RC 20778 03/20/97
Subject : Engineering and Technology

Research Report

Reliability and integration of ultra-thin gate dielectrics for advanced CMOS

D.A. Buchanan and S.-H. Lo
IBM Research Division
T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

LIMITED DISTRIBUTION NOTICE
This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties).

IBM **Research Division**
Almaden T.J. Watson Tokyo Zurich

Reliability and integration of ultra-thin gate dielectrics for advanced CMOS

D.A. Buchanan and S.-H. Lo

IBM Research Division, T.J. Watson Research Center,
PO Box 218, Yorktown Heights, NY 10598 USA

The rapid scaling of CMOS technology over the last few years has resulted in the reduction of gate dielectric thickness to only a few nanometers. Some estimates suggest that standard SiO_2 (or its nitrogen containing variants) will cease to be useful when the thickness is reduced to between 1.0-2.0 nm. The major causes for concern, assuming that such thin dielectrics may be grown with an acceptable yield, seem to be centered upon either process integration issues, (for example boron penetration), or upon the increasing gate leakage which lead to questions regarding dielectric integrity and reliability, or to standby power consumption. The future of CMOS down to and below the 0.1 μm gate lengths may well hinge on finding a replacement of silicon dioxide as the gate insulator. However, even for 2.0-4.0 nm dielectrics currently being used for the development of 0.1-0.25 μm technologies. There is only limited understanding of the issues pertaining to the reliability of the gate dielectric.

1. INTRODUCTION

Dielectric thickness' as thin as 1.5 nm have been used to fabricate 0.1 μm CMOS devices[1,2]. With the rapid reduction in the dielectric thickness, numerous difficulties have arisen regarding their growth, integration and the reliability. In this work, we will concentrate on issues relating specifically to the integration and reliability of these ultra-thin dielectrics, in particular with regard to CMOS device operation.

2. MODELING : THICKNESS EXTRACTION AND QUANTUM MECHANICAL TUNNELING

Currently one of the issues of greatest concern to those working in the area of thin gate dielectrics, is that of dielectric integrity. This issue comprises many areas including, but not limited to, wafer cleaning and dielectric film growth, characterization and modeling. The issue of modeling encompasses both the detailed calculations of the quantum mechanical tunneling current, as well as poly-depletion effects and the prediction of dielectric integrity as it pertains to CMOS device lifetime. But this modeling requires a fundamental physical understanding of the relationship between the applied bias, band bending in the silicon substrate as well as the gate leakage current that stems from direct quantum mechanical tunneling.

2.1 Dielectric Thickness Extraction

The modeling of the leakage currents for ultra-thin dielectrics requires obtaining an accurate value for the physical dielectric thickness. Most electrical extraction schemes are not able to separate the dielectric thickness, from poly-depletion effects or quantization effects in either inverted or accumulated surface layers.

It has been found that the capacitance of thin oxide MOS devices in strong accumulation (or inversion) does not saturate, as it would for thicker dielectric films, but changes as the gate bias is changed. This is in part due the quantization effects arising from the change in the effective thickness of the silicon surface accumulation (or inversion) layer.

This layer can be of the order of several angstroms[3] and may therefore have a surprising large effect, especially for the thinnest gate dielectrics. Using a one dimensional (1-D) Poisson - Schrödinger solver, a capacitance-voltage curve can be calculated accounting for the quantization of the silicon surface. In Figure 1, measured and calculated capacitance-voltage curves are shown for an n^+ gate/p-type substrate and a p^+ gate/n-type substrate capacitor. The oxide thickness was 2.5 nm. From this modeling work it was found that a

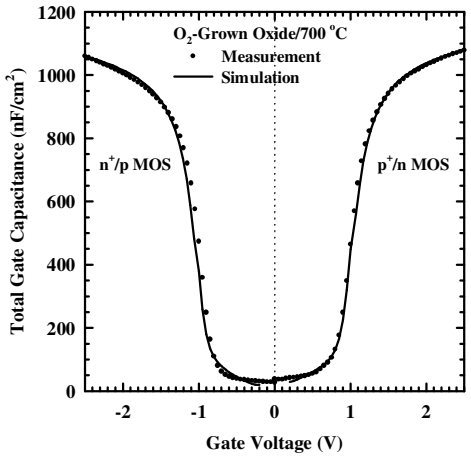

Figure 1 : Measured and simulated (QM) C-V curves of an n^+ gate/p-type substrate and a p^+ gate/n-type substrate capacitor.

