traps in the path [10] as for SiON. For the HfSiON of thickness 3.6nm, and assuming the same 'radius of conduction' for a trap of 0.45nm as predicted by the percolation model, it follows that 3 (perfectly aligned from anode to cathode) or more likely 4 traps are needed to form a breakdown path. Using a similar trap generation rate to SiO_2 would lead us to believe that the Weibull slope should be about 3 and not 1.12 as we measured. The reason for this once again lies in the problem with detecting the real breakdown. If we choose a small current jump as our breakdown criterion, it is possible that a conduction path has not opened up from anode to cathode, but that a small path of two or three traps has formed within the oxide, locally

lowering the gap for tunnelling carriers and causing an increase in the current magnitude and noise. The time to observe the 1st current step higher than a given magnitude (ΔI), is also Weibull distributed, but as small current steps can be triggered by conduction paths with less traps, the slope of this distribution ($\beta_{\Delta I}$) will be lower at low current step[8]. To observe the relationship between β and ΔI, we analysed the I-t traces (figure 5.5) of 30 identical samples and measured the development of β as the current increases.

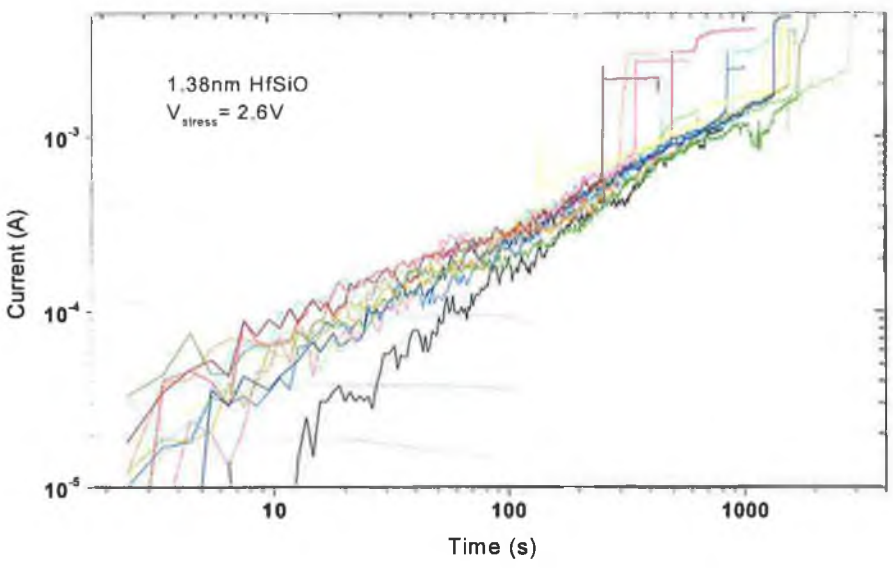


Figure 5.5: I-t traces used to monitor the development of β during a CVS. Only some of the 30 are shown, for clarity.

The results are shown in figure 5.6, which plots the calculated Weibull slope for a range of current steps. The Weibull slope starts at about 1 for very low current steps. It is difficult to measure the Weibull slope at steps lower than $1x10^{-5}$ A as the level of noise is too high. As the stress continues, the Weibull slope increases and paths with more than 1 trap begin to form and so on until a breakdown event occurs. Beyond this point, we are into the progressive wearout stage, and the measured Weibull slope begins to reflect the growth of the wearout path in a similar manner to SiON. It is evident then, that selecting the breakdown criteria is especially difficult in these physically thicker layers, as slopes can be underestimated when a low current trigger is used, and overestimated (like SiON) when an excessively high trigger is used. It is possible that the plateau regions in figure 5.6 correspond to the regions where 1 trap, 2 trap, and 3 trap

conduction paths dominate. When 4-trap paths are active, they will generally cause breakdown, and thus the plateau for 4-trap paths is very short, and the measured slope is in the progressive breakdown phase. Due to the error bars associated with extraction of Weibull slope, this measurement is not conclusive on this point. However, from the figure, the number of plateaus is consistent with the number of traps required to cause breakdown.

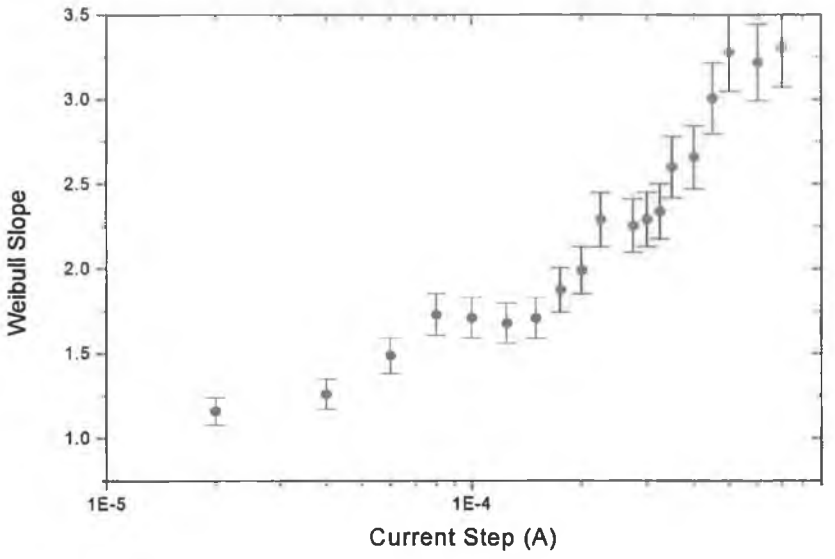


Figure 5.6. For increasing current step magnitude during a CVS, the Weibull slope rises in steps, that correspond to regimes where conduction paths of single traps, two traps, and so on, are dominant, until breakdown. Breakdown requires a 4-5 trap conduction path.

5.3.2 Voltage Acceleration

Figure 5.7 examines the voltage acceleration for all thicknesses studied. Very low values of voltage acceleration are found for HfSiON, indicating that the breakdown trigger (i.e. a significant rise in gate current) used may be in the post breakdown phase as was found with SiON layers. The thinnest layer, used in the experiment above, shows the highest voltage acceleration, as expected, because the measurements were made during the trap generation stage. The thickest layer shows γ close to the curve for SiO_2, and its Weibull slope is once again in the percolation regime. Only the layers of 1.53 and 1.67nm show unusually low γ for measurements taken with a very low current step. A possible explanation for this

may be the level of noise in the gate current, or simply the voltage acceleration curve (figure 5.7) is shifted to the left for this material.

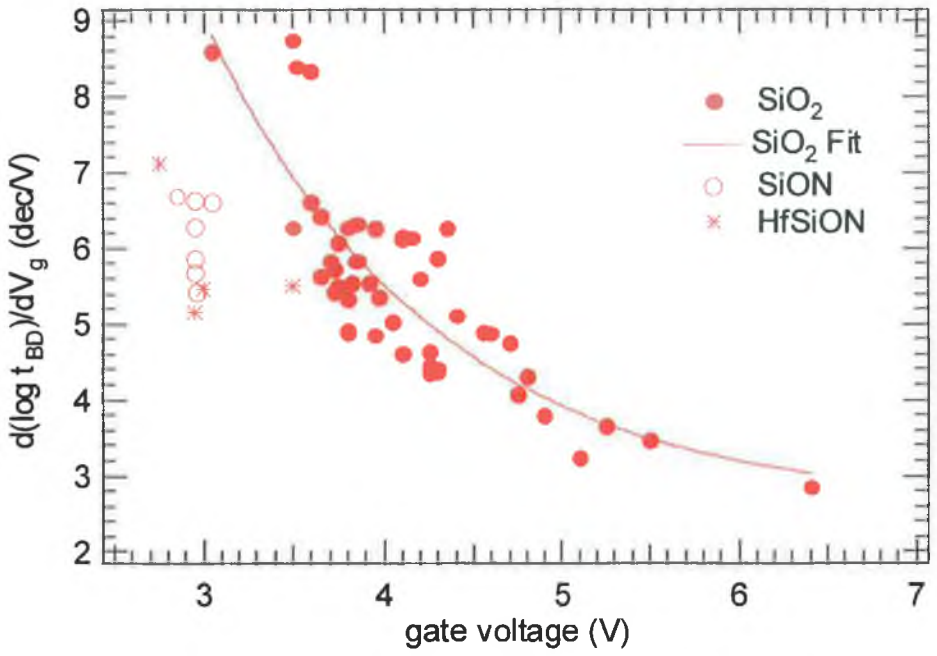


Figure 5.7: Comparison of γ for SiO_2 (closed circles) SiON (open circles) and HfSiON (stars) shows the HfSiON has similar voltage acceleration to SiON, much lower than for SiO_2.

For HfSiON, the t_{bd} measurement problem can be summarised as based on presented results. If the current trigger used is above that which really signifies a breakdown, we are measuring in the wearout phase and a high Weibull slope is observed. At high current step, the voltage acceleration values are found to be similar to those at low current step. Presumably, this means that the voltage acceleration of path wearout is roughly the same as that of trap creation. Calculated 10-year max operating voltages using the measured Weibull slopes and acceleration factors are shown in figure 5.8, at room temperature and up to 125°C. The lower time-to-breakdown and slightly reduced acceleration factors at high temperature means the reliability outlook suffers. Notably, the thickest layer suffers the greatest decrease in t_{bd} (and thus max operating voltage) at high temperature. The reason for this is examined in the next section on temperature accelerated breakdown. The results in figure 5.8 show that based on this method of evaluating reliability, none of the layers meet the requirement for ~0.8V at operating temperature.

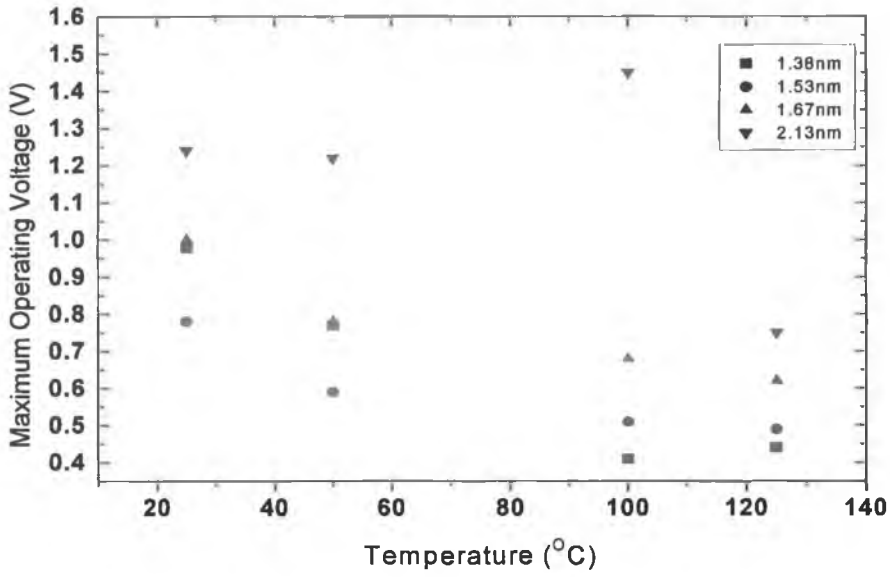


Figure 5.8: Maximum operating voltage, deduced from reliability extrapolations. At operating temperature using this reliability evaluation, none of the oxides meet the 0.8V specification.

5.4 Temperature Acceleration of Breakdown

As with SiON, elevated temperature causes an acceleration of the degradation and breakdown processes. It is of course important to understand the behaviour of the layers at operating temperature to examine how favourably they compare with SiO_2 and SiON. This section examines the results for the thickest (2.14nm) and thinnest (1.38nm) layers studied and a comparison with SiON is made.

Shown in figure 5.9 is the time to breakdown at a number of voltages for the mentioned layers at both 100°C and 200°C. As expected the t_{bd} scales with thickness, but the data for different temperatures seems to converge for decreasing thickness. For the 1.38nm HfSiON for example, there is almost no temperature dependence, as the two data sets are very close together.

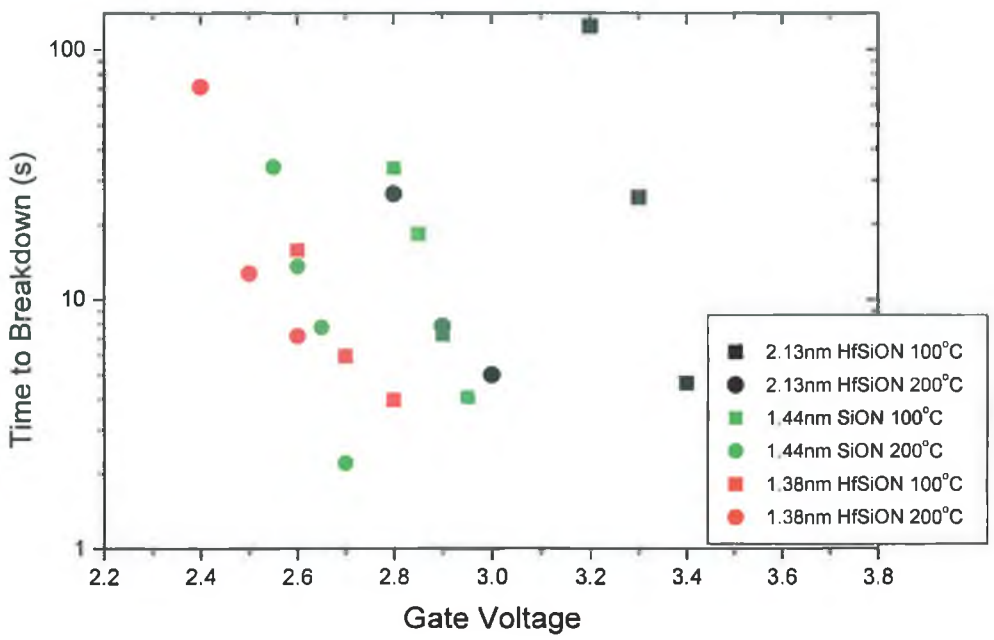


Figure 5.9: Temperature dependence of t_{BD} for SiON and HfSiON. The temperature dependence was found to be thickness dependent. In the thinnest layers t_{BD} is much less dependent on stress temperature.

Thus, the relationship between stress temperature and voltage described in chapter 4 seems incomplete. It was deduced in chapter 4 for SiON that a fixed breakdown time could be achieved for a number of temperature/voltage conditions, with an increase in temperature of ~40°C being equal to a voltage rise of 0.1V. However, on examining a number of thicknesses for HfSiON layers, we find that the constant linking the two is in fact thickness dependent.

Figure 5.10 illustrates the thickness dependence of the V/T constant. It can be seen that for increasing thickness, a change in temperature causes a rapid acceleration of the breakdown with a change of only 200°C being equivalent to a 1V change, whereas in the thinnest layer, a change of almost 600°C is required. Available data points for SiON and SiO_2 are also shown. A plausible explanation of this trend would be that the increased temperature causes more trapping in the bulk of the layer, and as the layer is made thinner with less bulk, the temperature dependence becomes less important. This would suggest that rather fortuitously, thinner layers are more suitable for higher temperature operation. The effect of elevated temperature on device operation is discussed later in this chapter.

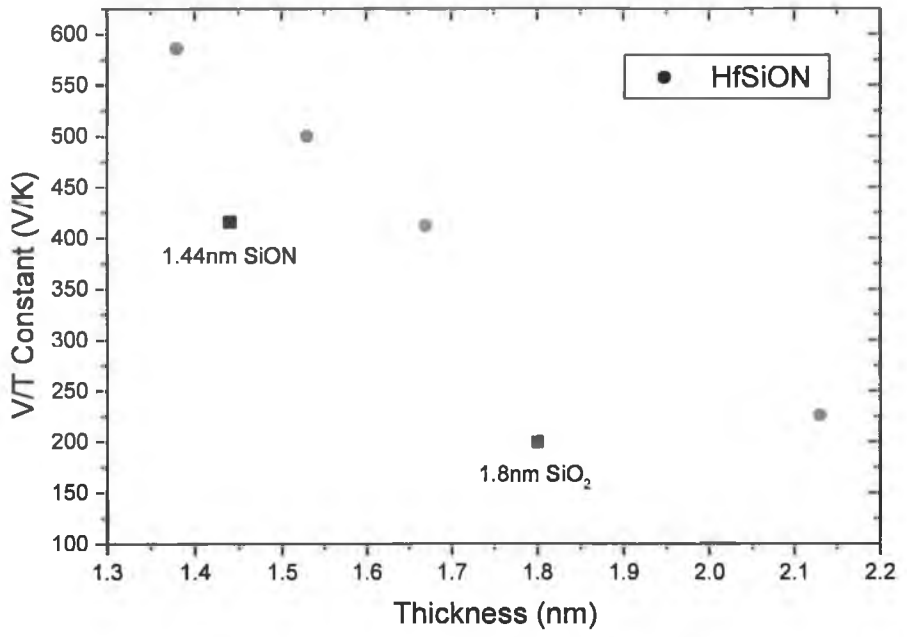


Figure 5.10: The V/T constant for varying thickness shows the thickness/temperature dependence of oxide breakdown.

5.5 Degradation of FET Parameters during stress

The reliability study shows that the hafnium silicate layers under consideration do not meet reliability requirements. However, if we assume that devices will continue to function satisfactorily in the post breakdown phase, then it may be still possible to use such layers. This section examines the degradation of FET parameters during the lifetime of the layer, as even if the insulating properties are maintained, other factors may cause device failure within this time frame.

To evaluate the functionality of n-MOSFETs during a CVS or CCS, transistors of various channel lengths (10μm-0.25μm) and widths (10μm-0.25μm) were studied. The study was similar to that described in chapter 3 for SiON, consisting of I_dV_g measurements and I_dV_d measurements at a variety of gate voltages and examining the ratio of on-state and off-state currents. From the set of I_dV_g curves (figure 5.11) taken as a function of increasingly long voltage stress times, the

change in threshold voltage, transconductance degradation and the sub-threshold swing (the change in gate voltage required to change the gate current by one order of magnitude, in the sub-threshold regime). By monitoring I_{on} and I_{off} the functionality of the device during oxide degradation into the post breakdown wear-out phase was assessed and the impact of device aspect ratio examined. The sample set of I_dV_g curves shown in figure 5.11 suggests a very large threshold shift, even for very low stress times.

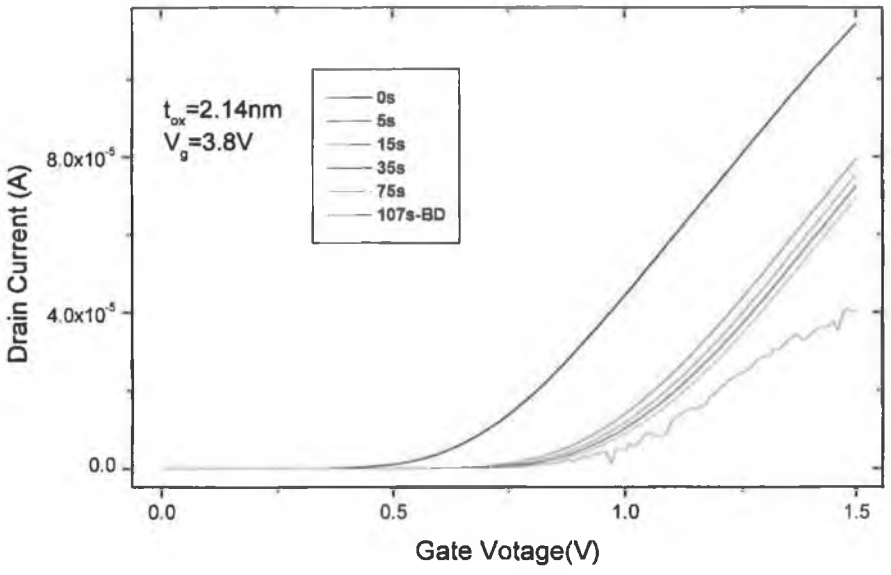


Figure 5.11: A set of I_dV_g curves during constant voltage stress. The curves are seen to shift to the right, leading to an increase in threshold voltage

5.5.1 Threshold Voltage

The threshold voltage (V_t) is one of the most important parameters in analog and mixed signal applications, which ideally require a constant V_t throughout device lifetime. The gate voltage however, creates trapped charge in the oxide, which over time, can lead to an increase in V_t, by shifting the I_dV_g curve along the V_g axis, and in some cases lowering the slope. The amount by which the threshold voltage shifts during oxide stress as a result of trapped charge can be calculated from[9]:

$$\Delta V_t = -\frac{\gamma \Delta Q_{ox}}{C_{ox}} - \frac{\Delta Q_{it}(\phi_s)}{C_{ox}} \qquad \text{Eq 5.1}$$

where, γ is the charge distribution factor, Q_{ox} is the trapped-oxide sheet charge density and Q_{it} the interface sheet charge density. ϕ_s is the surface potential, and C_{ox} is the oxide capacitance per unit area. Equation 5.1 indicates that trapped charge both at the interface and in the bulk oxide can cause a V_t shift. The experimental measurements illustrated in figure 5.12 support the idea that bulk trapped charge can affect the threshold. The V_t shifts for 3 oxide thicknesses as a function of increasing charge injection during CCS are shown. At low injection fields and low Q_{inj}, positive charge buildup dominates and there is a slight decrease in V_t. At higher values of Q_{inj}, V_t is seen to rise, the extent of which depends on the dielectric thickness. For the thickest layer investigated (2.13nm), the voltage shift reaches 1.35 V_{t0} corresponding to a 250mV shift. The thinnest layer (1.38nm) however shows very little V_t shift even at high Q_{inj}.

