THE DEGRADATION AND TIME-DEPENDENT BREAKDOWN OF P-TYPE MOSFETS WITH A HIGH-κ DIELECTRIC

THESIS

Presented to the Graduate Council of
Texas State University-San Marcos
in Partial Fulfillment
of the Requirements

for the Degree

Master of SCIENCE

by

Brian Yust, B.A.

San Marcos, Texas
May 2008

THE DEGRADATION AND TIME-DEPENDENT BREAKDOWN OF P-TYPE

MOSFETS WITH A HIGH-κ DIELECTRIC

Committee Members Approved:

_____
Terry Golding, Chair

_____
Carl Ventrice

_____
Gregory Spencer

Approved:

_____
J. Michael Willoughby
Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to thank SEMATECH for allowing me to use their samples and equipment for this study. Specifically, I would like to express my gratitude to Gennadi Bersuker, Chadwin Young, and Dawei Heh of the FEP division at SEMATECH for taking the time to further my understanding of these devices and offering their advice and suggestions.

I would also like to thank Dr. Golding for his support, guidance, and urging to explore various areas of research. I want to thank Dr. Ventrice and Dr. Spencer for being on my committee. I owe additional thanks to Dr. Ventrice for his advice and guidance in deciding where my career will go from here. I would also like to thank my fellow physics graduate students at Texas State University-San Marcos, for their own brand of moral support and inspiration.

This manuscript was submitted on April 9, 2008.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

ABSTRACT

THE DEGRADATION AND TIME-DEPENDENT BREAKDOWN OF P-TYPE MOSFETS WITH A HIGH-κ DIELECTRIC

by

Brian Yust, B.A.

Texas State Univeristy-San Marcos

May 2008

SUPERVISING PROFESSOR: TERRY GOLDING

In this study, the degradation and eventual breakdown of the gate stack due to trap generation in p-type metal-oxide-semiconductor devices comprised of SiO_2/HfO_2/TiN was investigated. Negative bias constant voltage stress was applied in conjunction with charge pumping, stress induced leakage current, and carrier separation measurements to examine the trap generation phenomena. Of interest in the study was location of the trap generation in the gate stack, whether or not the degradation was gradual, and if the degradation in PMOS occurred in the same manner as NMOS devices.

CHAPTER I

INTRODUCTION

The transistor, a basic building block for modern electronics and integrated circuits in particular, is a very important and well studied device. However, in order to keep with Moore's Law and reach the goals set by the ITRS (International Technology Roadmap for Semiconductors), the semiconductor industry has needed to scale these devices to smaller and smaller proportions. In doing so, phenomena either not initially seen or not significantly contributing to the physics of larger devices have begun to alter device functionality and reliability. To overcome these challenges, the semiconductor industry has investigated and adapted device fabrication processes, device geometry, and materials used. One such problem is that as devices are further scaled down towards smaller and smaller nodes, gate oxides in field effect transistors (FETs) are becoming thinner and thinner. As this insulating layer is scaled down, the gate leakage current comprised of charge carriers tunneling through the insulating layer has been found to increase exponentially.[1] In order to reduce this leakage current while scaling down the device, gate dielectric materials with a higher dielectric constant (high-κ dielectrics) such as HfO_2, HfAl-based oxides, and Hf-derived silicates have been proposed to replace the standard gate dielectric SiO_2. These materials, however, exhibit a high density of

intrinsic charge carrier traps, and the growth of Hf based oxides on a silicon substrate by any of the widely used deposition methods, chemical vapor deposition (CVD), atomic layer deposition (ALD), or physical vapor deposition (PVD), has proven to lead to a multilayer gate stack which includes a very thin SiO_2 layer between the Hf based film and the Si substrate.[1] Due to this as-yet unavoidable interfacial layer of SiO_2, it is important to study by electrical characterization the distribution of electrically active defects throughout the multilayer gate stack. Knowledge of the defect placement, growth rate, and contribution to device degradation is crucial to improving current device fabrication processes, and this manuscript aims to clarify these issues.

The Transistor

The transistor is a basic electronic device used to amplify an electronic signal or to switch an electronic signal from an "off" to an "on" state when used as a logic gate. The metal-oxide-semiconductor field effect transistor (MOSFET) is one of the most common of these devices, and was the subject of study for this project. The basic structure of a MOSFET includes a substrate, usually silicon, doped with an n- or p-type element, a source and drain region highly doped with the opposite type of dopant used in the substrate, and the gate oxide(s) with an electrode, usually poly-silicon. With the recent use of metal-derived oxides as dielectrics, metal gate electrodes have been used since there is less diffusion than was seen in silicon-derived oxides. All of this is

deposited on the substrate between the source and drain.

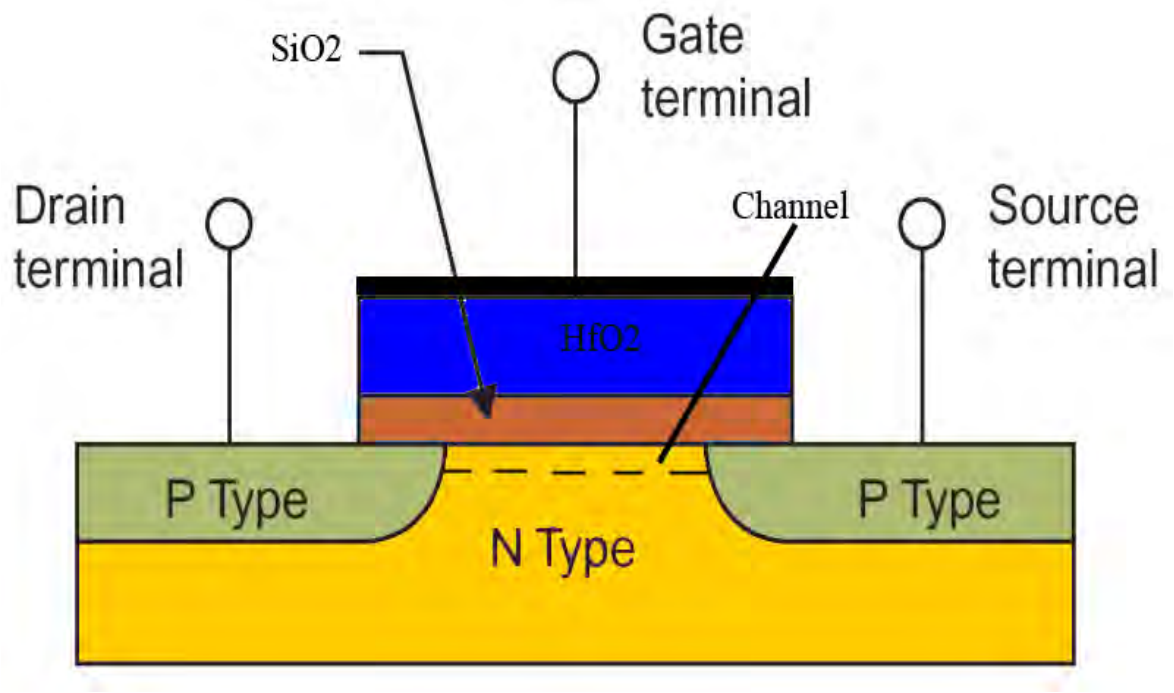

Figure 1: P-type MOSFET structure

The dopant in the source and drain region serves to provide extra charge carriers available for conduction when the device is turned “on”. Normal operation of a MOSFET is carried out by applying a voltage bias to the gate, which causes charge carriers from the source and drain to flow into the area directly beneath the gate called the channel or “inversion layer”; for this reason, a MOSFET is said to be in inversion for this mode of operation. Once the applied gate voltage reaches a sufficient bias, the threshold voltage, there are enough charge carriers in the channel to facilitate conduction from the source to the drain, resulting in a current through the device. By convention, when a current freely flows through a MOSFET in inversion, the device is said to be “on”. When the gate voltage is below the threshold value, the device is said to be “off”, though there may be a small leakage current from the source to the drain due to the channel being

lightly populated by charge carriers at low gate biases. For an n-channel MOSFET (NMOSFET) where the substrate is p-type (hole majority charge carrier) and the source, drain, and inversion channel are n-type (electron majority charge carrier), the gate voltage bias needed to put the device in inversion is positive; for a p-channel MOSFET (PMOSFET), the substrate is n-type and the source, drain, and inversion channel are p-type, so the gate voltage bias needed is negative. When the opposite voltage bias is applied to the gate (positive for PMOS and negative for NMOS), the channel is further populated by the majority carrier already present in the substrate; this regime is called "accumulation". While a transistor is in inversion, the charge carriers making up the channel are attracted there due to the electric field caused by the gate bias; with the gate stacks of today's devices being 10nm or thinner, there is now an increased chance that the carriers in the channel will move through the dielectric to the gate electrode via quantum-mechanic tunneling. This tunneling through the gate stack results in a leakage current to the gate electrode, thus diminishing the current through the device's channel.

In larger scale devices, this tunneling is usually carried out through a process called Fowler-Nordheim tunneling, in which the charge carriers tunnel from the conduction band of the Si substrate into the conduction band of the insulator and "hop" along the gate oxide to the electrode via trap-to-trap tunneling.[2] With devices being scaled down, the gate oxides have become so thin that charge carriers may tunnel directly through the gate, or directly from the substrate to the interface between the SiO_2 and HfO_2 layers and then from that interface to the gate electrode.[2] Apart from this leakage current decreasing the transistor's output, it poses another problem: prolonged tunneling through the gate stack can lead to catastrophic electrical breakdown of the oxide and,

ultimately, device failure. This phenomenon called time-dependent dielectric breakdown (TDDB), still not fully understood, is the cause of some controversy in the semiconductor community and is at the heart of the physical characteristics studied in this project.

CHAPTER II

DEVICES AND PROCEDURE

Devices

The samples investigated in this study were fabricated using a standard CMOS process. The gate stacks formed through atomic layer deposition (ALD) had either a 2.2nm or 3nm thick layer of HfO_2 on the substrate, with a subsequent 1.1nm thick interfacial layer of SiO_2, which is inadvertently created in the process. The gate electrode was TiN, and the device dimensions were 10x1μm (width x length). Dopant density of the substrate was 5.4×10^{17} cm^{-3}, and the dielectric constant of the HfO_2 film was found to be ~22. All samples were property of SEMATECH, and as such, specifics of fabrication have been withheld at their behest.

Procedure

Measurements were conducted with a four-point probe set up connected to a Keithley 4200 semiconductor analyzer tool. For the majority of the measurements taken, the "stress and sense" approach was employed in which one monitors the time progression of gate leakage current as a higher than normal bias, or voltage stress, is applied to the gate, and then either after a predetermined amount of time or at the users discretion, the stress is interrupted to run the "sense" measurements. Device tests were designed and run through Keithley's KITE program and involved constant voltage stress

(CVS) interspersed with sense measurements including fixed-amplitude, fixed-frequency charge pumping and stress induced leakage current (SILC). The experimental setup was also altered to include the carrier separation technique during CVS and SILC. CP measurements were enabled by utilizing a Keithley 708A switch matrix to reassign SMU's without needing to physically reconnect the cables to the Agilent 8110A pulse generator during the "sense" portion of measurements. For the CVS measurements performed, time evolution of the gate leakage current was monitored. Sense measurements were taken at predetermined times for certain trials, and the progression of the gate leakage was monitored manually for others. For the manually monitored trials, sense measurements were still taken periodically in order to provide a better sense of degradation; whenever the gate current exhibited drastic, sustained changes, sense measurements were also taken. The gate bias was varied between trials, and, in a few cases, within the same trial which will be denoted appropriately in the results and figures. Each trial was performed on a new device and is denoted by the stress bias (Ex: V_{stress} = -3.5V indicates the results for one device stressed at that bias).

Charge pumping is a technique useful in extracting charge trap densities and was applied at a high and low frequency during the "sense" portion of the experiments. While the source, drain, and substrate are grounded, a voltage pulse is applied to the gate such that it drives the channel to accumulation and inversion. At the base of the pulse, the channel is in accumulation; interface states and traps are filled with charge carriers present in the channel (holes for NMOS devices, electrons for PMOS devices). As the gate bias begins to change, trapped charge carriers are emitted at a rate governed by the rate of change of the surface potential seen at the gate stack/substrate interface (known as

steady-state hole/electron emission).[3] Once the gate bias reaches the flatband voltage, the surface potential is changing so rapidly that the trapped charge carriers cannot be emitted fast enough to keep up with the changing potential, since there is a much shorter time period during which they can be emitted.[3] At this point, trapped carriers are emitted at a rate governed by non-steady-state hole/electron emission, which is a slower process. For this reason, more carriers are captured by electrically active traps than emitted. Once the pulse reaches the threshold voltage, charge carriers of the opposite type than those trapped (electrons for NMOS, holes for PMOS) move in from the source and drain regions and recombine with the trapped carriers (from accumulation) which have yet to be emitted, thus becoming trapped themselves. As the gate pulse begins to fall back to the base voltage, a similar progression of events is seen, with steady-state emission of the source/drain carriers, then non-steady-state emission of the source/drain carriers, and finally trapping via recombination with the substrate carriers. Since there are more charge carriers flowing into the channel than out of it during the rise and fall of the pulse, a current, I_{cp}, which is proportional to the number of trapped charge carriers is seen at the substrate. Conventionally, when the base voltage of a pulse with a fixed amplitude is swept over a range including the flatband and threshold voltage, the charge pumping current reaches a maximum at a certain V_{base}; this can be expressed as $I_{CP,Max} = qfA_G N_{it}$, where q is the charge, f is the frequency of the pulse, A_G is the effective channel area and N_{it} is the density of traps within the detected energy and spatial regions.[1]

For our purposes, a fixed amplitude trapezoidal pulse with a rise and fall time of 100ns was used. The amplitude was set at -1.4V, chosen so that it would span both the flatband voltage and the threshold voltage. The base voltage, initially set at a positive

voltage, was stepped down towards zero with each successive pulse. Before each charge pumping measurement was taken, the device was subjected to a positive bias on the gate to discharge any traps still containing holes from CVS or other measurements. By applying a positive bias to the gate, the PMOS device is put in accumulation; thus, the channel is populated by electrons which can recombine with any trapped holes. This allows our charge pumping to reflect a more accurate defect density. It has also been shown previously that as CP frequency is reduced, electrically active traps further from the gate stack/substrate interface can be reached, giving an inverse correlation between frequency and probing depth.[4] Exactly how deep a fixed amplitude, fixed frequency pulse can probe is still under debate, however. After sweeping the device over a range of frequencies, an upper and lower value of 1MHz and 3kHz was chosen, for which the data was still reliable. Comparing these two measurements will give a better understanding to where the majority of traps are being generated: near the SiO_2/Si interface or the HfO_2/SiO_2 interface.

The SILC measurements were taken by a voltage sweep of the PMOS devices in inversion while recording the gate leakage current. These were then time stamped and compared to explore any time evolution of deterioration of the dielectric. One calculation found to be useful in correlating device degradation and gate leakage current for NMOS devices was the relative change with time in the gate current at a certain gate voltage from the measurement taken prior to stress.

$$SILC\left(V_g,t\right)=\frac{\left(I_g\left(V_g,t\right)-I_g\left(V_g,t=0\right)\right)}{I_g\left(V_g,t=0\right)}$$

Carrier separation techniques were added to the CVS and SILC measurements to help clarify the physical processes involved in oxide degradation. By measuring all four terminals (gate, substrate, source, and drain) during CVS and SILC and constructing band diagrams, it can be deduced what charge carriers are flowing where. When a PMOS device is in inversion, holes gather in the channel. Due to the electrode being metal, there are no free holes to be injected from it through the gate stack and into the substrate, but there are free electrons which may tunnel through to the substrate and result in a net current measured at the substrate. By summing the source and drain currents, how many holes from the source that are leaking through the gate stack instead of traveling through the channel and reaching the drain can be detected. The current measured at the gate should be equal to the sum of the source, drain, and substrate currents; however, it may not be exact due to recombination of electrons and holes, possible impact ionization from electrons injected from the gate, or any other leakage paths in the device. This scheme allows us to say that the source/drain current represents the flow of holes in our device, and the substrate current represents the flow of electrons.[5] By monitoring the time evolution of these currents during stress, as well as during the SILC gate voltage sweep, the device degradation can be better understood. It has also been noted that a strong correlation between changes in the gate current and bulk current indicate that breakdown path is occurring near the middle of the channel, as opposed to being near the source or drain regions.[6] To verify that other physical processes aren't causing the source/drain current, the source and drain currents were measured separately and summed afterwards as well as tied together physically with a triaxial t-connector joining the two raw current

signals together. There were no obvious differences seen in the results of the two different measuring techniques.

Energy band diagrams were constructed using the Dual Gate Dielectric Band Diagram Program developed by Dr. Knowlton and his research group at Boise State University.[7] Although the program could only give us an approximation since it does not take interface states or traps into account, it was very helpful in understanding the physical processes taking place, particularly for the SILC and CVS of SILC measurements.

CHAPTER III

RESULTS

Time Evolution During Constant Stress Voltage

CVS measurements were done on PMOS samples with a 3nm and 2.2nm HfO_2 layer. Gate leakage current, Ig, was measured as a function of time during the stress; it showed that there was very little indication as to when a device would breakdown critically and cease to function as intentioned. The time progression of gate leakage under stress showed slight variations while still functioning, then would suddenly increase by about three orders of magnitude once it reached a breakdown event. The time progression of gate leakage current at different biases is shown in Figs. 2-6.