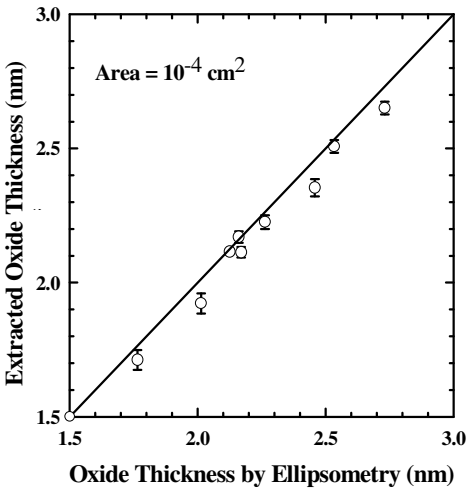

Figure 2 : Comparison between C-V extracted oxide thickness and measured ellipsometry thickness. The data represent averages of nine and 64 points for the ellipsometry and C-V data respectively. The solid line shown in Figure 2 has a slope of one is given as a guide to the eye and is not a fit to the data.

reasonably good agreement can be found between the oxide thickness extracted from the simulation (over the range of gate bias of interest) and that obtained using standard ellipsometry techniques. In Figure 2, a comparison between thickness' obtained by standard ellipsometry and those extracted from the simulation are shown. The data represent averages of nine and 64 points for the ellipsometry and C-V data, respectively. The error bars represent the mean square error of the dielectric thickness of the 64 capacitors measured across a 200 mm wafer and as such represent the thickness variation of the dielectric across the silicon substrate.

2.1 Quantum-Mechanical Tunneling Current

Electron tunneling current from the quantized inversion layer was calculated using a transverse-resonant method.[4] Figure 3 shows the excellent agreement between the calculated and the measured I_G–V_G characteristics of four nFETs from different lots with dielectric thickness' varying from 2.9–3.6 nm. Simulated I_G–V_G characteristics for oxides down to 1.5 nm are also shown in Figure 3. At a gate bias of 1.5 V, the current density increases by 10 orders of magnitude as the oxide thickness decreases from 3.6 nm to 1.5 nm. The tunneling current density calculated for a 1.5 nm oxide agrees reasonably with reported experimental data. [1]

While the gate leakage current may be at a level negligible compared with the on-state current of a silicon MOSFET, it may have a large effect on the chip standby power. If one assumes that the total active gate area per chip is of the order of 0.1 cm^2 for future generation technologies, the maximum tolerable gate current (for a standby power dissipation of 10 W) would be of the order of 1 A/cm^2 for a power supply of 1 V. Assuming that this

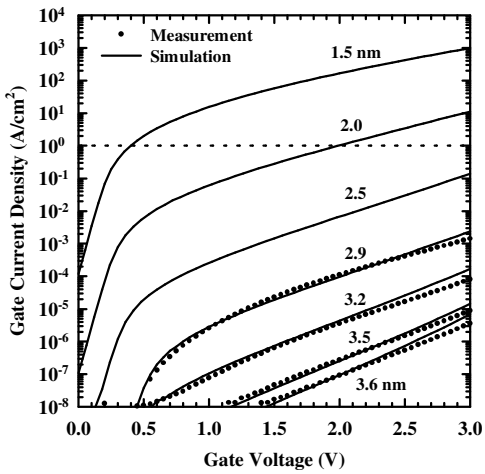

Figure 3 : Measured and simulated I_G-V_G characteristics under inversion conditions. The dotted line indicates a value of 1 A/cm^2 that may be the maximum operating gate leakage current (see text).