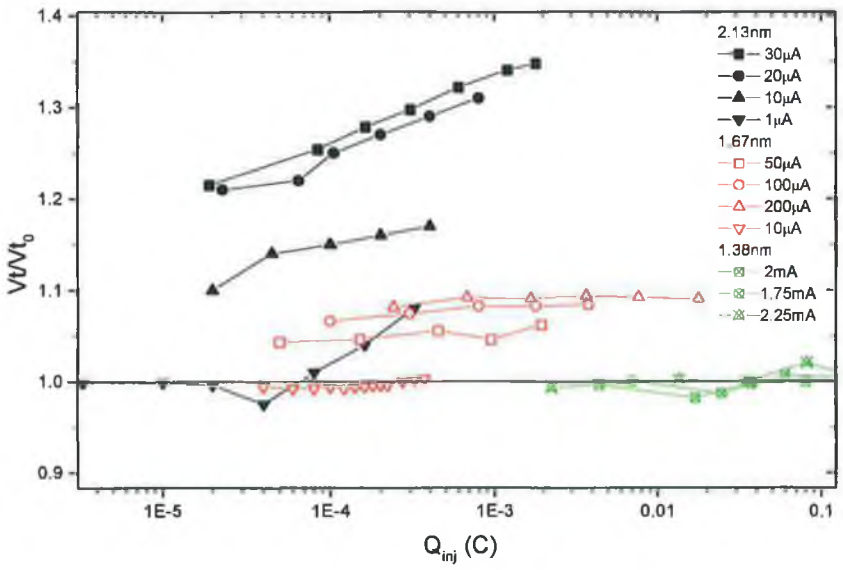


Figure 5.12: Saturation threshold voltage shifts at a number of stress currents for different dielectric thicknesses. The thickest oxide suffers the largest shift, suggesting that trapped charge in the film plays a role determining the magnitude of the shift.

If a threshold voltage shift tolerance of 30mV is defined, then an extrapolation can be made by measuring the time for a 30mV shift at high field and

extrapolating to the low-field range, in much the same was as done with time to breakdown statistics. The result is shown in figure 5.13 for a 2.14nm oxide, which displayed the most extreme shift.

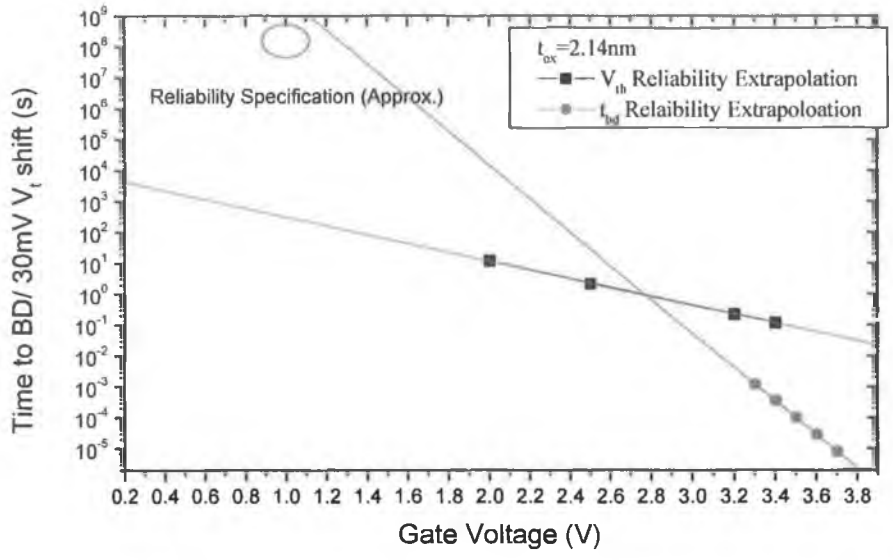


Figure 5.13: Reliability estimate for a 2.14nm oxide. Although it satisfies the t_{BD} requirement, the projection shows it will have undergone a large V_t shift much sooner than the required lifetime allows.

At operating voltage, the V_t shift will have risen above the 30mV tolerance after 300 seconds, and so even though the layer maintains its insulating properties well beyond the 10-year mark, the threshold voltage shift proves much too large for use in device applications.

The opposite is true for the thinner 1.38nm oxide. In this case the t_{bd} only just makes the specification (at room temperature), while the 30mV shift is impossible to measure as the breakdown always occurs before a shift of this magnitude is reached. The oxide of thickness 1.67nm was found to fall just short of the reliability condition on both counts.

Some information about the processes leading to threshold voltage shift can also be gained from considering gate injection stress in addition to substrate injection stress. The results of these measurements for both conventional SiON and HfSiON are shown in figure 5.14. The charge was injected at the same oxide field for both gate and substrate injection. It is clear that the stress polarity affects

the two layers quite similarly. With the HfSiON layers, the onset of the turnaround from negative to positive V_t shift (i.e positive to negative charge trapping at the interface) occurs at very high Q_{inj}. This suggests that gate injection results in less charge trapping than substrate injection for similar fluence.

A possible explanation for this is that gate injection leads to the creation of positive species at the substrate interface. This positive species passes back through the oxide but is first incident on the nitrogen rich area of the layer. The generation of traps is somehow suppressed due to the incorporation of nitrogen at the interface in agreement with Lee et al[10], and so we see only a slight V_t shift at high Q_{inj}. When substrate injection is used to stress the oxide, the same process leads to the positive species being released into the less nitrided region and so traps are created more quickly, leading to a substantial V_t shift, even at low Q_{inj}.

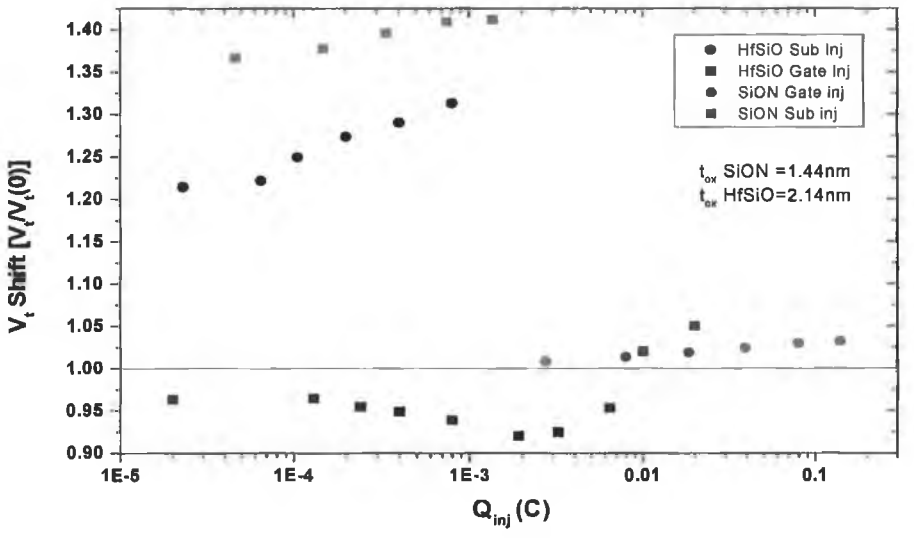


Figure 5.14: Threshold voltage shift behaviour is quite different depending on the stress polarity used, and the trend is different for SiON and HfSiON. This allows us to make some conclusions about trap generation in the layers

5.5.2 Transconductance

As well as threshold voltage shifts, MOSFETs subject to electrical stress display a reduction in maximum transconductance (g_m), and a shifting of the transconductance curve along the V_g axis.

In the linear region i.e. before the drain saturation current is reached the transconductance is given by [11]:

$$g_m = \frac{\delta I_d}{\delta V_G}_{V_{D=const}} = \frac{W}{L}\mu_n C_{ox} V_d \qquad \text{Eq 5.2}$$

where W and L are channel width and length respectively, μ_n is channel mobility, and C_{ox} the capacitance per unit area. So, clearly a decline in transconductance can be attributed to a reduction in channel mobility, possibly caused by scattering from interface charge sites.

Figure 5.15 shows the transconductance curves before and after stress for SiON and HfSiON of various thicknesses (all nMOSFETs). Lateral shifting along the V_g axis is similar to that for SiON except for the thickest layer investigated (2.14nm), which is indicative of less bulk charge trapping in the thinner films in agreement with the V_t shift. Maximum transconductance is reduced slightly with increasing thickness as expected. For thinner layers, the transconductance curves become noisier after stress. It has been proposed that this could be due to changes in mobility by local Coulomb scattering off interface states [8]. The more interface states present, the more scattering and thus the noisier the g_m curve.

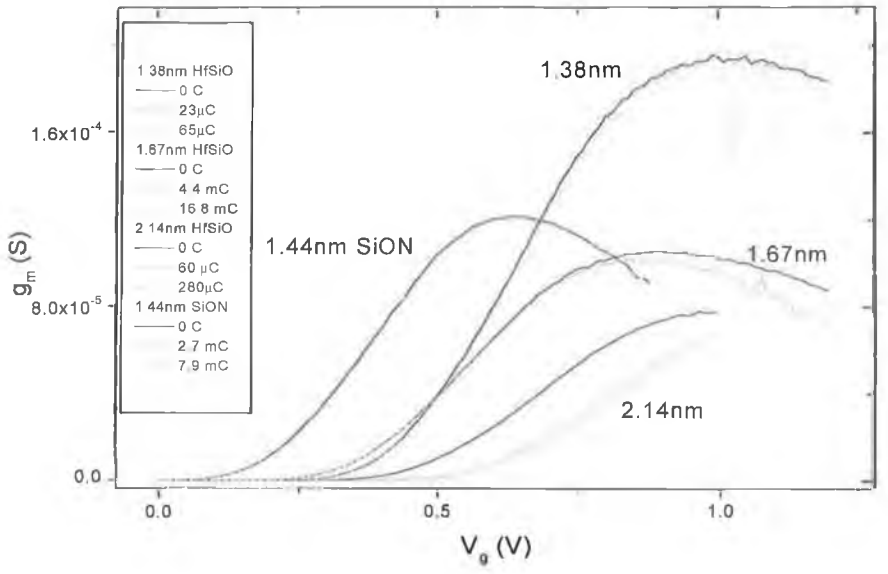


Figure 5.15: Transconductance curves are shifted further for thicker oxides, furthering the idea that they store more bulk-trapped charge. Reduction in max-transconductance is relatively independent of thickness, because it is primarily interface-dependent. V_D=1.2V Area(HfSiON)=1x10^{-6}cm^2 Area(SiON)=1x10^{-8}cm^2

The relationship between the number of interface traps and reduction in maximum transconductance is shown in figure 5.16 for $V_D>V_{Dsat}$ (i.e. a drain voltage greater than the drain saturation voltage). The graph displays the transconductance at time t normalised to the transconductance before stress plotted against the number of interface traps at time (t) normalised to the number before stress. There is a slight thickness dependence indicating that bulk traps play a small part in the degradation process. To obtain the data for this figure, the number of interface traps was measured by a 1MHz charge pumping measurement on W=10µm L=0.5µm devices, and an I_dV_g curve was acquired. Then the oxide was stressed at a fixed high field (16MV/cm) for 5s and the charge pumping measurement was repeated. The measurement sequence was then repeated for increasingly long stress intervals, to show the dependence of g_m on N_{it}. The 1.53nm where bulk trapping is least significant shows the lowest g_m degradation for a given number of interface traps, whereas the thickest layer shows slightly more degradation. In the early stages of the stress, the thicker layer shows about a 7% extra reduction in g_m when compared with the 1.53nm

layer, indicating that the bulk contribution is small compared with interface trap contribution.

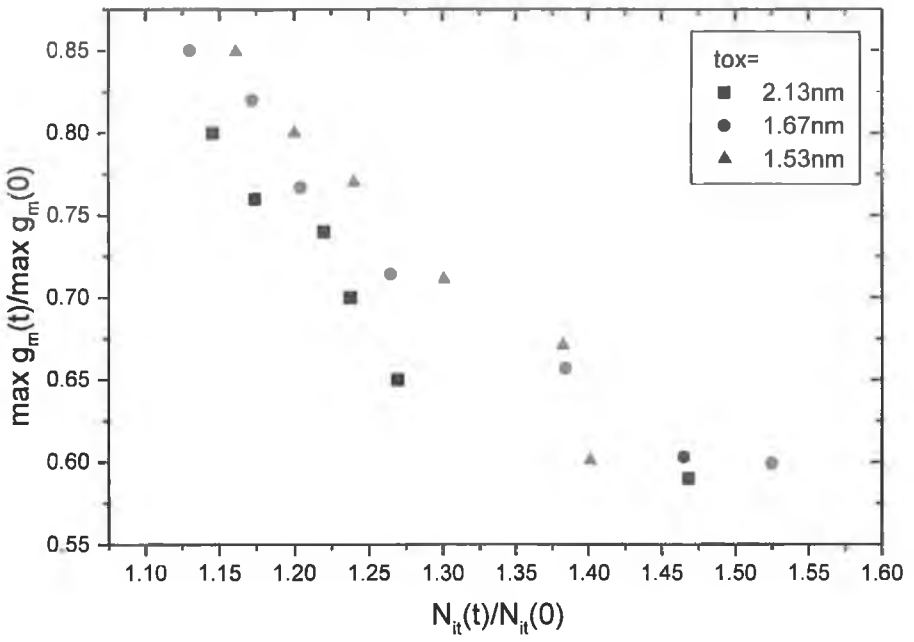


Figure 5.16 Reduction in max-transconductance vs. increase in number of interface traps. Thicker layers show a greater degradation for the same N_{it} indicating that bulk trapping also degrades transconductance slightly. A=5x10^{-6}cm^2. V_D=0.3V

The reduction in maximum transconductance during stress, attributed to mobility degradation is shown in figure 5.17, and is relatively independent of thickness as interface traps dominate. With both materials, gate injection causes a greater reduction. Because the transconductance degradation is more interface sensitive than V_t shift, we can conclude from this that in both materials, gate injection creates more interface traps than substrate injection in agreement with Degraeve et al[6]. The data for the 1.38nm oxide, at higher Q_{inj}, can be seen to continue the trend of the data for the 2.14nm sample, furthering the idea that mainly interface states affect g_m and thickness is largely irrelevant. The fact that gate injection creates more interface traps is confirmed in figure 5.18, where the number of interface traps (from charge pumping measurements) is plotted versus time during a constant voltage stress. The slope is higher for gate injection indicating that the trap generation rate is higher.

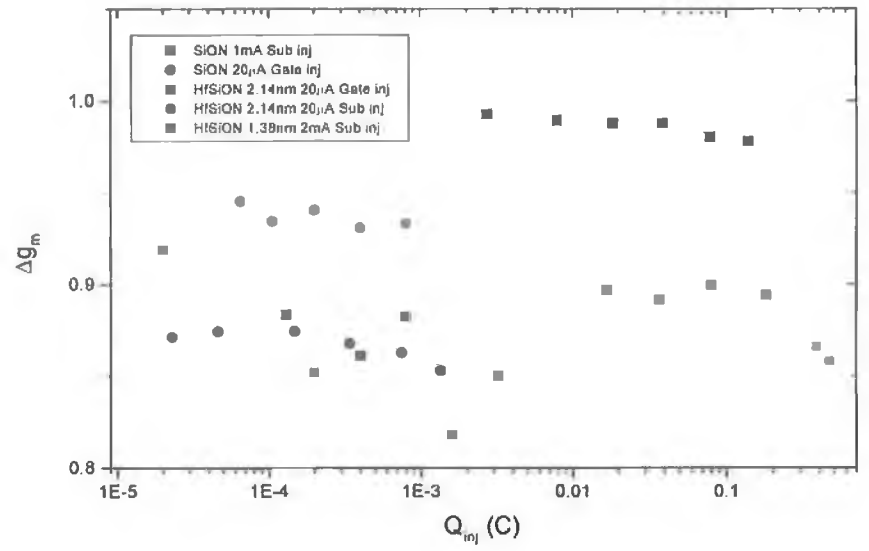


Figure 5.17 Reduction in max-transconductance shows gate injection causes more of reduction, and thus creates more interface traps than substrate injection

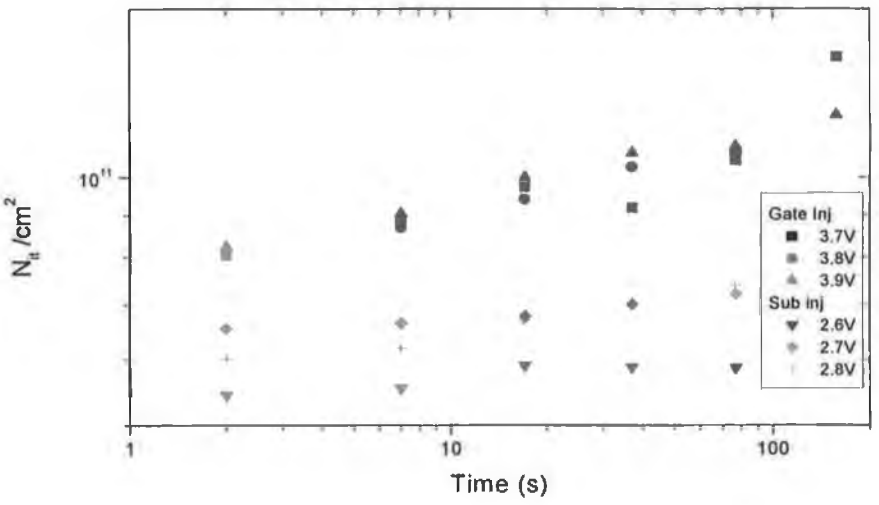


Figure 5.18: Charge pumping measurements confirm that interface trap generation rate is higher when using gate injection stress.

5.5.3 Sub-threshold Swing

Another quantity that can be used to quantify the degradation of the oxide is the sub-threshold swing (S), that is, the change in gate voltage required to change the gate current by one order of magnitude, in the sub-threshold regime. The value of S rises during stress due to the creation of interface traps, and the curve is shifted along the V_g axis due to bulk oxide charge trapping. The value of ΔS ($S_{\text{after stress}} - S_{\text{before stress}}$) can be related to the number of interface traps created by[14]:

$$\Delta D_{it} = \frac{C_{ox}\Delta S}{2.3kT}\left[1-\left(\frac{C_D}{C_{FB}}\right)^2\right] \qquad \text{Eq 5.3}$$

An example is shown in figure 5.19 for the 2.14nm oxide. The shift along the V_g axis is clearly evident, though the change in S (the slope of the line) is more difficult to detect. The change in S for two HfSiON oxides at a number of injection currents is shown in figure 5.20. The rise in S during a CCS can be seen more clearly.

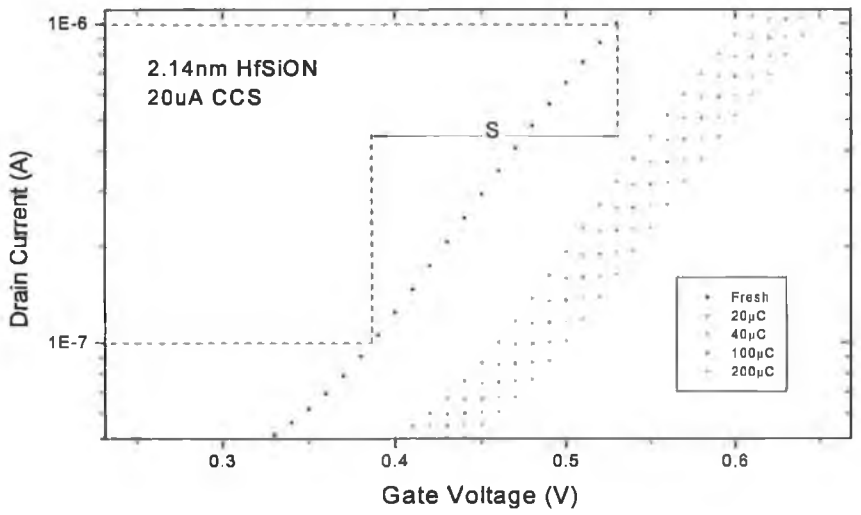


Figure 5.19: Sub-threshold I_dV_g curve. A shifting of the curve along the x-axis during the stress is observed which is an indication of oxide-trapped charge.

At the lowest injection current of 1μA, S is seen to take a negative turn at low Q_{inj}, indicating initial positive charge trapping. As the stress continues however, negative trapping dominates. The case for a thinner oxide is shown also, and indicates that interface trap build-up occurs at higher Q_{inj} for thinner oxides.

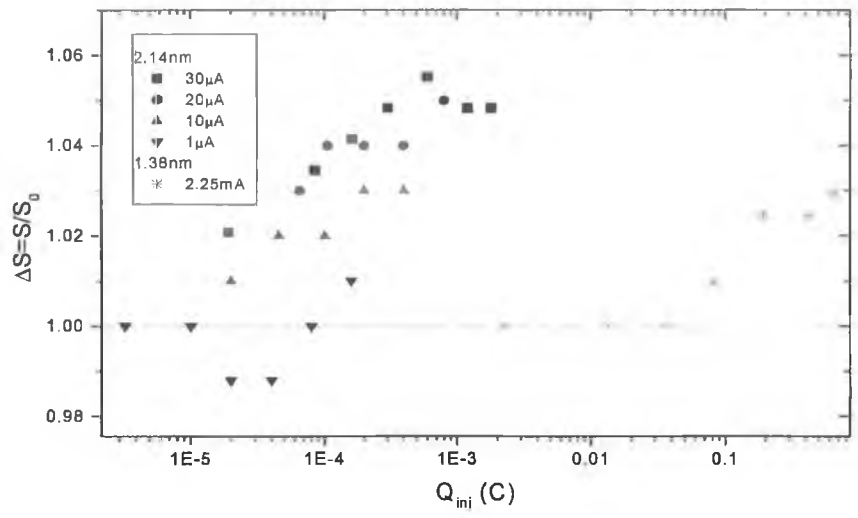


Figure 5.20: Change in sub-threshold swing magnitude during stress. Interface charge buildup occurs at higher Q_{inj} for thinner layers

5.5.4 Device Operation

In terms of device operation, pre- and post-stress, the layers were found to broadly follow the trends of SiON in that when a gate-substrate breakdown occurs, wide devices continue to function long after breakdown, in the progressive breakdown stage, as breakdown only inhibits transistor performance in the region of the breakdown path. Conduction from source to drain in the on-state is still possible across most of the device as an inversion layer can be maintained. Gate to source (drain) breakdowns cause almost instantaneous catastrophic change of the on (off) current.