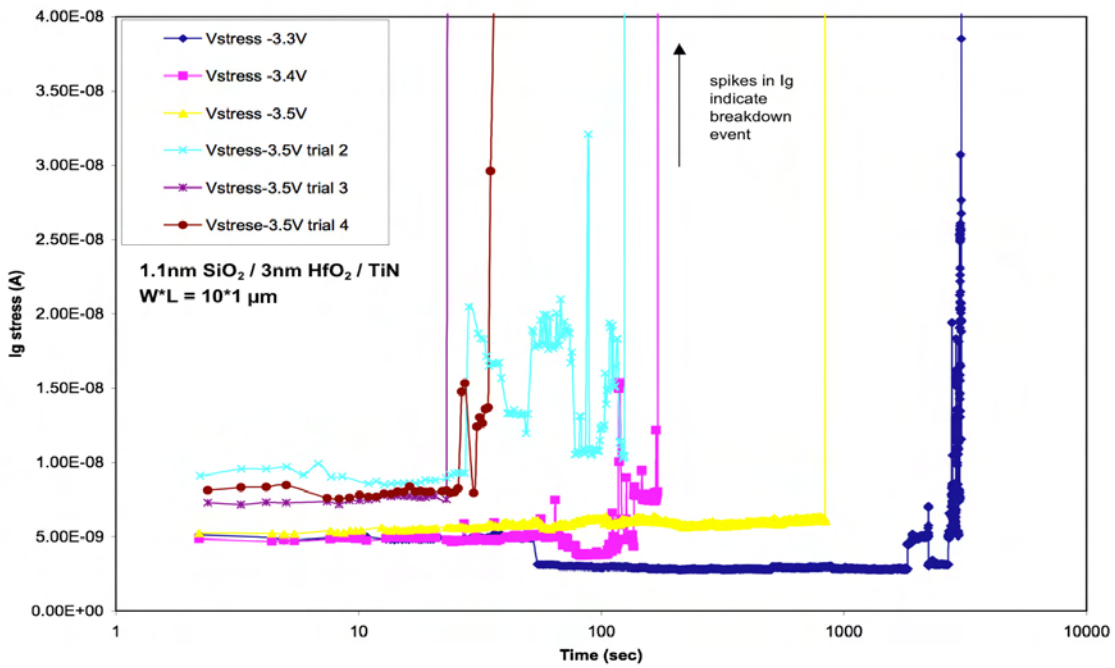

Figure 2: Ig vs log(time) for 3nm gate stacks before breakdown

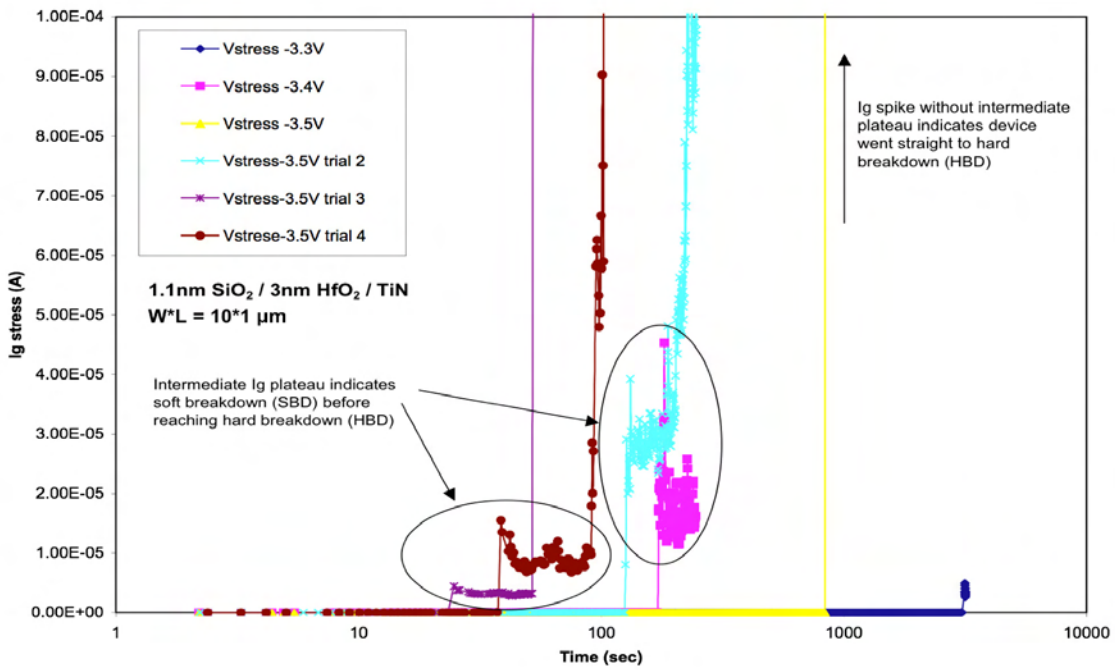

Figure 3: Ig vs log(time) for 3nm gate stacks after soft breakdown

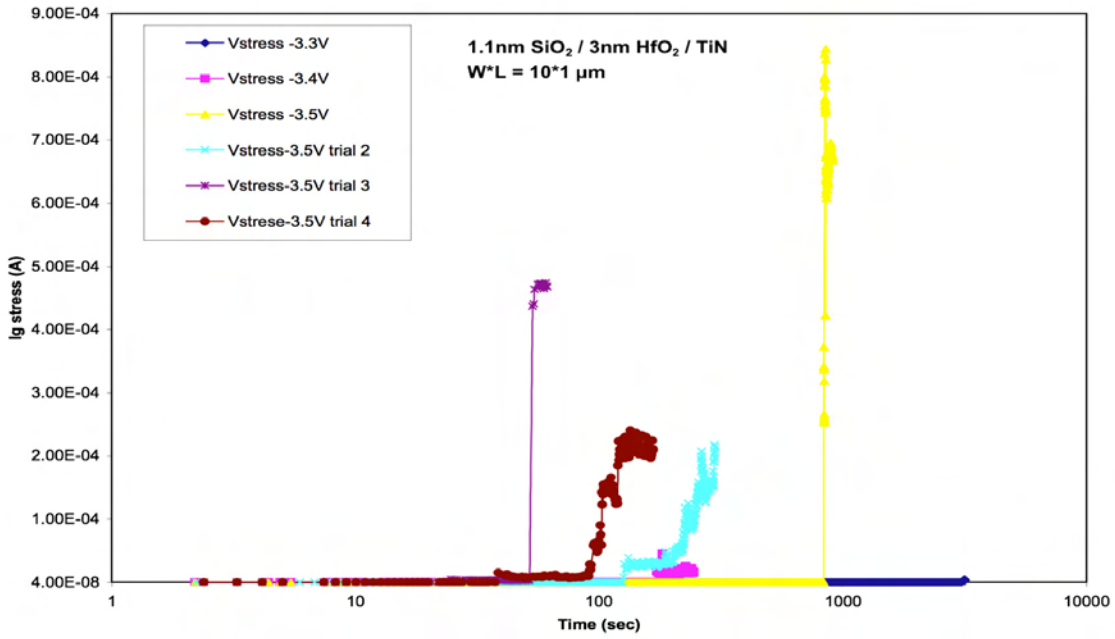

Figure 4: Ig vs log(time) for 3nm gate stacks after hard breakdown

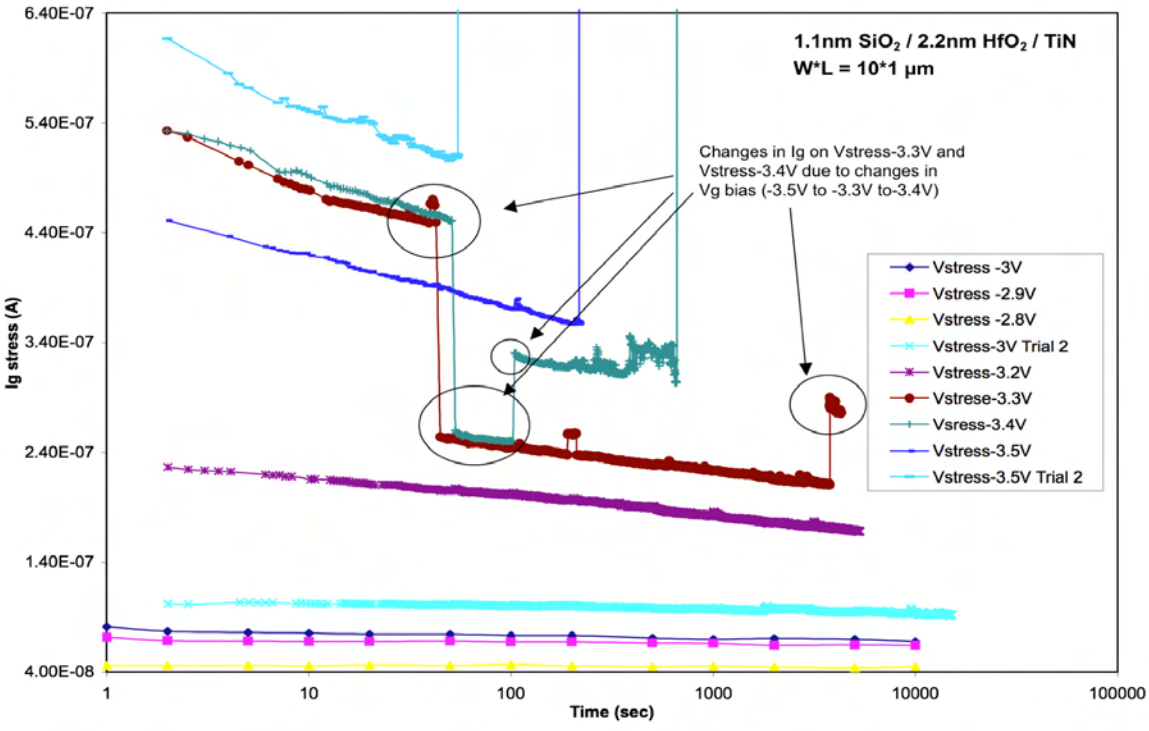

Figure 5: Ig vs log(time) for 2.2nm gate stacks before breakdown

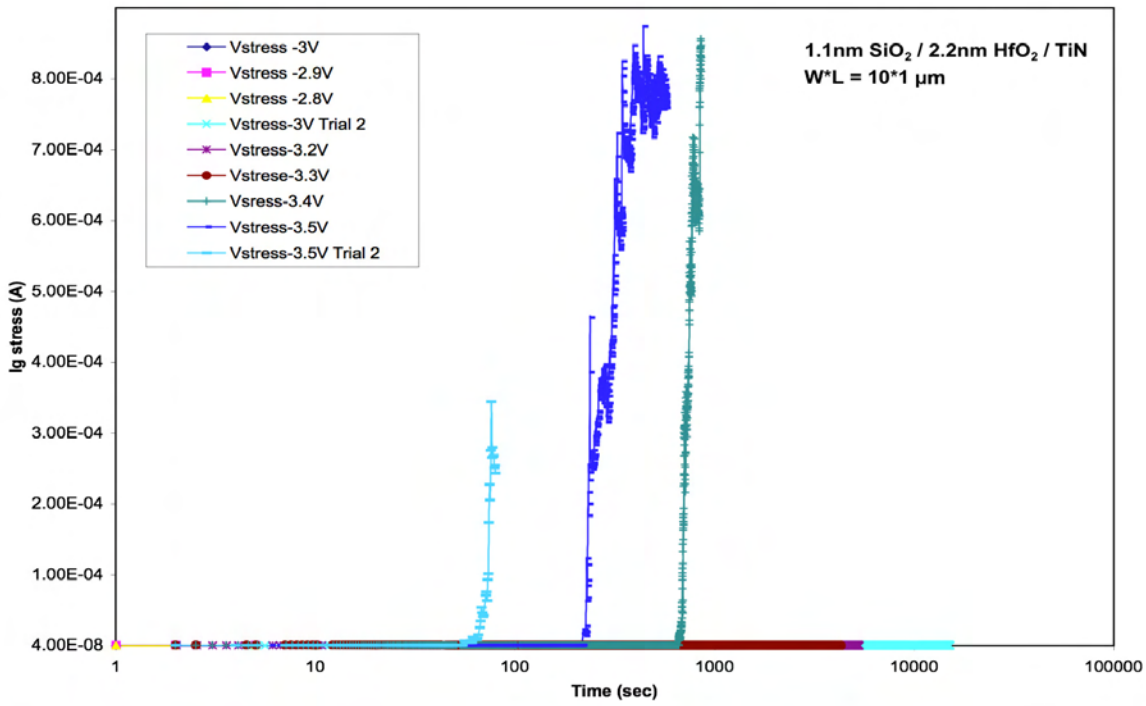

Figure 6: Ig vs log(time) for 2.2nm gate stacks after breakdown

At the higher gate biases for the 2.2nm stack, the gate leakage showed a significant decrease over time prior to breakdown (Fig. 5). This is likely due to a build-up of traps at the substrate interface, effectively blocking some gate leakage and changing the potential seen there, and warrants further investigation beyond the scope of this paper. The stair-step pattern in which the gate current increases suggests that the devices are being subjected to soft breakdown (SBD) until reaching a point of hard breakdown (HBD) (see Fig. 3). Hard breakdown occurs when a percolation path forms through the gate dielectric, effectively shorting the gate to the substrate, and the current moves through this path with enough energy to permanently damage the structure of the dielectric; HBD is characterized by a single drastic increase in gate leakage. In contrast, soft breakdown (SBD) occurs when percolation path(s) form, at least partially, through the dielectric, but the current traveling through them does not have enough energy to permanently damage

the dielectric. For this reason, multiple SBD events may be seen before the device breaks critically. These multiple SBD events show up as steps in the gate leakage current over time. Our devices showed characteristics of both SBD, as is evidenced by the gate currents which stepped up such as the -3.4V and -3.5V trials on the 3nm stack (see Fig. 3), as well as HBD, as is characterized by the non-stepping currents such as -3.5V trial on the 3nm stack and the 2.2nm trials (see Figs. 3 & 6).

Using Dr. Knowlton's energy band diagram program, an approximation of the device under stress was constructed.[7] Due to the band bending in the Si substrate, there is an accumulation of holes directly under the gate which can tunnel to the gate electrode. At the same time, the electrode bias is large enough that electrons may be injected into the Si substrate, since they have an energy higher than that of the substrate's potential barrier.

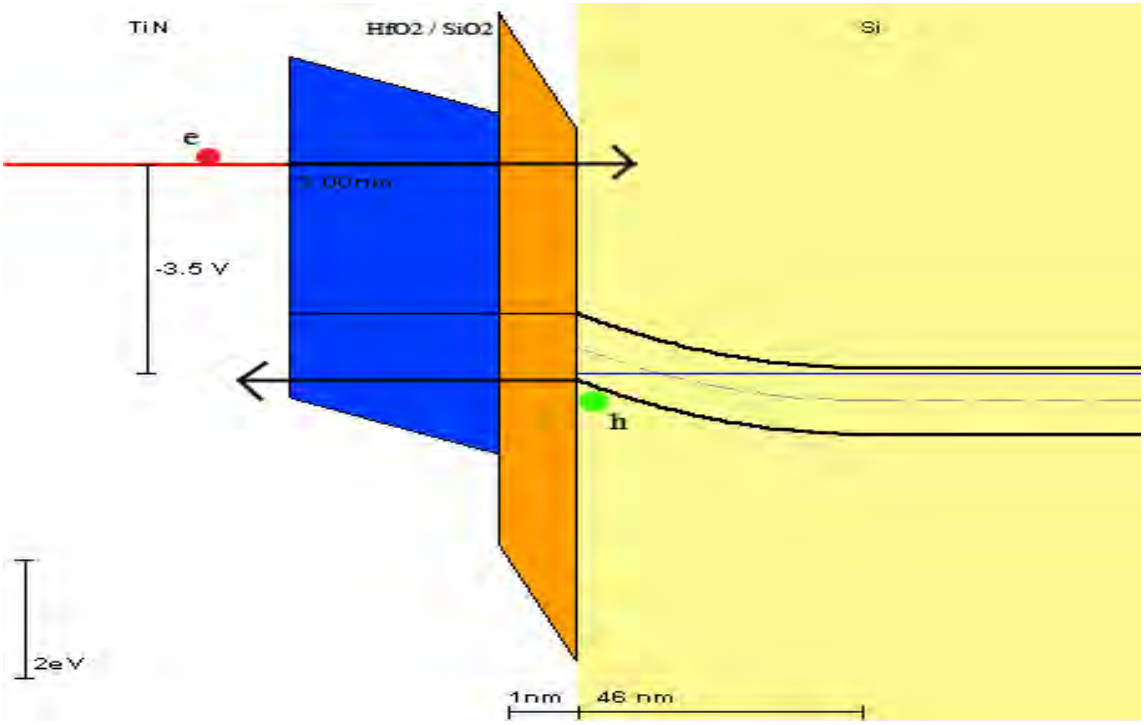

Figure 7: Band diagram for 3nm gate stack at Vg = -3.5V

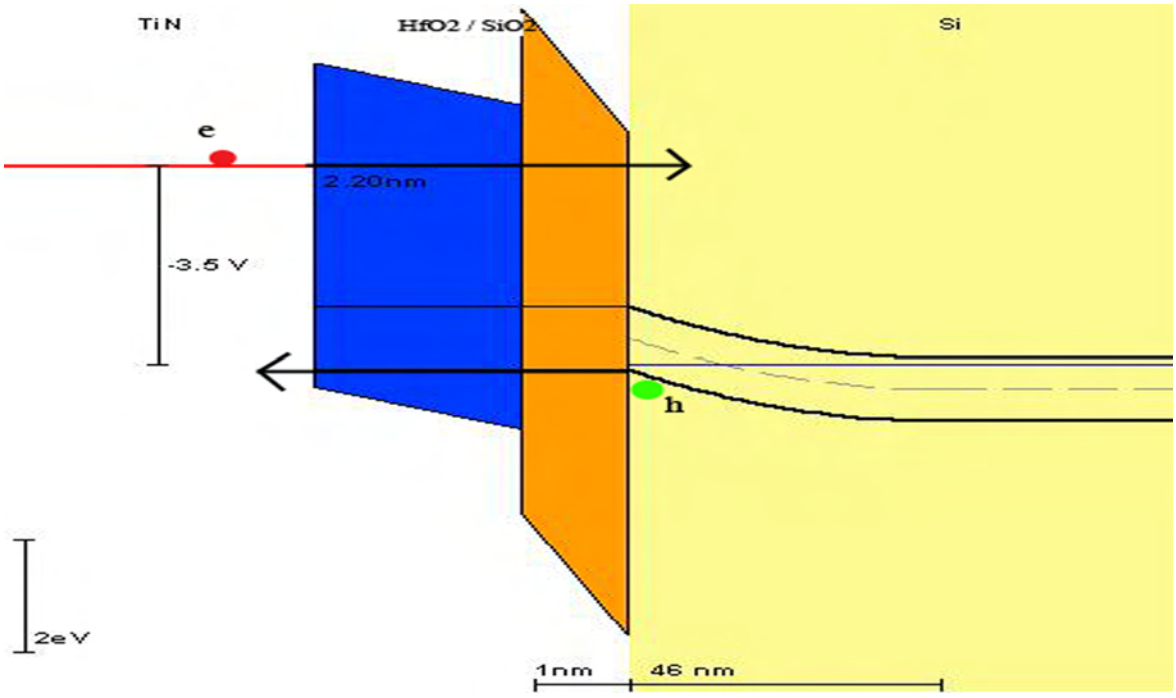

Figure 8: Band diagram for 2.2nm gate stack at Vg = -3.5V

The carrier separation technique was used during CVS of some 3nm samples. In all cases, prior to breakdown, the source/drain current dominated that of the bulk. After the initial breakdown, the bulk current tends to increase significantly and dominate the source/drain current.

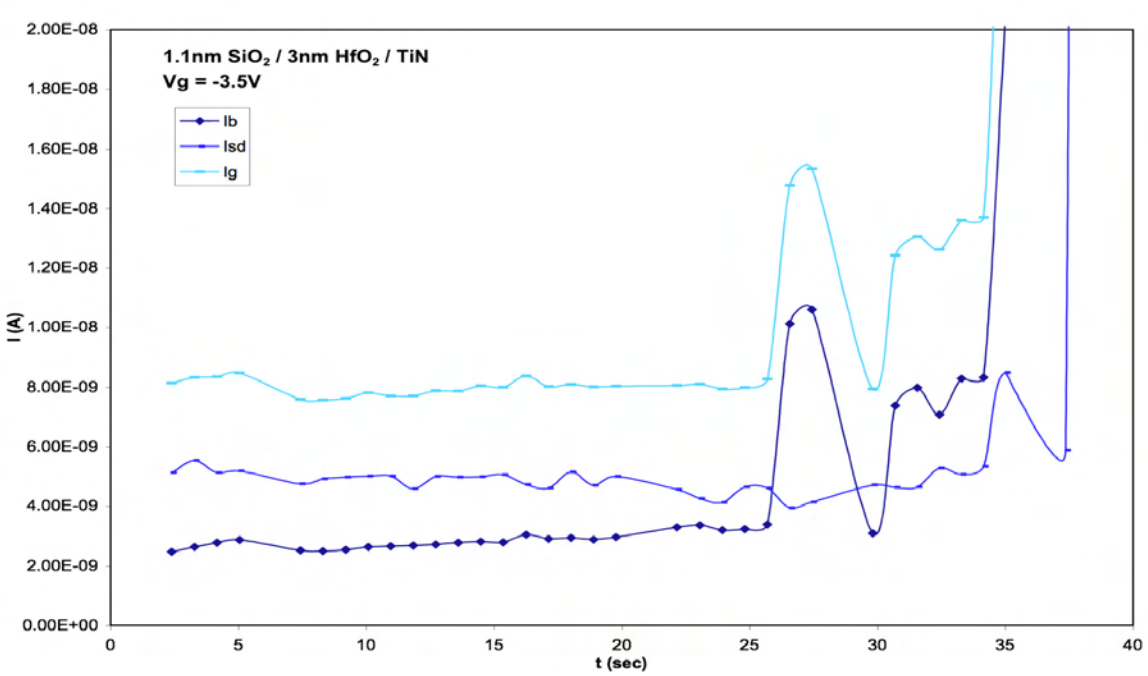

Figure 9: Carrier Separation of CVS without t-connector before BD on a 3nm sample at Vg = -3.5V

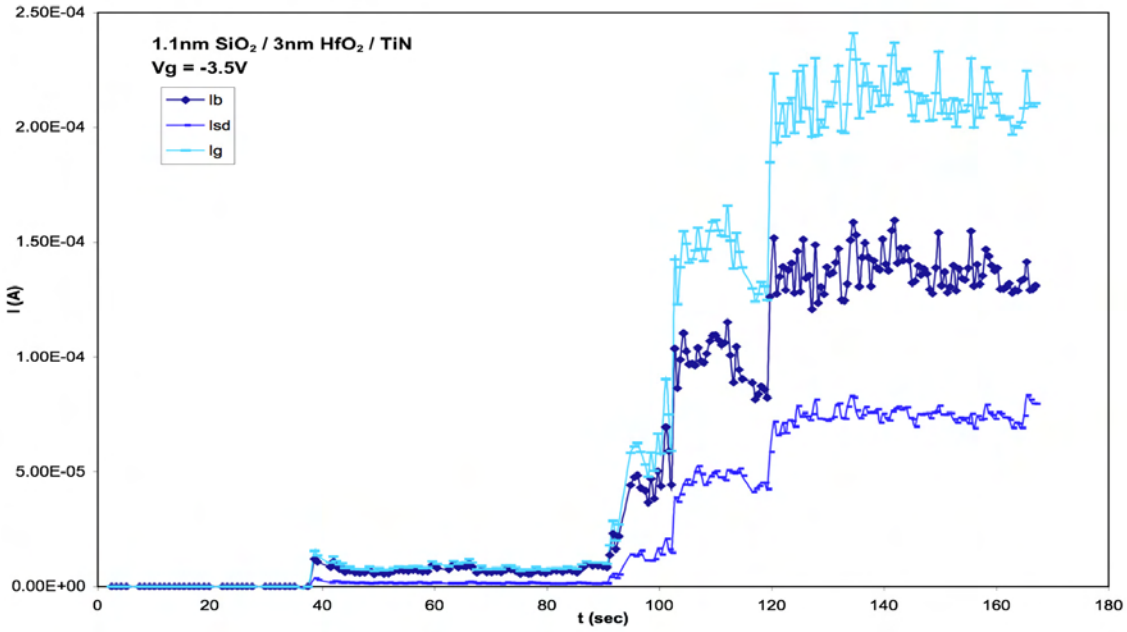

Figure 10: Carrier Separation of CVS without t-connector after BD on a 3nm sample at Vg = -3.5V

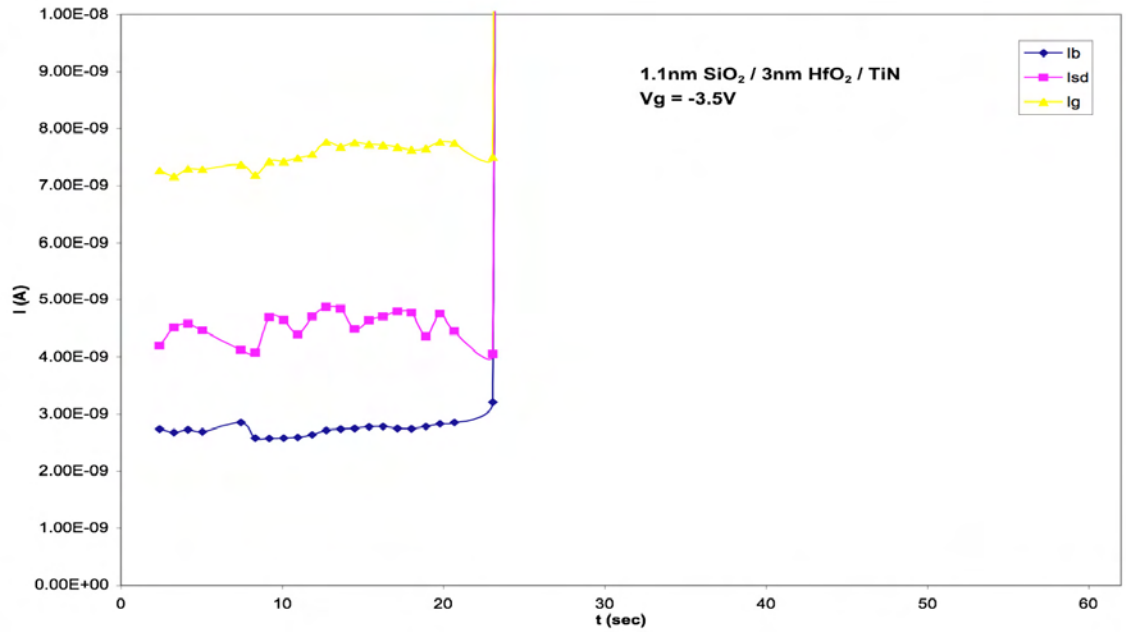

Figure 11: Carrier separation of CVS without t-connector before BD on a 3nm sample at Vg = -3.5V

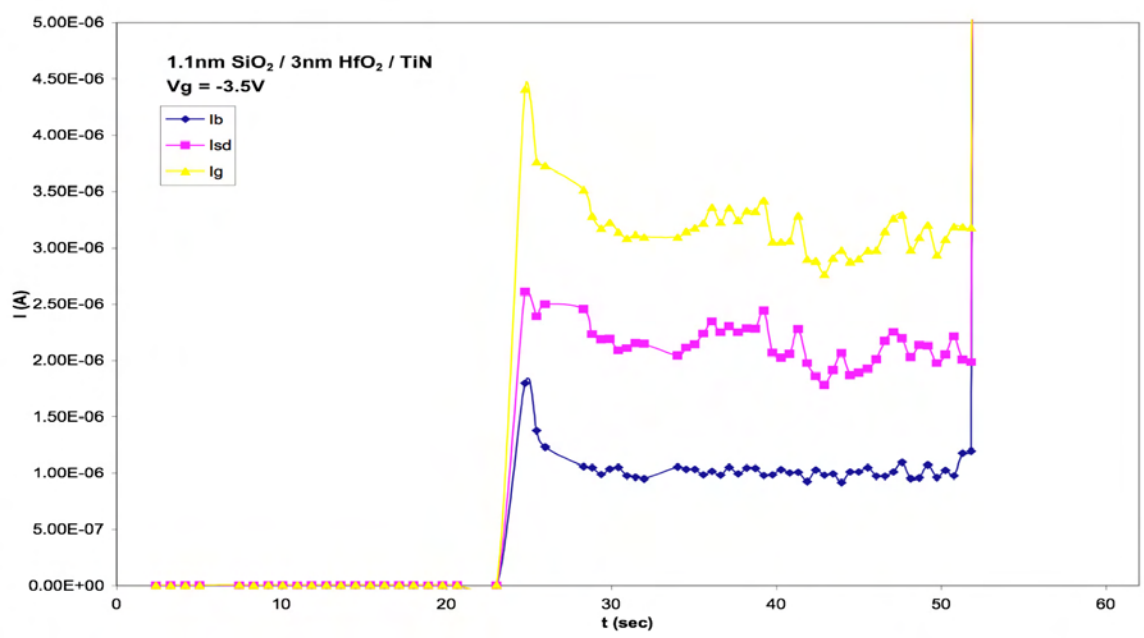

Figure 12: Carrier separation of CVS without t-connector after SBD on a 3nm sample at Vg = -3.5V

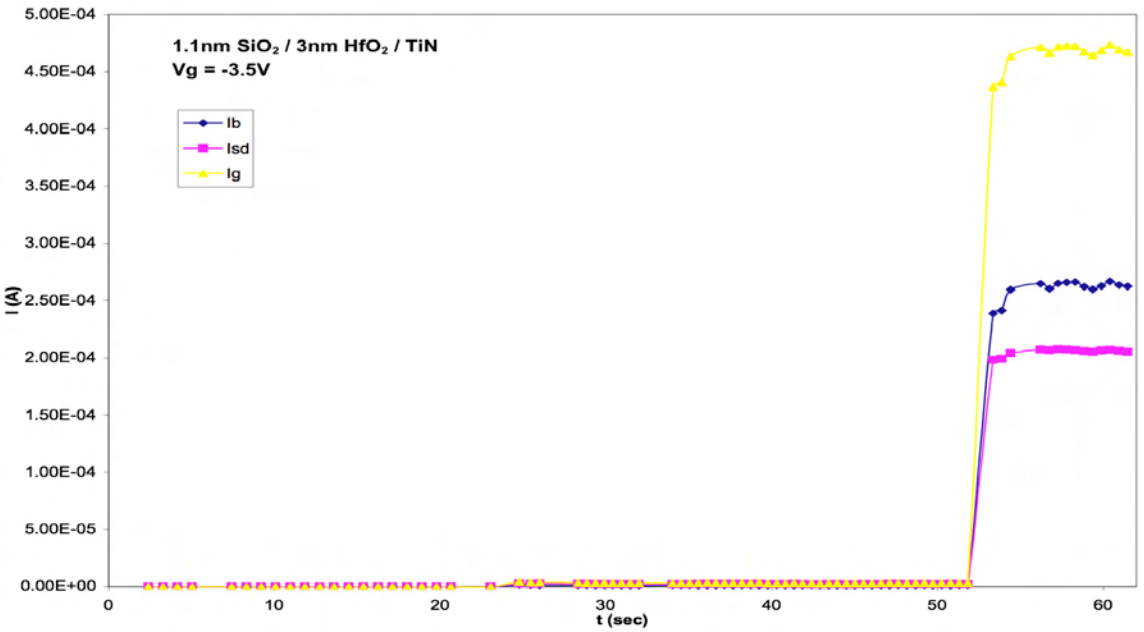

Figure 13: Carrier separation of CVS without t-connector after HBD on a 3nm sample at Vg = -3.5V

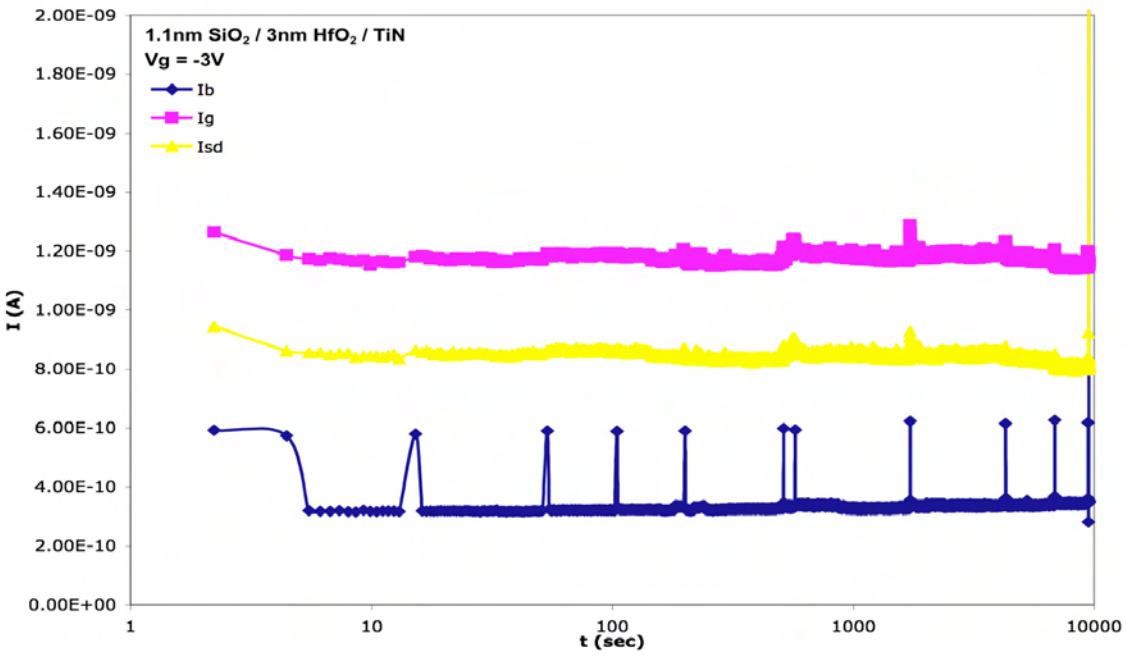

Figure 14: Carrier Separation of CVS with t-connector before BD on a 3nm sample at Vg = -3V