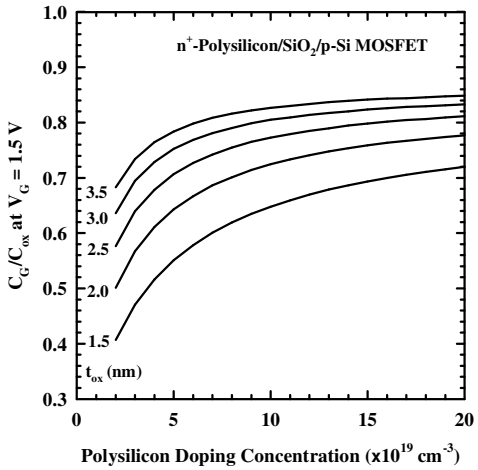

Figure 4 : The ratio of gate capacitance to oxide capacitance versus the effective poly-silicon doping concentration for oxide thickness from 1.5 to 3.5 nm.

standby power is the limiting factor, the minimum gate silicon dioxide dielectric thickness would be 2.0 nm as can been seen from Figure 3..

3. POLY-SILICON DEPLETION AND BORON PENETRATION

Poly-depletion effects are also related to the dielectric scaling since the operating bias is not being scaled at the same rate as the dielectric thickness. Therefore the average surface field has increased. The increasing surface field and the reduction in the effective poly-silicon doping concentration leads to a substantial poly-silicon depletion effect.

The adverse properties of a reduction in the effective poly-silicon doping concentration, as demonstrated by the data in Figure 4, the gate oxide capacitance ratio is calculated as a function of the effective poly dopant concentration for dielectric thickness from 1.5 to 3.5 nm. For very thin films, (i.e. 2.0 nm or less), the poly depletion effect can result in a reduction of the effective gate capacitance by almost a factor of 50%, if the effective gate doping level reaches values as low as 2×10^{19} cm^3. Typically the effective doping concentration has been found to be in the range from $0.5\text{-}1.0\times10^{20}$ cm^3 for total implant doses in the range $1.5\text{-}2.0\times10^{20}$ cm^3. Therefore the effective gate capacitance typically reaches about 65-75% of the dielectric capacitance, C_{ox}. The reduction in effective capacitance may be translated into a substantial loss in silicon surface field and therefore drive current, resulting in devices that perform outside of operating specifications. It might seem straightforward to solve this problem by simply increasing the dose level for the dopants in the poly-silicon gate or to anneal the devices at higher temperatures for longer times. Unfortunately this is not as simple as it may seem. The thermal budgets for most sub 0.25μm technologies are extremely stringent. For example, increasing time and/or temperature will adversely affect the source and drain profiles. Dopant penetration, specifically boron, also becomes problematic as dopant implantation doses and/or thermal anneal cycles are increased. To overcome these difficulties, it has become the *"responsibility"* of the dielectric to block the penetration of the gate dopants. For current technologies, the thermal cycles and implant doses following the p^+-polysilicon implant need be optimized in order to avoid boron penetration.

To demonstrate the resistance of oxides to boron penetration, Figure 5 compares the flat-band voltages of 2.1-nm O_2 and N_2O-grown oxide capacitors following a post-implantation RTA at 950–1050 °C for 5 s. From the change in the flat band voltage as a function of the anneal time, it is obvious that the N_2O grown oxides exhibit a superior resistance to boron penetration as well as