Figure 5.21 displays the consequences of detecting the actual breakdown incorrectly. In this case, no large rise in gate current is observed, but the off-current is seen to rise significantly (indicating that a gate/drain breakdown path has been formed). The significant rise in gate current is only observed many

seconds after the real breakdown has occurred, and using the point where the current rises significantly as the t_{bd} will lead to overestimation of the device reliability.

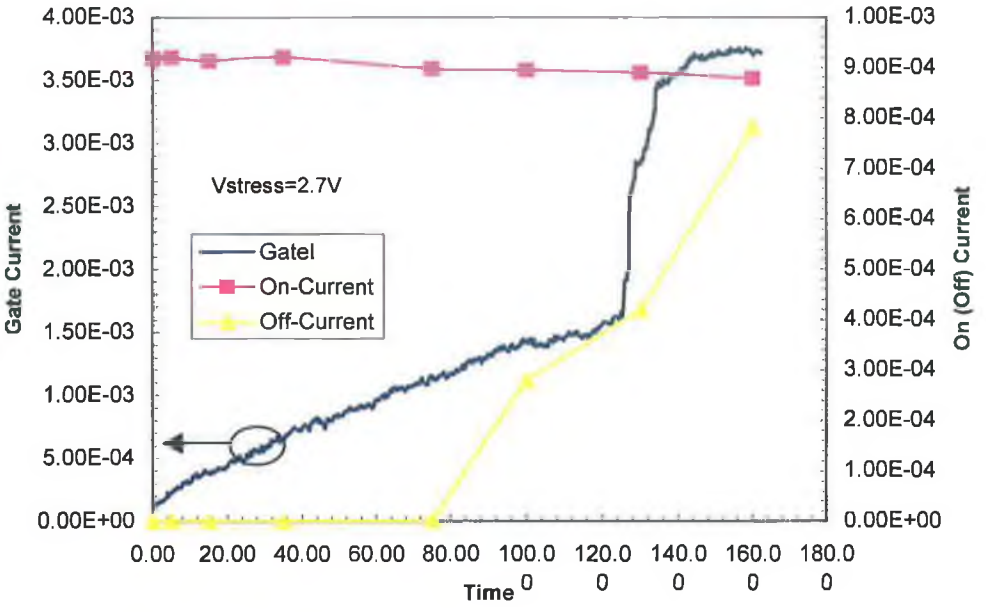


Figure 5.21: Off-current is seen to rise significantly before a significant rise in gate current, indicating that using a significant current increase in breakdown detection will lead to an overestimate in device reliability.

As the channel width is reduced, the breakdown path affects a higher fraction of the channel area until the channel width is in the sub-micron regime, when a single breakdown becomes fatal for the device. SiON based devices consistently followed this trend but exceptions are seen with the HfSiON layers, whereby sometimes, a narrow device will work long into the progressive breakdown stage, and sometimes a wide device will cease to function at the onset of breakdown.

The most notable difference between the two dielectric layers is found when monitoring the on-current, during a stress. With SiON, the on-current was found to stay constant throughout the duration of the stress, until breakdown, at which time the position of the breakdown in the channel determined by how much it changed. With HfSiON however, the on-current falls constantly during the stress, though more rapidly in the early part, and can be less than half its original value by the time breakdown occurs. Figure 5.22 shows the effect for a range of device

areas. This significant reduction in on-current is area independent and is likely caused by trapped charge building up during the stress causing local electric fields in the channel. Similar to the V_t shift, the effect is less dramatic in thinner layers, where oxide charge trapping is less prevalent.

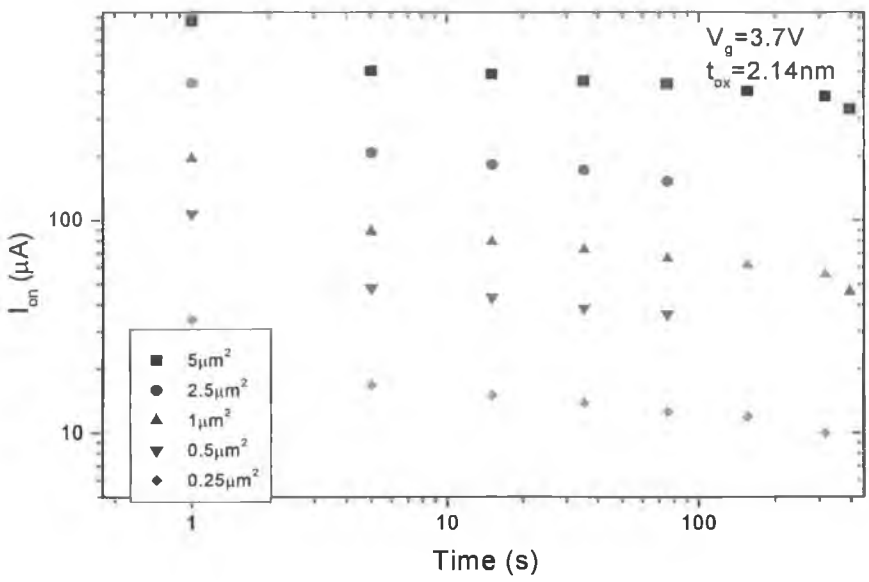


Figure 5.22: On-current reduces dramatically during stress. The rate of change is independent of area. In sub-100nm devices where the ratio I_{on}/I_{off} is quite small even for a fresh sample, this could cause problems with distinguishing on and off states.

The main point with respect to device operation is that if the breakdown occurs over source/drain extension regions, the device will in most cases cease to function soon after breakdown occurs. If the breakdown occurs over the substrate, the device in some cases continues to function normally even after multiple breakdowns and even at low channel width, and in some cases the device ceases to function even in very large devices after 1 breakdown event. Because of this uncertainty, we cannot realistically take the progressive phase as an extension to lifetime with these devices in the same manner as is possible with large aspect ratio SiON layers, where post breakdown behaviour is more predictable. The constant degradation in on-current is also a possible threat to reliability, as are the large threshold voltage shifts, and degradation in transconductance. All of the above mean that the layers under consideration are not suitable for large scale integration, and further 'fine-tuning' is required to engineer a layer with lower charge trapping, and better electrical reliability.

5.6 Stress Induced Leakage Current

Stress induced leakage current (SILC) is an important reliability problem in modern devices, causing power waste in MOSFETs and resulting in the loss of stored charge and thus information in non-volatile memory devices[12]. In some cases a critical level of SILC can be reached long before oxide breakdown occurs, and therefore it is sometimes more of a reliability concern than time-to-breakdown[13]. When an oxide is stressed at high-field, an I-V characteristic can appear quite different post stress to the corresponding pre-stress curve as shown in figure 5.23. SILC is generally characterised by carrying out a high field stress and then interrupting the stress to measure the current at operating voltage. The reason for this approach is outlined in figure 5.24. Even though there is no significant increase in leakage current at the stress voltage, there is a significant increase in SILC.

SILC has been widely attributed to the build-up of neutral electron traps within the oxide during the stress[14], however, the physical mechanism linking SILC to these traps is still uncertain. There is a general consensus that it arises from trap-assisted-tunnelling, using the neutral electron traps as 'stepping stones', which make it easier for electrons to tunnel from anode to cathode. Some groups attribute the SILC process to interface traps [12], while others attribute it to bulk trapping[7]. Our results show for both conventional SiON layers and 'high-k' HfSiON, that the SILC can be predominantly related to bulk oxide traps.

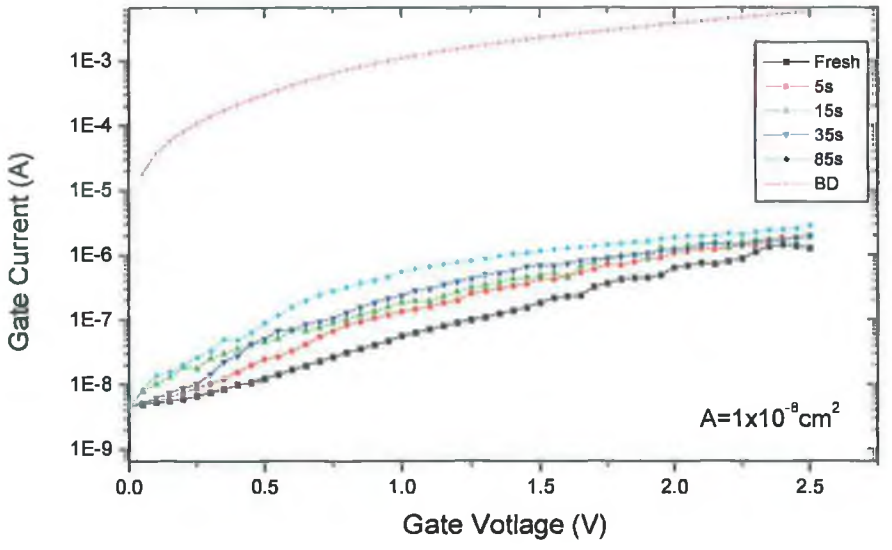


Figure 5.23: Stress induced leakage current can be measured from the low voltage I-V curve. The effect disappears at higher voltage.

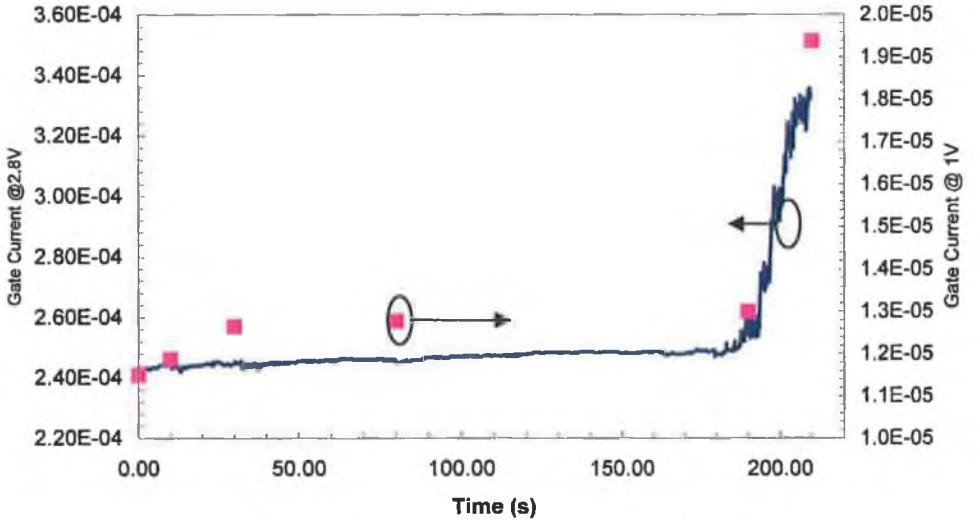


Figure 5.24: During a constant voltage stress, the high field I-t trace masks the SILC effect. When measured at a lower V_g, the SILC is evident. Right axis displays I@1V, left axis displays I@2V

A series of I-V sweeps after increasing stress time for an nMOSFET (with p-type substrate) device with EOT 1.4nm and W=L==1μm is shown in figure 5.25. SILC is evident across the voltage range. Many works measure SILC at a set low voltage e.g. 1V however figure 5.26 shows this approach is incomplete. The figure shows the level of SILC for increasingly long stress times, measured at a continuum of gate voltages as opposed to just one. The measured SILC is

actually voltage dependent and a peak is observed at a voltage that is slightly more negative than the flat band voltage. This has been observed in several previous works [12,15].

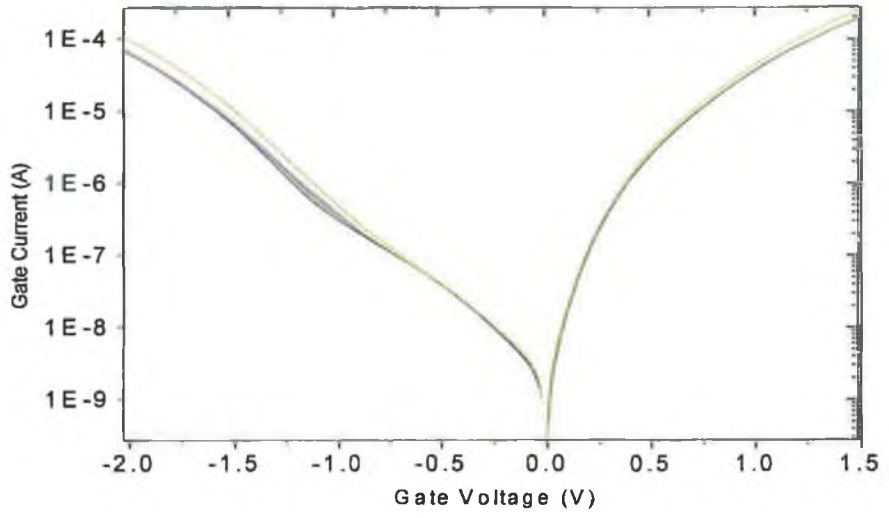


Figure 5.25: Stress induced leakage current can be measured from the low voltage I-V curve. The effect disappears at higher voltage

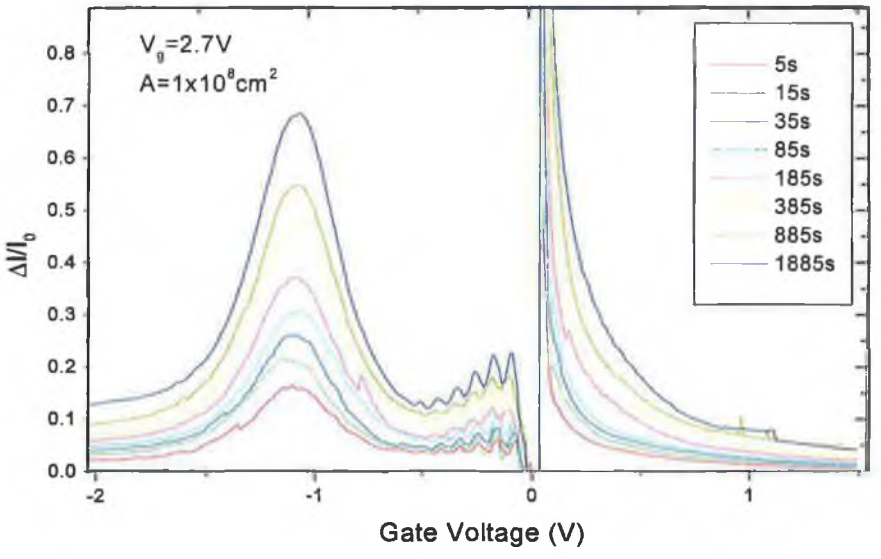


Figure 5.26: During a constant voltage stress, the high field I-t trace masks the SILC effect. When measured at a lower V_g, the SILC is evident.

If we compare the SILC measurements with the actual gate current from I-V curves as in figure 5.25, we can draw a number of conclusions. In the negative

gate voltage range, the tunnel current shows two components[16]. Above the flatband (0>Vg>Vfb) voltage, the current is mainly due to conduction between the gate and the source/drain extension (SDE) regions. As the voltage becomes more negative, the current is dominated by conduction between the accumulated substrate and the gate. The peak in SILC is at the point where the two cross. This means that if we measure the SILC increase above V_{fb}, we can get a measure of the contribution of traps above the SDE region, and measuring the SILC below V_{fb}, the contribution of the traps above the channel. In all cases the channel contribution is greater as expected, due to the small size of the SDE relative to the channel. The position of the breakdown can also be determined from the I-V curves. If the breakdown occurs over the SDE, the current is seen to rise above V_{fb}, and if the breakdown is above the channel, as in figure 5.25, the gate current rises significantly when $Vg<V_{fb}$.

The origin of the peak around V_{fb} has been discussed by Nicollian et al[13]. They suggested that in oxides less than 3.5nm and at stress voltages of <5V SILC is due to tunnelling through interface traps only and that bulk traps would no longer play a part. Our results contradict this however. Figure 5.27 shows SILC vs. Stress time for two oxide thicknesses at a number of voltages. The thicker oxide shows a greater SILC generation rate suggesting that bulk traps could account for the difference between the two.

The logarithmic SILC generation rate for the layers under consideration is ~0.35, independent of substrate bias, doping, and stress voltage. This value is similar to those found in the literature[17].

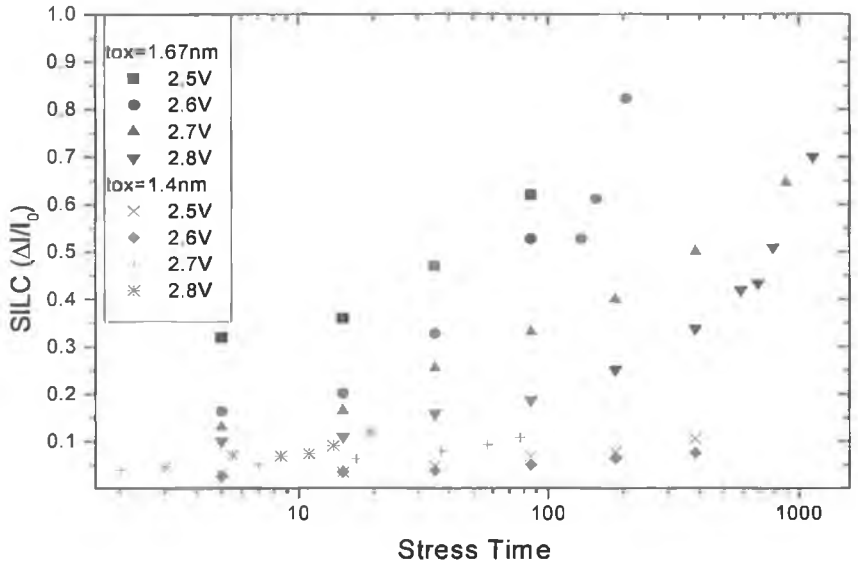


Figure 5.27: SILC shows a thickness dependence in ultrathin oxides, indicating that bulk traps do play a part in SILC in ultrathin layers, contrary to some reports in the literature.

The evidence that bulk trapping plays a part in low voltage SILC means the model of Nicollian et al requires extending and the origin of the peak at V_{fb} may not be due to interface traps. Their model proposes that the peak originates solely from interface traps with favourable energy with respect to the band structure. At voltages that are +/- 1V from V_{fb}, direct tunnelling dominates.

In this work, we suggest a refinement of the model. Due to the number of interface traps present, in all likelihood, there will *always* be interface traps meeting the required energy condition. This is possibly true even before any stress. Indeed many models treat the interfaces as an infinite sheet of traps that assist in tunnelling (e.g. the percolation model, described in chapter 3) . Charge pumping measurements show that N_{it} at breakdown can be as little as $1.2N_{it}(0)$.

So, we suggest that it should be bulk traps of a particular energy that are created, which cause the peak at V_{fb}. This is outlined in figure 5.28.

At $V_g<V_{fb}$ (figure 5.28a) electrons are injected from the gate via direct tunnelling. This process swamps the trap assisted tunnelling (TAT) into the substrate conduction band, and thus only a small SILC rise is detected in this voltage range. As V_g is made more negative FN tunnelling dominates, and SILC disappears almost completely (not seen for these thin layers). When $V_g>V_{fb}$ (figure 5.28b), substrate injection direct tunnelling dominates and again, SILC

(from TAT from substrate valence band) is swamped. At $V_g \sim V_{fb}$ (figure 5.28c) where there is little direct tunnelling, the increase in bulk traps in the correct energy range assists trap assisted tunnelling between states in the anode and cathode bandgaps, resulting in an increased SILC measured at this voltage.