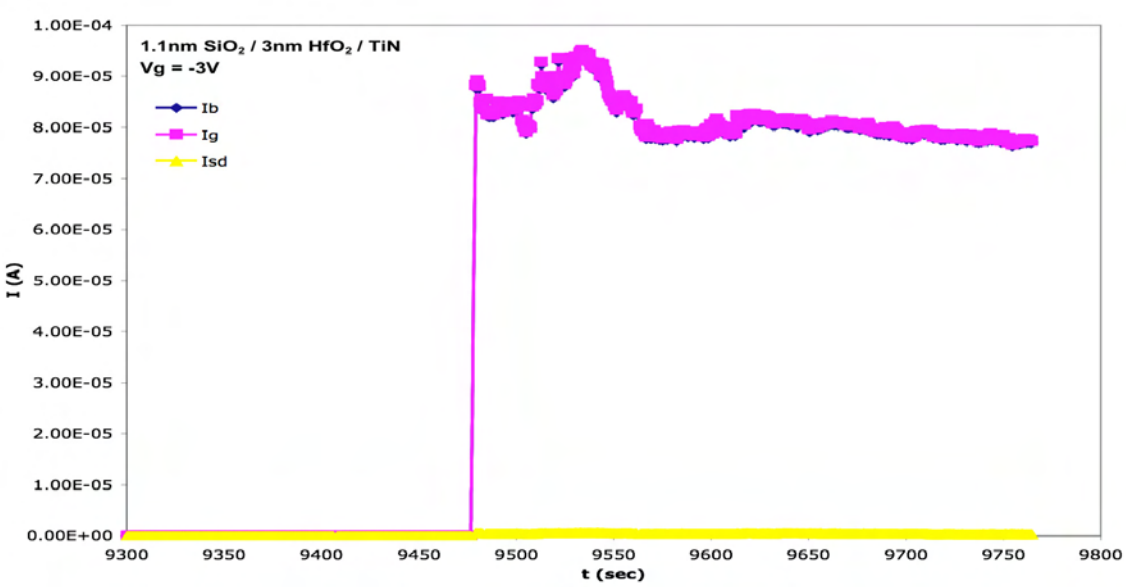

Figure 15: Carrier Separation of CVS with t-connector after BD on a 3nm sample at Vg = -3V

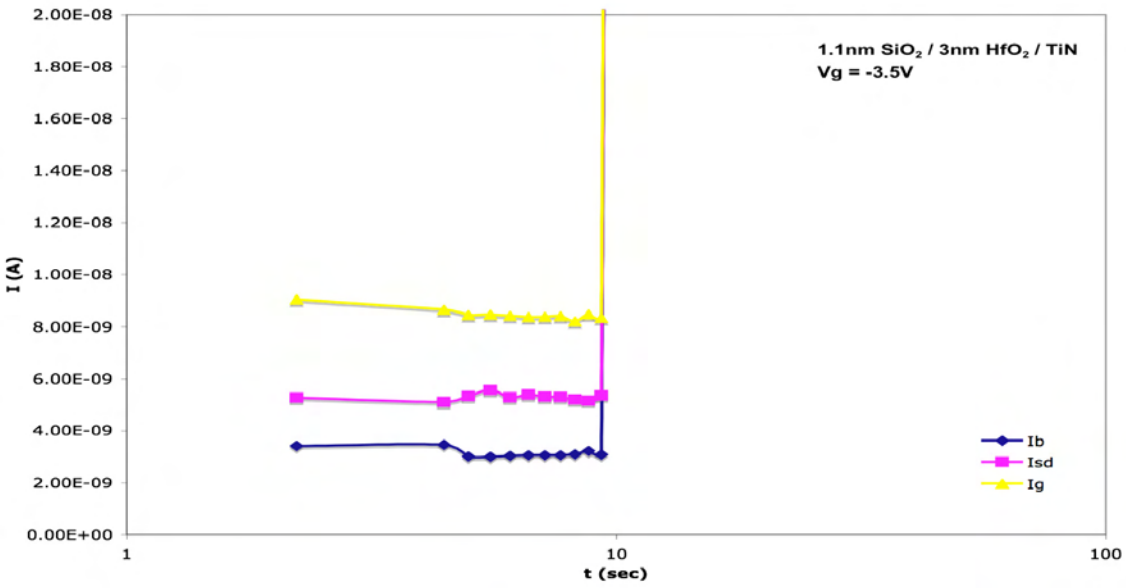

Figure 16: Carrier Separation of CVS with t-connector before BD on a 3nm sample at Vg = -3.5V

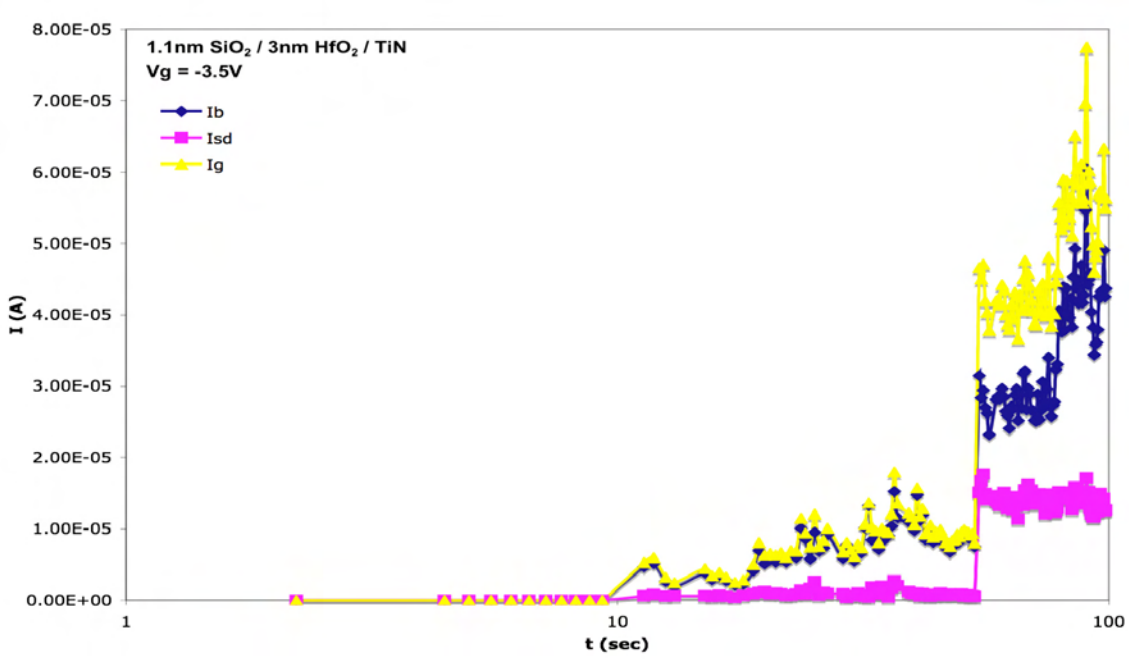

Figure 17: Carrier Separation of CVS with t-connector after BD on a 3nm sample at Vg = -3.5V

Note that in all cases except one (the Vg=-3.5V trial in Fig. 11-13), soft breakdown resulted in the bulk current dominating that of the source/drain. The one trial that showed a different result may have had a SBD pathway nearer to the source or drain region. The bulk current did dominate after the next breakdown event, though. It's clear from the various trials in the previous figures, that some of the devices were subjected to SBD before succumbing to critical breakdown, while others went directly to HBD. No apparent correlation between the different conditions used in CVS and whether the device went through SBD or not could be found. For cases of higher stress voltages (Figs. 11, 12, 13, 16, 17) both SBD and straight to HBD cases were seen; for cases of lower stress voltage (Figs. 14, 15) SBD is expected due to the lower energy involved, yet the device went straight to HBD. The range of stress voltages used seems to be high enough to cause HBD with some frequency. Given more time and a large amount of devices, one

may find that there is a statistical distribution to how many devices at a certain stress gate bias will go through the SBD process before reaching HBD and how many will go directly to HBD; this, however, was not possible given the time and resources for this project, nor was it the aim. The monitoring of gate leakage current was mainly used in this experiment as a way to watch for any breakdown events, such as SBD.

From close inspection of the previous figures, it is apparent that the gate leakage current tends to follow any jumps or dips in the bulk current after breakdown (see Figs. 10, 13, 15, 17). From this the location of the breakdown paths can be inferred to be near the middle of the channel of the device, and not close to the source or drain regions. If the percolation paths were towards the edges of the device, the gate leakage current and the source/drain current should mirror each others' increases and decreases more.[6]

Charge Pumping

Before proceeding with fixed-amplitude, fixed-frequency CP measurements during the "sense" periods interspersed throughout the constant voltage stress, a charge pumping frequency sweep was taken after brief periods of stress. From the earlier stated equation $I_{CP,Max} = qfA_G N_{it}$, the density of traps can be derived as $N_{it} = \frac{I_{cp}}{qfA_G}$, where I_{CP} is measured directly as described earlier. The results showed more trap generation at higher frequencies for PMOS, corresponding spatially to the area closer to the SiO_2/Si interface. This is contrary to what was seen in the NMOS devices, where the greater trap generation was seen at the lower frequencies corresponding to the HfO_2/SiO_2 interface (see Figs. 18 & 19).

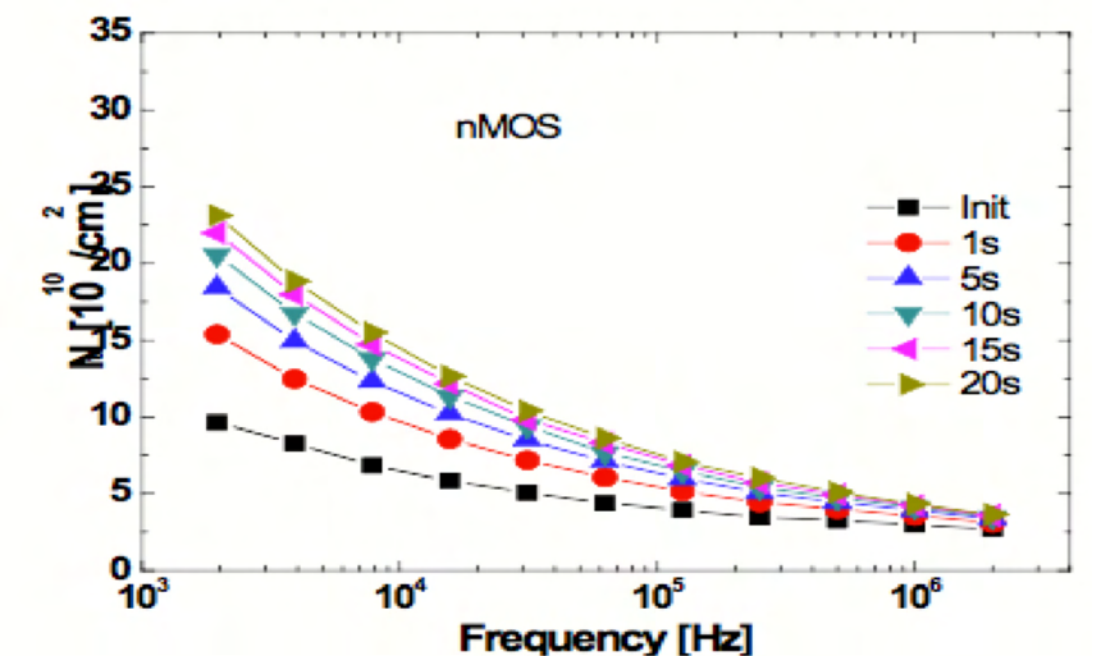

Figure 18: NMOS CP frequency sweep time evolution on 3nm gate stack

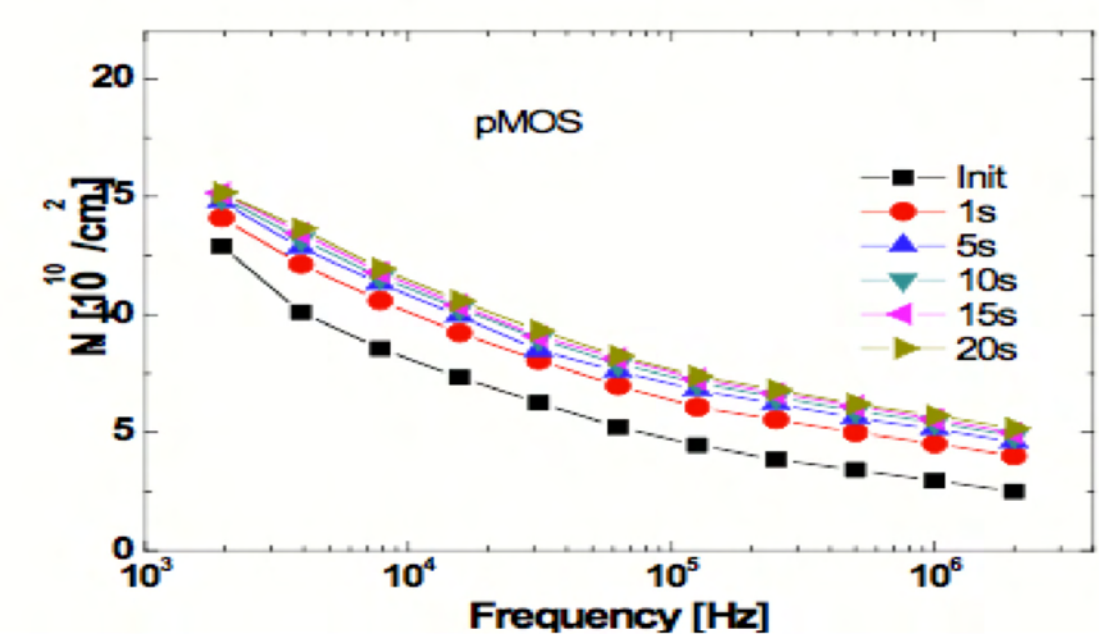

Figure 19: PMOS CP frequency sweep time evolution on 3nm gate stack

Using this as a starting point, 1MHz and 3kHz were chosen as our high and low frequencies to use in the CP measurements; in particular, 3kHz seemed to be the lowest

frequency which would yield reliable data due to the large amount of leakage seen at even lower frequencies. CP measurements at both frequencies were then taken intermittently during the CVS measurements. A discharge step was executed, as described earlier, to evacuate any holes from the traps prior to taking each CP measurement. The pulses used had a fixed amplitude of -1.4V and a rise/fall time of 100ns. Due to the fact that charge pumping is a cumulative measurement, meaning that the pulse causes electrically active traps from the SiO_2/Si interface to be filled up to it's maximum probing depth, the lower frequency measurement should include those traps detected by the high frequency measurement since it can probe deeper into the gate stack. In order to adjust for this, the high frequency defect density curves were subtracted from the low frequency defect density curves to give a new, corrected curve for each "sense" point that should better reflect the actual defect density. The time evolution of the density of traps, Nit, as the base voltage, V_b, was swept for the 3nm samples under various gate voltage biases is shown in Figs. 20 - 27. The graphs only show data until the last "sense" point prior to breakdown, SBD or HBD, since afterwards the direct leakage pathway to the gate makes the data unreliable. The same graphs for the 2.2nm samples are shown in Figs. 28-45. In these cases, the 3kHz CP measurement turned out to be more "leaky" than that of the 3nm samples; due to this, the usual peak in the N_{it}-V_b graph expected to be seen was less apparent. This extra leakage is likely due to the fact that the 3kHz CP pulse is activating traps deep enough within the gate stack that charge carriers trapped there may be tunneling to the gate electrode before the CP measurement was completed, since the dielectric was so thin on those devices. A higher frequency would reduce the leakage during CP, but also change the probing depth. Once it was realized that the low

frequency CP was leaky, it was decided that continuing with a frequency of 3kHz was the best course of action in order to compare with the 3nm devices already measured. Despite this minor issue with the low frequency CP, the data gathered is still useful, as the proportional increase of traps with stress is of more interest than definitive trap density numbers.