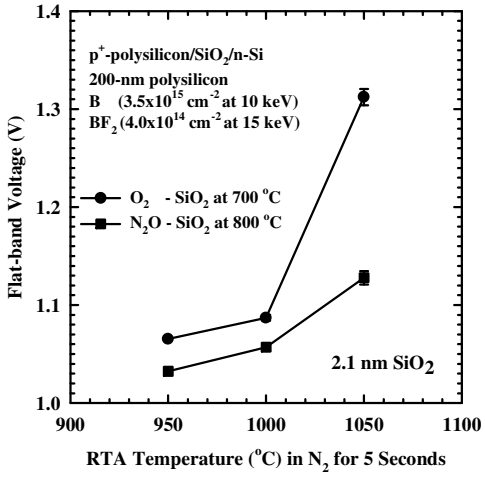

Figure 5 : Flat-band voltage versus rapid thermal anneal (RTA) temperature for ~ 2.1 nm O_2- and N_2O-grown oxide capacitors. Boron and BF_2 implants were activated for 5 s.

an acceptable thickness variation (≤ 1% over a 200 mm silicon wafer).

However, one should also note that for the lowest anneal conditions, (where boron penetration might be assumed to be negligible), the flat band voltage for the N_2O grown oxide has a more negative value than that for the O_2 grown oxides. It has been observed that while the nitrogen within the oxide film may result in less boron penetration, it also appears to contain more positive charge, the density of which, increases with increasing nitrogen concentration. The resulting shift in the threshold voltages translates into a mobility or transconductance loss that is typically found to be 10-20% for the 0.1-1.0 atomic nitrogen concentrations typically used. Therefore, although the use of nitrided oxides may be required to limit the boron penetration, there is still a need for an optimization between the implant dose, energy and species and the thermal cycles that may be tolerated for any given level of technology.

4. GATE DIELECTRIC RELIABILITY - BREAKDOWN AND DEVICE LIFETIME PREDICTION

With the scaling of the dielectric thickness, many fundamental difficulties have arisen regarding our ability to predict the lifetime of CMOS devices. The lifetime of these devices has a direct implication on the ability to mass produce large quantities of usable circuits at a reasonable cost. Setting aside for the moment the issue of manufacturing yield (which of course, may be the ultimate detraction is using ultra-thin dielectrics), let us concentrate on what is required for the prediction of, for example, a 10 year lifetime.

If one assumes that the gate leakage current, and all the damage that it causes, will ultimately cause the devices failure, then there is a need to understand the relationship between the applied bias, the gate leakage current and the type of damage and the rate at which this degradation occurs. Previous breakdown data presented in the literature has been shown as a function of either gate bias or injected current density as a function of either the charge injected (Q_{DB}) or the time to breakdown, (T_{DB}). Both of these types of measurements show how the dielectric breaks down and the data are usually shown as a function of either gate bias or oxide field. The absolute values of Q_{DB} and T_{DB} depend not only upon the details of the measurement technique (for example on the criterion used to describe breakdown, or whether constant current or constant voltage techniques are used) but also on the details of the process that was used to manufacture or process the MOS devices. It is therefore extremely difficult to compare results emanating from different laboratories. However, the trends should at least be similar.

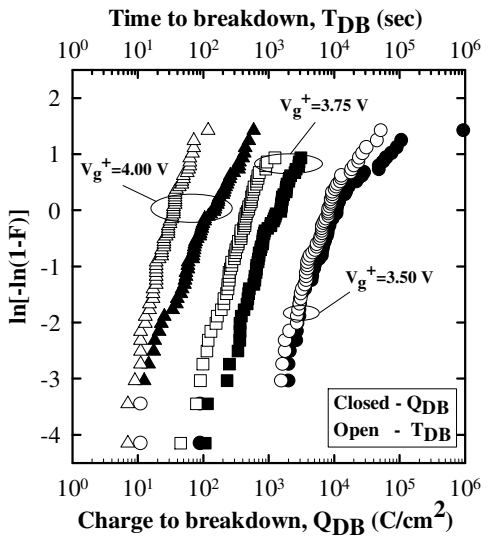

Figure 6 : The critical failure (Weibull) distributions of charge and time to breakdown for a series of 64 1×10^{-4} cm^2 capacitors as a function of the stress bias condition, (positive bias in this case). The closed and open symbols represent Q_{DB} and T_{DB}, respectively. Note the change in the Q_{DB} distribution as the stress bias is increased. Samples were p+ gates on an n-type substrate. The dielectric thickness is $t_{ox} \approx 1.9$ nm