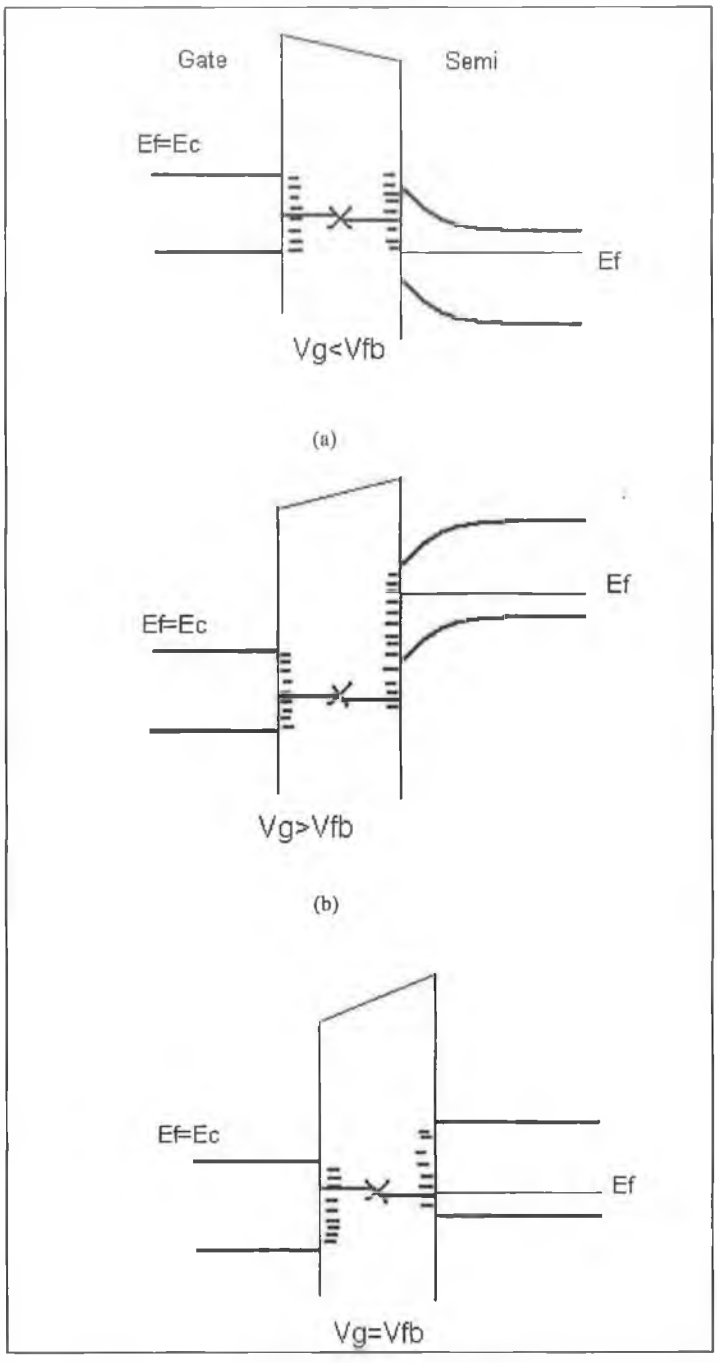


Figure 5.28. At voltages above and below V_{fb}, the effect of bulk trap assisted tunnelling is reduced as tunnelling takes place between anode and cathode conduction and valence bands. Around V_{fb}, tunnelling is between anode and cathode bandgaps, through traps of the correct energy. At this sense voltage a higher level of SILC is measured.

When the bands are bent sufficiently such that the bandgap of the anode aligns with either the conduction band or valence band of the cathode the TAT subsides. This argument is consistent with the width of the peak, which is approximately equal to the Si band gap (1.1eV). Similar results are observed for pmos structures, where $V_{fb}>0$.

Stress induced leakage current was found to be a much larger problem in the HfSiON layers studied. The difference between SILC for SiON and HfSiON is shown in figure 5.29. Layers of similar EOT stressed at comparable voltages are found to have SILC around 100 times that of the SiON layers. The majority of the additional SILC is created in the first few seconds of the stress, after which the SILC generation steadies to a rate similar to that in SiON. The resulting effect of the high SILC is that any benefit that arises from using the high-k layer is quickly lost as shown in figure 5.30. The absolute leakage current density (J) at +1V is plotted against stress time for two layers stressed at 2.8V. The absolute rise in J over time for the SiON is very small, when compared with that for the HfSiON and thus after <20000s the leakage at operating voltage will be greater in the HfSiON.

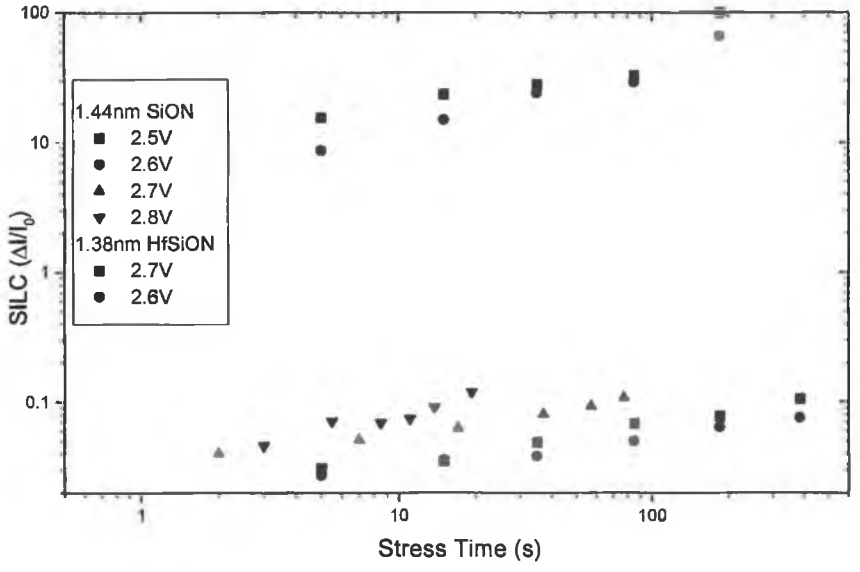


Figure 5.29. SILC measurements for HfSiON and SiON of comparable thickness show that SILC is a much greater problem in high-k layers than in SiON.

The sense voltage response in HfSiON was very similar to that in SiON though the peak around V_{fb} shifted, presumably due to V_{fb} shifts during the stress, attributed to the very high levels of charge trapped in the stacks. In thicker layers, extraction of the SILC proved difficult for the same reason.

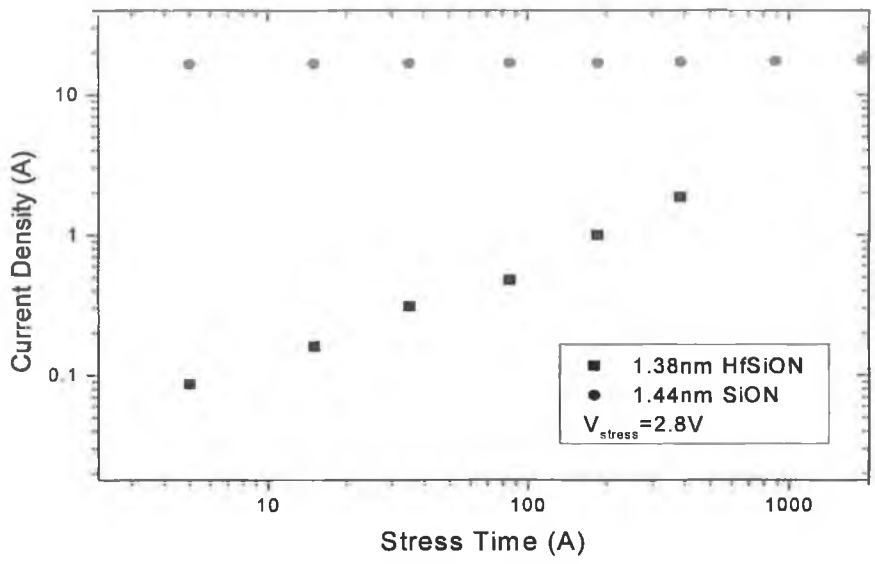


Figure 5.30: Gate current density during constant voltage stress shows that even though the initial leakage current is over 100 times lower for HfSiON, the effect of increased SILC means that the leakage quickly becomes comparable to that of SiON.

The HfSiON layers were found to have a significant transient component in the SILC measurement. Shown in figure 5.31 is the SILC for two different thicknesses of HfSiON, measured from I-V curves at various time intervals following a constant voltage stress. There is no change with time for SiON. This transient component has been attributed the emptying of slow electron trap after the stress[18]. As the transient component disappears after 600s, the steady state SILC is lower for the thicker layer as expected, as the probability of trap assisted tunnelling is lower in the thicker layer. The thick layer shows quite a large transient component indicating that slow traps, further into the bulk of the layer may also assist the process.

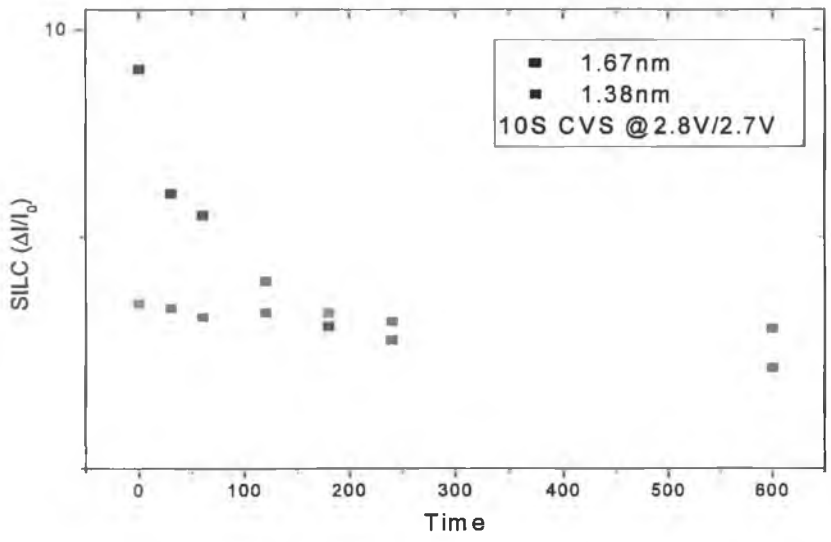


Figure 5.31: In HfSiON layers, SILC shows a transient component. The transient component is greater for thicker layers, as they have more bulk to store slow-emptying traps. The steady state SILC is lower for the thicker layer, as the tunnelling probability is lower.

As well as the power loss problem associated with leakage current through the gate, SILC causes a reduction in drive current in the SiON, where carriers that are being transported in the channel are lost to the gate via the trap assisted tunnelling process. The effect is shown in figure 5.32. In the case of the SiON layer, the drive current is reduced by ~5% during the device lifetime and the relationship between SILC and reduction in drive current is clear. The drive current vs. SILC for HfSiON is also shown in figure 5.32. The effect of positive fixed charge trapping during the stress is evident, as the drive current is seen to increase. One would expect SILC to eventually dominate and cause a reduction in the drive current before breakdown. This however was not observed even for extremely low stress voltages. The increase in I_d is not beneficial however, as it is accompanied by high threshold voltage shifts, and increased off-current, as the fixed charge attracts carriers to the channel even with no applied gate bias.

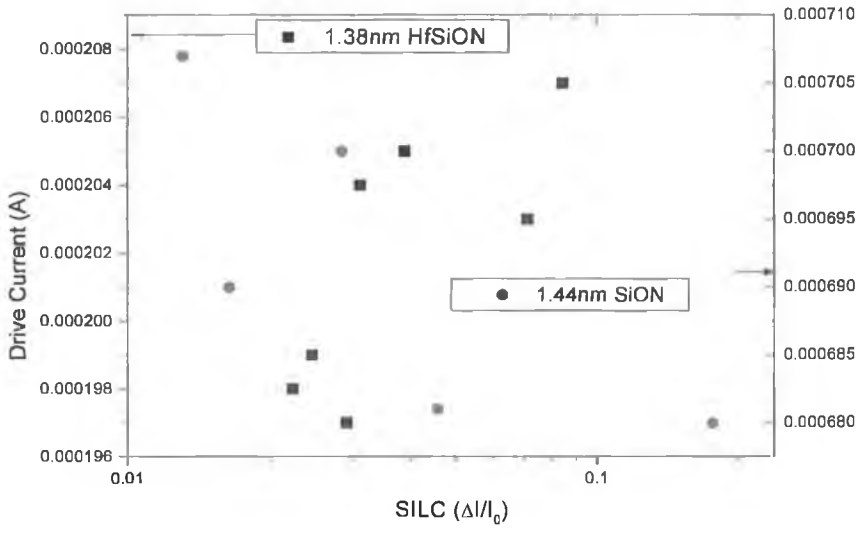


Figure 5.32: Change in drive current vs. SILC. HfSiON shows a rise due to positive charge trapping. SiON layers were stressed at 2.8V and I_d was measured at $V_g=V_D=1.2V$. HfSiON layers were stressed at 2.25V and measured at $V_g=V_D=1V$.

5.8 Summary

This chapter has outlined the important electrical properties of a possible candidate for replacing SiON as gate dielectric for future CMOS technologies. The leakage current was found to be significantly lower than for SiON. The 10-year maximum operating voltage was found to fall short of requirements, particularly at high temperature. However, the results also show the shortcomings of the method used to determine oxide breakdown. The high-temperature behaviour of the oxides was also found to have thickness dependence, with thinner layers showing less temperature acceleration than thicker ones. We have also examined the degradation of MOSFETs with HfSiON gate dielectrics during stress, and compared the results with those for SiON dielectrics. Our findings allowed us to make some tentative conclusions about trap generation within the layers. For both materials, substrate injection causes the creation of negative traps in the dielectric layers. The effect is more dramatic for HfSiON, confirmed by large V_t shifts, a larger degradation in g_m, a larger change in sub-threshold swing, and a continuous degradation in the on-current. The effect is thickness dependent, with thicker layers, storing more charge,

further degrading the above-mentioned parameters. At extremely low values of Q_{inj}, and low injection currents, the effects of positive charge trapping are visible. These positive traps may be created by anode hole injection, but are only visible at low fluence. For gate injection, positive charge trapping is more prevalent. This again, could be due to carriers releasing holes at the anode. However, over time, the effects are swamped by a build-up of negative traps. Negative charge trapping occurs very quickly in SiON and so negative V_t shifts are not observed.

We noted from transconductance measurements that gate injection stress causes more interface traps than substrate injection, and this was confirmed with charge pumping measurements.

Finally stress induced leakage current was examined for both SiON and HfSiON layers. SILC was confirmed to be sense voltage dependent, with a peak around the flatband voltage of the device in question. Bulk traps were found to play a part in SILC in ultrathin layers, in contradiction with some reports. The model of Nicollian et al was extended to include the presence of bulk traps. SILC was found to be a greater problem in the HfSiON layers, and even though the initial leakage current is lower, the high level of SILC, quickly brings the leakage on a par with SiON layers. SILC was also found to have a transient component in the high-k layers.

In summary, devices with thin HfSiON oxides, which suffer least from V_t shifts and mobility degradation, are most likely to fall short of t_{bd} reliability criteria, and vice versa for thicker layers, and thus the future for these HfSiON layers is uncertain, unless reliability can be improved.

5.8 References

[1] S. Mathew, L.K. Bera, N. Balasubramanian, M.S. Joo, B.J. Cho, Channel mobility degradation and charge trapping in high-k/metal gate nMOSFETs, Thin Solid Films, 462-463, (2004) pp. 11-14.

[2] J Robertson, Band offsets of wide-band-gap oxides and implications for future electronic devices, J.Vac. Sci. Technol. B 18, 1785 (2000).

[3] S.A. Campbell, D.C. Gilmer, X. Wang, M.T. Hsich, H.S. Kim, W.L. Gladfelter, J.H. Yan, MOSFET transistors fabricated with high permittivity TiO_2 dielectrics, IEEE Trans. Electron Devices 44, 104 (1997).

[4] Jang-Hyuk Hong, Tae-Hyoung Moon, Jae-Min Myoung, Microstructure and characteristics of the HfO_2 dielectric layers grown by metalorganic molecular beam epitaxy, Microelectron. Eng. 75 (3) pp 263-268 (2004).

[5] G.D. Wilk, R.M. Wallace, Electrical properties of hafnium silicate gate dielectrics deposited directly on silicon, Appl. Phys. Lett. 74, pp. 2854-2856 (1999)

[6] D.C. Gilmer et al. Compatibility of silicon gates with hafnium-based gate dielectrics, Microelec. Eng. 69 (2-4) (2003) pp. 138-144.

[7]R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, H.E. Maes, A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides, IEDM Tech. Dig., pp. 863-866, 1995.

[8] R. Degraeve, B. Kaczer, F. Schuler, M. Lorenzini, D. Wellekens, P. Hendrickx, J. Van Houdt, L. Haspeslagh, G. Tempel and G. Groeseneken , Statistical model for SILC and pre-breakdown current jumps in ultra-thin oxide layers, IEDM Techn. Dig. (2001) pp. 121-124.

[9] Y.B. Park, D.K. Schroder, Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM, IEEE Trans. Electron. Devices, 45 (6) pp. 1361-1368 (1998).

[10] Y.M. Lee, Y. Wu, C. Bae, J.G. Hong, G. Lucovsky, Structural dependence of breakdown characteristics and electrical degradation in ultrathin RPECVD oxide/nitride gate dielectrics under constant voltage stress, Solid-State Electron. 47, (2003), pp 71-76

[11] Sze, Physics of Semiconductor Devices, pg 441, Wiley, 1981.

[12] M.L. Green, E.P. Gusev, R. Degraeve, E.L. Garfunkel, Ultrathin (<4nm) SiO_2 and Si-O-N gate dielectric layers for silicon microelectronics: Understanding the processing, structure, and physical and electrical limits, J. Appl. Phys. 90 (5) (2001) pp. 2057-2121.

[13] P.E. Nicollian, M. Rodder, D.T. Grider, P. Chen, R. M. Wallace, S.V. Hattangady, Low voltage stress-induced-leakage-current in ultrathin gate oxides, Proceedings of IRPS 1999 pp. 400-404.

[14] J. DeBlauwe, R.Degraeve,R. Bellens, J.V. Houdt, P. Roussel, G. Groeseneken, H.E. Maes, Proceedings of ESSDERC 1996 pp. 316-364.

[15] C. Petit, A. Meinertzhagen, D. Zander, O. Simonetti, M. Fadlallah, T. Maurel, Low voltage SILC and P- and N-MOSFET gate oxide reliability, Microelectron. Reliab, In Press (2004)

[16] N. Yang, J.J. Wortman, A study of the effects of tunnelling currents and reliability of sub-2nm gate oxides on scaled n-MOSFETs, Microelectron. Reliab. 41 (2001) pp. 37-46.

[17] M. Fadlallah, G. Ghibaudo, M. Bidaud, O. Simonetti, F. Guyader, Stress-induced leakage current at low field in NMOS and PMOS devices with ultra-thin nitrided gate oxide, Microelecton. Eng 72 (2004) pp. 241-246.

[18] R. Moazzami, C. Hu, Stress-induced current in thin silicon dioxide films, IEDM Tech. Dig. 139 (1992) pp. 1-3

Chapter 6: Chemical Characterisation of Blanket SiON and HfSiON Gate Dielectrics Layers

6.1 Introduction

This chapter examines some of the important material properties of silicon oxynitride and hafnium silicate dielectric layers, using the techniques outlined in chapter 2. The first part examines the SiON layers. The initial XPS study examines the chemical composition of the elements present, and the effect of varying sample preparation method on elemental concentration. The ratio of the Si2p oxide peak to the Si2p substrate peak is used to make an estimate of the layer thickness. The ARXPS study is used to determine the depth profile of the elements and prove that the detected carbon is surface localised. The depth profiles are then compared with sample preparation conditions. The synchrotron study allowed analysis of the valence band spectra and calculation of the valence band offset for the layers. The $Si2p^{4+}$ peak width was also studied and compared with findings in the literature. Finally, SIMS was used to confirm depth profile information. The preparation methods for the SiON layers studied are summarised in Table 1. SiON samples had a base oxide of SiO_2 grown by in-situ steam generation, and the nitridation was carried out by an rf decoupled plasma at a frequency of 10kHz, a pressure of 10mTorr, 100% N_2, and for 15 seconds. The only variations were base oxide thickness, plasma effective power and post nitridation anneal.

The aim of decoupled plasma nitration (DPN) is to specifically incorporate a high level of nitrogen preferentially onto the top surface of an ultra thin gate oxide[1]. More conventional methods of nitration, namely N_2O or NO thermal nitration have been reported to result in the build up of nitrogen at the oxide/substrate interface. In order to limit boron penetration from the polysilicon gate in p-channel MOSFETs it is more advantageous to have the nitrogen preferentially at the oxide/polysilicon interface rather than the oxide/substrate interface. Plasma nitration processes allow a high degree of control of the concentration of nitrogen incorporated into the layer and SIMS profiles of the

nitrogen concentration have confirmed that the nitrogen is predominantly surface localised[2].

Sample	Base Ox (nm)	Power (W)	PNA	Optical th. (nm)
1	0.8	1125	1000°C, 0.5 Torr, O_2, 15s.	1.36
2	1.0	1750	1000°C, 0.5 Torr, O_2, 15s.	1.47
3	1.2	2375	1000°C, 0.5 Torr, O_2, 15s.	1.56
4	1	1250	1000°C, 0.5 Torr, O_2, 15s, +900°C, 100 Torr, H_2, 60s	1.40

Table 6.1: Growth conditions for SiON samples.

The HfSiON layers underwent a similar analysis. The synchrotron study also included analysis of the Hf4f peak, and X-ray absorption measurements.
The samples were prepared by carrying out an IMEC clean. Following this, the HfO_2 and SiO_2 were simultaneously deposited by MOCVD at 485°C. Subsequently one of the samples was annealed in NH_3, one in N_2, one sample was not nitrided. A summary of the resulting layers is shown in Table 6.2. Thus, the main focus of the study was the effect of nitridation on the layers.