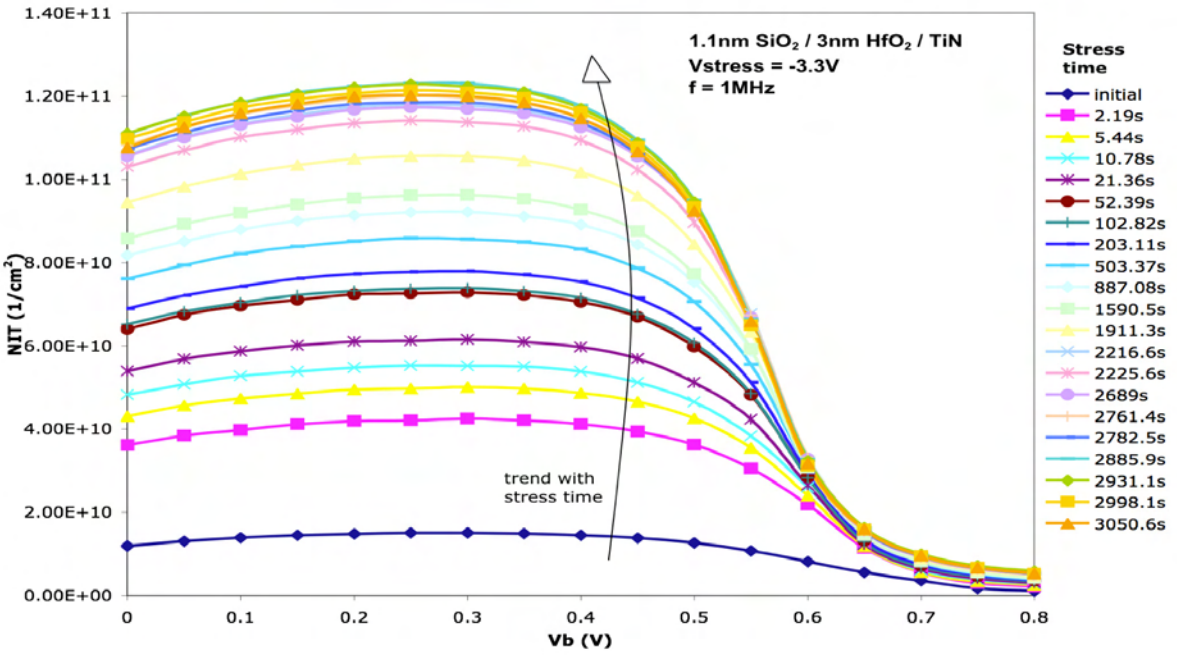

Figure 20: 1MHz Nit – Vb for a 3nm sample at Vg = -3.3V

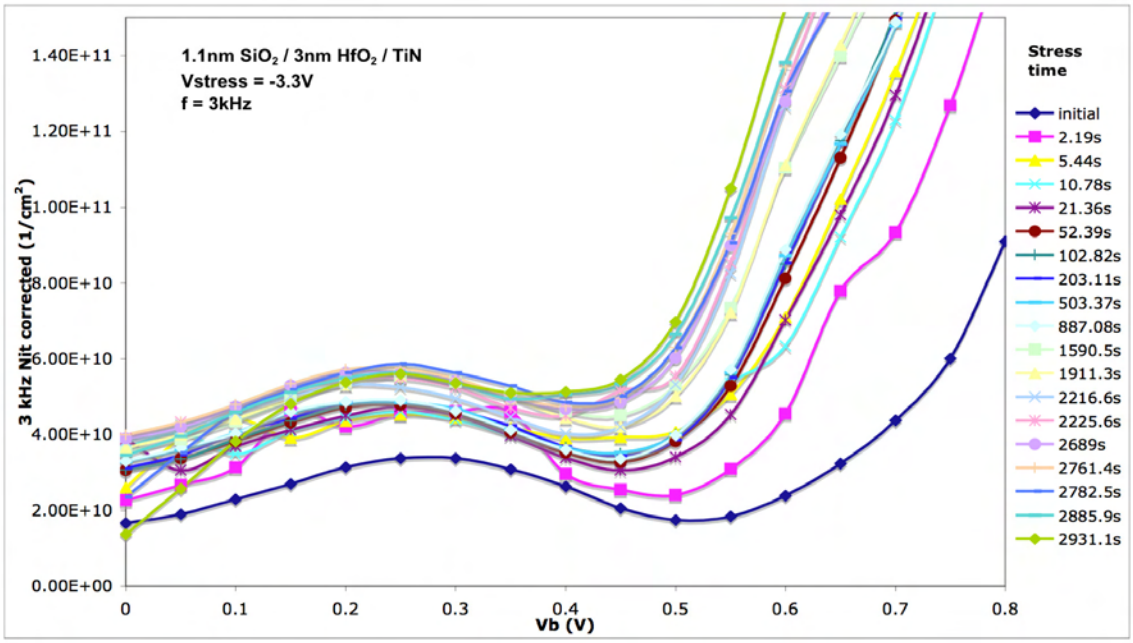

Figure 21: 3kHz Nit – Vb for a 3nm sample at Vg = -3.3V

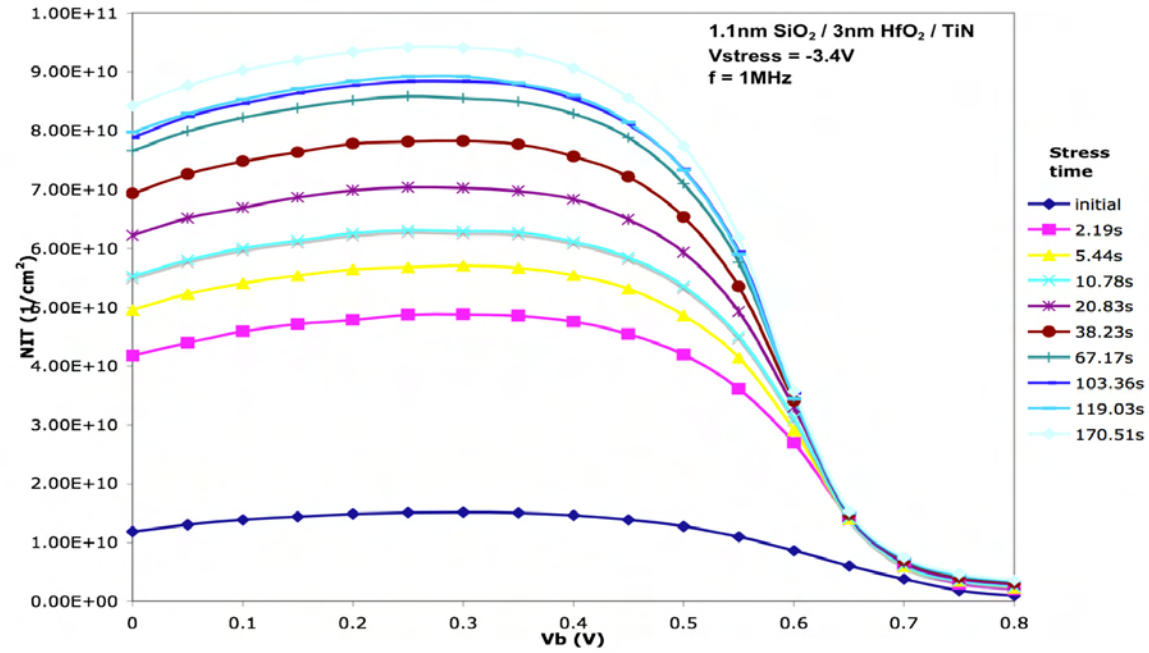

Figure 22: 1MHz Nit – Vb for a 3nm sample at Vg = -3.4V

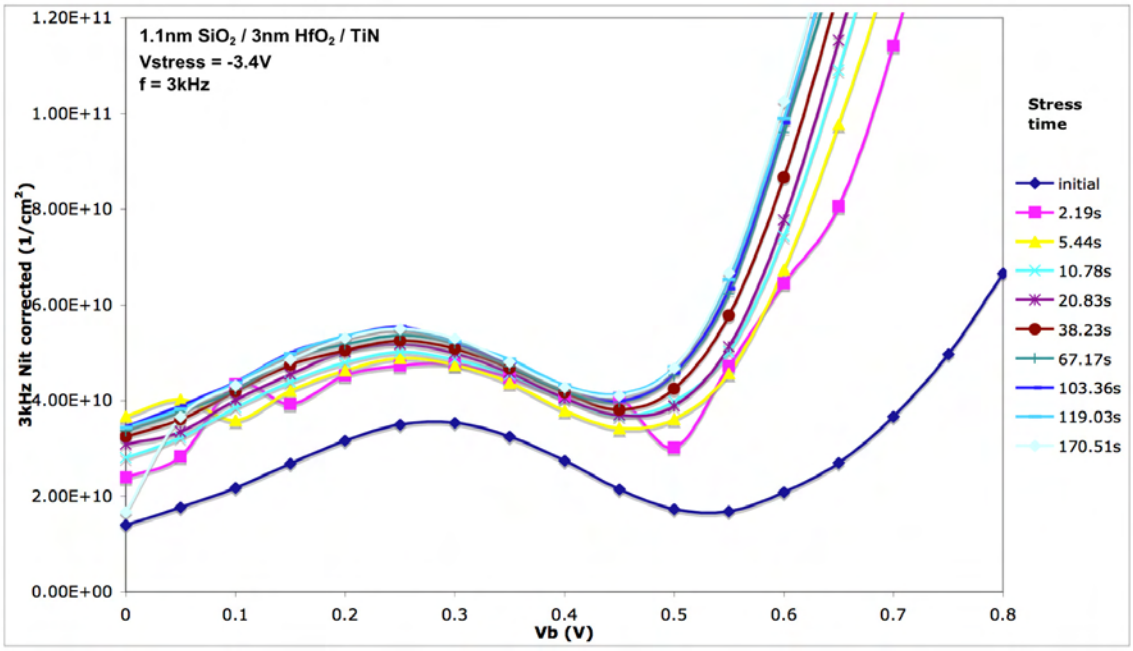

Figure 23: 3kHz Nit – Vb for a 3nm sample at Vg = -3.4V

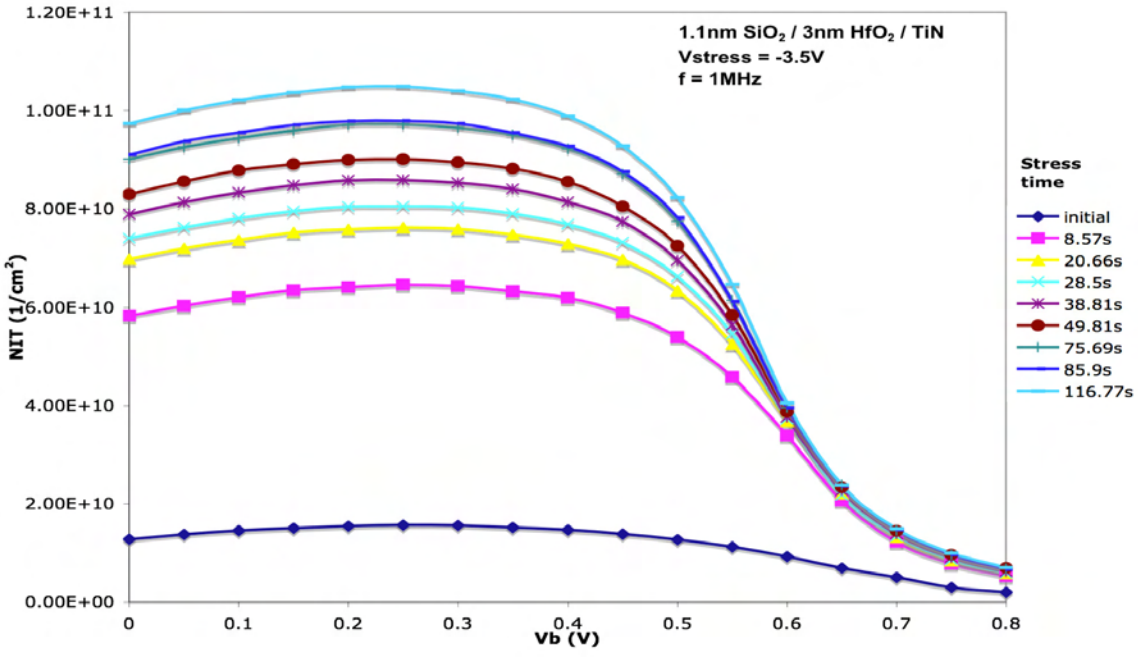

Figure 24: 1MHz Nit – Vb for a 3nm sample at Vg = -3.5V

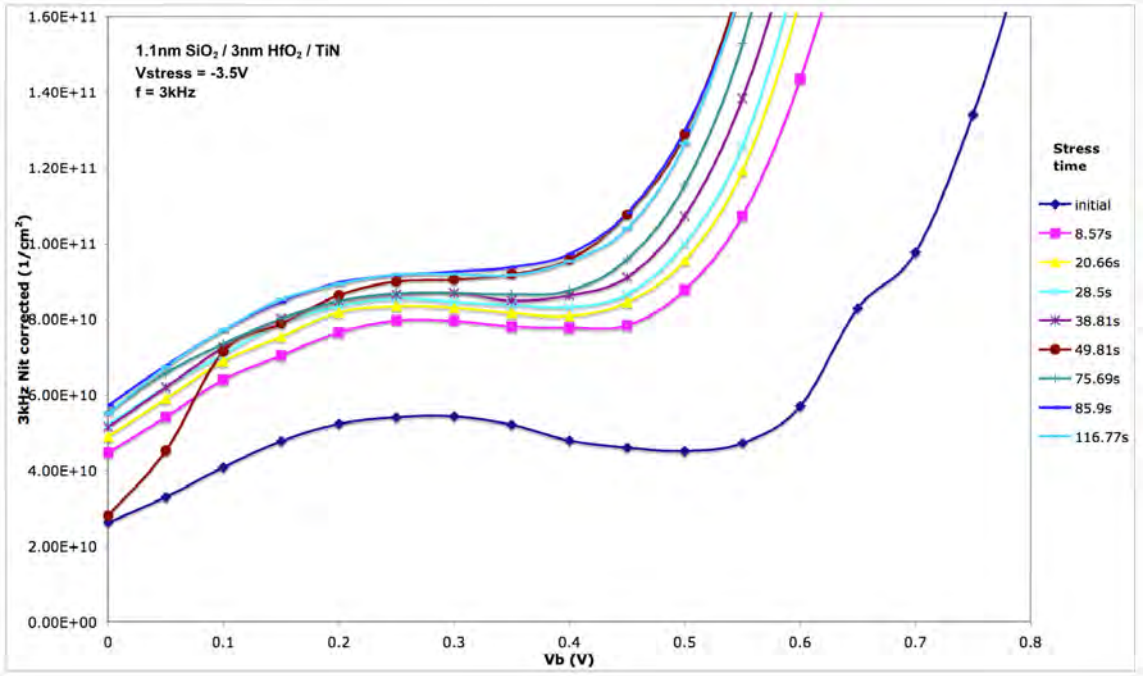

Figure 25: 3kHz Nit – Vb for a 3nm sample at Vg = -3.5V

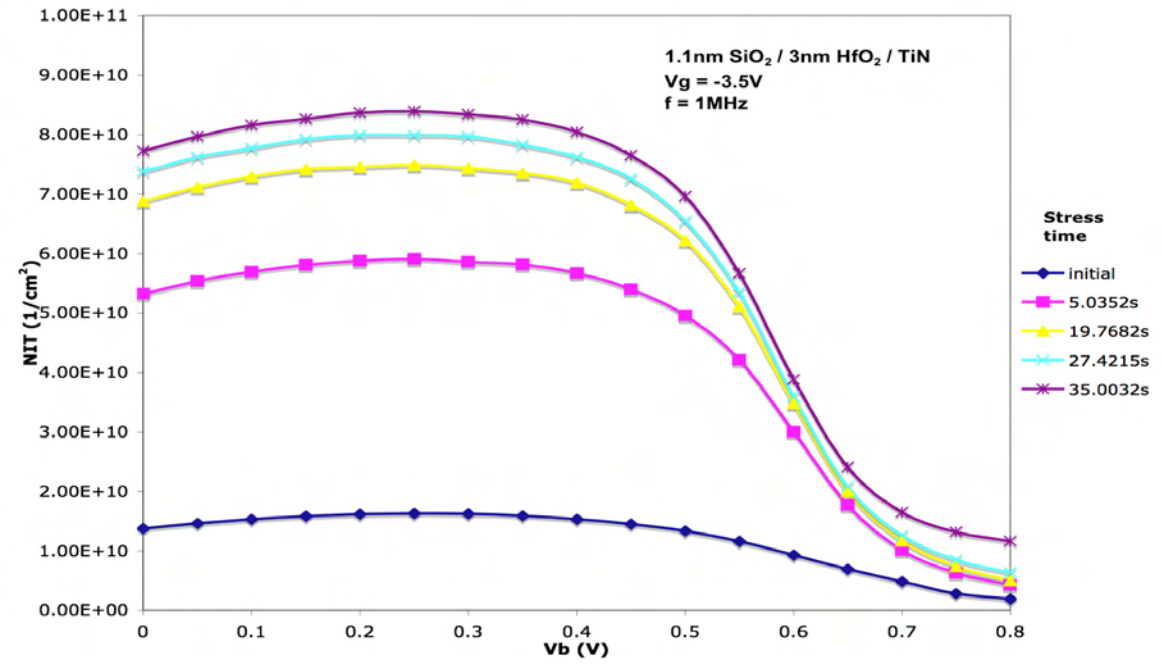

Figure 26: 1MHz Nit – Vb for a 3nm sample at Vg = -3.5V

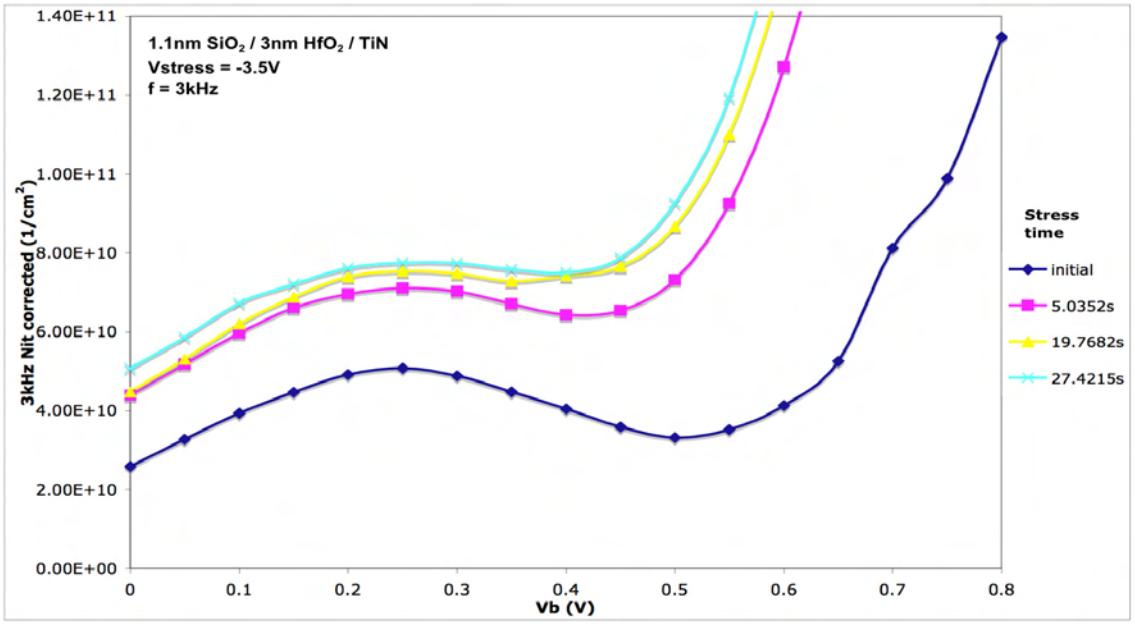

Figure 27: 3kHz Nit – Vb for a 3nm sample at Vg = -3.5V

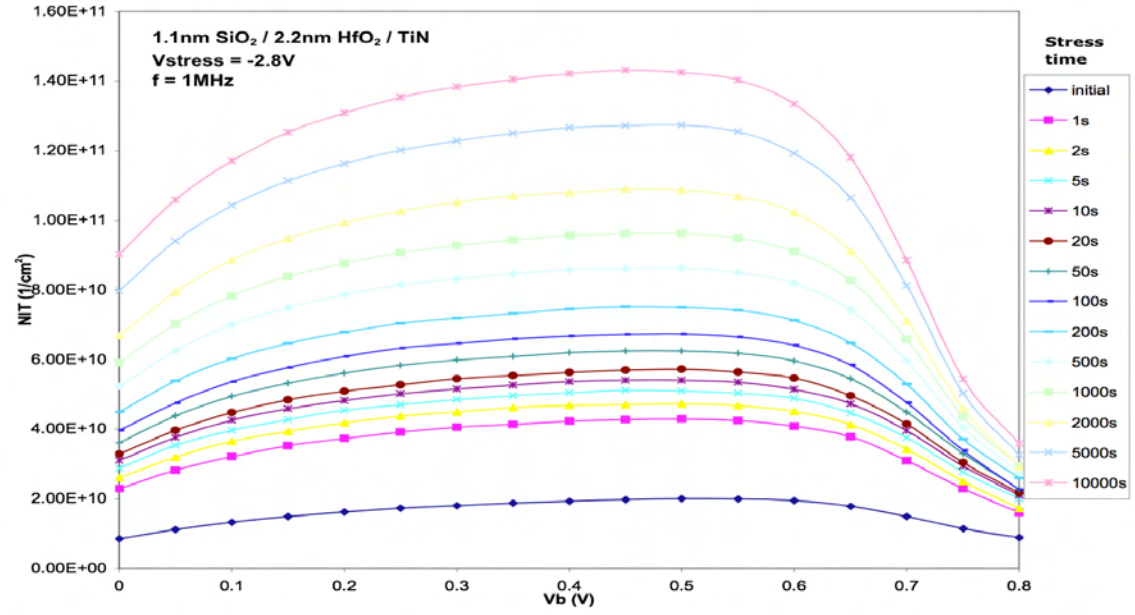

Figure 28: 1MHz Nit – Vb for a 2.2nm sample at Vg = -2.8V

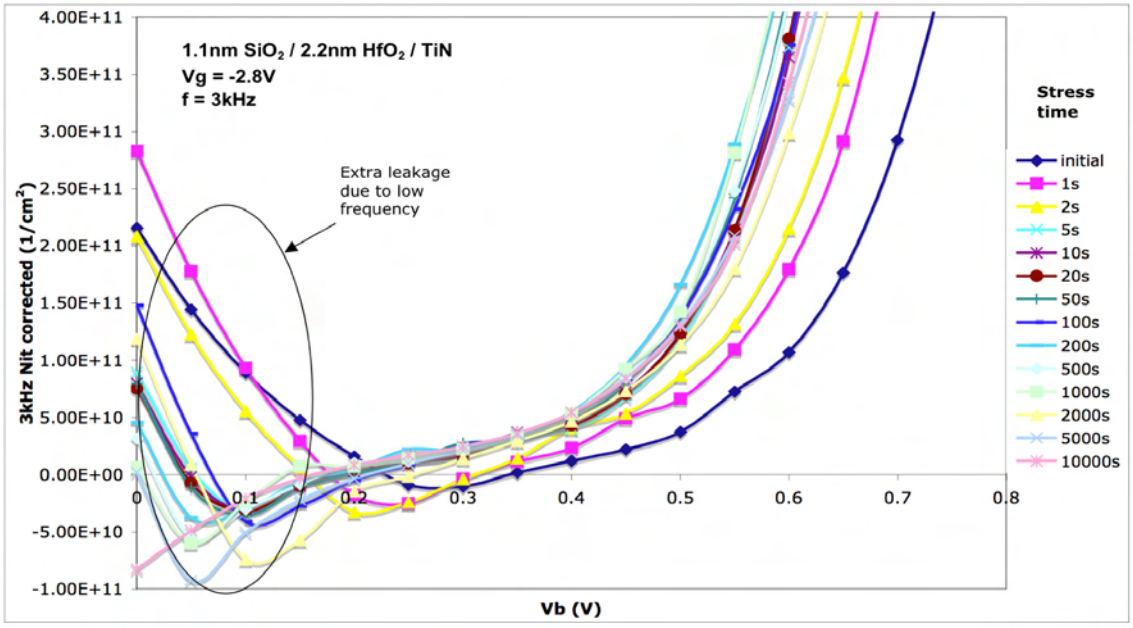

Figure 29: 3kHz Nit – Vb for a 2.2nm sample at Vg = -2.8V

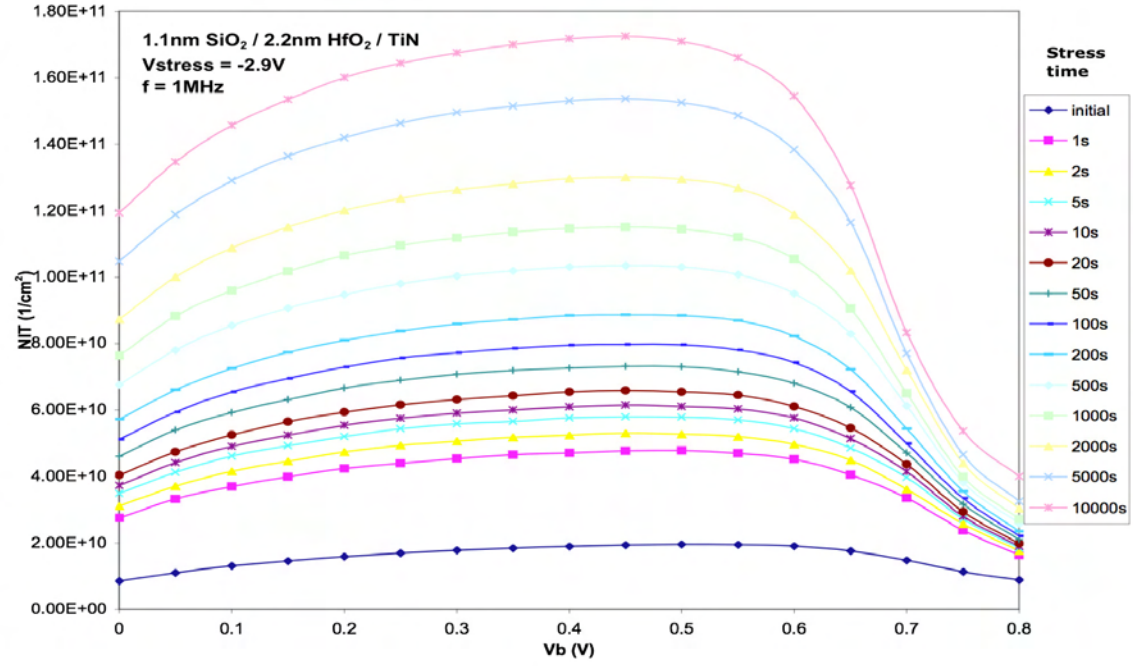

Figure 30: 1MHz Nit – Vb for a 2.2nm sample at Vg = -2.9V

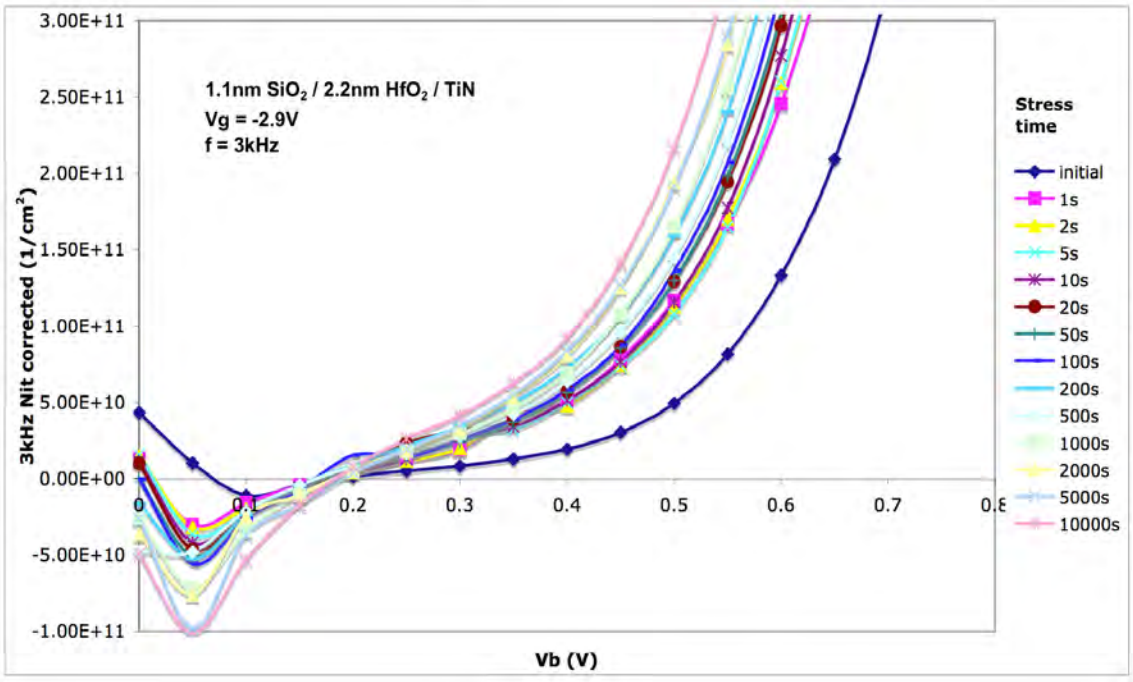

Figure 31: 3kHz Nit – Vb for a 2.2nm sample at Vg = -2.9V

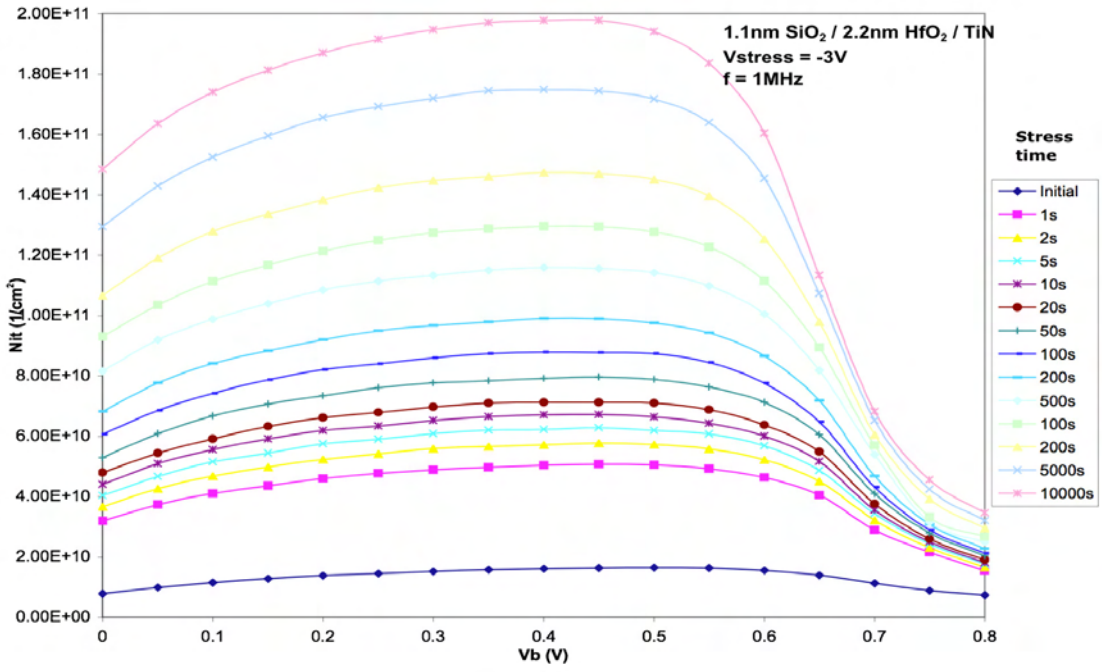

Figure 32: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3V

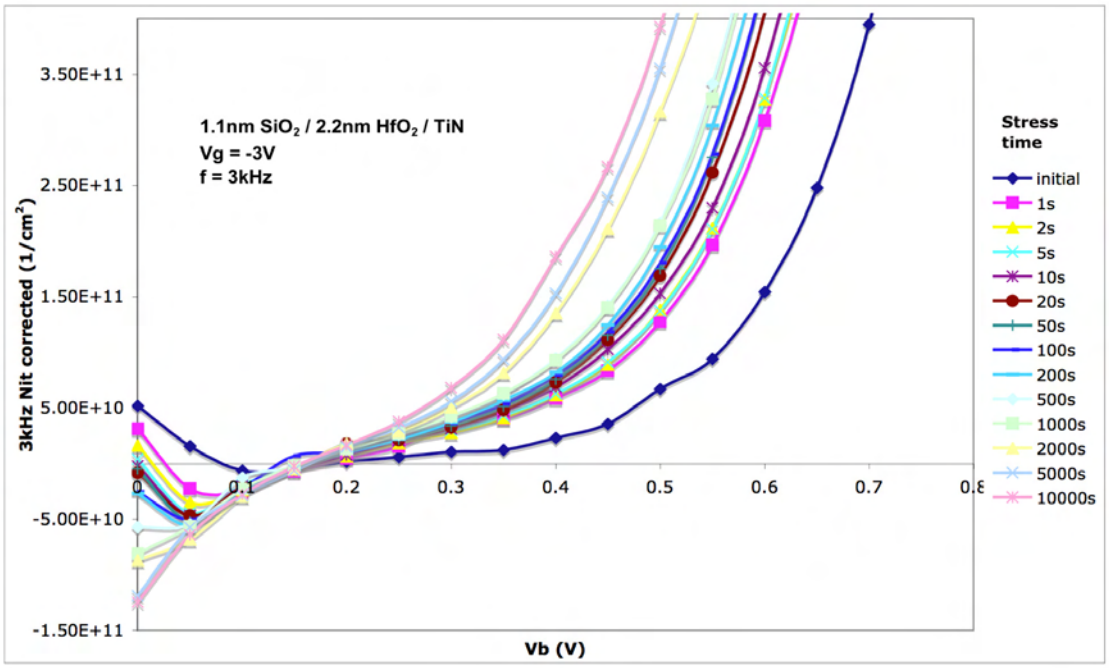

Figure 33: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3V

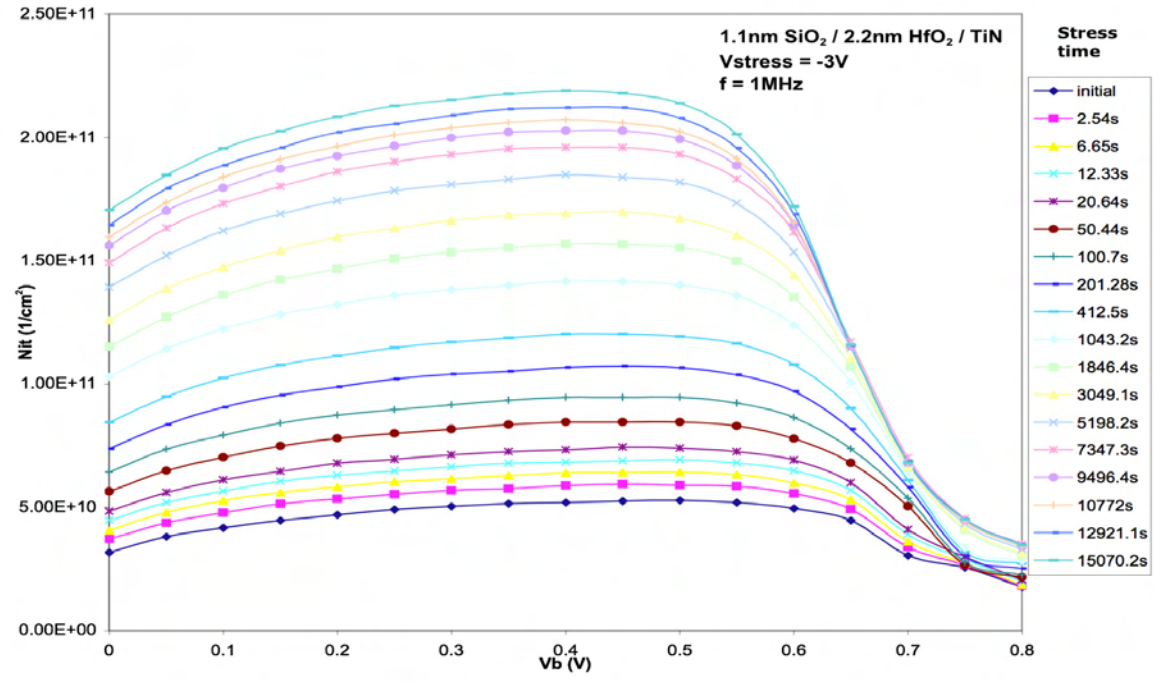

Figure 34: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3V

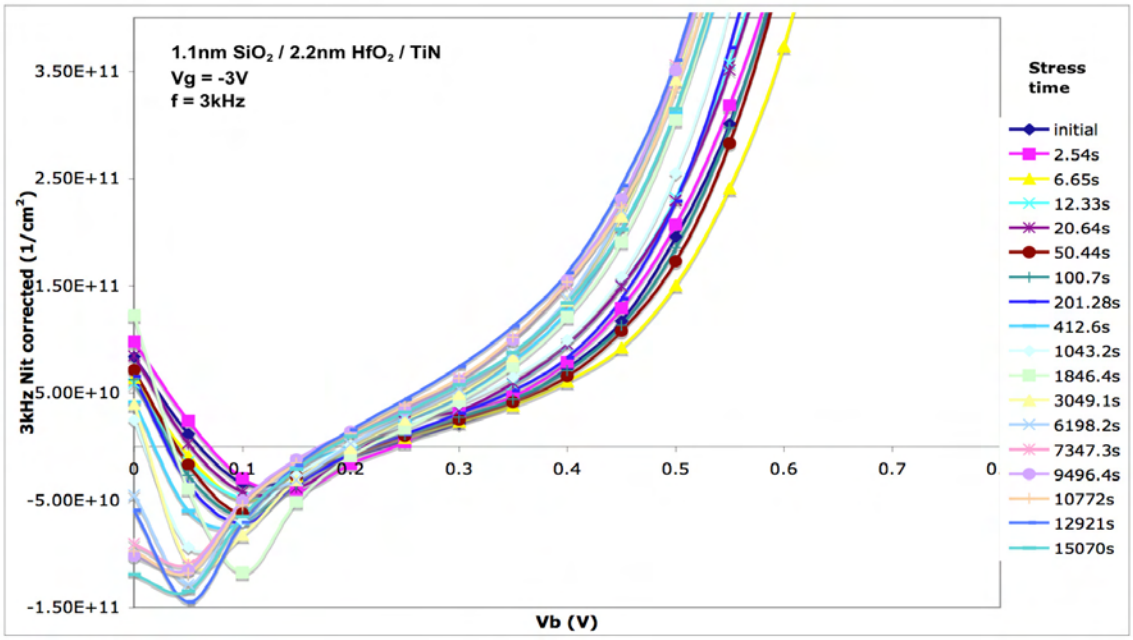

Figure 35: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3V

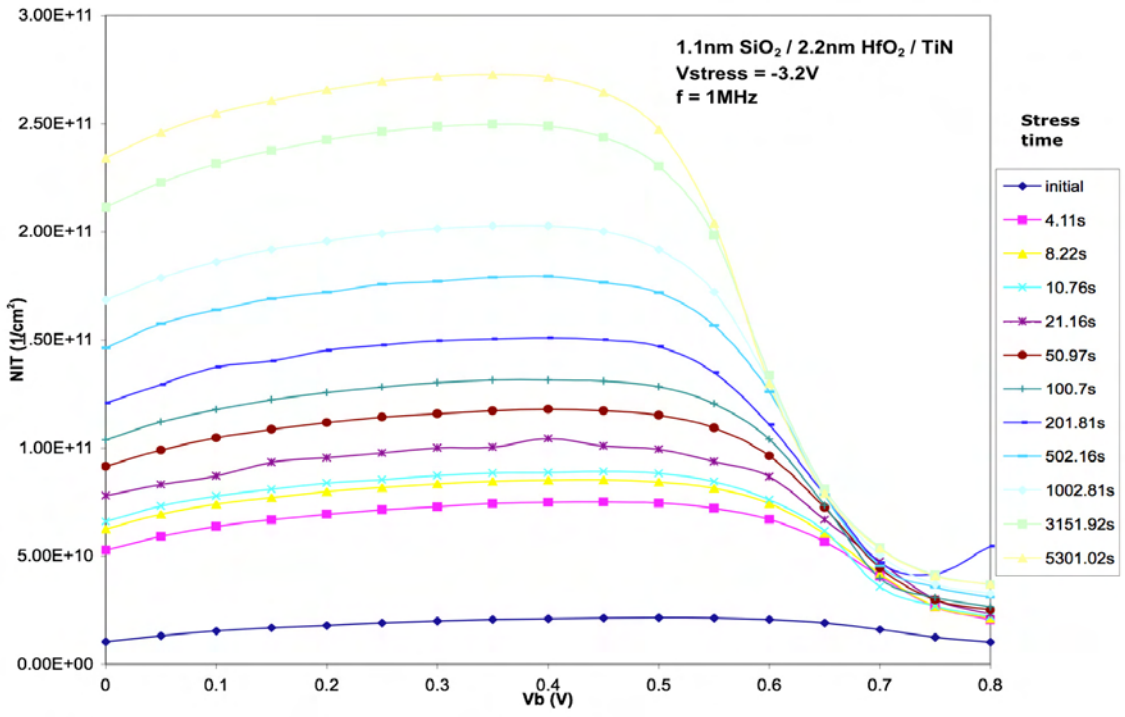

Figure 36: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3.2V

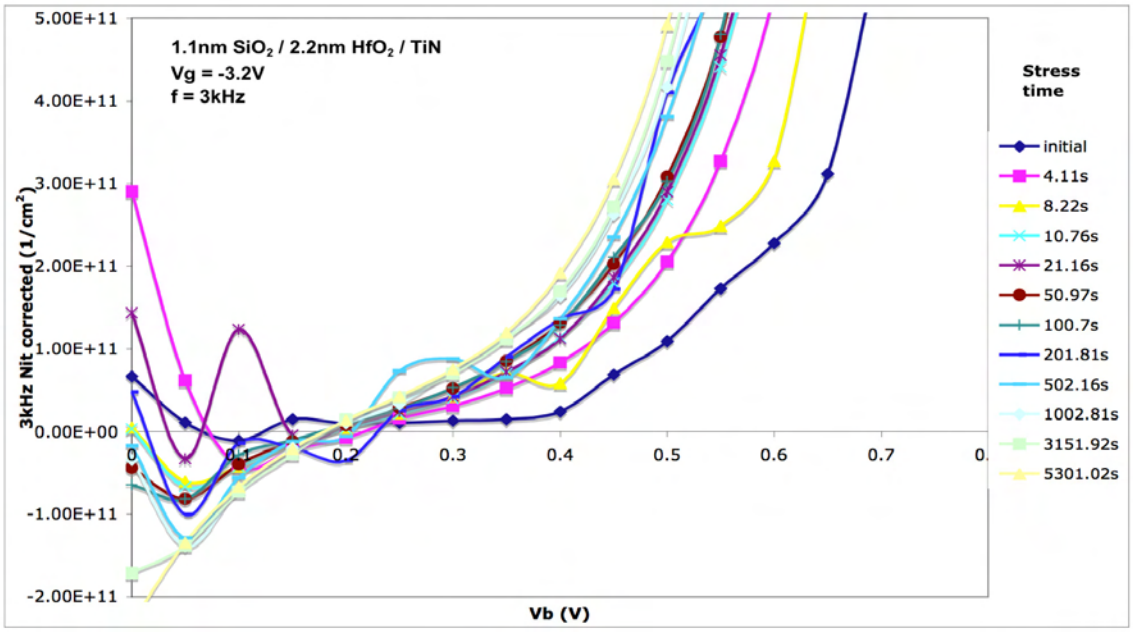

Figure 37: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3.2V

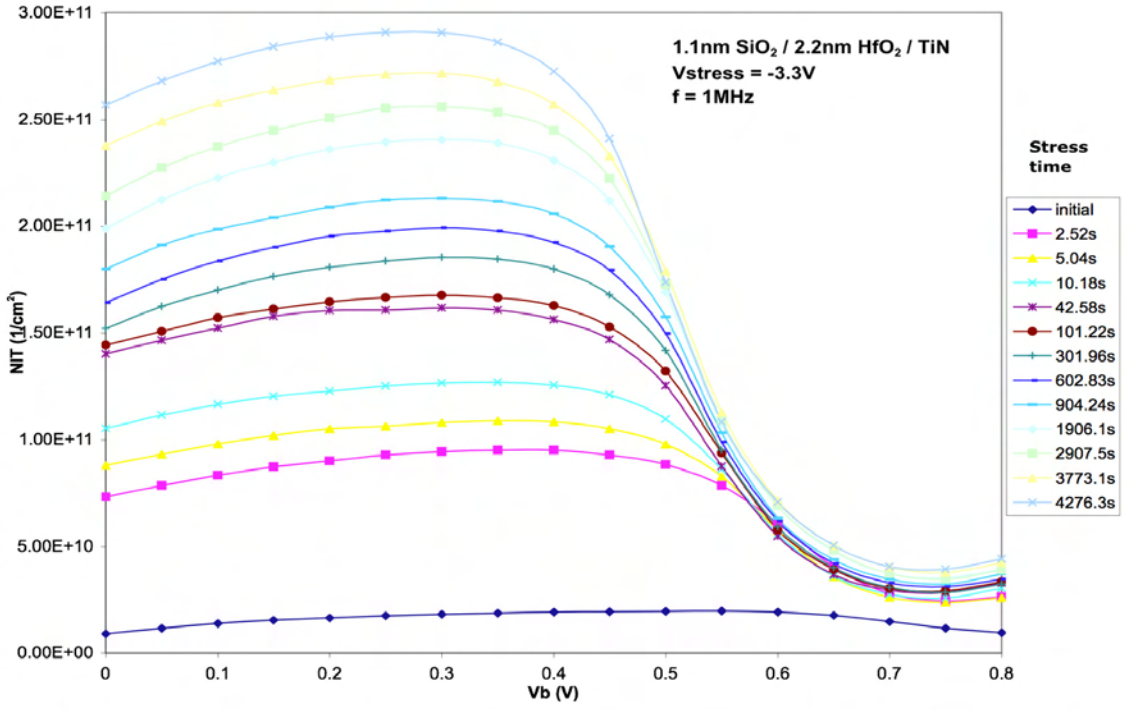

Figure 38: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3.3V