In Figure 6, an example of the dielectric breakdown are shown for a series of N_2O grown oxides that were approximately 1.9 nm thick. The critical failure (Weibull) distributions of charge and time to breakdown are shown as a function of the stress bias condition, (positive bias in this case).
The closed and open symbols represent Q_{DB} and T_{DB}, respectively. Note the change in the Q_{DB} distribution as the stress bias is increased. Samples were p+ gates on an n-type substrate. This data is typical of what may be found in the literature.[5-8] As the stress bias is decreased, T_{DB} and Q_{DB} increase.

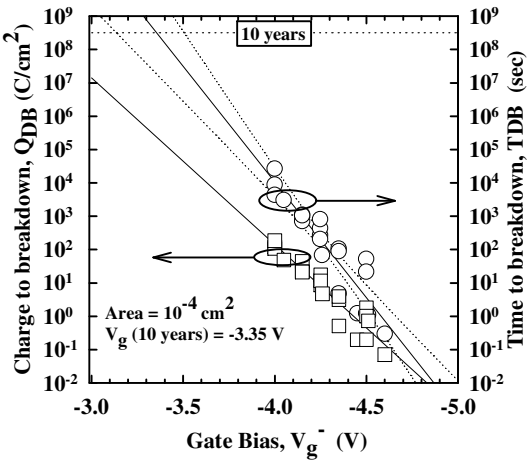

Figure 7 : Charge and time to breakdown as a function of the applied gate bias. These devices are n^+ polysilicon gate capacitors with p-type substrates. (The Q_{DB} and T_{DB} were taken at the 50 % of a distribution of 64 devices across an 200 mm wafer. All measurements were performed at room temperature.) These data include oxide films with thickness' ranging from 1.9-2.6 nm, some of which were grown in dry O_2 while others were exposed to various nitrogen treatments. Extrapolation (over 4 orders of magnitude) of this data to a time of 10 years yields a value of -3.35 V as a maximum operating gate bias. (The dashed lines represent the 95% confidence levels for the regression analysis). Above this bias, these devices would, on average, fail prior to their ten year lifetime.

In fact, these values increase so dramatically with the reduction in the applied bias, that to perform these measurements anywhere near the expected operating condition (for example between 1.0 and 1.5 volts) would take and unacceptably long period of time. As a result we attempt to accelerate these measurements by performing them at either higher biases and/or temperatures and attempt to extrapolate back to reasonable temperatures and gate bias values. However, as noted in Figure 7 above, it becomes necessary to extrapolate over many orders of magnitude to extract any "realistic" value for a lifetime or maximum operating voltage. The reader should also note that the data shown here are from measurements performed using capacitors structures. To relate such data to the reliability of MOSFETs, the flat band voltage (V_{FB}) must be subtracted.

The maximum operating gate bias that one could use (based on a 10 year lifetime for this particular set of capacitors) would therefore be approximately -3.35 V (approximately 2.35 V across the gate dielectric). The highest T_{DB} shown in Figure 7, is ~5×10^4 sec. The dashed lines in Figure 7, represent the 95% confidence levels for the regression analysis used to fit the data. The spread of this data suggest that we may only have a reasonable estimate of the maximum bias to within 0.3-0.4 V or an estimate of the lifetime (at -3.5) within two orders of magnitude in time); hardly what one might consider conclusive.

Given a lifetime of 10 years (3.15×10^8 sec) we extrapolate over *four orders of magnitude* (in time) to obtain a maximum operating bias. In doing so, we implicitly assume that the physics of degradation (i.e. the rate at which defects are produced) remains constant as bias is reduced. Based upon past results for thicker films[8], this may not be a reasonable assumption. One might reasonably ask whether any of these T_{DB} and Q_{DB} measurements are able (with any degree of accuracy) to predict the failure of *"real"* devices, operating at substantially different bias conditions. From the data presented above this would appear unlikely, at least with our present knowledge.