Sample	HfO_2/SiO_2	Anneal
1	60%/40%	N_2@1000°C 60Sec
2	47%/53%	NH_3@800°C 60Sec
3	47%/53%	No anneal

Table 6.2: Growth conditions for HfSiON samples

6.2 Characterisation of SiON Layers

6.2.1 XPS

The initial survey scans for the SiON layers are shown in figure 6.1. The spectra display peaks associated with Si2p, Si 2s, O1s, C1s and N1s as expected. Valence band peaks can just be made out at high kinetic energy. Only one nitrogen peak is observed, indicating that there is one type of nitrogen present. Other studies[3] have shown more than one N1s peak for different methods of nitridation, which resulted in the N being concentrated at the oxide/Si interface. The N1s peak is at similar kinetic energy as that found in Si_3N_4. This suggests that the nitrogen is preferentially bonded to Si atoms. The nitrogen concentrations for the layers were found to vary from 12.07% to 15.4%, which is quite high compared with other studies, which use more conventional nitration methods[4]. It appears then, the pulsed DPN method is capable of incorporating a large amount of nitrogen in the layer. The highest concentration was achieved for sample 3, with the highest rf-power. The calculated elemental concentrations for the four layers is shown in Table 6.3.

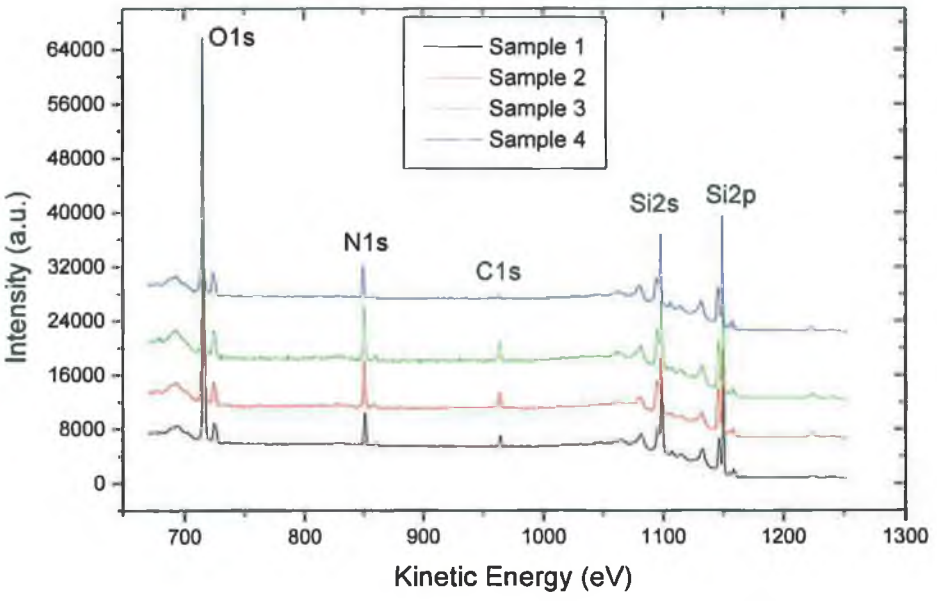


Figure 6.1: Wide energy window survey scans for SiON samples show the presence of all expected elements.

Sample	Silicon (%)	Oxygen (%)	Nitrogen (%)	Carbon (%)
1	22.4	57.2	13.0	7.4
2	22.6	54.4	14.9	8.1
3	23.3	57.1	15.4	4.2
4	23.7	61.7	12.1	2.5

Sample	Silicon (%)	Oxygen (%)	Nitrogen (%)
1	24.2	61.6	14.2
2	24.6	59.2	16.2
3	24.3	59.6	16.1
4	24.3	63.3	12.4

Table 6.3: (a) Elemental concentrations for all four SiON samples as calculated by XPS. (b) The same data, with carbon assumed to be surface localised and thus removed, make a more quantitative analysis possible.

The elemental concentrations show that increasing the plasma power increases the nitrogen concentration, with the exception of the layer that has a subsequent anneal in H_2. Table 6.3b shows the data with the carbon contribution removed, as it is assumed to be surface localised. This is confirmed later in this chapter. Removal of the carbon contribution shows that there is very little difference in the silicon concentration independent of preparation method. Any increase in nitrogen is associated with a decrease in oxygen, furthering the idea that the nitrogen bonds primarily to Si.

6.2.2 ARXPS Study

As outlined in Chapter 2, further information can be obtained by varying the electron take off angle. Measurements were performed from 0 ^{o} to 60^{o}, at 5^{o} intervals. The carbon peak was confirmed to be due to surface localised contamination by this method. The increase in carbon concentration as a function of angle is shown in figure 6.2. For all samples, the detected concentration rises linearly with cos θ, showing that the carbon is located as a layer on the surface. The carbon can thus be modelled as a thin layer that resides on top of the dielectric layer, which in turn lies on the Si substrate.

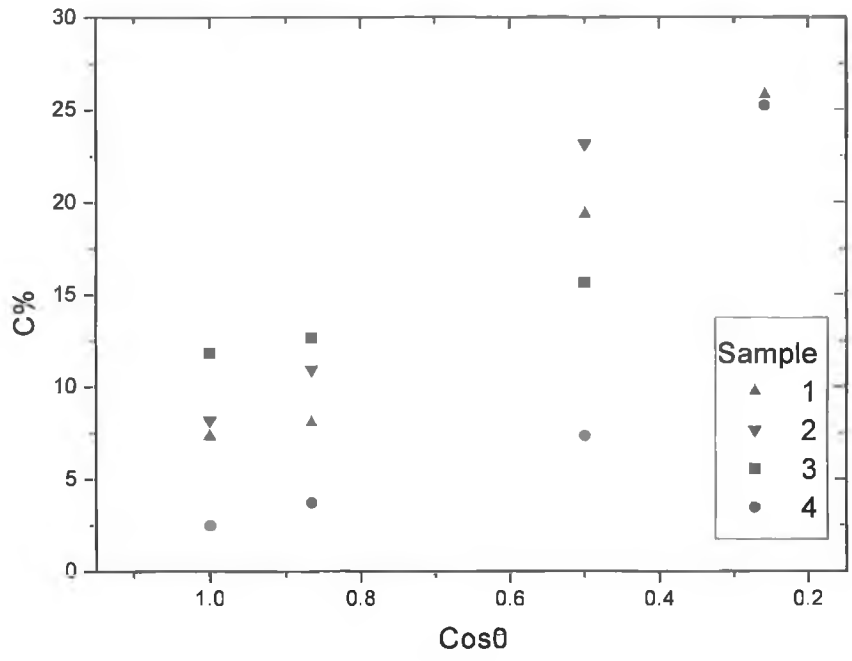


Figure 6.2: Variation of carbon signal as a function of cos θ. The linear increase in carbon concentration shows that it is a surface localised layer.

Figure 6.3 shows the Si2p peak for angles between 0 and 40°. The peak at 1148.5eV kinetic energy is from the silicon atoms in the substrate, while the peak at 1144.7 is due to the silicon atoms in the overlayer. The intensity of the overlayer silicon peak becomes greater with increasing angle as expected, as the technique becomes more sensitive to the surface layers. Also, as angle is increased, the oxide peak is drawn closer to the substrate peak, that is, the chemical shift[5] is reduced. This indicates a tendency toward bonds with a lower binding energy than Si-O bonds, i.e. more Si-N bond formation, and is a first indication that nitrogen is primarily located at the top surface of the layers. The shift is ~0.3eV at 40°. On examining the N1s peak (figure 6.4), it too is seen to shift to lower binding energy at increased angle. The position of the N1s peak relative to the Si2p substrate peak was monitored to confirm that the shift is a chemical one, and not due to oxide charging effects. It has been noted that the N1s peak shifts to higher binding energies for increased thickness[6] and so N atoms at the interface may indeed have different bonding properties to those in the bulk.

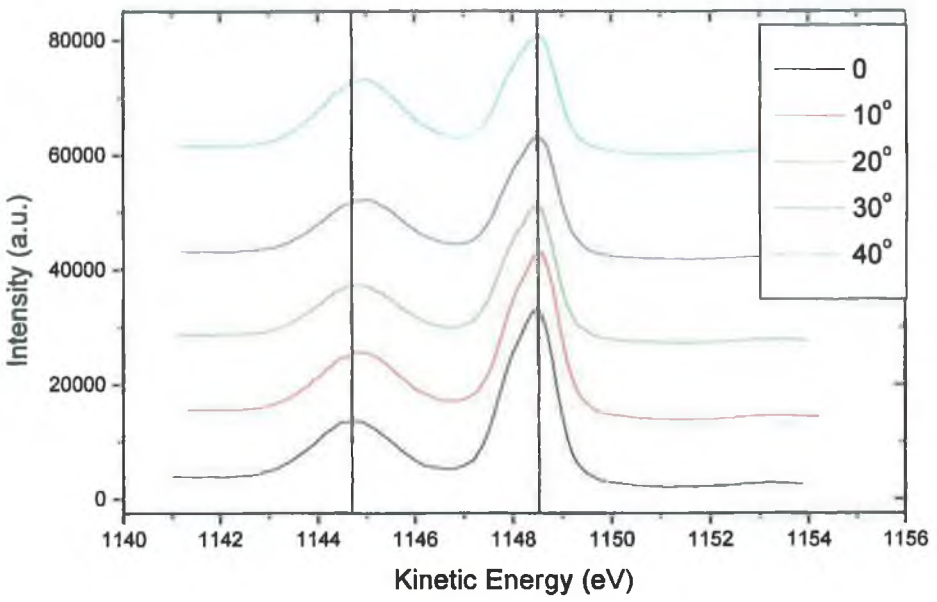


Figure 6.3 Si2p peaks for a number of take off angles. As take off angle increases, the oxide peak shifts to lower binding energies with respect to the substrate peak, indicating that the overlayer Si is involved in lower energy bonds at the surface.

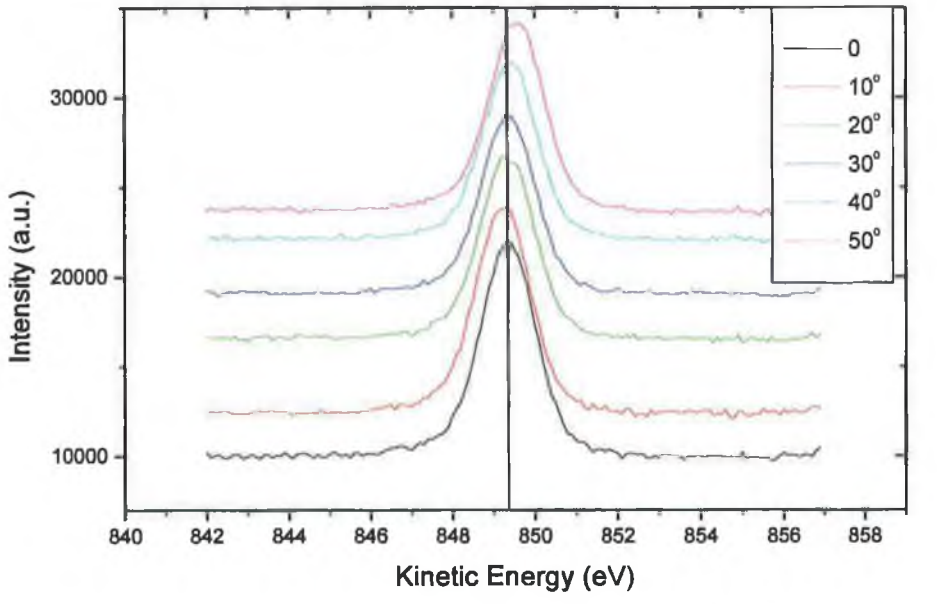


Figure 6.4 The N1s peak also shifts to lower binding energies, and so is involved in lower energy bonds at the surface.

The interfacial component is generally thought to result from N-Si_3 bonds and the bulk from some configuration that yields a higher binding energy[7]. As the

angle is increased, photoemission from the N-Si_3 bonds dominates, and so the peak shifts to lower binding energy. The bulk configuration is not well understood and various possibilities (N-Si_2O, N-SiO_2) have been proposed[8]. Takahashi et al[9] have carried out a detailed study showing that the N bonding tendencies are dictated by the temperature of the plasma nitridation. Our results agree in that high temperature nitridation yields a single N peak. In Takahashi's study[9], lower temperature processes were shown to have more than one nitrogen peak.

This indication that the nitrogen is preferentially located at the Si/SiON interface was further investigated, firstly by examining in the N1s peak height as a function of Cos θ. If a linear relationship is observed, then the N is uniformly distributed throughout the layer, whereas if the signal falls off more slowly than Cos θ, the N is located nearer the Si/SiON interface. Results for all samples are shown in figure 6.5.

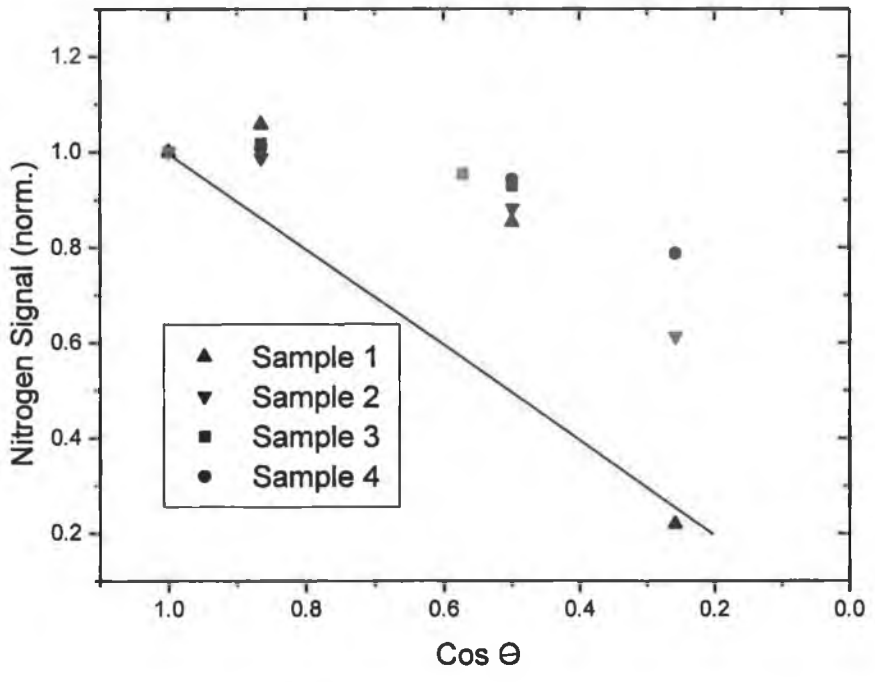


Figure 6.5: Variation of nitrogen signal as a function of cos θ. Data are normalised to peak height at normal emission

Angles up to 75° are shown. The straight line represents the linear dependence of a uniformly distributed element. In all cases the nitrogen can be seen to be surface localised. Differences between the samples begin to emerge at extreme angles, whereby the signal in sample 1 (low plasma power) indicates that the nitrogen is not as concentrated at the upper surface, whereas sample 4 (extra anneal) shows the nitrogen signal to be remain significantly strong, even at extreme take off angles, indicating that the nitrogen is strongly concentrated at the surface. It is possible that the extra anneal causes this effect. Further proof that mainly Si-N bonds exist at the upper surface of the dielectric layer is shown in figure 6.6, where the ratio of the N1s peak to the Si2p/O1s peaks is plotted vs. cos (θ). Close to normal emission where the sampling depth is a maximum into the material, the ratio N/O is quite small, indicating an abundance of O atoms. As the angle increases, the N signal intensity increases with respect to the O (while decreasing with respect to Si), indicating that N atoms replace O atoms at the surface.

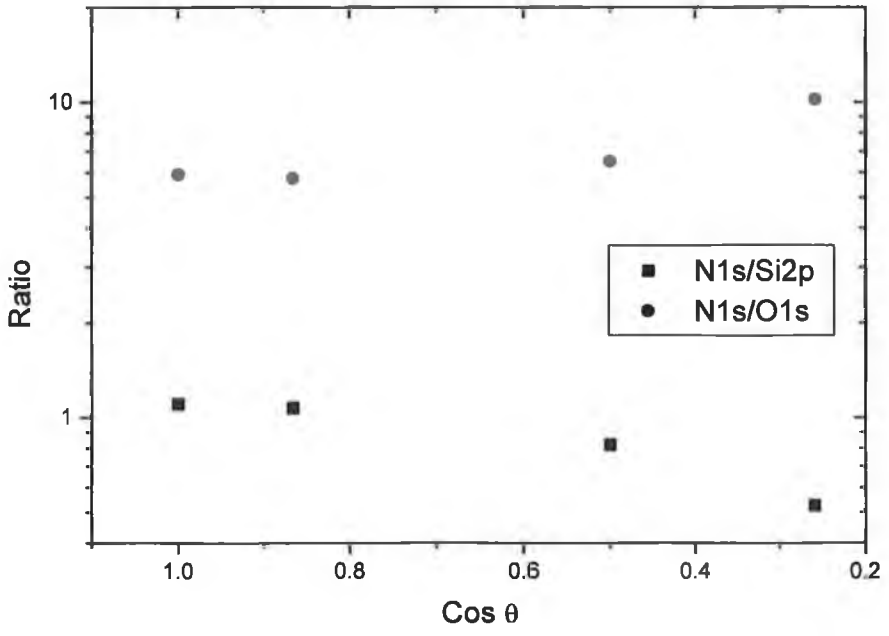


Figure 6.6: Ratio of N1s to Si2p/O1s peak heights for varying angle. The increase of the N/O ratio at high angle, where the technique is most surface sensitive, shows that N atoms replace O atoms near the surface, forming a Si_3N_4 like chemistry at the top of the dielectric layer.

An ARXPS program[10] was used to model the layer in this way, in an attempt to determine the thickness of the carbon contamination layer. The software calculates the optimum thickness to fit the N1s signal as a function of angle for the 3 layers system, with the only variable being the thickness of the carbon layer. Figure 6.7 shows the fit calculated for sample 1 (7.4%) carbon. The software obtains this best fit for a 0.26nm layer of carbon covering 92% of the surface.

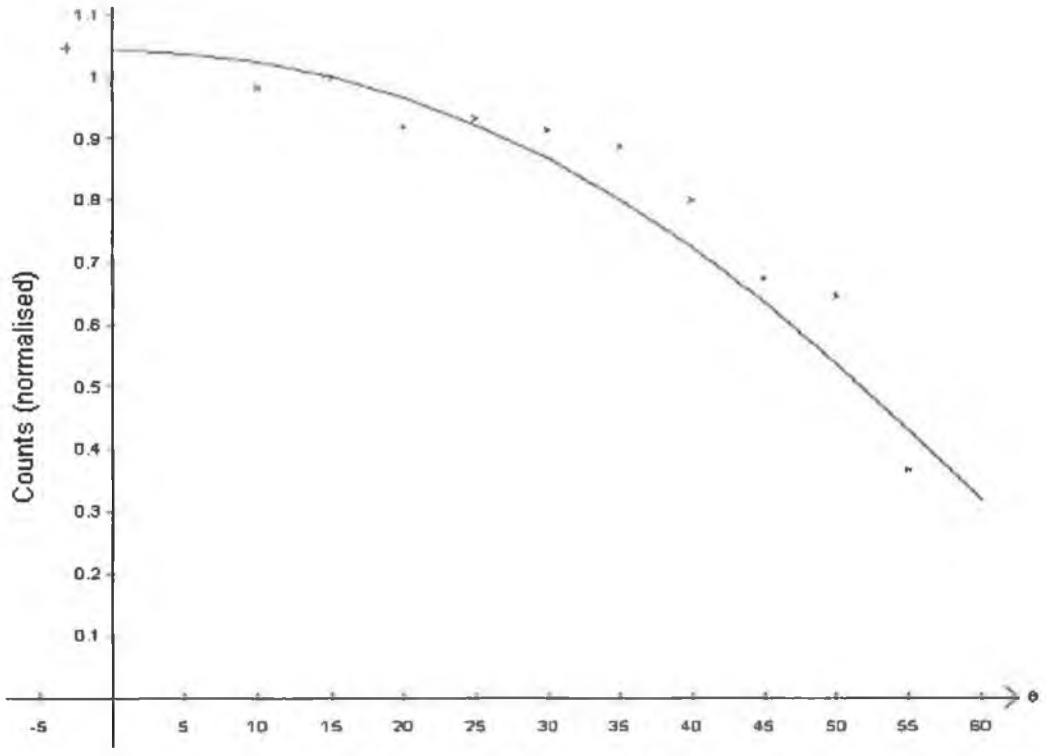


Figure 6.7: Calculated best fit for N1s peak height as a function of angle based on a 3-layer system give a carbon layer thickness of 0.26nm.

6.2.3 Soft XPS Using Synchrotron Radiation

This section outlines the results obtained for the same SiON layers, on the U4A beamline at the NSLS. As mentioned, the benefits of synchrotron radiation over conventional XPS include tuneable photon energy, and very high resolution. Figure 6.8 shows the $Si2p^{4+}$ spectra for our SiON samples taken at a photon energy of 130eV. The inherent surface sensitivity of the technique is apparent in the fact that the substrate silicon peak at 26.7eV kinetic energy is only just observable. The peaks show no dependence on how the samples were prepared.

In order to simplify the peak shape that inherently contains the Si 2p ½ and Si 2p 3/2 component, a spin-orbit stripping program[11] was used to remove the Si2p ½ component. The data was then fitted using a single Gaussian profile in order to extract the full width at half maximum (FWHM) for comparison with other studies. The values for the FWHM were between 1.38eV and 1.55eV. The narrowest peak corresponds to a layer which underwent an additional post nitration anneal (PNA) in H_2. The widest peak is found in the layer grown with the highest plasma effective power indicating a wider distribution of chemical bonding environments in the layer.