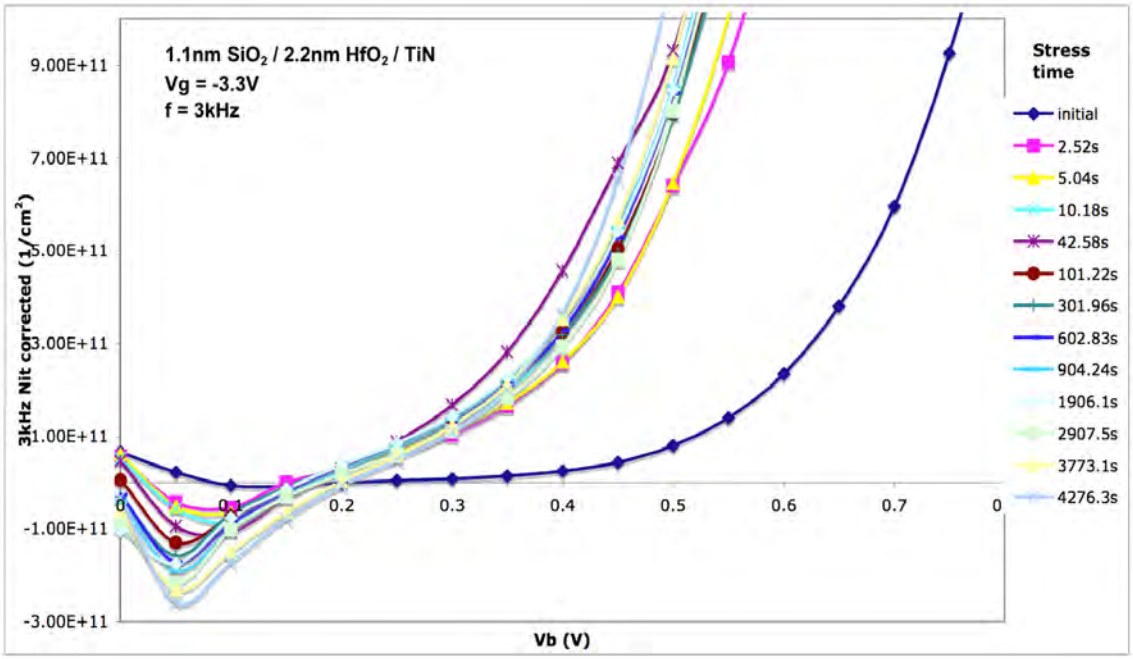

Figure 39: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3.3V

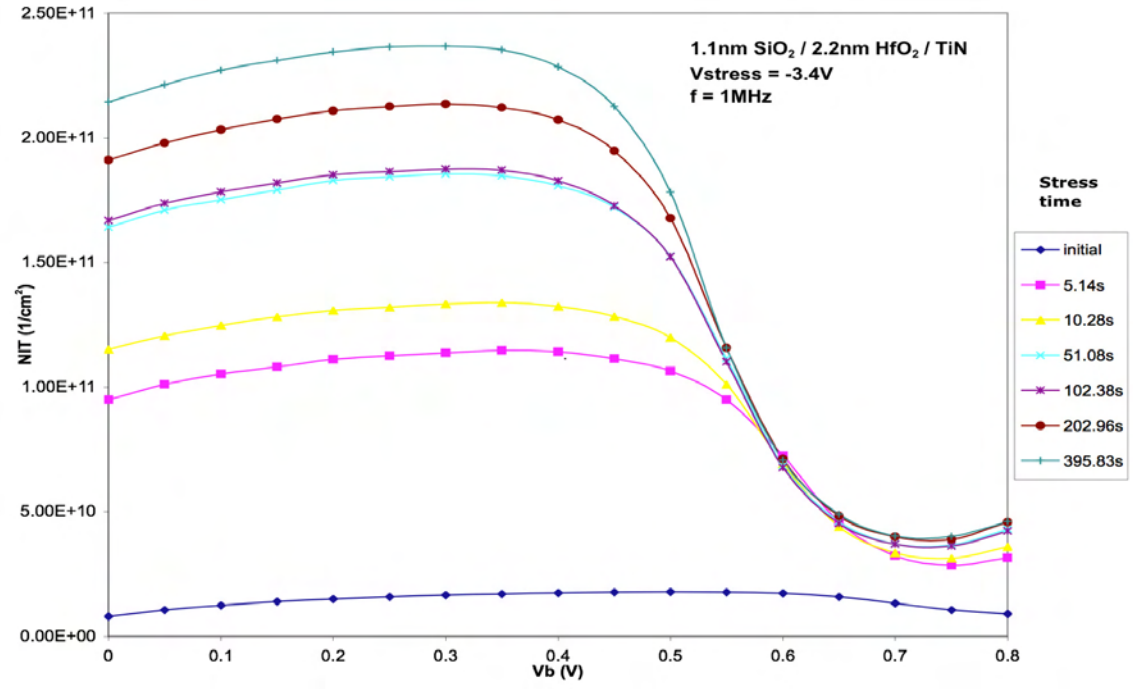

Figure 40: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3.4V

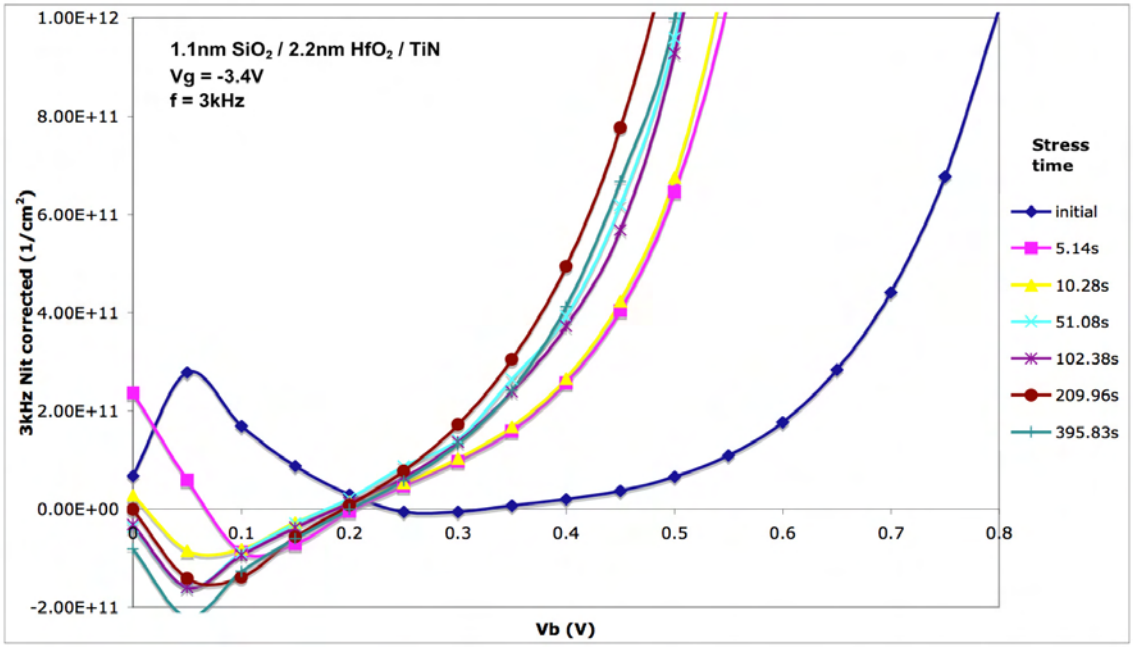

Figure 41: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3.4V

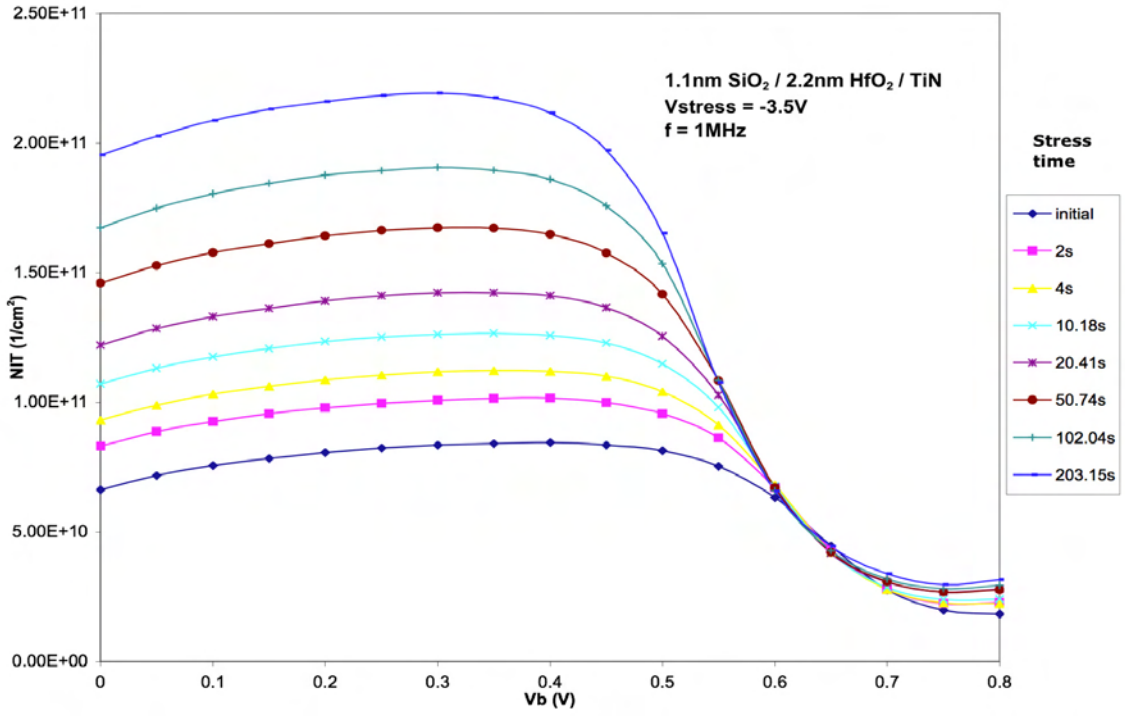

Figure 42: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3.5V

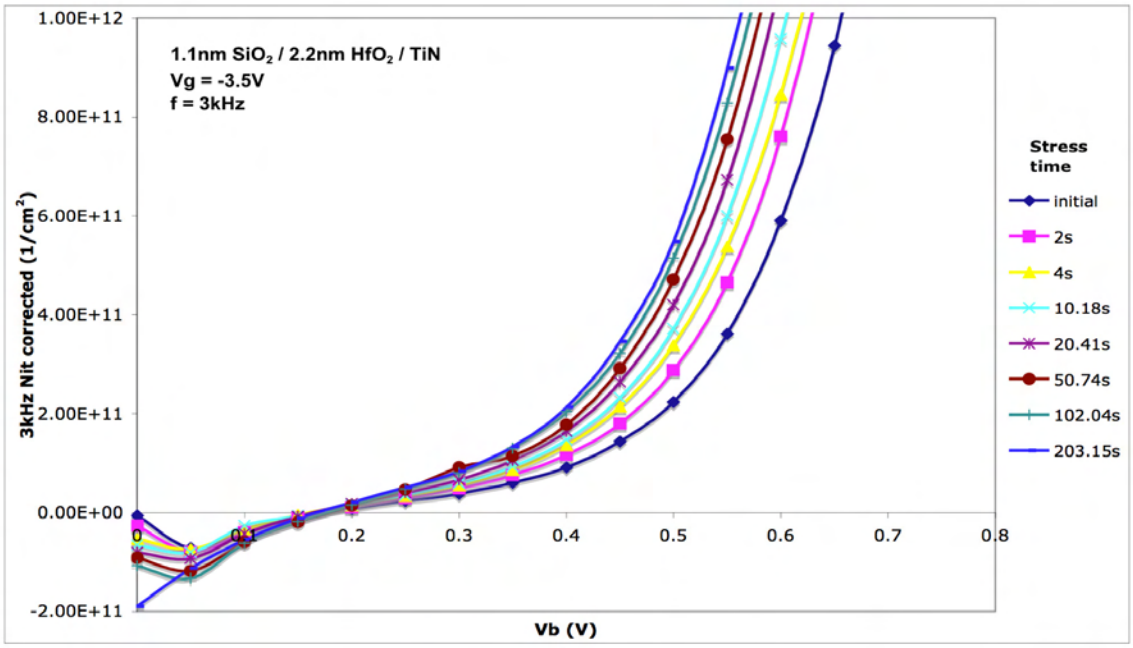

Figure 43: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3.5V

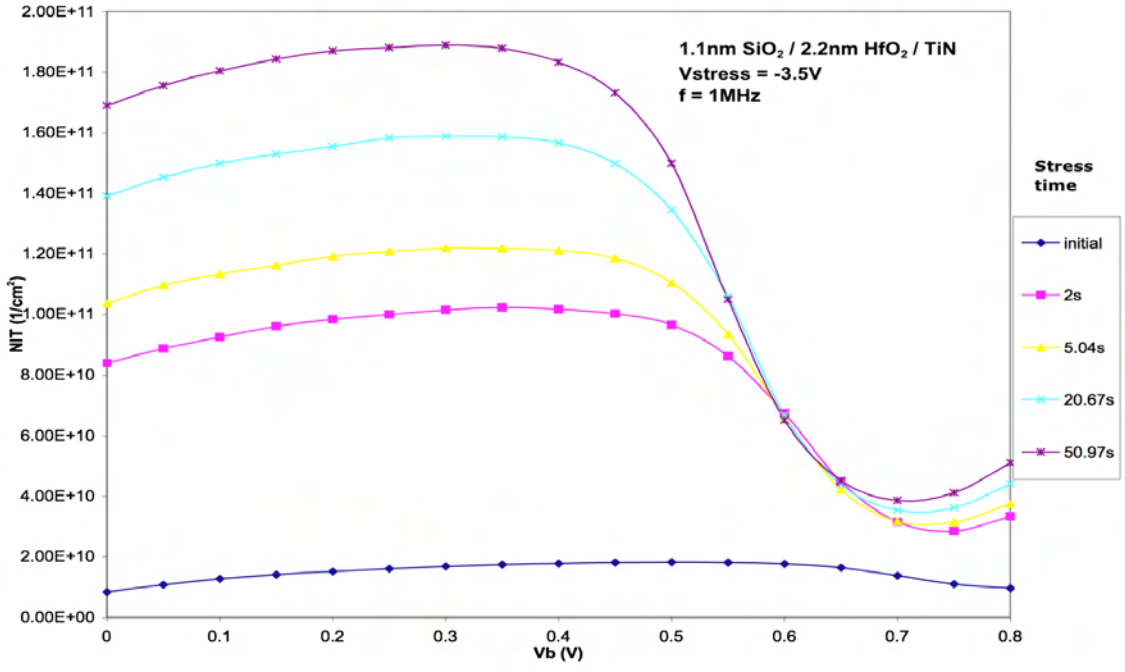

Figure 44: 1MHz Nit – Vb for a 2.2nm sample at Vg = -3.5V

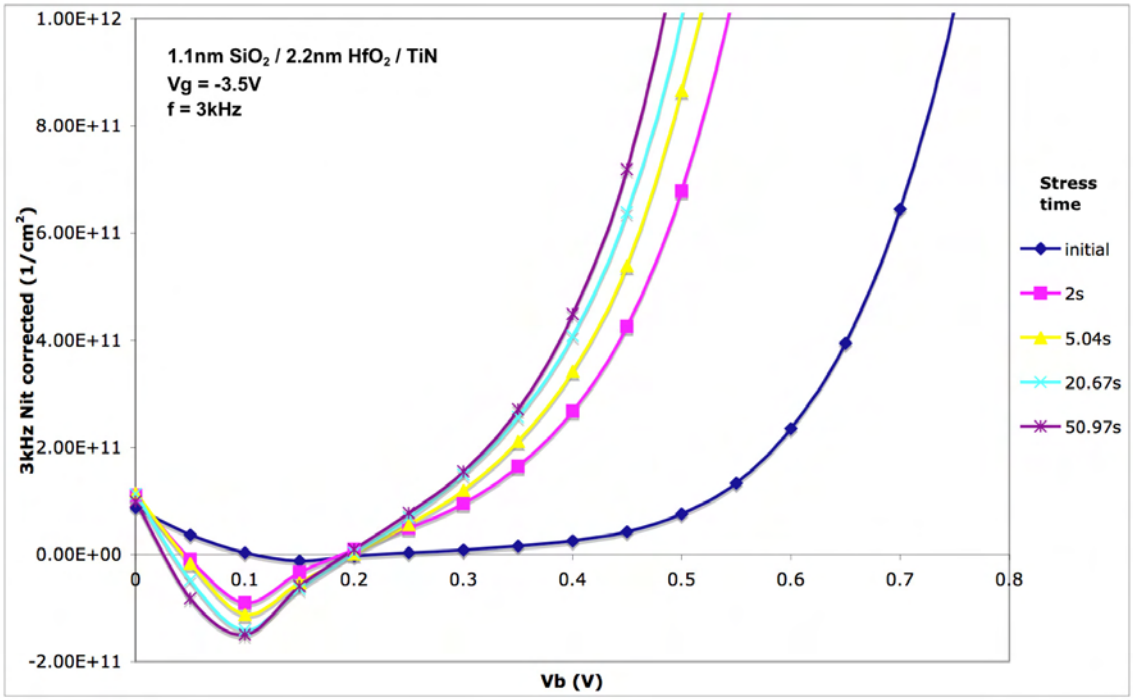

Figure 45: 3kHz Nit – Vb for a 2.2nm sample at Vg = -3.5V

It is apparent from the previous figures that the local maximum of N_{it} for both frequencies was seen at a base voltage of about 0.25V for the 3nm samples. For the 2.2nm samples, the maximum N_{it} is seen to occur between a base voltage of 0.35-0.45V for the high frequency measurements. It is expected that the peak for the low frequency occurs at the same place, so the data used in the calculations were taken from those corresponding V_b values. The slight variance in base voltage at which N_{it} max was seen in the 2.2nm samples between trials is due to charging of the interface states, which change the effective potential seen there. This can also be thought of as a small shift in the threshold voltage. The shift can also be seen in the longer trials, where the peak of the CP curve shifts slightly with long periods of stress time, though not enough to noticeably change the value of N_{it} when taken at a single V_b for the whole trial. By inspection of the measurements taken, it is clear that the low frequency CP showed a

larger initial density of traps than the high frequency, indicating that there is a larger number of intrinsic defects near the HfO_2/SiO_2 interface than the SiO_2/Si interface.