Aside from using higher bias conditions to accelerate testing, higher device test temperatures are commonly used, with temperatures usually in the range from 40-160 ^oC.
It is well know the higher temperatures accelerate the degradation, at least for thicker dielectric films when the bias conditions are well into the Fowler-Nordheim tunneling regime. For the capacitors structures shown in Figure 7, once the gate bias is below approximately 4.0 V, the electron current injected into the dielectric film no longer is injected into the conduction band of the gate insulator. The current transport changes from Fowler-Nordheim tunneling to direct quantum-mechanical tunneling. Is it reasonable to assume that the degradation mechanism would remain the same when the transport is physically much different? This would appear unlikely .

If we may at all relate the QDB and/or TDB measurements to the lifetime of real devices our understanding of the degradation mechanism must be explored in much more detail. What is required is a fuller understanding of how this degradation changes once the transport of the electrons moves into the direct tunneling regime.

We must also discover how this degradation changes with the applied bias and temperature in an effort to understand and model device reliability. If there is to be any impact from these efforts to understand the degradation mechanisms one must be able to accurately predict some reasonably accurate device lifetimes given some form of accelerated testing.

There is some evidence that the degradation mechanisms that leads to dielectric breakdown may be described in terms of two factors: a defect generation rate and a critical defect density. The critical defect density would be dependent upon processing, i.e. the details of the growth and integration of the dielectric into a CMOS process while the generation rate may be dependent upon processing details but is also strongly dependent upon the stress conditions (i.e. the applied bias and sample temperature). Therefore we may be able to make some progress in the understanding of these degradation mechanism (and lifetime predictions) since these two details may be de-coupled from each other.

4.1 Breakdown and stress-induced leakage currents (SILC)

In an attempt to discover a relationship between dielectric breakdown and the critical defect density, stress induced leakage currents (SILC) have been used to monitor the generation of defects as a function of the applied bias for ultra-thin dielectrics in capacitor structures. These stress-induced leakage currents generated during high field stress, were measured in the direct tunneling region and have been related to defects produced within SiO_2 films[5]. The measured defects are thought to be predominantly neutral electron traps. The density of these defects (and therefore SILC) increases until a critical density is obtained and catastrophic breakdown occurs. Mathematical models of this phenomenon have been published[9,10,11] which suggest that when the critical density is obtained, breakdown occurs through the creation of a continuous defect link across the dielectric. Catastrophic breakdown, typically measured using charge to breakdown (Q_{DB}) or time to breakdown (T_{DB}) is related not only to the defect density but also to the rate at which such defects are produced[5]. This *"defect generation rate"* is said to be inversely proportional to Q_{DB}, and therefore dependent upon the electron energy and applied bias[5]. On the other hand, the ***SILC at breakdown*** represents a measure of the defect density and therefore is not necessarily dependent upon the electron energy or applied bias, but rather, is a measure of a ***critical defect density.***

The statistical distribution of the SILC as a function of the stress condition was measured and related to the catastrophic breakdown. The stress induced leakage currents (SILC) in 1.9 nm N_2O grown oxide capacitors have been measured at 3.5, 3.75 and 4.0 V. The initial leakage currents are measured at 1.5, 2.0 and 2.5 V, prior to application of the stress bias. After a preset time interval, the stress bias is removed and the leakage currents are measured. This was repeated until catastrophic breakdown occurs as defined by an *instantaneous* doubling of the stress current.

A series of 64 capacitors, each with an area of 1×10^{-4} cm^2, was measured to obtain statistical information regarding the distributions of the charge to breakdown (Q_{DB}), time to breakdown (T_{DB}) and the relative change in the stress induced leakage currents ($\Delta J/J_o$), where $\Delta J=J_t- J_o$ and J_t is the current density at some time t, and J_o in the initial leakage current at a given bias. Bias stress and leakage measurements were performed under positive and negative bias for samples with p^+ gates/n-type substrates and n^+ gates/ p-type substrates, respectively.