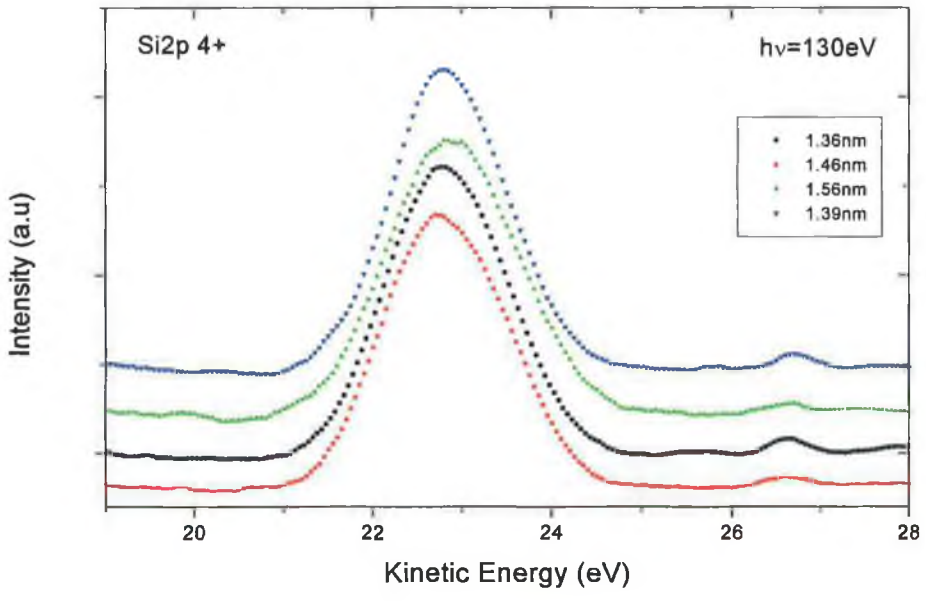


Figure 6.8: Si2p 4+ spectra for SiON samples with different recipes show no notable differences. Layers are just thin enough so as that a substrate signal can be seen at 26.8 eV

These values for the FWHM are significantly larger than the value for SiO_2 (1.1eV) in agreement with Keister et al[12], confirming the fact that a more complex chemical environment exists in SiON layers compared to pure SiO_2.
The valence band spectrum for an SiON layer illustrated in figure 6.9 shows a large N(2s) peak due to the high nitrogen concentrations attained by using a pulsed plasma. Once again, spectra for all preparation methods are identical and appear similar to those in reported in the literature [9].

Valence band offsets for the SiO_xN_y/Si system can be estimated by looking at the valence band edge that appears between 72 and 77eV in figure 6.10. The valence band maximum (VBM) for a SiO_2 dielectric layer has been established to be 4.4eV below the silicon VBM giving a 4.4eV band offset. In the SiON samples there is also a valence band maximum associated with nitrided layer and so the energy band diagram of the interface is graded because of the presence of nitrogen. The valence band maximum of the Si substrate occurs when the Si-substrate valence band signal drops to zero. The difference between the two is equal to the valence band offset. However, because the Si-substrate valence band peak is not well defined it is quite difficult to estimate the offsets. Shown in figure 6.10 is the valence band edge that appears between 72 and 77eV in figure 6.9, at lower photon energy for our SiON samples. From this it is clear that there are 2 offsets visible: The SiO_2 offset at 4.4eV, the SiON offset at 2.8eV. Note that the Si offset is estimated with respect to the valence band offset of SiO_2 [$E_V(SiO_2)$] in accordance with the findings of Keister et al[13]. Thus the SiO_2/SiON offset is found to be 1.6eV, and the SiON/Si offset is 2.8eV. The differences from Keister et al[13] may be attributable to the fact that their study involved the deposition of a Si_3N_4 interfacial layer.

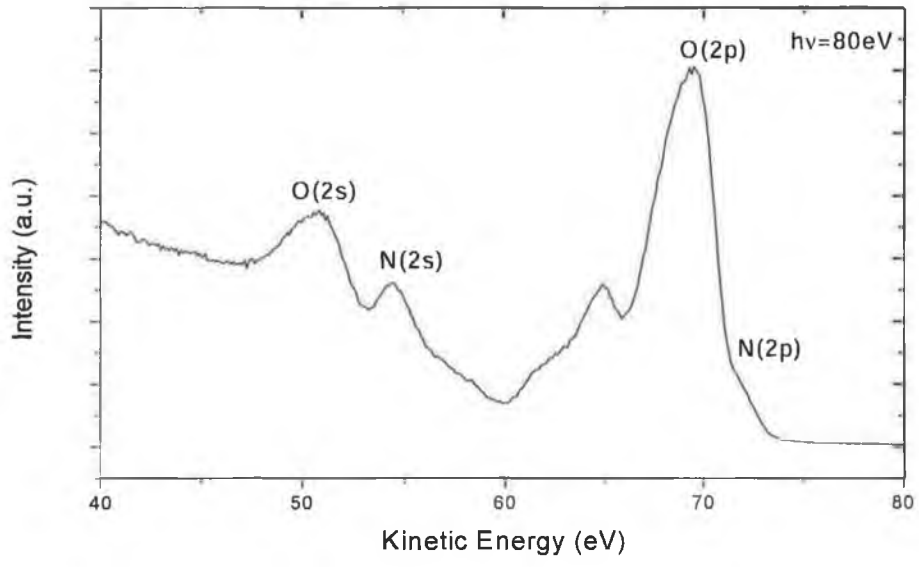


Figure 6.9: Valence band spectrum at a photon energy of 80eV.

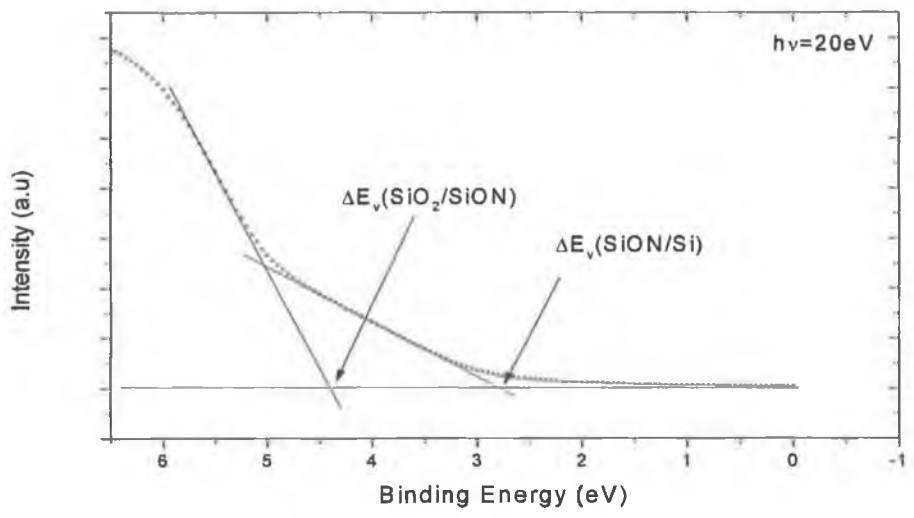


Figure 6.10: Valence band edge at an energy of 20eV. Two offsets are clearly visible.

Using the information about the valence band offsets and data from the literature[13], a schematic diagram of the band structure can be constructed. This is presented in figure 6.11

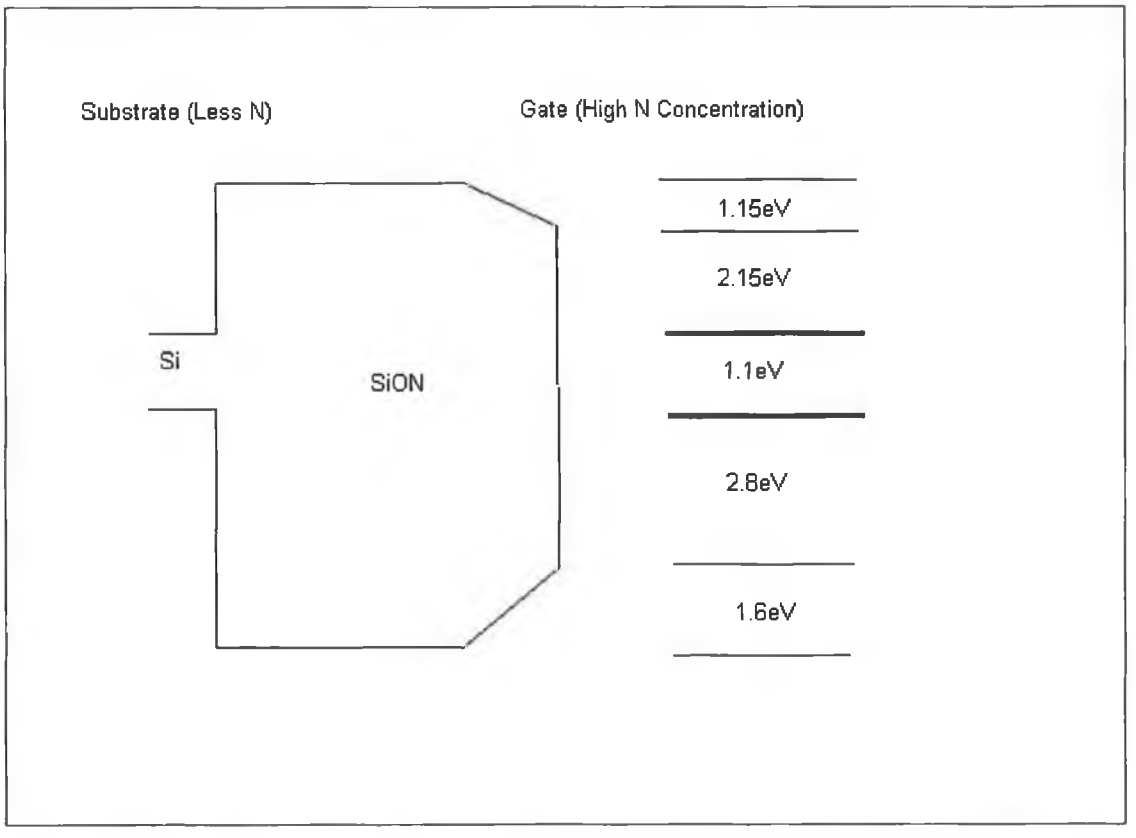


Figure 6.11: Schematic diagram of the oxide/Si band structure. The conduction band data are taken from Lucovsky et al. The valence band data are based on our results.

6.2.4 SIMS

All samples were depth profiled using SIMS as detailed in Chapter 2. A typical positive SIMS scan showing C, N, Si, and Si-O compounds is shown in figure 6.12. The signal from the surface localised carbon, and the nitrogen that resides predominantly at the surface are seen to drop off very quickly as expected.

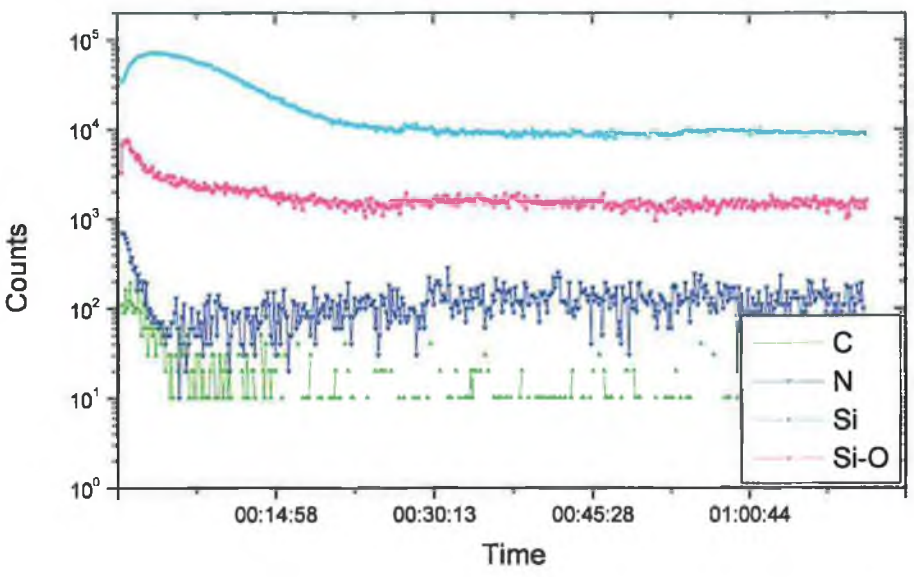


Figure 6.12: SIMS depth profile showing the rapid fall in nitrogen signal at the upper surface. Carbon also falls quickly, indicating that it is surface localised.

The SIMS spectra correlate well with results from ARXPS and further analysis was not undertaken.

6.3 Characterisation of HfSiO and HfSiON Layers

6.3.1 XPS and ARXPS Study

The initial XPS study of the HfSiO layer without a post deposition anneal revealed concentrations as follows: O: 67.7% Si: 18.2% Hf: 14.1% when the surface carbon signal is excluded. Stoichiometric SiO_2 has a 2:1 ratio of oxygen: silicon, so it is clear that within the layer, the Hf atoms take the place of the Si.

With SiO_2 and SiON, it is possible to estimate the thickness of the layer using XPS, by comparing the peak heights of the Si substrate peak, and the Si oxide peak (figure 6.13),using the equation $x=\lambda_f \cos\theta \ln(1-1/Q)$, where $Q=(I_{sub}/I_{overlayer})$ x 0.7.

This method proved quite successful in estimating the thickness of the SiON layers, however, this is only possible when the density of silicon atoms is close to that of SiO_2. This is not the case in HfSiON and so the thickness of the layers cannot be estimated in this way, as the concentration of silicon in the dielectric layer is not accurately known.

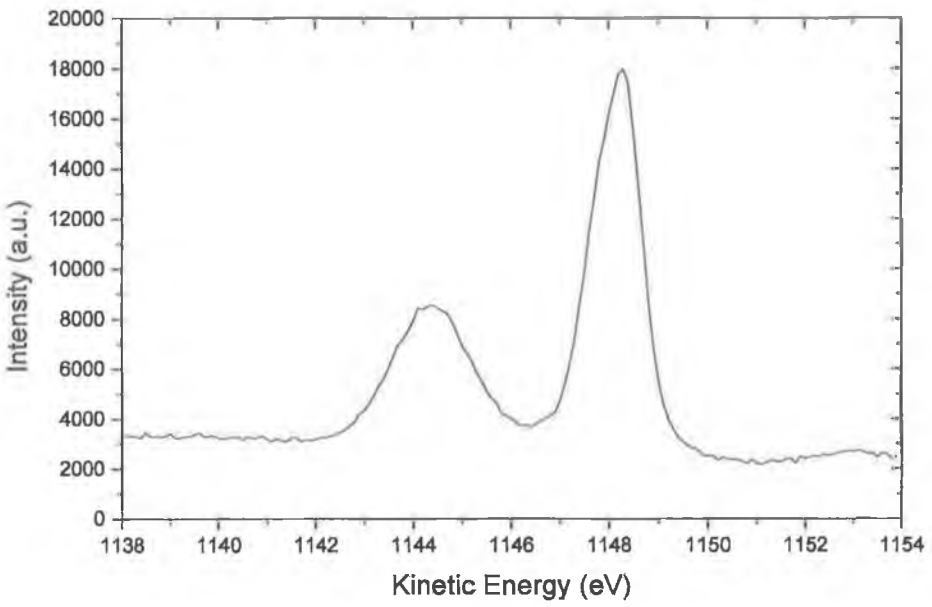


Figure 6.13: Si2p peak for a HfSiON sample. The layer thickness cannot be estimated from the relative peak heights, as it can with SiO_2.

Figure 6.14 shows that at increased angle, the ratio Si:O falls and the ratio Hf: O rises. This suggests that the dielectric layer is HfO_2-like at the upper surface, and SiO_2-like at the Si interface. The oxygen 1s peak can also be used to confirm this. The peak is made up of two components. One component is from oxygen in HfO_2 and the other is from oxygen in SiO_2. The oxygen peak can be fitted as such to give the percentage of oxygen attributed to each, as shown in figure 6.15. In this case, the ratio is 53.5% SiO_2, 46.5% HfO_2.

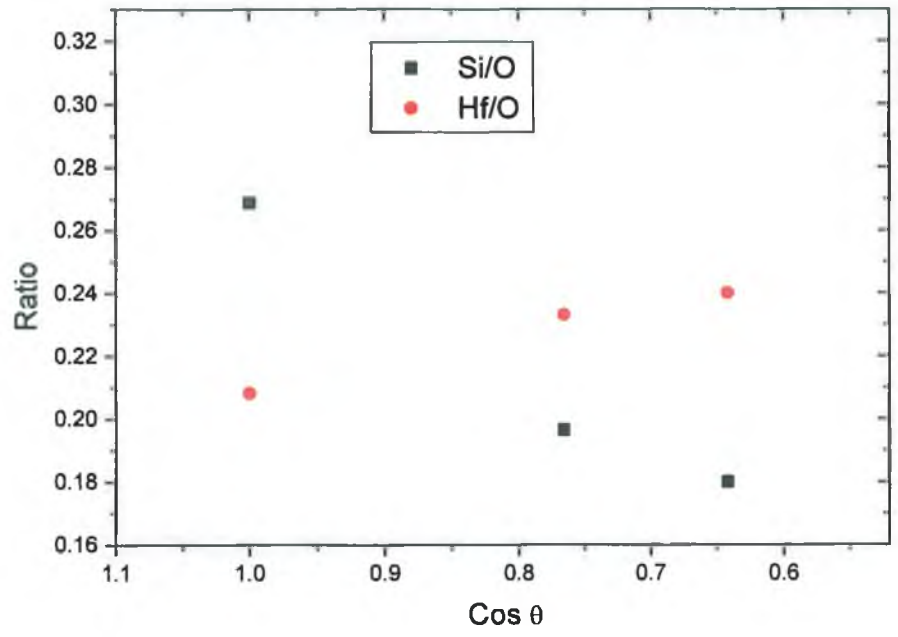


Figure 6.14: The ratio of the intensities of the Si and Hf peaks to the O peaks at various angles indicated that the Hf is located in the upper part of the layer.

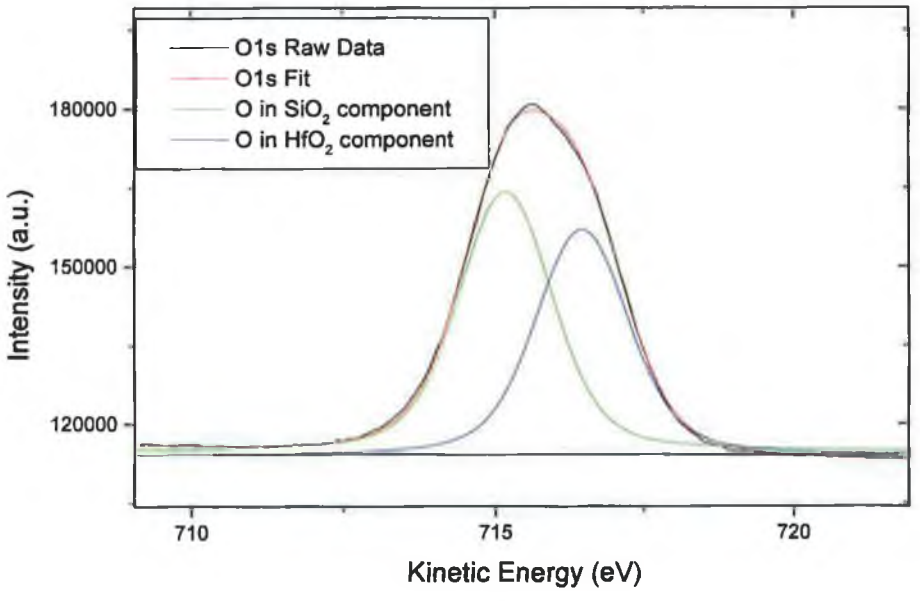


Figure 6.15: The O1s peak can be fitted with two components. One from oxygen in a HfO_2 matrix and from a SiO_2 matrix.

Results for the layer annealed in N_2 were similar to those presented for the non-annealed sample, with a very slight amount (<3%) of N incorporated into the layer. The small amount of N present made depth profiling difficult, and results differed little from those with no anneal, with a HfO_2-like layer sitting on top of a SiO_2 like layer. The N content was too small to accurately determine how it varied with angle. A high level of surface carbon contamination further complicated the angular study.

However, the sample annealed in NH_3 showed N concentration ~10.5% and so even though similar problems with carbon contamination were present, some interesting depth profile results (fig. 6.16) were obtained. The carbon was again confirmed to be surface localised.

The measured concentrations for the elements in the layer are shown in Table 6.4.

Element	Normal	15°	30°	45°	60°
Silicon	19.2%	20%	20.5%	20.6%	21.1%
Nitrogen	10.5%	11%	10.2%	9.5%	7.1%
Oxygen	59.2%	58.4%	57.1%	56.3%	56.7%
Hafnium	11.1%	11.6%	12.2%	13.6%	15.1%

Table 6.4: Measured elemental concentrations for NH_3 annealed HfSiON sample, as a function of take off angle.

From the summary in Table 6.4, the concentration of silicon detected appears relatively constant as a function of angle, indicating that it is uniformly distributed through the film. The oxygen content drops slightly at increased angle, along with the nitrogen signal which shows a large drop (relative to the concentration at normal emission). This indicates that the nitrogen is very much concentrated at the oxide/Si interface, and the oxygen is slightly concentrated in the bottom half of the layer. The hafnium on the other hand increases significantly at increased angle, and so it is concentrated at the upper surface. This information allows us to speculate that the film is structured as a layer with SiON-like chemistry at the substrate interface, and a layer with HfSiO-like

chemistry at the upper surface, with a graded transition region in between. The ratios of the Hf4f peak heights and the N1s peak heights for the NH_3 annealed sample as a function of angle are plotted in figure 6.16.