Taking the peak value of the N_{it}-V_b curves at high and low frequency, the time evolution of generated traps can be compared (see Figs. 46, 47, 50, 51). The relative change of the trap density was also investigated by calculating $\Delta N_{it} = \frac{N_{it}(t) - N_{it}(t=0)}{N_{it}(t=0)}$ and noting it's progression with time (see Figs. 48, 49, 52, 53), as well as the trap generation rate extrapolated by taking the best fit curve (of the form $c + \alpha x^{\beta}$) of the N_{it} – t graphs and using the power β as the rate of generation, since the graph is straight in log – log scale. (see Fig. 54).

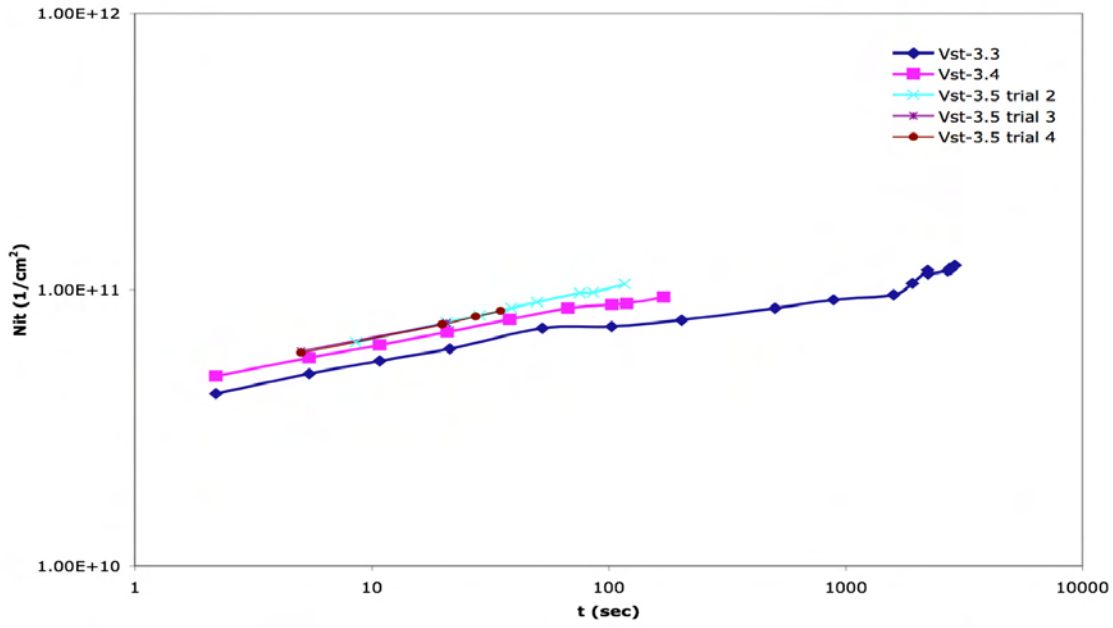

Figure 46: Time evolution of 1 MHz Nit max for 3nm gate stacks

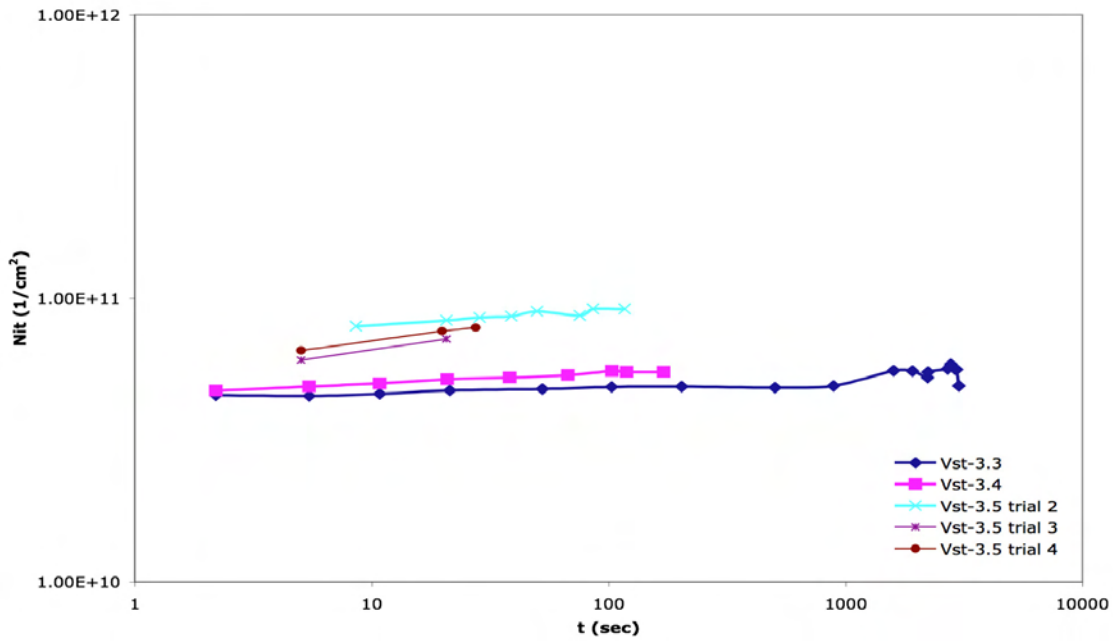

Figure 47: Time evolution of 3kHz Nit max for 3nm gate stacks

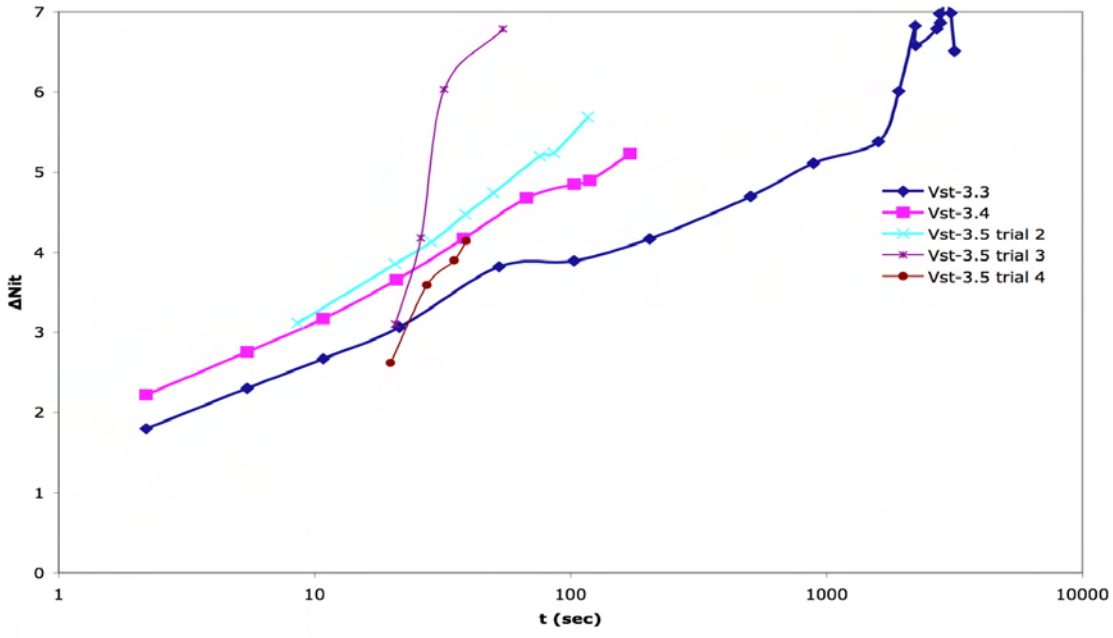

Figure 48: Time evolution of 1 MHz relative ΔNit 3nm gate stacks

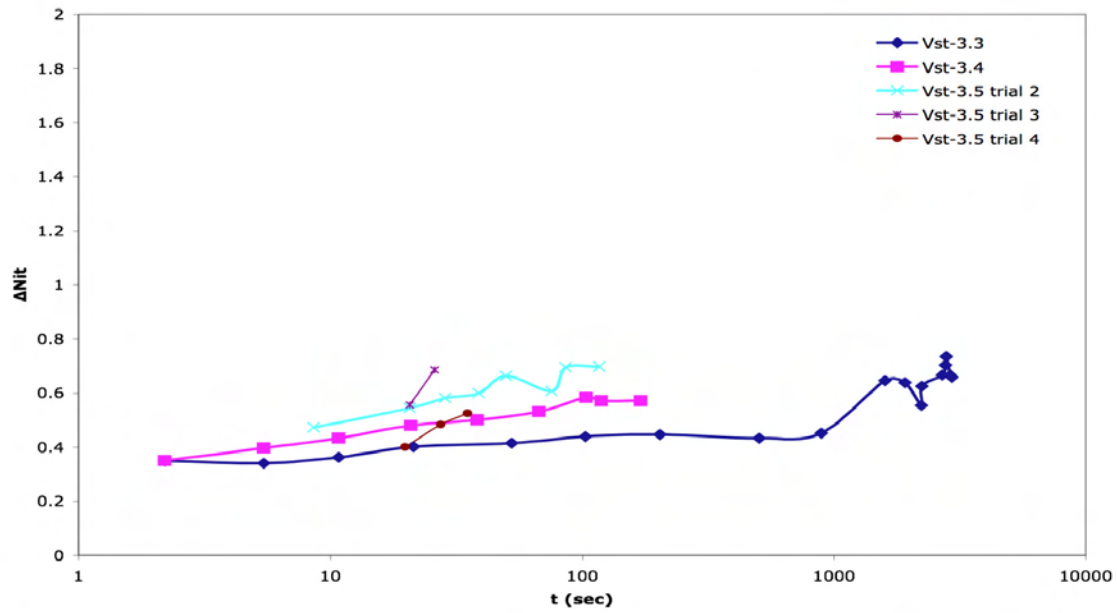

Figure 49: Time evolution of 3kHz relative ΔNit for 3nm gate stacks

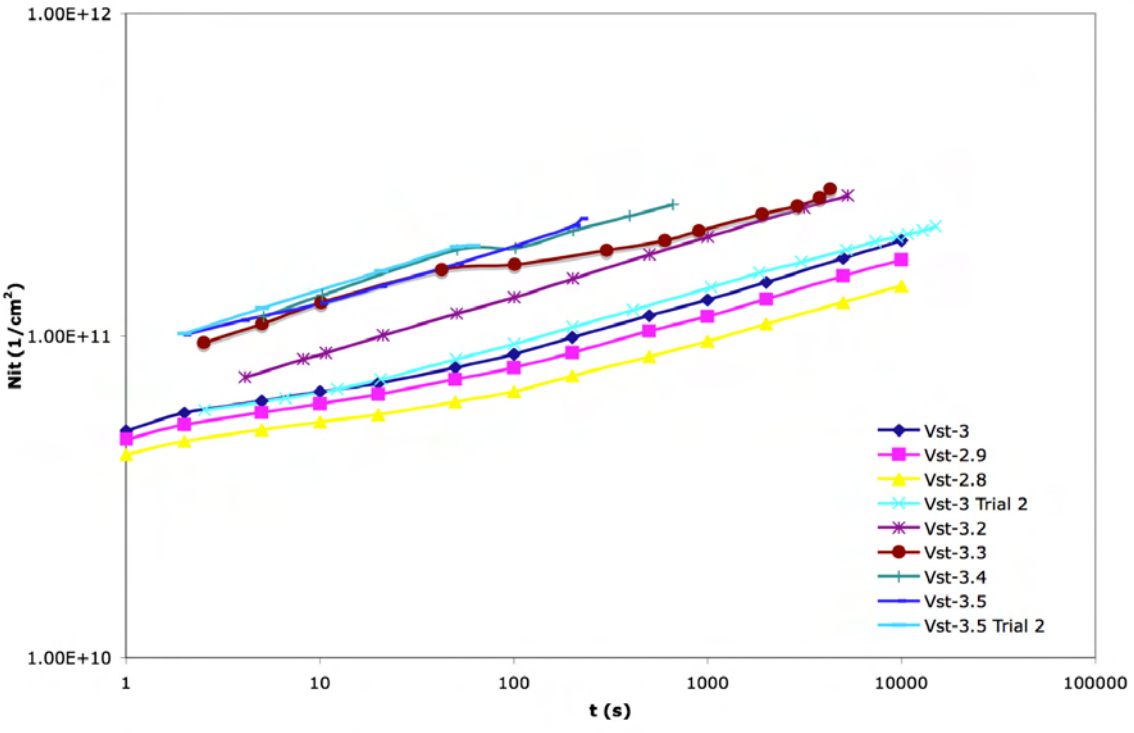

Figure 50: Time evolution of 1MHz Nit max for 2.2nm gate stacks

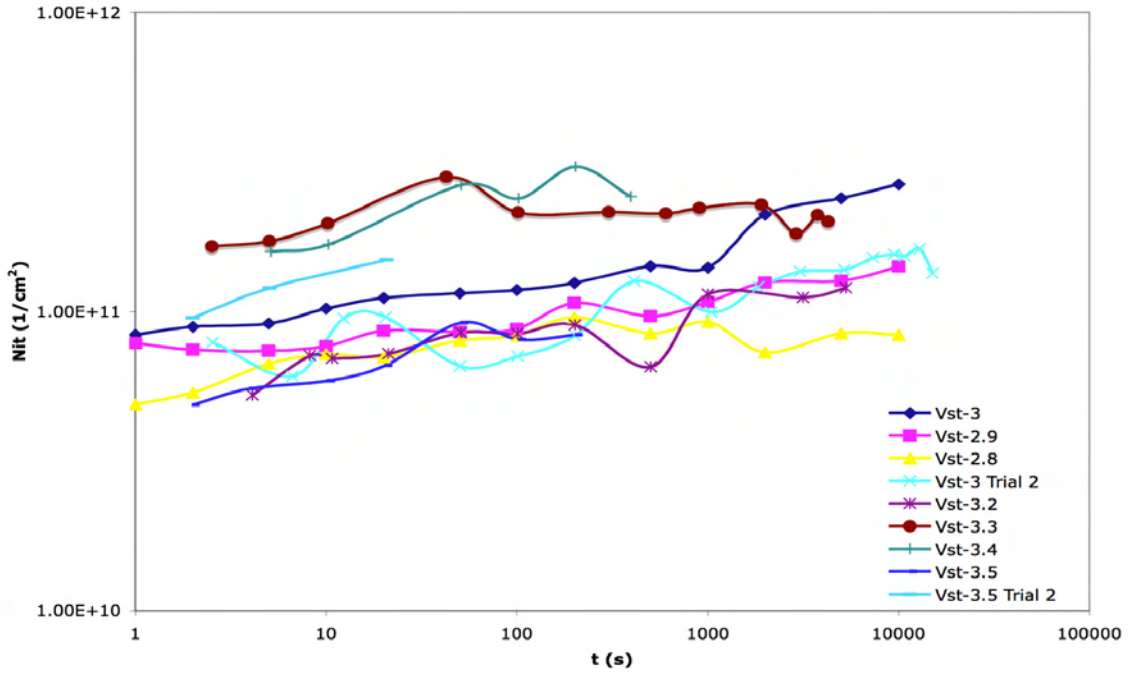

Figure 51: Time evolution of 3kHz Nit max for 2.2nm gate stacks

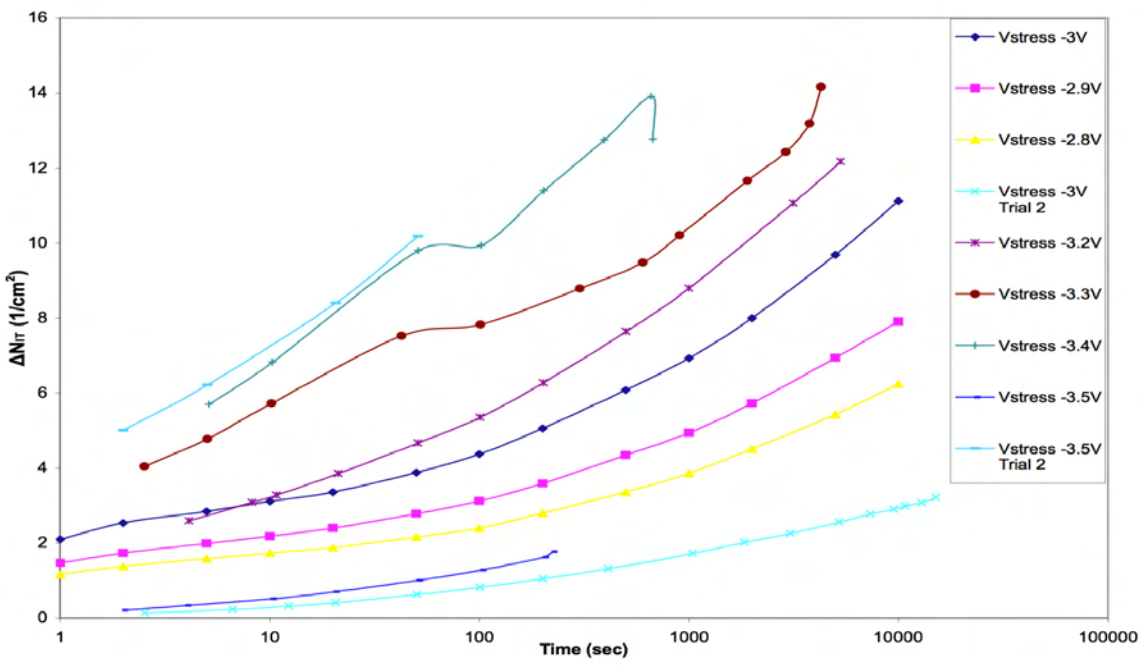

Figure 52: Time evolution of 1MHz relative ΔNit for 2.2nm gate stacks

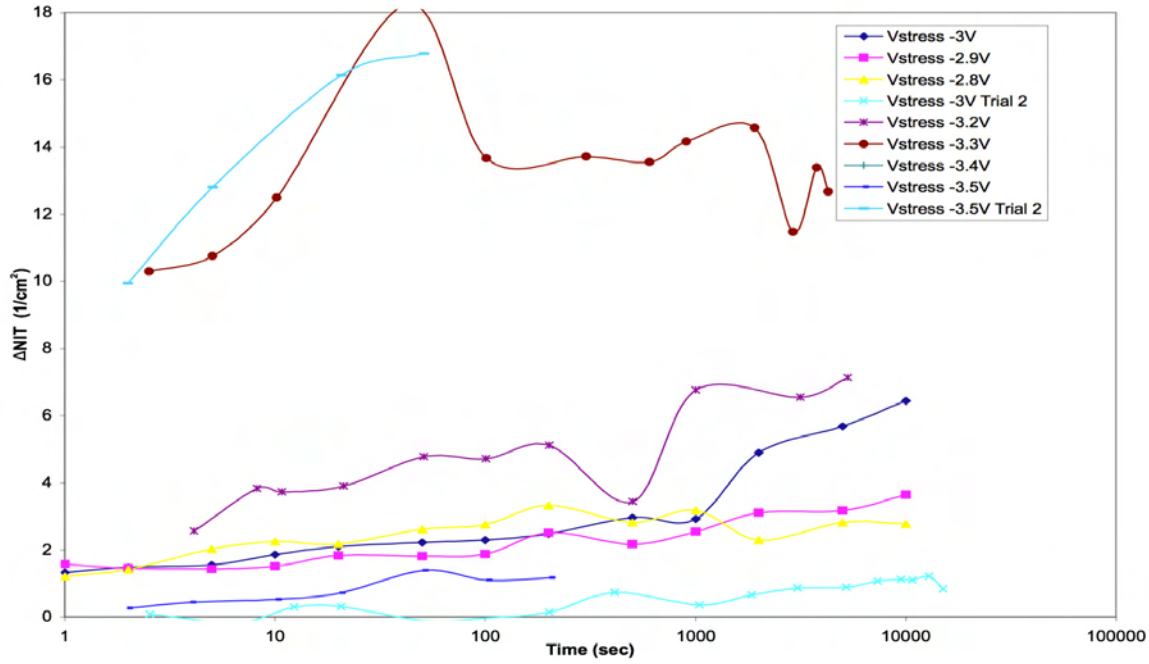

Figure 53: Time evolution of 3kHz relative ΔNit for 2.2nm gate stacks

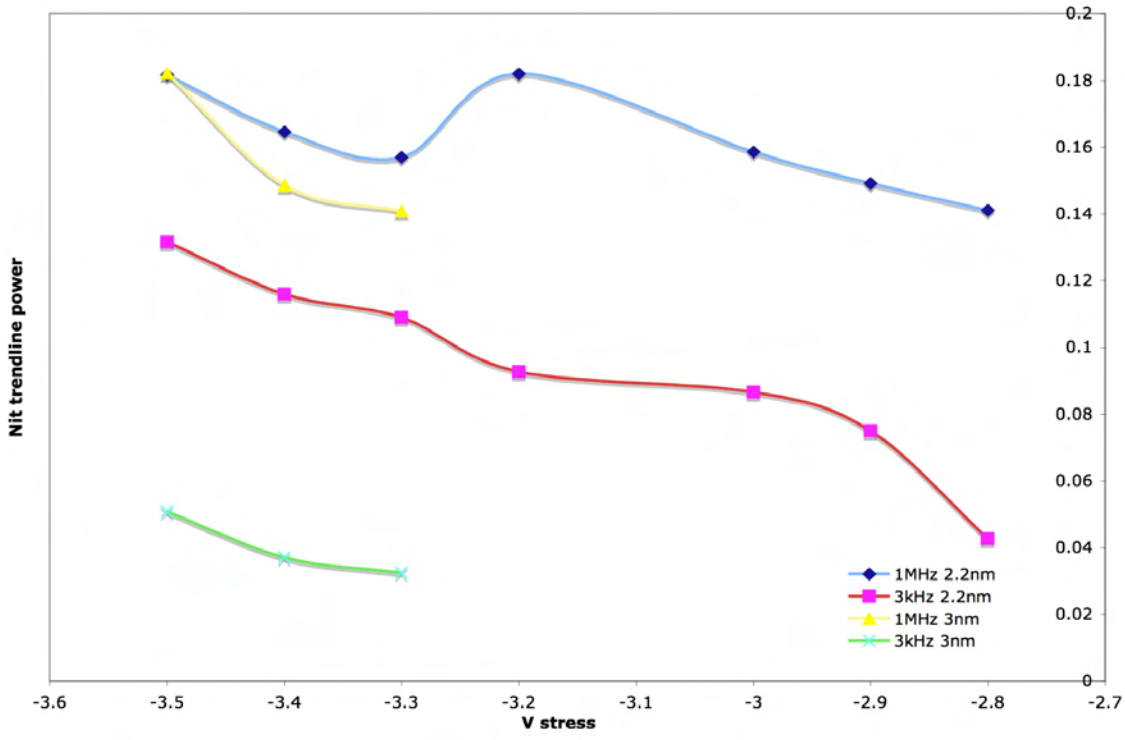

Figure 54: Power of best fit curve (Nit - t) vs. Vstress at high and low frequencies for both thicknesses

As time increases, more traps are being generated near the SiO_2/Si interface (see Figs. 46 & 47). By comparing Figs. 48 & 49, it can be seen that the relative increase of high frequency CP traps was much greater than that of the low frequency traps in the 3nm samples. This indicates that even though there is a larger number of initial traps near the HfO_2/SiO_2 interface, more trap sites activated due to stress lie closer to the SiO_2/Si interface. It is also apparent from the power law behavior of the trap density that more traps are generated earlier during stress. Comparing the generation rates for the 1MHz and 3kHz traps (Fig. 54), it is seen that the generation rate is greater for the high frequency traps in the 3nm samples, and there is a trend for the generation rate to increase with greater stress voltage bias.