The critical failure (Weibull) distributions for time to breakdown, $T_{DB,}$ and the charge to breakdown, Q_{DB} for stress voltages of +3.5, +3.75 and +4.0 V were shown in Figure 6. One should note the strong stress voltage dependence of these distributions. From highest to lowest stress bias the distributions shift by 3-4 orders of magnitude, proceeding, of course, to lower T_{DB} and Q_{DB} for higher bias conditions. The differences in these distributions between positive and negative (not shown) bias (for the same absolute value) was found to be minimal. The critical failure (Weibull) distributions for $\Delta J/J_o$, ***at breakdown***, are shown in Figure 8, as a function of the stress bias. Two series of distributions are shown for SILC measured at +1.5 and +2.5 V.

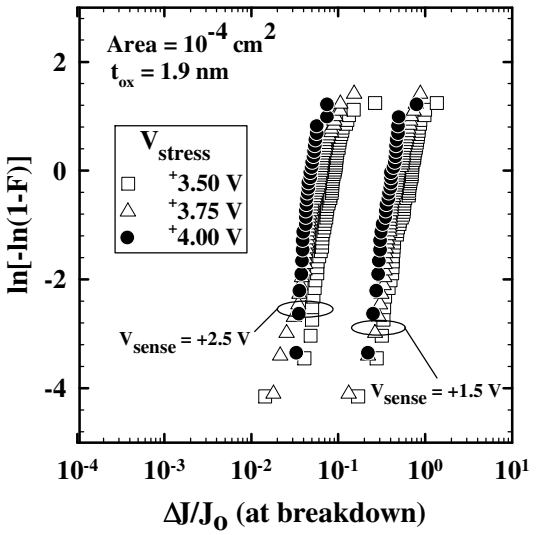

Figure 8 : Critical failure (Weibull) distributions of $\Delta J/J_o$(at breakdown) for different stress and sense conditions. Note the curves are relatively insensitive to V_{stress} (positive in this case). The distribution does, however shift to lower values with increasing V_{sense}.

From the data shown in Figure 6, both T_{DB} and Q_{DB} demonstrate a large voltage dependence. Similar dependencies have been found for much thicker dielectrics where Q_{DB} has been shown to be caused by hot electron induced *"trap creation"* and/or impact ionization[5,8]. For the data shown here, the dielectrics are substantially thinner and the electron energies during stress are below the *"threshold"* for trap creation. However, catastrophic breakdown and defect generation still occurs, suggesting that perhaps similar (although not identical) mechanisms of defect generation are occurring in these ultra-thin dielectrics even at the lower bias conditions. The measurement of $\Delta J/J_o$ at breakdown from Figure 8, show little (if any) stress voltage dependence. This implies that ***$\Delta J/J_o$ at breakdown*** is a measure of the ***critical defect density***; this being the precursor to catastrophic breakdown. Therefore, the measurement of T_{DB} (Q_{DB}) represents the time (charge) required to produce catastrophic breakdown and is ultimately dependent upon the rate at which the defects are created and the critical defect density. It is the defect generation rate, and not the critical defect density, that has a strong bias dependence. It is for this reason that this bias dependence is manifested in the T_{DB} (Q_{DB}) measurements.

From other measurements where electrons were injected from either the gate or from the substrate, little difference was noted. We found that the values $\Delta J/J_o$ (at breakdown) were independent of the bias polarity but did depend upon the magnitude of the sense voltage. Such a dependence has already been reported in 4.0-6.0 nm oxides, using sense voltages from 1.2-3.4 V.[5] This dependence may in part be caused by the specific shape of the current voltage characteristics and the shape of the SILC. Alternately, it may represent the energy distribution of the generated defects within the gate dielectric. For different sense voltages, different portions of the defect distribution would be probed by trap-assisted tunneling. If this applies, as the bias is increased, and tunneling occurs closer to the SiO_2 conduction band edge and the average value $\Delta J/J_o$ decreases. Therefore, one might conclude that the density of the generated defects within the dielectric increases towards mid-gap of the SiO_2.