The ratio varies linearly with cos θ, indicating the N replaces Hf in a Si-O environment deeper into the layer. This agrees with the findings of Lysaght et al[14], who found that the N replaces Hf in their layers that were prepared in a similar fashion, nitrided in NH_3 at 800°C. When the temperature is raised to 900°C however, they find that the N exchanges more with O.

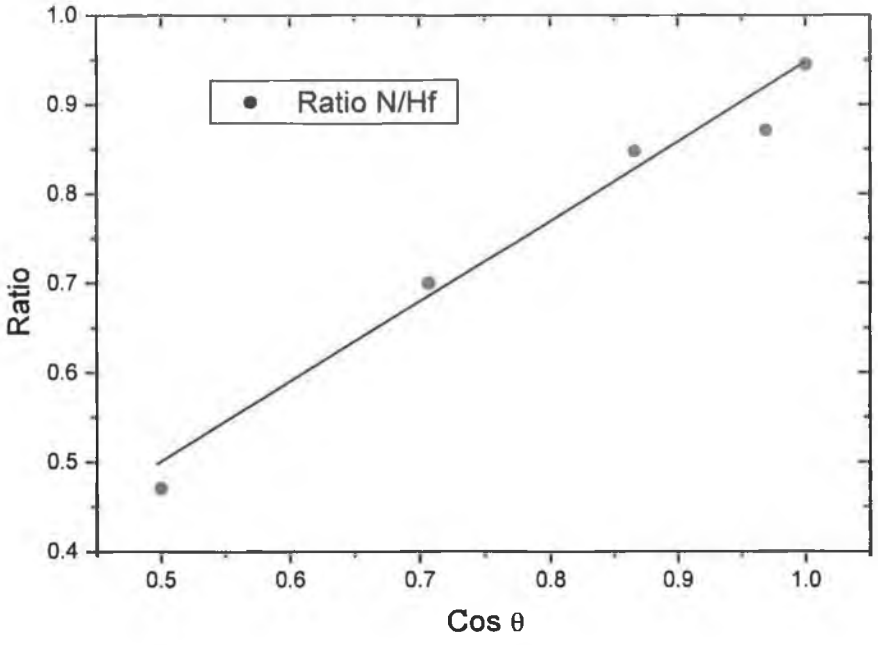


Figure 6.16: Ratio of N1s to Hf4f peak heights. The ratio changes linearly with cos θ. This indicates a transition from HfSiO to SiON deeper into the layer.

6.3.2 SIMS Results

SIMS results show the effect of nitrogen incorporation on the Hf depth profile. Figure 6.17 shows the normalised SIMS depth profile for Hf in the layers under consideration. The profiles are very similar for the samples regardless of whether or not nitrogen is present. Both samples show a high concentration of Hf at the upper interface, the only difference being that hafnium is slightly more concentrated at the upper interface (i.e. the SIMS signal falls more rapidly) for the sample with no anneal. This indicates that the incorporation of the N may

cause the Hf to diffuse through the layer slightly, giving a more graded film. The two layers are drawn schematically in figure 6.18 based on the ARXPS and SIMS results. The main difference between the two is that in the nitrided layer, silicon is found to be distributed uniformly throughout the layer. As the nitrogen has a higher concentration deeper into the layer, we can speculate that the N atoms replace Si, which is brought to the surface.

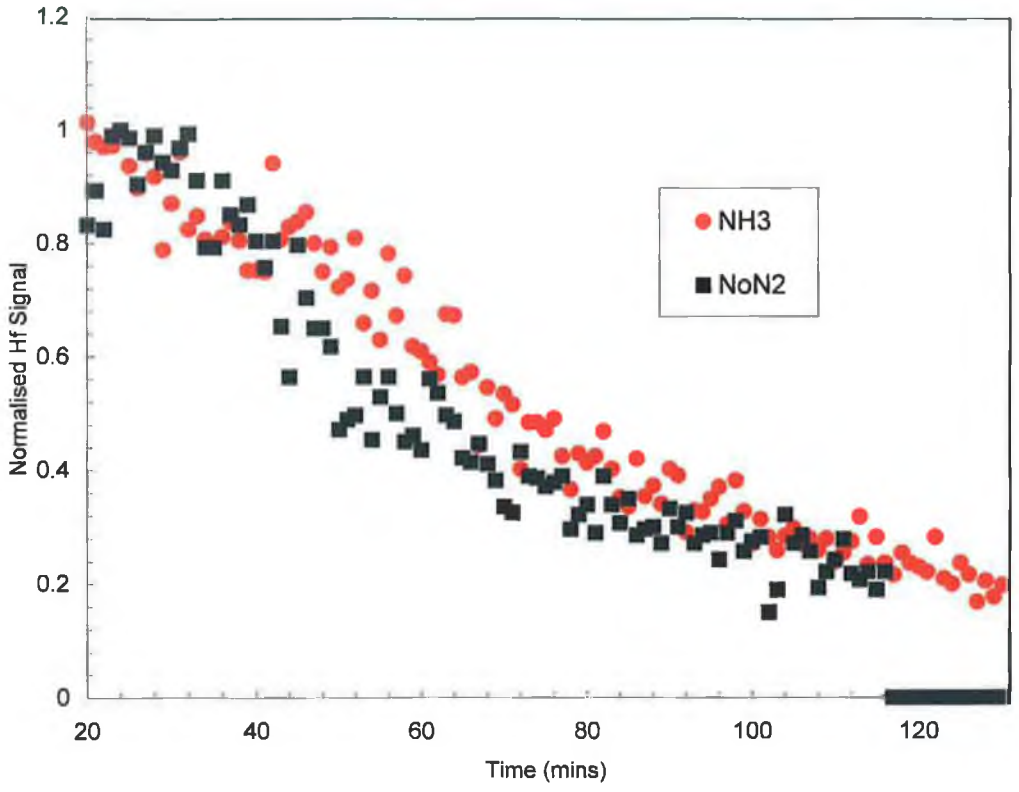


Figure 6.17: SIMS plot normalised to maximum intensity shows that the signal falls slightly faster for the sample with no nitridation, indicating the hafnium is more concentrated at the interface.

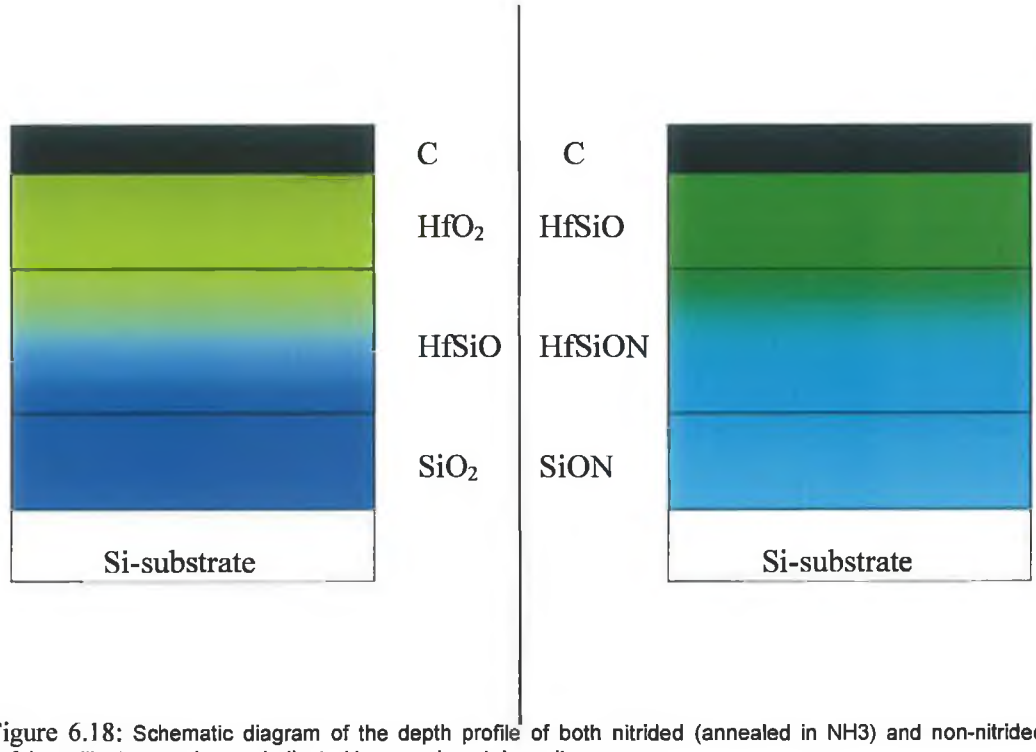


Figure 6.18: Schematic diagram of the depth profile of both nitrided (annealed in NH3) and non-nitrided hafnium silicate samples, as indicated by experimental results

6.3.3 Soft XPS using Synchrotron Radiation

The HfSiO samples underwent a similar set of experiments to the SiON at the NSLS. The HfSiON $Si2p^{4+}$ spectrum was acquired for all samples, spin-orbit stripped and curve fitted, and the FWHM was extracted. The fit for one sample is shown in figure 6.19 at a photon energy of 130eV. From the scan it is clear that the HfSiON layers are thicker than the sampling depth of the technique as no evidence of the substrate peak can be detected. Even at high photon energies of 180eV, which have greater sampling depth into the material, no substrate signal could be detected.

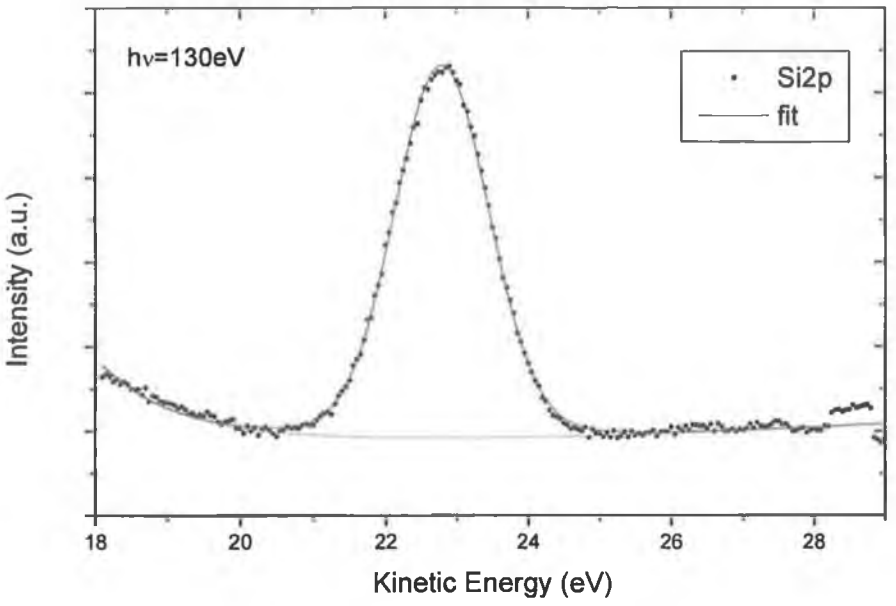


Figure 6.19: HfSiO Si2p 4+ for a sample with no PDA, showing how the FWHM was extracted.

The extracted Si $2p^{4+}$ FWHM for the 3 samples, one with no PDA, one with PDA in N_2 and one with PDA in NH_3 were 1.46eV, 1.43eV and 1.47eV respectively, indicating that PDA has no measurable effect on the Si2p $^{4+}$ width. These widths agree well with results for thick $(HfO_2)_{0.25}(SiO_2)_{0.75}$ films reported by Ulrich et al[15]. Note also that our peak widths are very similar to those obtained for SiON samples.

In addition to the Si2p peak widths, the Hf4f peaks (figure 6.20 for the HfSiON layers were also examined. The average energy splitting (the energy gap between the two peaks) was 1.66eV and the branching ratio (relative height of the two peaks) was 0.71. This is slightly lower than expected, perhaps due to the background fitting technique. The Lorentzian width was fixed at 0.15eV and the Gaussian FWHM was measured.

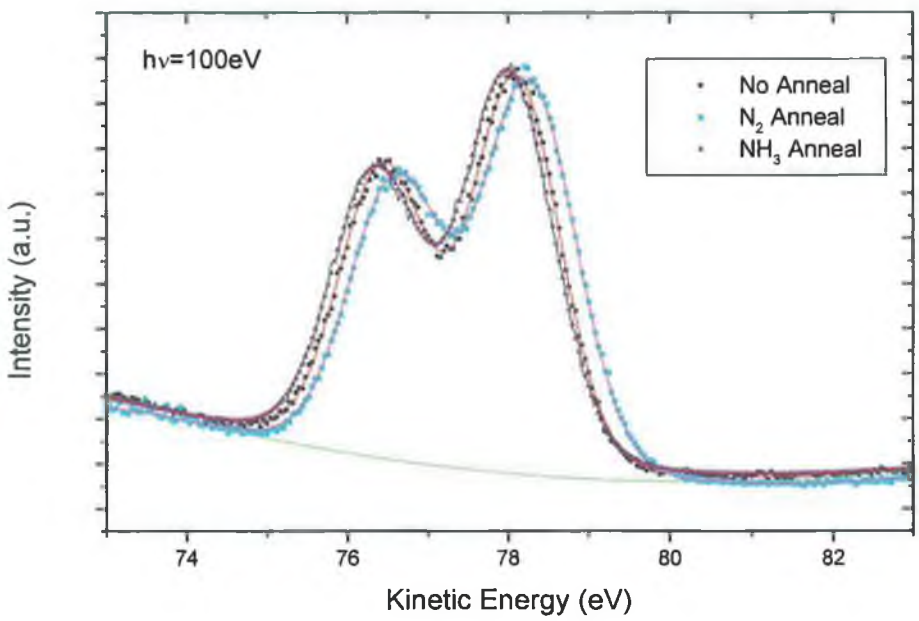


Figure 6.20: Hf4f peaks for all samples studied fit with a Gaussian/Lorenzian peak shape. Post deposition annealing affects the peak widths slightly.

The fits showed that the sample with no anneal had the narrowest width at 1.2eV, and those with an anneal were both significantly wider at 1.38eV. This indicates that the annealing process induces a more complex chemical environment. The value for the non-annealed film is similar to the obtained by Ulrich et al [11] for a pure HfO_2 layer, however the peaks for the other films are significantly broader than Ulrich's findings.

The valence band spectra acquired at 80eV for two of the HfSiO layers are shown alongside a SiON valence band spectrum in figure 6.21. The spectra are similar, however the N peaks are much stronger for the SiON sample. The N2p peak at 66eV appears as a shoulder on O2p peak, and the N2s peak at ~52eV is barely visible, even though the N concentrations are similar in the NH_3 annealed HfSiON layer and the SiON. A possible reason the peak intensity is so low is that as we have seen from XPS, the N in the HfSiON samples is concentrated deep into the layer. As the sampling depth of the SXPS equipment is less than the oxide thickness, it is possible that much of the N is not being probed by the technique, and so only very small peaks are observed.

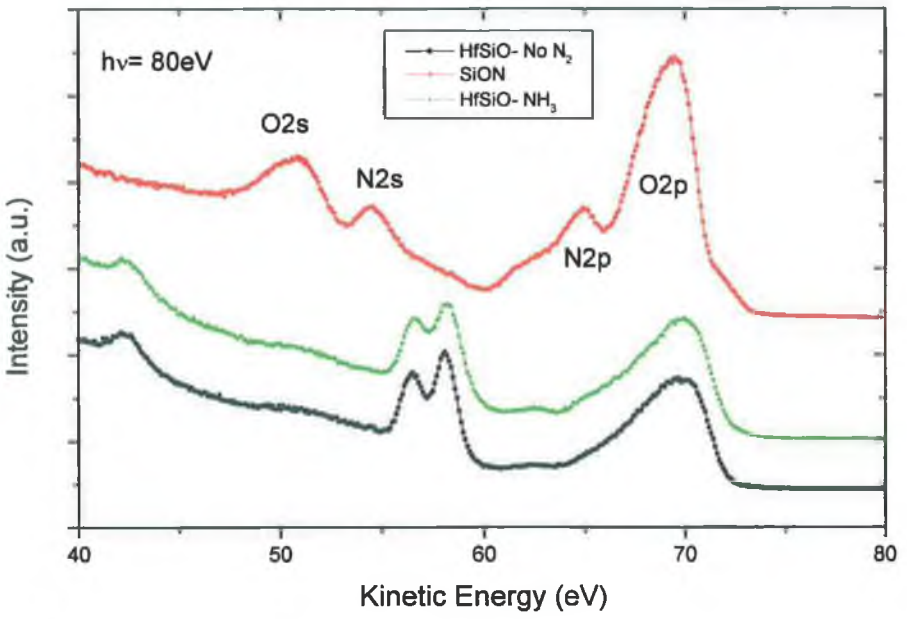


Figure 6.21: Valence band spectra for 2 HfSiO layer. A SiON layer is shown for comparison. The N peaks are not visible for the NH_3 annealed HfSiON because the N_2 is located beyond the sampling depth.

These findings are in agreement with the work of Renault et al[16] on stacked HfO_2/SiO_2 layers. The angular XPS study has shown that our silicate layers also have SiON preferentially located at the interface, and the attenuation of 65eV N peak and broadening of the 68eV O peak could be caused by the presence of a HfO_2-like layer on top.

The band offsets for a HfSiON layers is shown in figure 6.22 and again compared with the results for SiON. The valence band offset is shifted for the HfSiON layers with respect to the SiON, in agreement with the findings of Renault et al[12]. Knowing that the valence band offset for the SiO_2 system is 4.4eV we can calculate that the offsets of the silicate layers are 3.62eV and 3.43eV for the layers with no anneal and NH_3 anneal respectively.

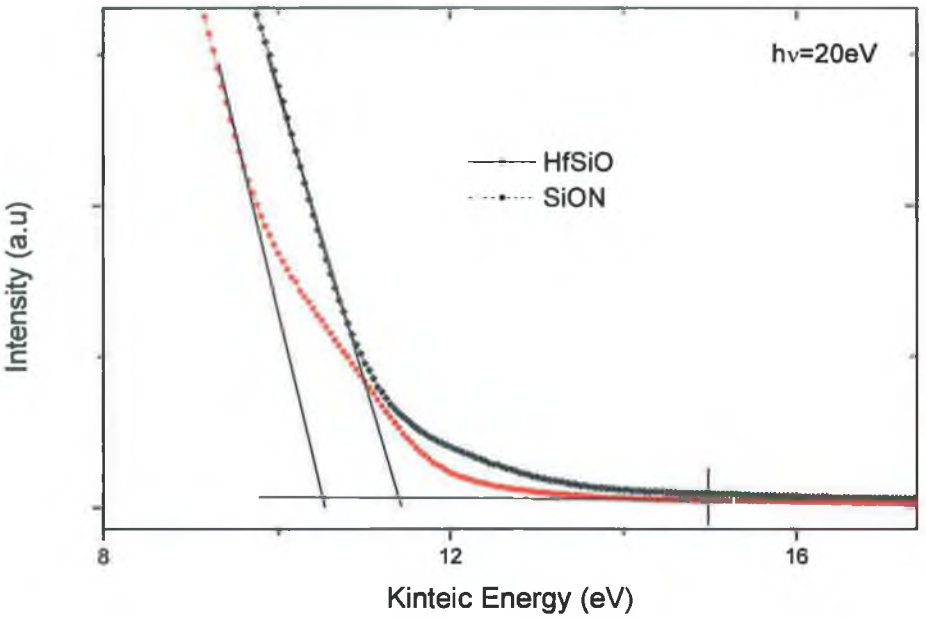


Figure 6.22: Valence band offsets for a HfSiON sample. A SiON sample is shown for comparison

In both cases, the offset is not very clear and there is still significant signal at kinetic energies above the extrapolated valence band offsets. This may be due to interface states within the band gap. Referring back to figure 5.1, the calculated offsets are very similar to those for pure HfO_2. The gap between the SiON valence band offset and the HfSiON offset is ~1eV. This agrees well with literature values[12].

6.3.4 X-ray Absorption Spectroscopy

The Soft-XPS results discussed above give information about the valence band offset in the layers. As described in chapter 2, the conduction band offset can be extracted from XAS measurements. The HfSiON samples were measured by XAS and the results for the oxygen absorption edge are shown in figure 6.23. Also shown for reference are XAS O edge spectra for NiO (used as a reference metal-oxide), and HfO_2 powder.

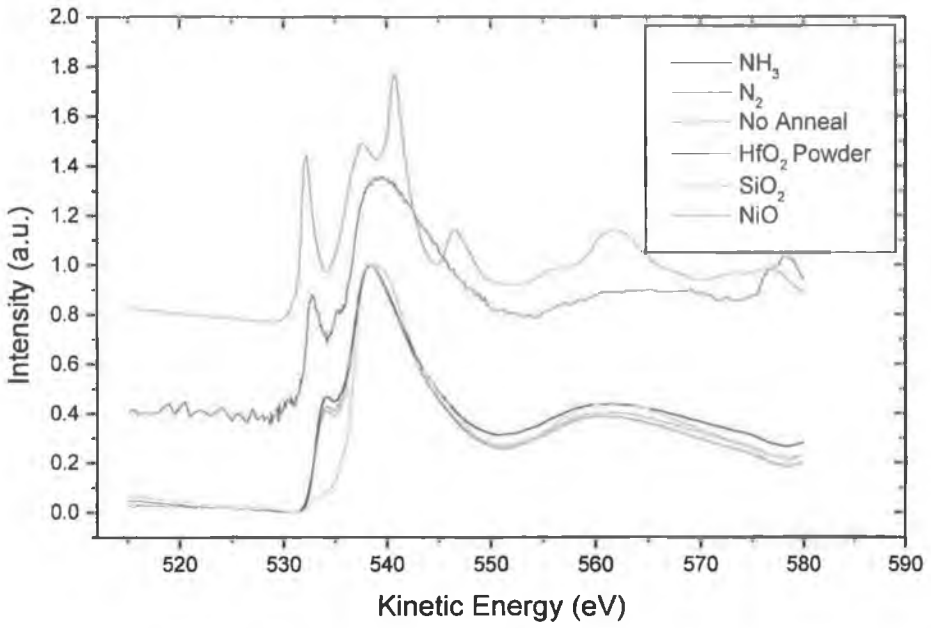


Figure 6.23: Oxygen K edge XAS spectrum for HfSiON samples. Conduction band offsets can be measured by comparing the absorption edges with that of SiO_2.