The 2.2nm samples showed a similar story, although the low frequency data was somewhat suspect, particularly when looking at definite values for defect densities. The relative increase in traps for low and high frequencies were very similar, with a tendency for the high frequency data to be slightly larger. The generation rates were closer to each other than in the 3nm samples, and also showed an increasing trend with higher stress voltage bias. The higher number of traps being generated near the HfO_2/SiO_2 interface for the 2.2nm samples is to be expected, since the HfO_2 is thinner. This provides a thinner potential barrier for electrons tunneling from the gate electrode to the interface, and also increases the effective field seen there.

In both samples, the greatest increase in the trap density near both interfaces happens immediately after a stress voltage is applied. Afterwards, the trap generation rate decreases rapidly, as a power law with time, and continues until breakdown. This corroborates the theory that traps are generated at weaker points within the oxide

structure. After this large initial jump in trap density within the gate stack, traps are generated at a much slower rate, though it is still not clear why. One hypothesis is that since traps are created at weak points in the oxide structure, this initial application of stress activates the majority of the available defect sites. Because there is now a much reduced number of sites within the gate stack available to be activated as traps, the reduction in the generation of traps as time progresses is seen. Another approach is to view the total population of possible traps as being of two types: fast traps and slow traps. Most of the fast traps would be activated during the first few moments of stress, although a few of the slower traps could also be activated during this time. These fast traps are thought to be the weaker points in the oxide, so they respond very quickly to stress; as such, nearly all of the fast trap population becomes activated after a relatively short time. The slow traps are generated at a much slower rate, so that their effects are seen more during the latter part of the stress measurements. Both theories would explain the power series behavior of the trap densities, with N_{it} increasing drastically during the first few moments then falling off to a slower increase for the rest of the trial, but which one holds true is still being investigated.

It has also been noted that during the manufacture of the devices, the HfO_2 layer has a tendency to siphon oxygen from the SiO_2 layer, creating the large number of initial defects seen near that interface. Looking at the generation rates (Fig. 54) from samples of both thicknesses, it is clear that the rate for trap generation nearer to the SiO_2/Si interface is very similar between the two thicknesses. This indicates that there is little difference in how quickly traps are generated near the substrate interface. The low frequency generation rate, however, shows a large difference, with the 2.2nm samples

having a much higher rate than the 3nm samples. In the context of HfO_2 weakening the SiO_2 layer during manufacture, this makes sense because the thicker HfO_2 layer will take more oxygen from the SiO_2 layer, creating more precursory traps and leaving fewer weak sites available to be activated as traps during stress. Thus, a larger generation rate of traps is seen near the HfO_2/SiO_2 interface for the samples with a thinner HfO_2 layer because fewer traps were generated through the manufacturing process.

Stress Induced Leakage Current (SILC)

As part of the "sense" portion of the experiments, gate leakage currents were measured as the gate voltage was swept. SILC calculations, using $SILC\left(V_g,t\right)=\frac{\left(I_g\left(V_g,t\right)-I_g\left(V_g,t=0\right)\right)}{I_g\left(V_g,t=0\right)}$, were then done to investigate any correlation between relative increase of the gate leakage current and the gate voltage prior to breakdown. The gate leakage current rate of change, $rate\ of\ I_g\left(V_g,t\right)=\frac{I_g\left(V_g,t\right)-I_g\left(V_g,t=0\right)}{t-\left(t=0\right)}$, was also investigated. In Figs 55 - 64, I_g and SILC curves are shown for the 3nm samples; In Figs. 65 – 82, I_g and SILC curves for the 2.2nm samples are shown.