From these data we find that both charge to breakdown (Q_{DB}) and time to breakdown (T_{DB}) are strong functions of the applied bias, effectively demonstrating the bias dependence of the defect generation rate. The change in the statistical distribution of the stress induced leakage ($\Delta J/J_o$) at breakdown is found to be relatively independent of the magnitude and polarity of the bias during stress, suggesting that ($\Delta J/J_o$) at breakdown is a measure of a critical defect density required to cause catastrophic breakdown in ultra-thin dielectrics. This critical density (for a given sense bias) is not dependent upon the electron energy during stress. However, the dependence of ($\Delta J/J_o$) at breakdown on the sense bias suggests that the generated defect density, as measured by the SILC, is distributed in energy, if one assumes that the SILC increases as a result of an increase in trap-assisted tunneling.

4. CONCLUSIONS

The continued scaling of the gate dielectric has precipitated the need for a greater and more detailed understating of the issues pertaining to integration and reliability. In this work we have attempted to outline and discuss the more critical

issues. These included the physics of the electron transport, including quantization and poly-depletion effects. As device scaling continues difficulties arising from increased leakage currents, poly-depletion and quantization effects will continue to restrict the use of standard SiO_2 based dielectrics. From the viewpoint of standby power, a dielectric limit appears around the 2.0 nm mark. However, if the device designers tolerate such higher leakage currents and power dissipation thickness' may be even further reduced. However, upon thickness reduction, other processing problems will more certainly arise. Boron penetration will continue to be a primary concern unless a better dielectrics (or composite) can be found. Finally the issue of dielectric reliability still remains foremost in our minds as a detractor for thinning the gate dielectric. However, the present data suggests that even for these ultra-thin dielectrics, as the bias is reduced, the dielectric lifetime increases rapidly. The true issue, is one attempting to find reasonable measurement techniques that allow accurate prediction of the lifetime for ***real devices***. At present, a substantial gap still exists between the currents accelerated testing techniques and a reasonable prediction of lifetime. This gap, however, does appear to be getting smaller.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Advanced Silicon Technology Center at IBM-Microelectronics in East Fishkill, NY for the fabrication of these devices. In particular, for the assistance of W. Ma we are extremely grateful. The many discussions with J. Stathis, E. Cartier and D.J. DiMaria are gratefully acknowledged.

REFERENCES

[1] H.S. Momose, M. Ono, T. Yoshitimi, T. Ohguro, S. Nakamura, M. Saito and H. Iwai, IEDM Tech. Dig., 593, (1994).

[2] H.S. Momose, M. Ono, T. Yoshitimi, T. Ohguro, S. Nakamura, M. Saito and H. Iwai, IEEE Trans. on Elect.Dev., **43,** 1233, (1995).

[3] F. Stern, Phys. Rev B, p.4891, 1972.

[4] R. E. Collin, *Field Theory of Guided Waves*, 2nd ed., *IEEE* Press, 1991

[5] D.J. DiMaria and E. Cartier, J. Appl. Phys., **78,** 3883, (1995).

[6] D. Dumin, S.K. Mopuri, S. Vanchinathan, R.S. Scott, S. Subamoniam and Terrill G. Lewis, IEEE Trans. on Elect.Dev., **42,** 1233, (1995).

[7] D.J. DiMaria, IEEE Elect.Dev. Lett., **16,** 184, (1995).

[8] D.J. DiMaria, E. Cartier and D. Arnold, J. Appl. Phys.., **73,** 3367, (1993).

[9] J. Suñe, I. Placencia, E. Barniol, E. Farrés, F. Martin and X. Aymerish, Thin Solid Films, **185,** 347, (1990).

[10] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas and H. Maes, IEDM Technical Digest, 863, (1995).

[11] J.H. Stathis, INFOS 1997 - this publication (1997).