The absorption edge is visible at 532eV for the samples. When compared with the O absorption edge for SiO_2, a difference of between 2.8eV and 3eV is observed with the sample with no anneal having the lowest difference, and the NH_3 annealed sample having the largest offset. The conduction band offsets are thus between 2.8 and 3eV lower than for SiO_2, which has a widely accepted value of 3.3eV, giving offsets of between 0.3eV and 0.5eV. These values are much lower than the 1.5eV reported for pure HfO_2, and do not represent a significant barrier to tunnelling electrons.

By combining the conduction band and valence band offsets from photoemission and x-ray absorption measurements respectively, figure 6.24 presents a schematic illustration of the band structure of the layers with and without NH_3 anneals. Interestingly, the total bandgap of the two; that is $(E_g(Si)+\Delta E_V+\Delta E_C)$ are almost equal at slightly above 5eV. The sample with no anneal has a slightly lower conduction band offset and higher valence band offset than the annealed sample. The interface is shown as slightly graded for the non-annealed sample in accordance with valence band offset measurements, presumably due to interface

states. The annealed sample shows a more severe grading because of the combined effect of interface states and nitrogen incorporation.

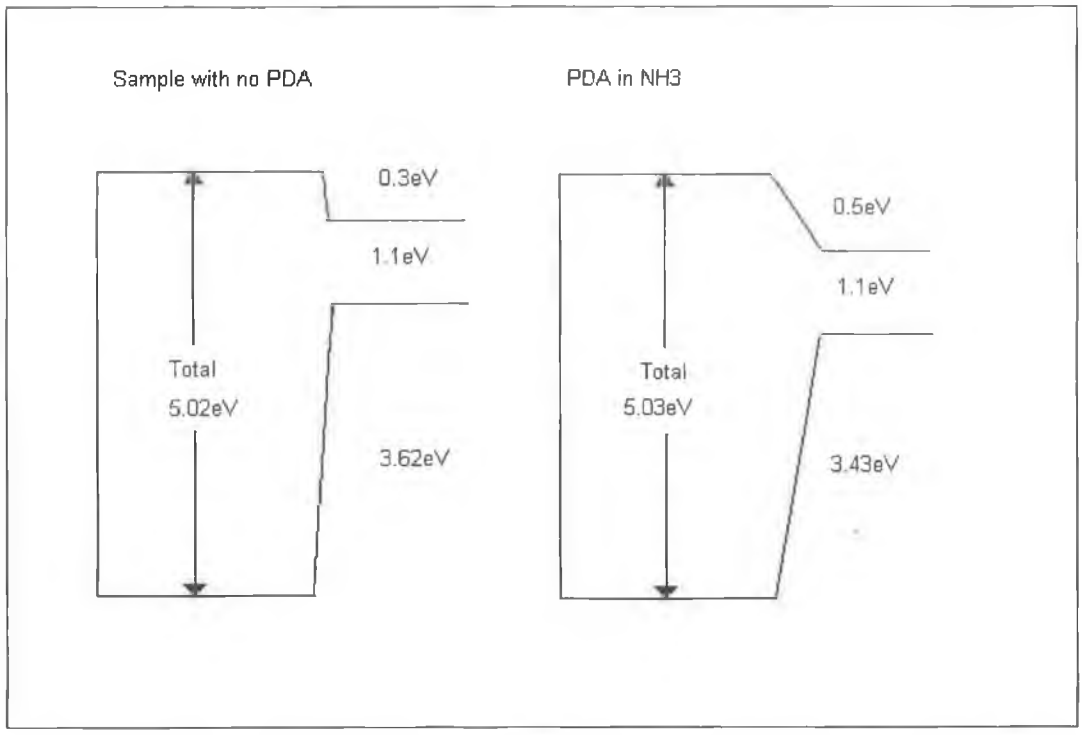


Figure 6.24: Schematic diagrams of HfSiON band structures based on synchrotron results. The band gaps are identical, however the conduction band and valence band offsets differ.

6.4 Summary

This chapter presents some interesting results on the chemistry of ultrathin dielectric layers, using fundamental surface analysis techniques. Conventional ultrathin SiON layers nitrided by pulsed rf-DPN were found to have a high N content, which is beneficial in suppressing boron penetration, reducing interface states and increasing the dielectric constant. ARXPS showed that the N was concentrated near the surface of the layer particularly when a high rf-plasma power or a PNA in H_2 were used. The shift in the kinetic energy position of the N1s peak at high take off angles, and SIMS depth profile measurements further confirmed this N profile. Synchrotron measurements showed that the Si2p peak width was wider than for SiO_2 representing a broader range of chemical states as would be expected. The peak widths were similar to those found in literature. Valence band spectra were also similar to those previously reported, and the

valence band offset was found to have 2 components, one relating to the SiO_2/SiON transition region and one from the SiON/Si interface. This results in a stepped valence band offset.

The hafnium silicate samples underwent a similar analysis. In the sample that had no PDA, the SiO_2 and HfO_2 was found to form a layered type structure, with a HfO_2 rich layer above a SiO_2 rich layer. For the sample with a PDA in N_2, the Si and O were found to be distributed uniformly through the layer, with Hf present in the upper part of the layer and N replacing it further down, towards the substrate interface. This layer can be roughly described as a layer of HfSiO, grading into SiON at the interface. The synchrotron measurements showed wider Si2p peaks than the SiON as expected and Hf4f peak widths that were very similar to a $(HfO_2)_{0.25}(SiO_2)_{0.75}$ mixture studied by Ulrich et al[16]. Valence band spectra were affected by the low sampling depth of the technique, and because the N was located deep in the layer, the N peaks were attenuated. Valence band offsets showed only one component and a significant number of electrons were detected beyond the extrapolated offset, possibly due to interface states, or nitrogen incorporation. XAS measurements showed the layers to have very low conduction band offsets (0.3eV-0.5eV). Such low conduction band offsets mean that any benefit of using a physically thicker layer may be outweighed by the increasing tunnel current caused by the low barrier to tunnelling electrons.The total bandgap of the layers was found to be ~5eV. This is much lower than the ~7eV that Takeuchi et al[17] found for films with a similar ratio of Hf/Si+Hf.

6.5 References

[1] C.M.Lek, B.J.Cho, C.H.Ang, S.S.Tan, W.Y.Loh, J.Z.Zhen and C.Lap, Semicond. Sci. Technol. 17 (2002) L25-L28.

[2] R.Kraft, T.P.Scgneider, W.W.Dostalik and S.Hattangady, J.Vac.SciTechnol. B15(4) (1997) 967-970.

[3] J. W. Kim, H. W. Yeom, Y. D. Chung, K. Jeong, and C. N. Whang, Chemical configuration of nitrogen in ultrathin Si oxynitride on Si(100), Phys. Rev. B 66 (035312) (2002)

[4] S. Miyazaki, T. Tamura, M. Ogasawara, H. Itokawa, H. Murakami and M. Hirose, Influence of nitrogen incorporation in ultrathin SiO_2 on the structure and electronic states of the SiO_2/Si(100) interface, Appl. Surf. Sci. 159-160 (2000) pp. 75-82.

[5] M. P. Seah, W. A. Dench, Quantitative electron spectroscopy of surfaces: A standard data base for electron inelastic mean free paths in solids, Surf. Interface Anal 1 (1979) pp. 2-11.

[6] G.-M. Rignanese, Alfredo Pasquarello, J.-C. Charlier, X. Gonze, and Roberto Car, Nitrogen Incorporation at Si(001)-SiO2 Interfaces: Relation between N *1s* Core-Level Shifts and Microscopic Structure, Phys. Rev. Lett. 79, (1997) pp. 5174-5177

[7] Z.H. Lu, S.P. Tay, R. Cao, P. Pianetta, The effect of rapid thermal N_2O nitridation on the oxide/Si(100) interface structure, Appl. Phys Lett. 67 (19) (1995) pp.2836-2838.

[8] R.I. Hegde, P.J. Tobin, K. G. Reid, B. Maiti, and S.A. Ajuria, Growth and surface chemistry of oxynitride gate dielectric using nitric oxide, Appl. Phys. Lett. 66 (21) (1995) pp. 2882-2884

[9] M. Takahashi, T. Mizokuro, Y. Nishioka, H. Kobayashi, Experimental and theoretical studies on N1s levels of silicon oxynitride films, Surf. Science, 518 (2002) pp. 72-80.

[10] QUASES ARXPS. S. Tougaard, T.S. Lassen. Odense, Denmark.

[11] J. Roche, PhD Thesis, Pg 71. Dublin City University (2002)

[12] J.W. Keister, J.E. Rowe, J.J. Kolodziej, H. Niimi, T.E. Madey, G. Lucovsky, Band offsets for ultrathin SiO_2 and Si_3N_4 films on Si(111) and Si(100) for photoemission spectroscopy, J. Vac. Sci. Technol. B, 17 (4) (1999) pp. 1831-1835.

[13] H. Niimi, H. Yang, G. Lucovsky, J. W. Keister and J. E. Rowe, Barrier layer model determined by XPS data for tunneling current reductions at monolayer nitrided Si–SiO_2 interfaces, *Appl. Surf. Sci., 166 (1-4), (2000) pp. 485-491*

[14] P. Lysaght, B. Foran, S. Stemmer, G. Bersuker, J. Bennett, R. Tichy, L. Larson, H.R. Huff, Thermal response of MOCVD hafnium silicate, Microelec. Eng. 69 (2-4) (2003) pp.182-189.

[15] M.D. Ulrich, J.G. Hong, J.E. Rowe, G. Lucovsky, A. S.–Y. Chan, T.E. Madey, Soft x-ray photoelectron spectroscopy of $(HfO_2)_x(SiO_2)_{1-x}$ high-k gate-dielectric structures, J. Vac. Sci. Technol. B 21 (4) (2003) pp. 1777-1782.

[16] O. Renault, N.T. Barrett, D. Samour and S. Quiais-Marthon, Electronics of the SiO_2/HfO_2 interface by soft X-ray photoemission spectroscopy, Surface Science 566-568 (1) (2004) pp. 526-531

[17] H. Takeuchi, T.J. King, Scaling limits of hafnium silicate films for gate dielectric applications, Appl. Phys. Lett. 83 (4) (2003) pp. 788-790.

Chapter 7: Conclusions and Future Directions

This chapter contains a brief summary of the key results from this work, and the important conclusions that can be drawn from the measurements. The second section of the chapter deals with future directions of CMOS technology with an emphasis on the gate dielectric.

7.1 Conclusions of this Work

7.1.1 Electrical Measurements

Chapters 4 and 5 detail the experimental results for electrical testing of ultra-thin oxide layers, using conventional testing techniques. Although the methods used are conventional, a number of issues arise when dealing with layers in the sub-two nanometre thickness range, and such issues are the main focus of the first part of chapter 4. The first point is that because the layers are so thin, and the leakage current so high, results for a number of techniques are very difficult, if not impossible to comprehend. Only the smallest test structures can be used, as series resistance becomes a problem with larger devices. Charge pumping currents appear as a bump on the leakage current profile and in the thinnest of layers (with the highest leakage current) the charge pumping current is undetectable. Neutral electron trap densities can no longer be measured by filling the traps and measuring current as, in ultrathin layers, the traps empty much too quickly.

Of course, the most important point when measuring on ultrathin layers is the difficulty in making a reliability prediction using conventional methods. The determination of t_{bd} becomes difficult, as the current rise associated with t_{bd} is progressive. Using such methods results in very high Weibull slopes, and low voltage acceleration, and a relationship between the two, which has not been observed before. After some further analysis, it was found that measuring the t_{bd} as the time to a significant current increase is not correct. The breakdown path actually forms long before this, but only has a very slight current rise associated

with it. Thus, measuring t_{bd} at the point where the current begins to rise, is actually a measure of the resistance of the breakdown path to wearout, and this is the reason for the high Weibull slopes and low voltage acceleration.

Even with this knowledge, making an accurate reliability projection proves very difficult. When the current begins to rise after breakdown, it does so quite slowly, and some FETs can in fact continue to function long after the oxide has broken down. It would therefore be illogical to categorise an oxide as broken, when the device into which it is fabricated continues to function. The ability of the devices to function after breakdown is dictated by several parameters; the location of the breakdown, the device aspect ratio, whether the device is nmos or pmos, amongst others. Because the location of the breakdown path is random, it is then almost impossible to make a lifetime prediction based on (t_{bd} + post bd time), because every device will have a different post bd time.

The high temperature study was also complicated by the difficulty in breakdown detection. As the breakdown becomes 'softer' at high temperature, using a high current trigger results in measuring the t_{bd} even further into the progressive wearout regime, and so relating data acquired at room temperature and elevated temperature becomes difficult.

The main conclusions of Chapter 4 are ultimately based on the methodology of time dependent dielectric breakdown in ultra-thin layers, and not on the actual performance of the particular SiON layers used in the study. The initial idea was to evaluate the layers in comparison with conventional SiO_2. However, using the same measurement methods was not comparing 'like-with-like' due to the behaviour of ultrathin layers, and so the focus switched to addressing the measurement problems.

The high-k layers, studied in chapter 5 show similar measurement behaviour to the SiON, in that the actual t_{bd} is quite difficult to detect, particularly when the I-t trace contains many features that appear as breakdowns but shortly afterwards the current and noise reduce to their original level. Using conventional t_{bd} triggers (i.e. current level) results in a much lower than expected Weibull slope. This is because breakdown can be triggered by a path within the oxide that consists of fewer traps, and thus has a lower associated Weibull slope. So with physically thicker high-k layers, it is possible to overestimate the Weibull slope as with the SiON layers but also to underestimate it. The low voltage acceleration

values measured could have been due to a measurement artefact, or it is possible that trap generation in HfSiON has lower voltage acceleration than in SiON.

The high temperature work shows the temperature dependence of t_{bd} also has a thickness dependence, in that the t_{bd} at two different temperatures for a thin layer will be much closer together than for a thick layer. This shows that bulk trapping is responsible for the temperature dependence, and the more bulk, the more temperature dependence observed.

The problem of charge trapping in HfSiON was found to be a much greater threat to devices than in SiON and this was investigated in detail. There were excessive threshold voltage shifts in thicker layers, even after only a few hundred seconds at operating voltage, attributed to bulk charge trapping. This trapped charge coupled with interface charge trapping led to severe transconductance degradation and sub-threshold swing. The charge trapping was investigated in terms of gate injection vs. substrate injection stress, and it was demonstrated that substrate injection caused more bulk trapping, while interface trapping was more severe when using gate injection.

In terms of device operation, the large charge trapping causes a constant degradation in the on-current of FET during device lifetime. Behaviour in the post breakdown regime was erratic, whereby even on devices of the same aspect ratio and with breakdown in the same location, some devices would continue to function for quite some time, and others would cease to function immediately.

Stress induced leakage current was found to be sense voltage dependent, and this dependence was attributed to bulk traps which assisted tunnelling at gate voltages approximately equal to the device flatband voltage. The initial trapping in HfSiON leads to a large SILC level compared with SiON, and thus any leakage current reduction from the high-k material is quickly cancelled out.

7.1.2 Surface Analysis Results

The surface analysis study consisted of 3 main sections: the determination of elemental concentrations from XPS, the depth profile of the elements within the layer using angular XPS, and determination of the valence band offsets from soft

x-ray photoemission. The depth profile information from ARXPS was confirmed with SIMS measurements.

For the SiON, XPS showed layers had ~24% Si independent of preparation and nitrogen content. The remaining 76% was divided between N and O depending on the amount of N incorporated. This showed that N takes the place of O atoms when it is incorporated. The angular results showed that the oxide Si2p peak was drawn to lower binding energies at high angle, where the technique is most surface sensitive. This was also true of the N1s peak. This indicated that the N was located in the upper part of the layer. Analysing the N1s peak height as a function of angle supported this. The layers were Si_3N_4 like in the upper part and SiO_2 like in the lower part at the interface. This profile was confirmed with SIMS.

Synchrotron measurements showed Si2p peak widths of 1.38eV-1.55eV, wider than SiO_2 as expected. Valence band spectra were used to calculate the valence band offset. When looking at thin SiON layers there are actually two offsets to consider, as the nitrogen alters the band structure. The SiO_2/SiON VB offset was calculated as 1.6eV and the SiON/Si offset as 2.8eV.

The HfSiON study consisted of layers with and without nitrogen. The layer with no N showed concentrations of O~68% Si~18% and Hf~14%, indicating the Hf takes the place of Si, when compared with SiO_2. The angular results for the layer shows that it is not uniform, but that HfO_2 preferentially forms at the top surface and SiO_2 at the bottom. Fitting the O1s peak with 2 components at increasing angle confirms this. For the NH_3 annealed layer, concentrations of Si~19% O~59% Hf~11% N~11% were observed. At increasing angle, the ratio of N/Hf was found to fall indicating that the Hf was located at the upper interface, and the N at the lower interface. The Si and O concentrations were found to be relatively constant as a function of angle. The layer is thus HfSiO-like at the top surface, and becomes SiON-like at the interface. This profile was also confirmed with SIMS measurements.

Soft XPS yielded Si2p widths of ~1.45eV for all samples, which agree with results in literature. The Hf4f width was 1.2eV for the sample with no anneal and 1.38eV for both samples that underwent an anneal. Valence band spectra were similar to those for SiON. However, they showed almost no N was present. This

was presumed to be because the N was located deep in the layer, beyond the sampling depth of the technique.

The valence band offset was difficult to extract, as there was no clear cut off, so it was extrapolated and measured as being between 3.4-3.6eV. These values are similar to those reported for HfO_2. The conduction band offset was also measured for the HfSiON using x-ray absorption spectroscopy. When compared with SiO_2 differences of between 2.8ev and 3eV were observed, giving offsets of 0.3eV and 0.5eV depending on whether nitrogen was present. Taking results from both valence band and conduction band offsets, the bandgap of the layers was calculated as ~5eV.

7.2 The Future for High-k Dielectrics[1]

As the results in this thesis show, furthering the properties of SiON may prove difficult and even if further advances are made, reliability will more than likely remain uncertain. Replacing SiON is certainly not a simple matter of finding a material with a high dielectric constant and 'plugging it in'. The replacement material must meet all of a number of conditions, such as having a sufficient barrier height. The HfSiON layers were found have a conduction band offset as low as 0.3eV, which is unacceptably low. Further to this, the material must have good thermodynamic stability on Si. Almost all materials analysed to date form a parasitic SiO_2-like layer at the interface. This has the effect of lowering the dielectric constant and thus the overall capacitance. Our HfSiO layer showed phase separation into SiO_2 and HfO_2.

The interface quality is also very important. The target is to have an interface quality as close to SiO_2 as possible, with $<2x10^{10}$ states/cm^2. Our layers show that this goal is achievable with ~5-6x10^{10}/cm^2, possibly due to the formation of SiO_2 at the interface. For this reason, stacks of high-k material on SiO_2 are being investigated, to take advantage of the high quality interface of the SiO_2 and the high-k of the metal oxide.

The crystallisation temperature of the high-k layers also needs to be considered. If the processing takes place at temperatures such that the layer begins to

crystallise, this can have serious consequences for the electrical properties of the layers. Grain boundaries can form and act as leakage paths, even before the layer undergoes an electrical stress.

Ideally, the gate dielectric should be compatible with poly-Si gates, so as to allow control of the gate dopant density, and thus allow control of the threshold voltage. However, almost all high-k materials to date require metal gates, to avoid reaction between the dielectric and the poly-Si at the gate interface, this problem further lowers the capacitance. To date, silicates have shown promise in this area.

The most serious roadblock for the high-k materials found in this work is the problem of charge trapping. As the goal is to have as high a 'k' value as possible, the physical thickness that can be used to achieve the same capacitance value will go up accordingly. However, the fact that these materials are good insulators and that there will always be some current flow, no matter how small, means that there is an increased risk of charge trapping in the film. As our results have shown, for the thicker high-k layers investigated in this study significant issues need to be addressed before they are in a position to replace sub-2nm SiO_2 based dielectric layers.

Thinner conventional dielectric layers have an advantage in that there is no "bulk" for the trapped charge to lodge in, and so, this problem may be very difficult to resolve.

The challenges facing the semiconductor industry from the scaling of the gate dielectric are clear, and they need to be solved in line with an acceptable timeframe, which makes extensive research more difficult, and therefore full confidence in the chosen materials may never be attained. However, it is likely that materials will prove at least satisfactory for the duration of conventional CMOS technology, which is expected to be phased out due to other limitation in less than 15 years.

---

[1] This section has been written using reference to 'High-k gate dielectrics: Current status and materials properties considerations' by G.D. Wilk, R.M. Wallace, and J.M. Anthony, appearing in: Appl. Phys. Rev. 89 (10) (2001) pp. 5243-5275