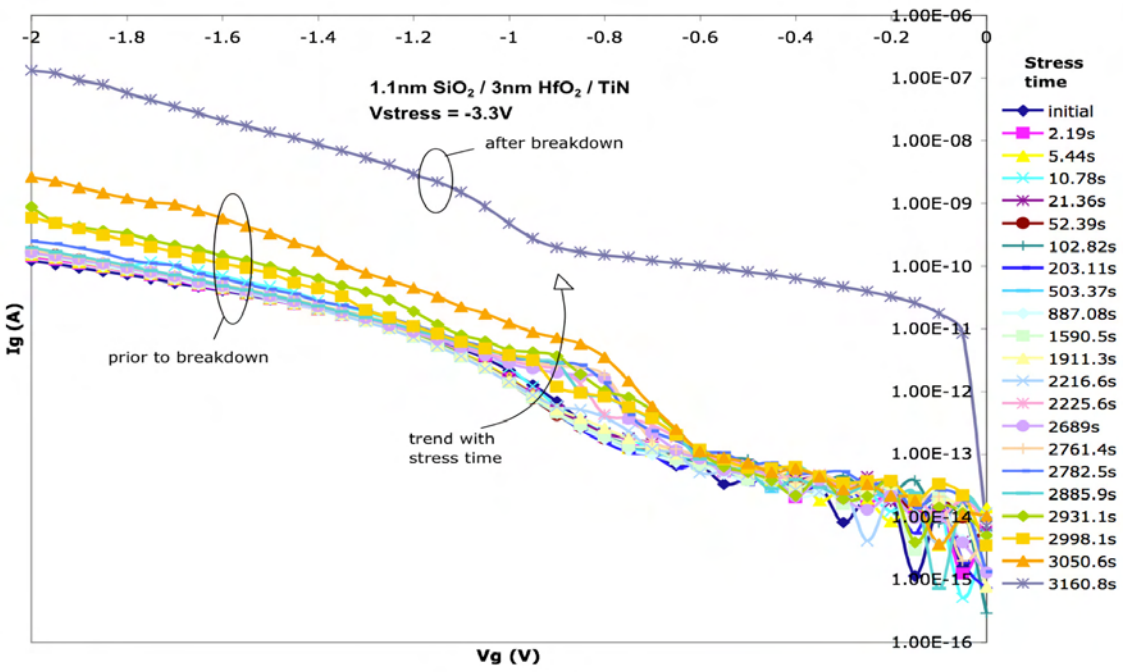

Figure 55: Time evolution of Ig - Vg curves for a 3nm sample at Vg = -3.3V

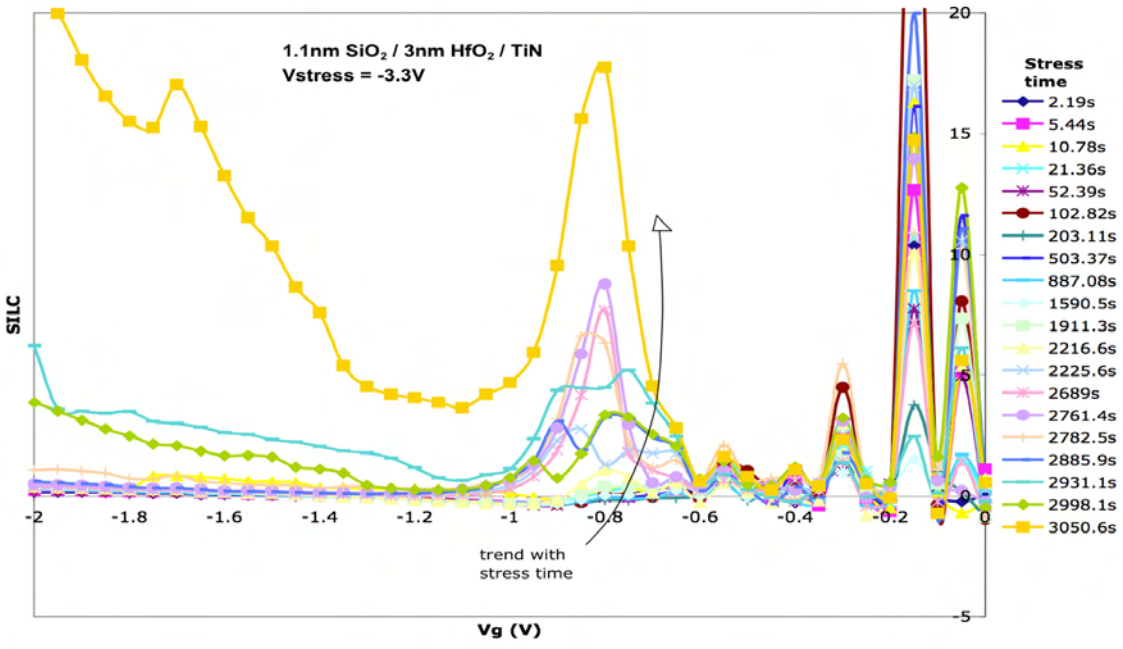

Figure 56: Time evolution of SILC curves for a 3nm sample at Vg = -3.3V

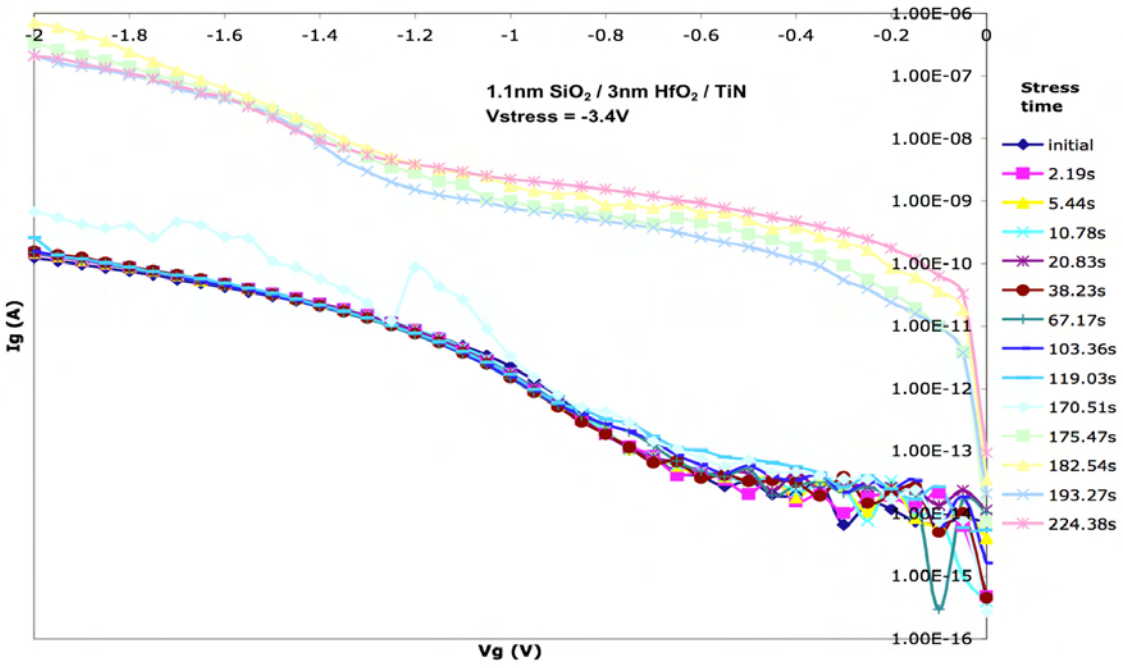

Figure 57: Time evolution of Ig - Vg curves for a 3nm sample at Vg = -3.4V

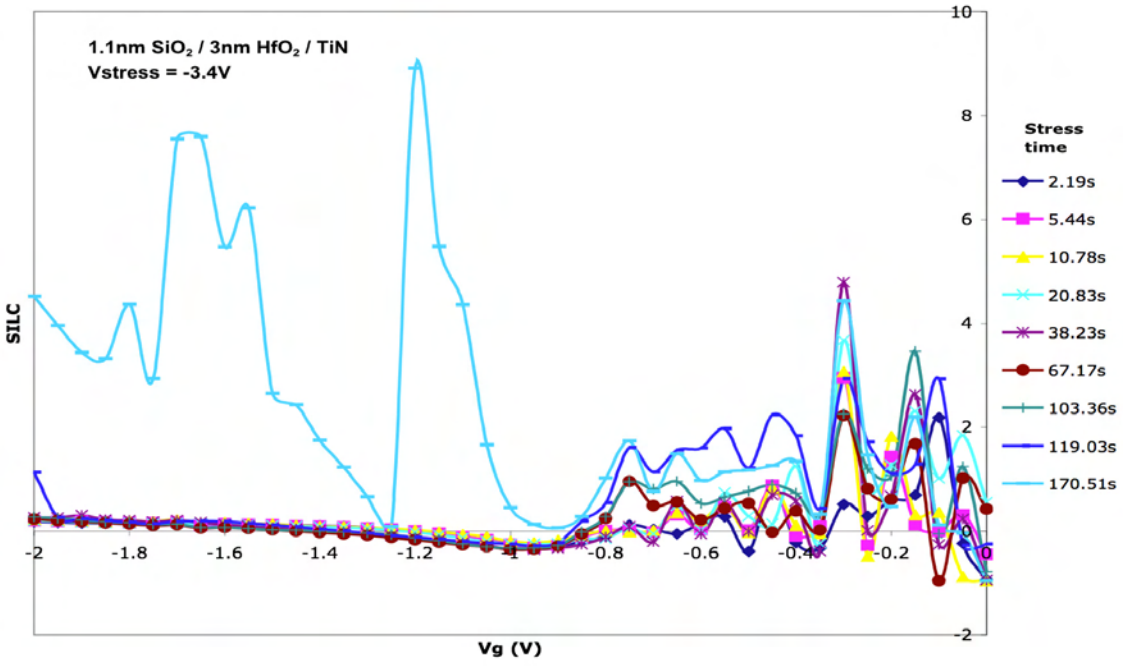

Figure 58: Time evolution of SILC curves for a 3nm sample at Vg = -3.4V

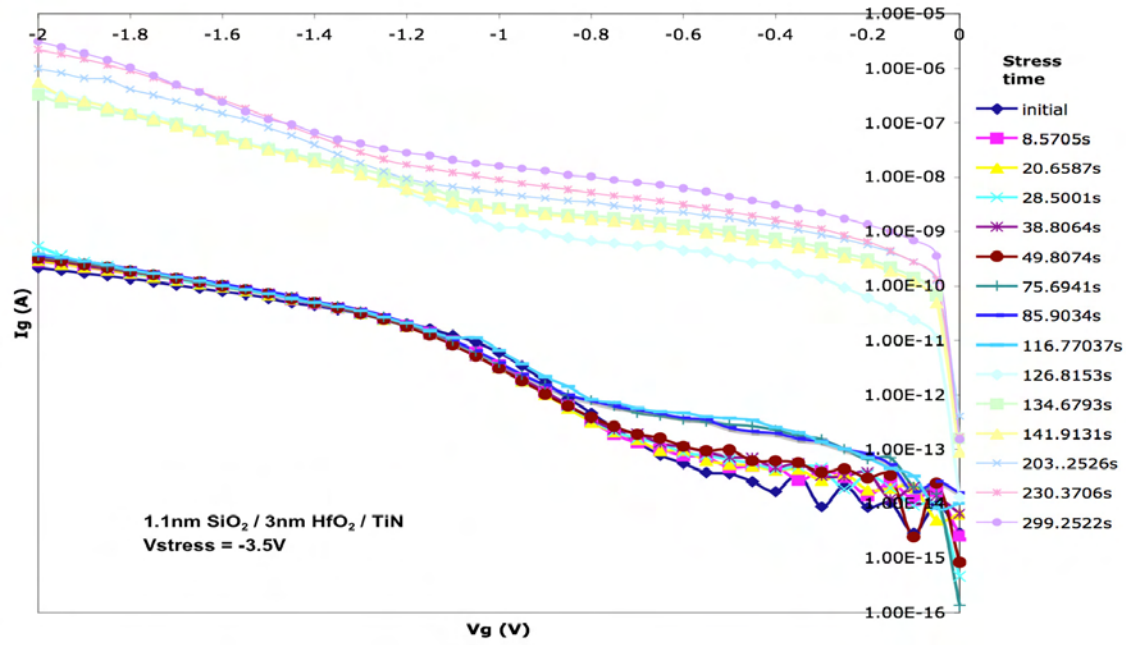

Figure 59: Time evolution of Ig – Vg curves for a 3nm sample at Vg = -3.5V

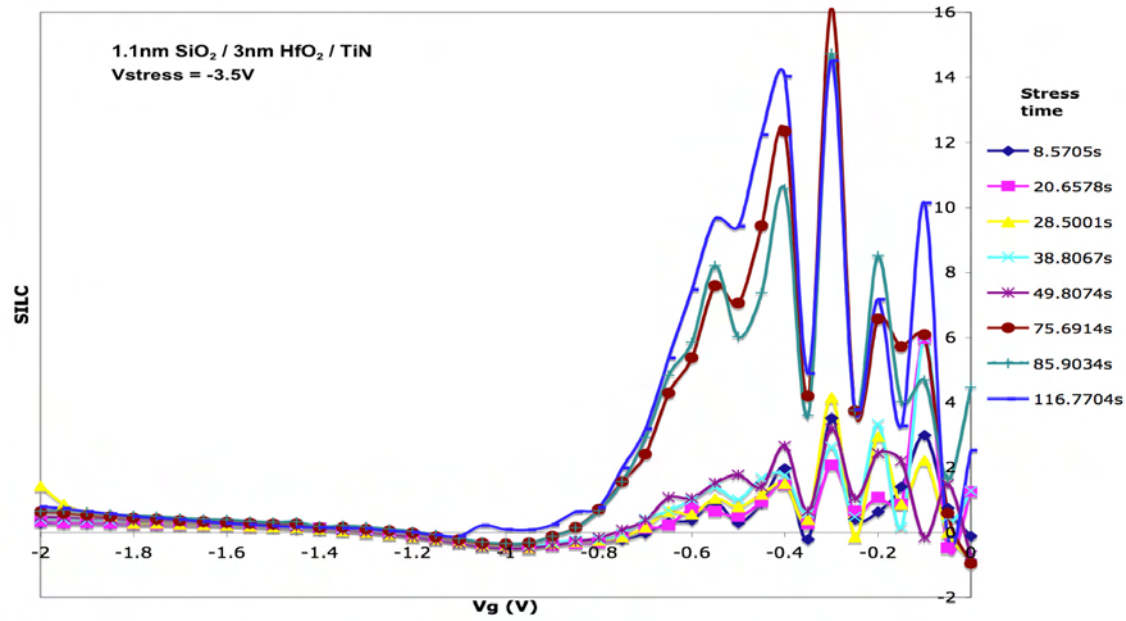

Figure 60: Time evolution of SILC curves for a 3nm sample at Vg = -3.5V

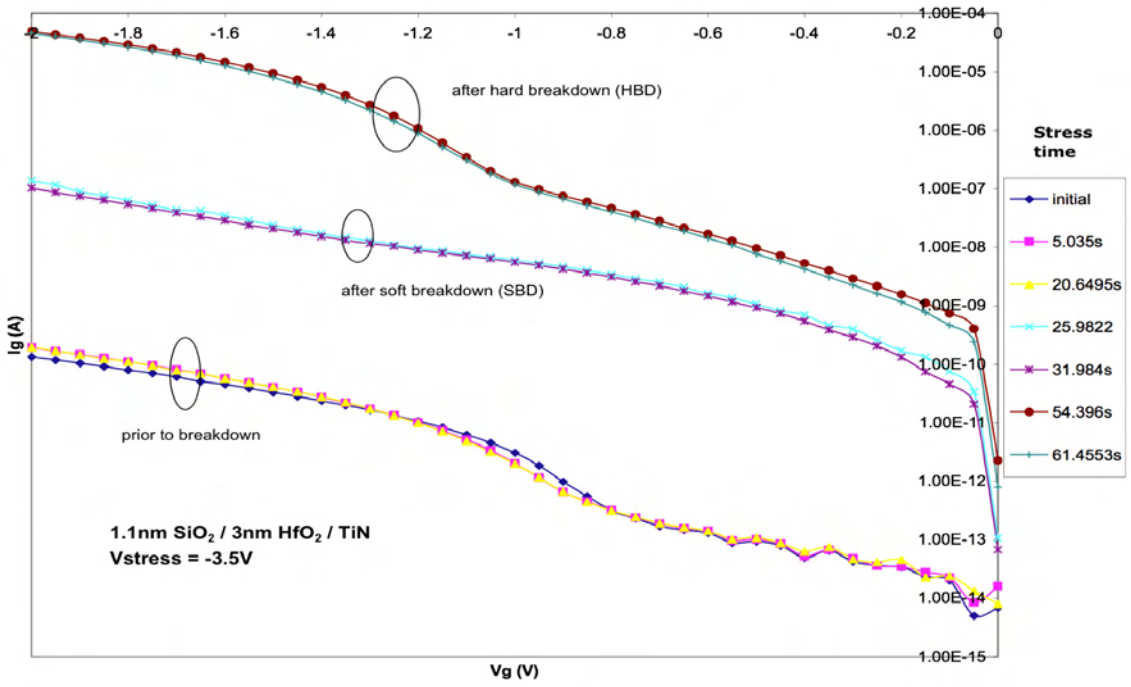

Figure 61: Time evolution of Ig – Vg curves for a 3nm sample at Vg = -3.5V

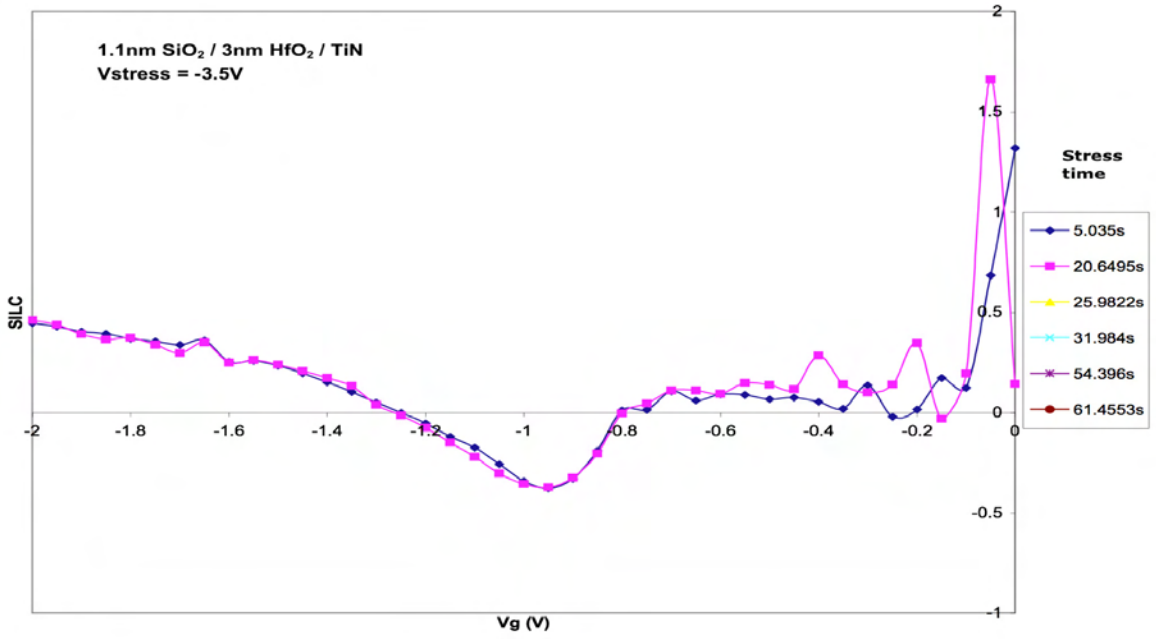

Figure 62: Time evolution of SILC curves for a 3nm sample at Vg = -3.5V

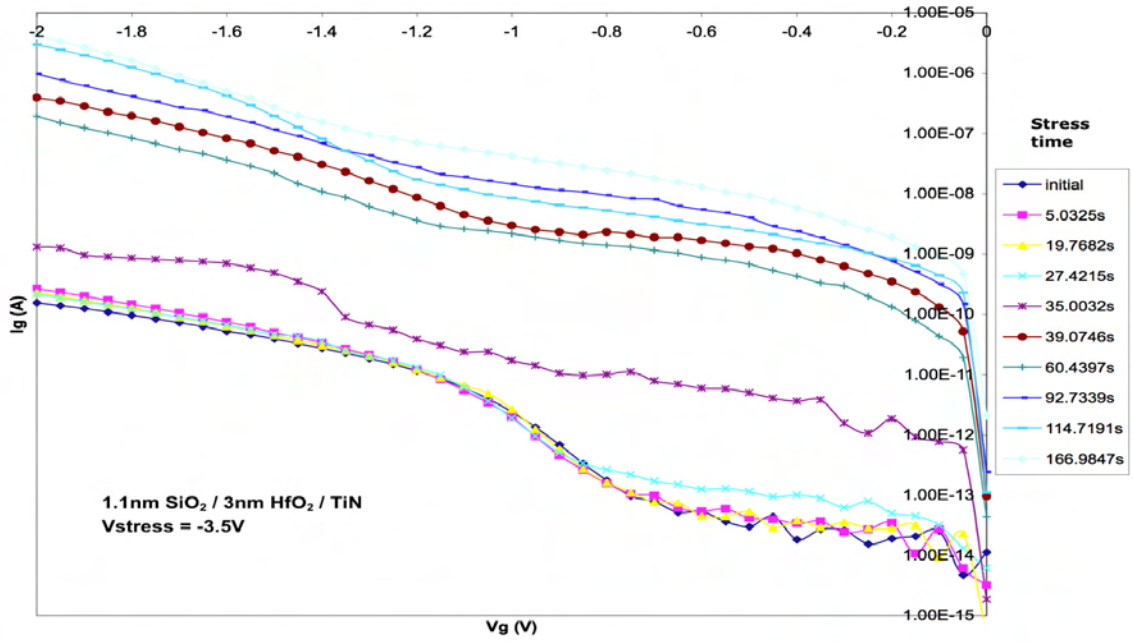

Figure 63: Time evolution of Ig – Vg curves for a 3nm sample at Vg = -3.5V

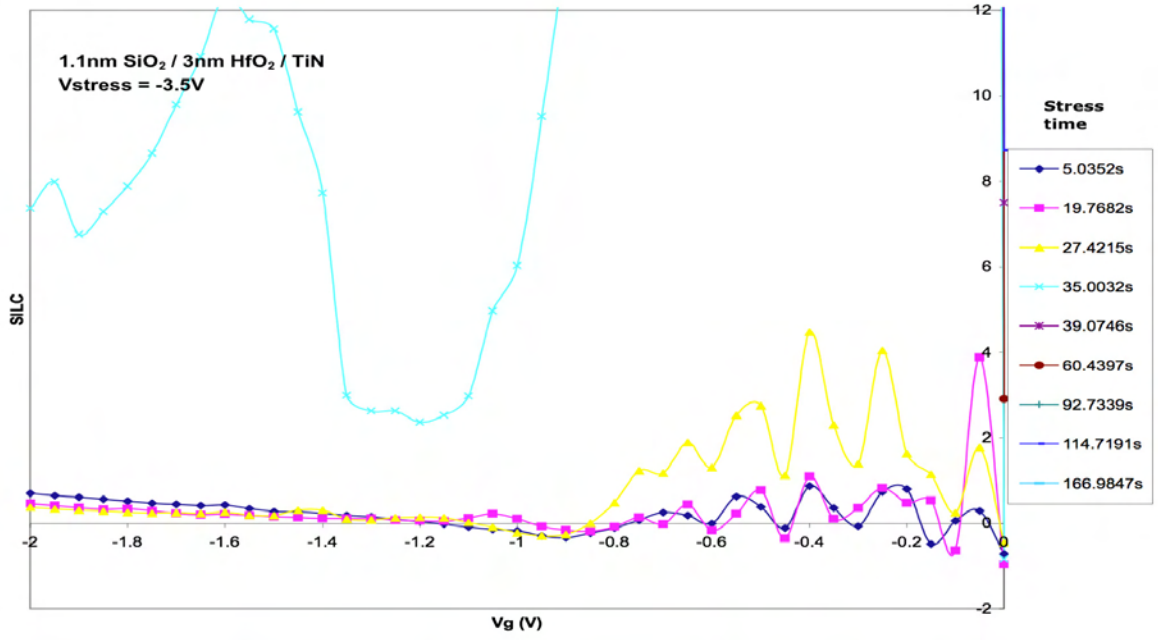

Figure 64: Time evolution of SILC curves for a 3nm sample at Vg = -3.5V

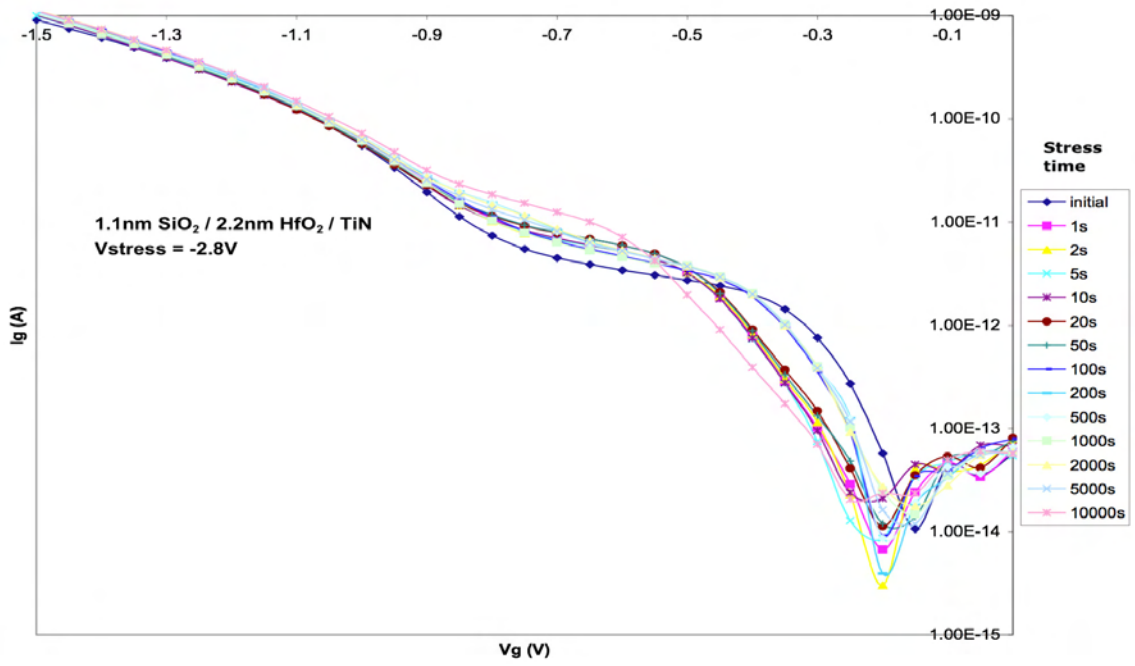

Figure 65: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -2.8V

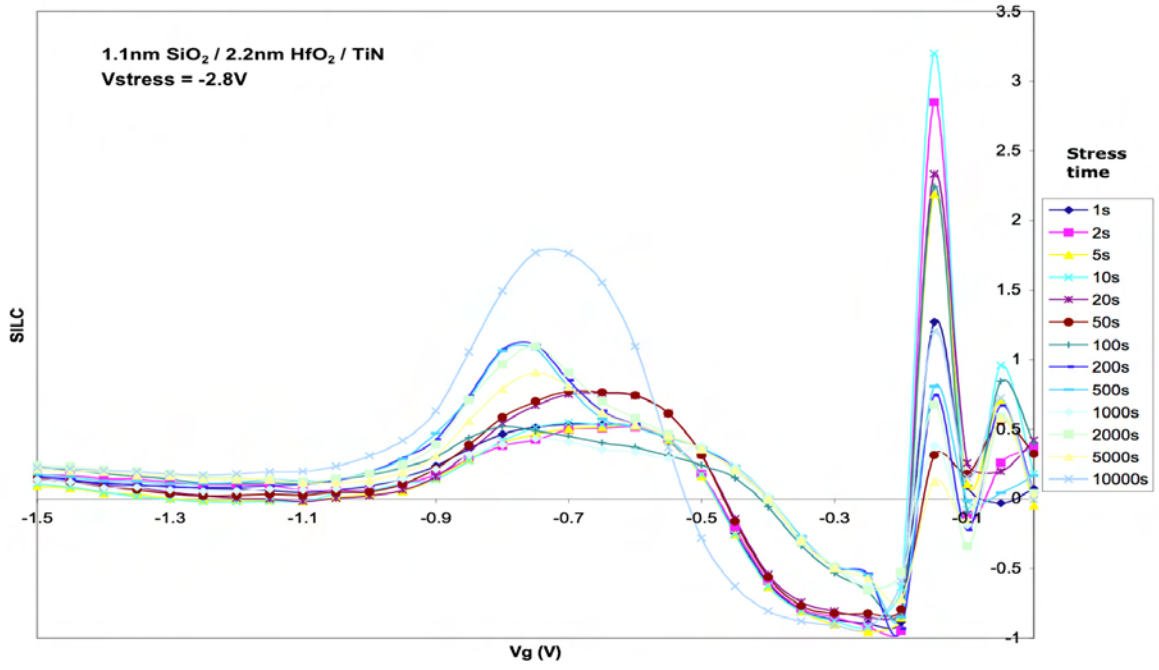

Figure 66: Time evolution of SILC curves for a 2.2nm sample at Vg = -2.8V

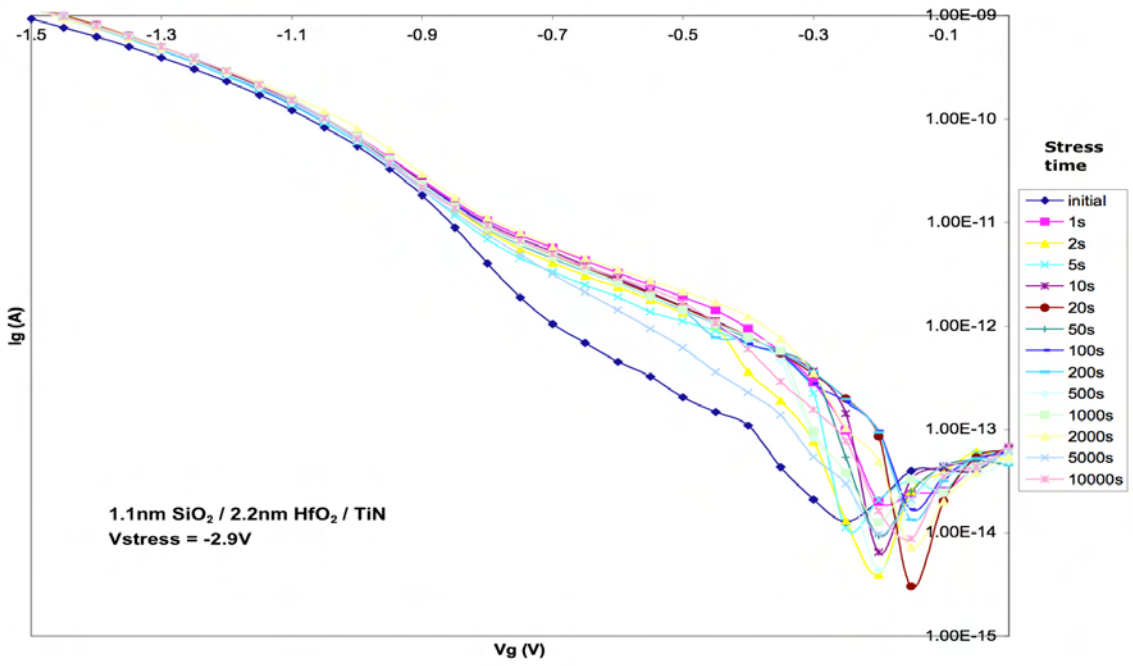

Figure 67: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -2.9V

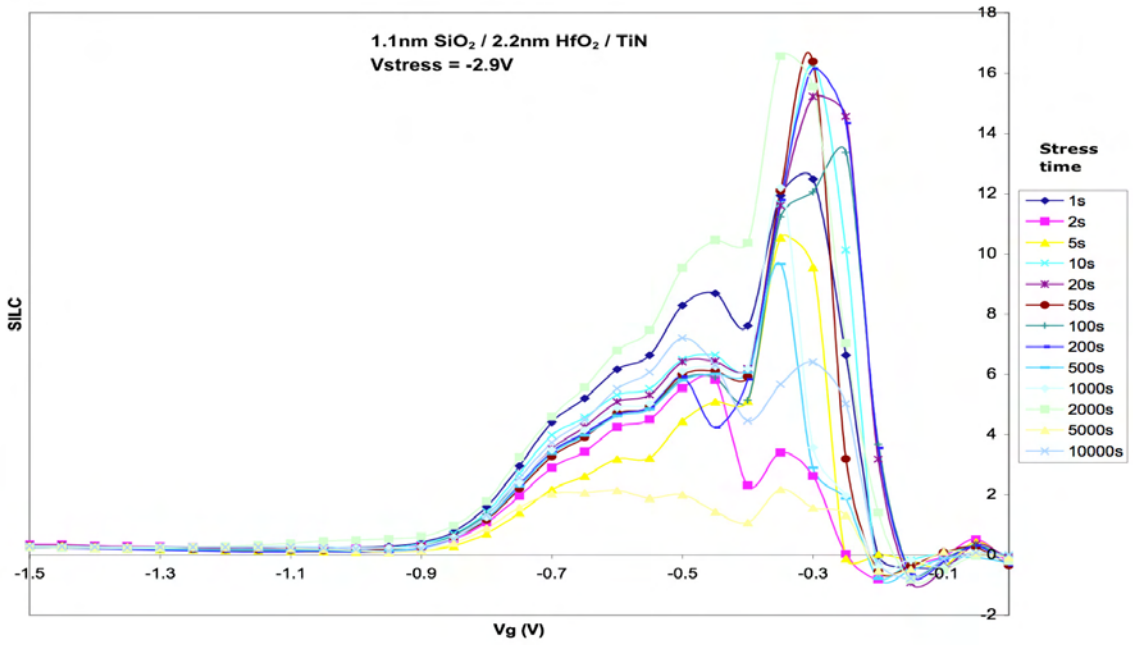

Figure 68: Time evolution of SILC curves for a 2.2nm sample at Vg = -2.9V

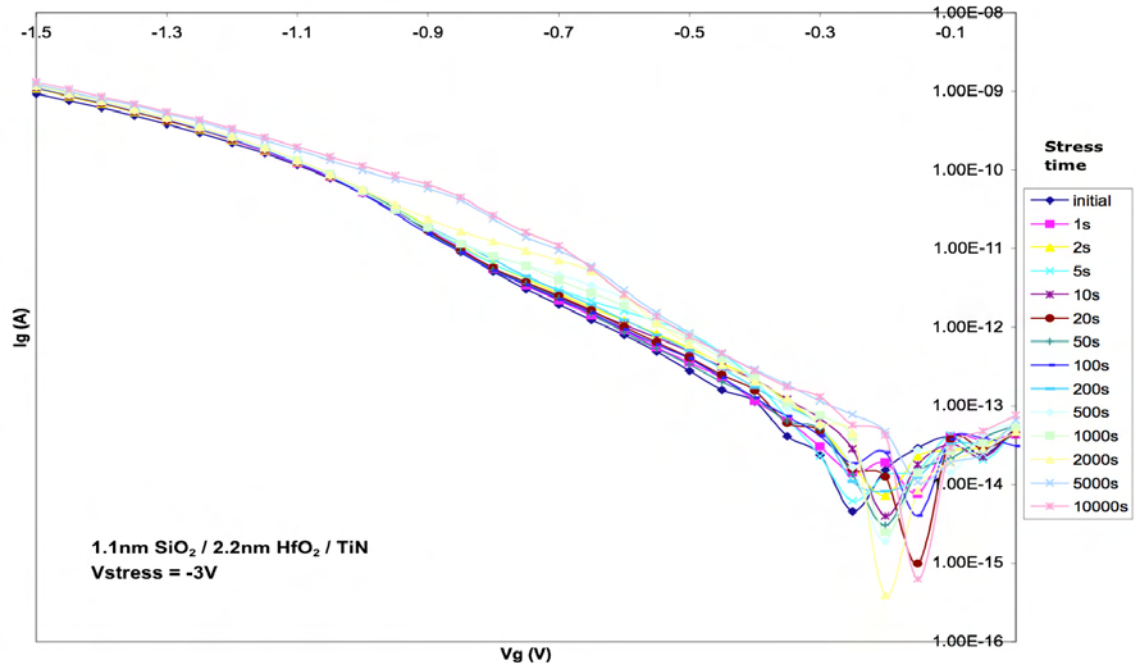

Figure 69: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3V

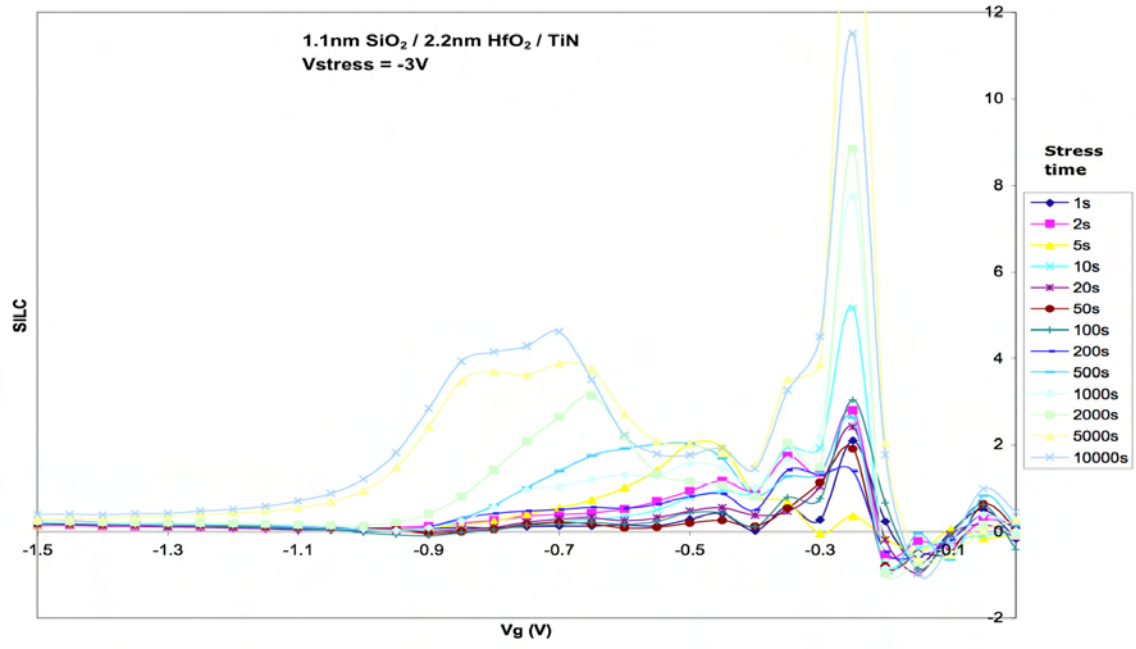

Figure 70: Time evolution of SILC curves for a 2.2nm sample at Vg = -3V

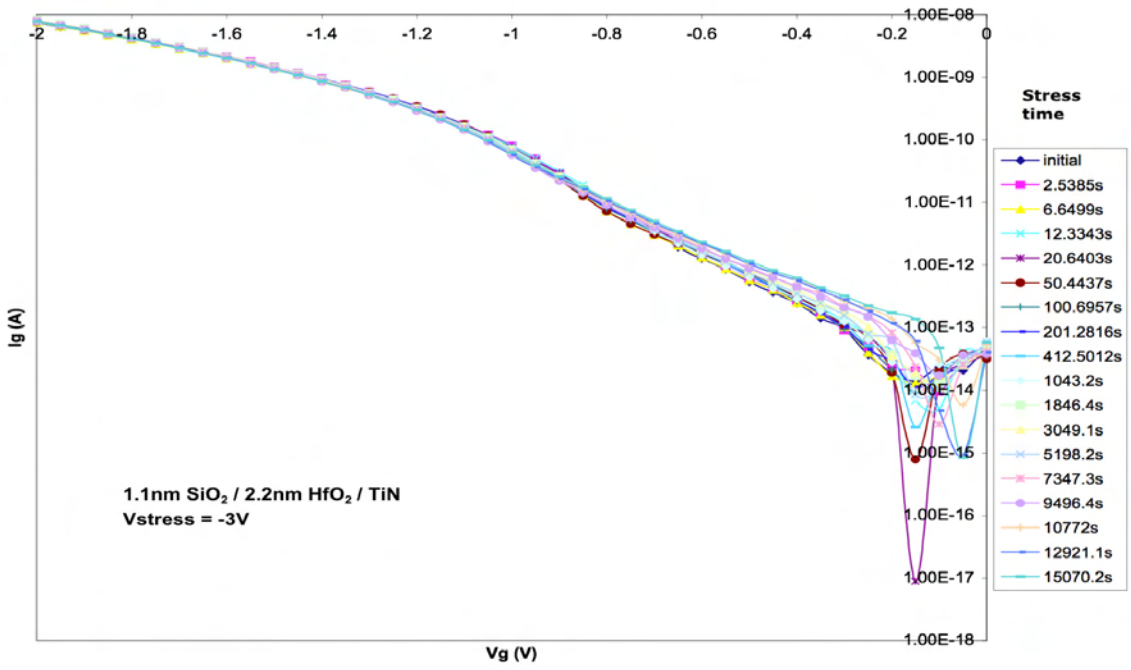

Figure 71: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3V

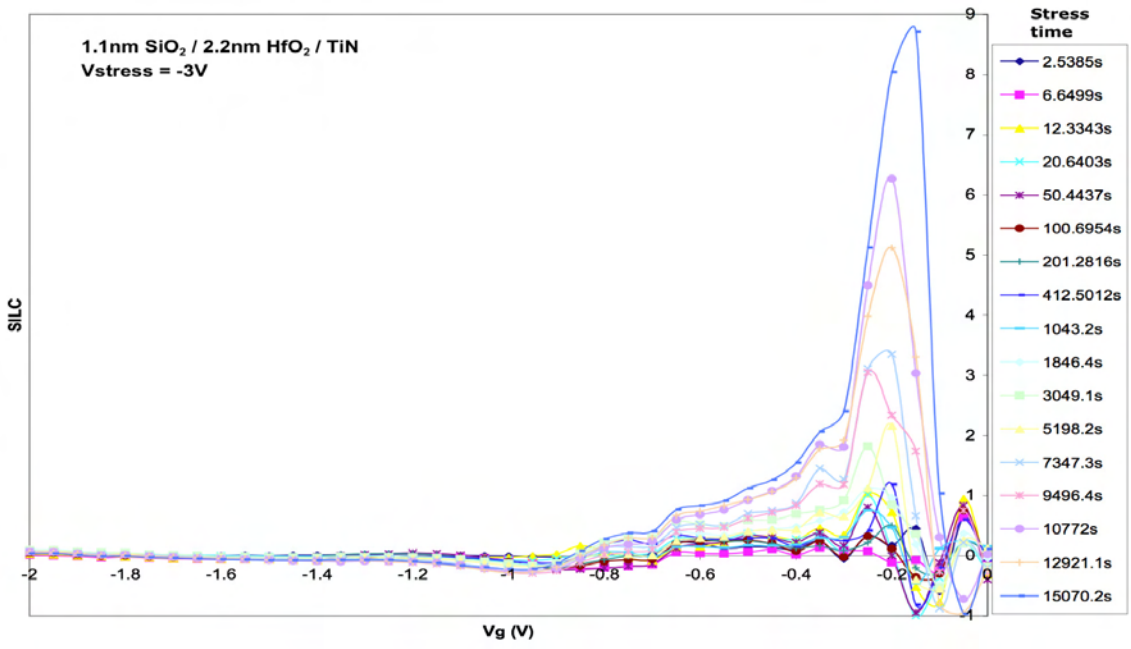

Figure 72: Time evolution of SILC curves for a 2.2nm sample at Vg = -3V

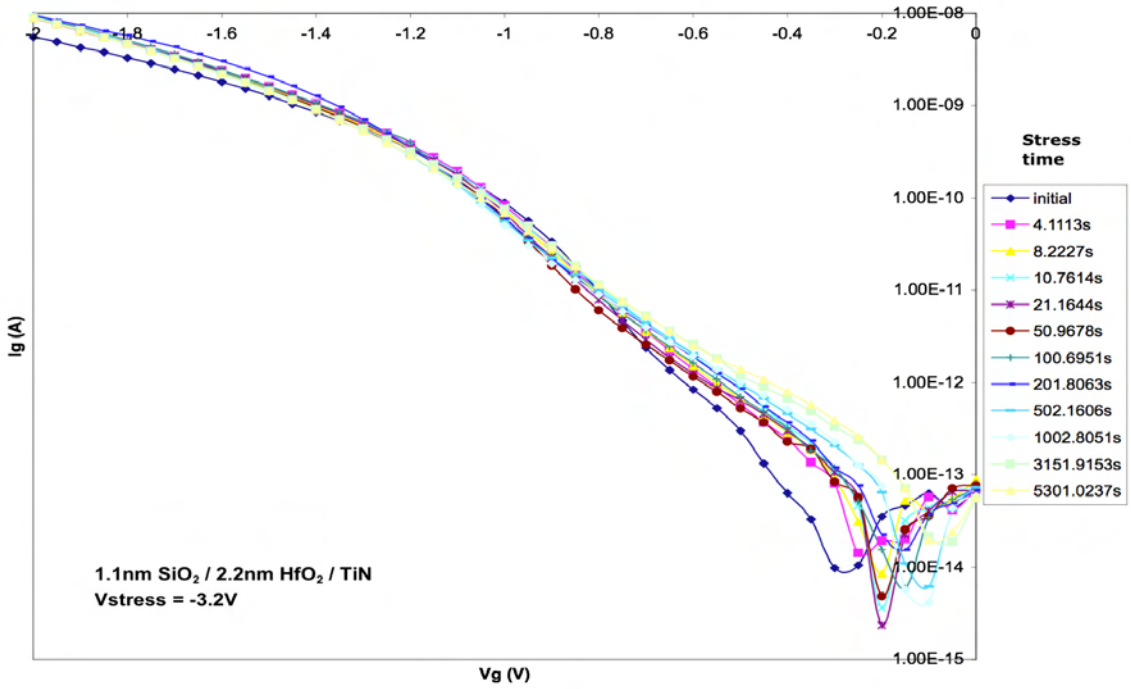

Figure 73: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3.2V

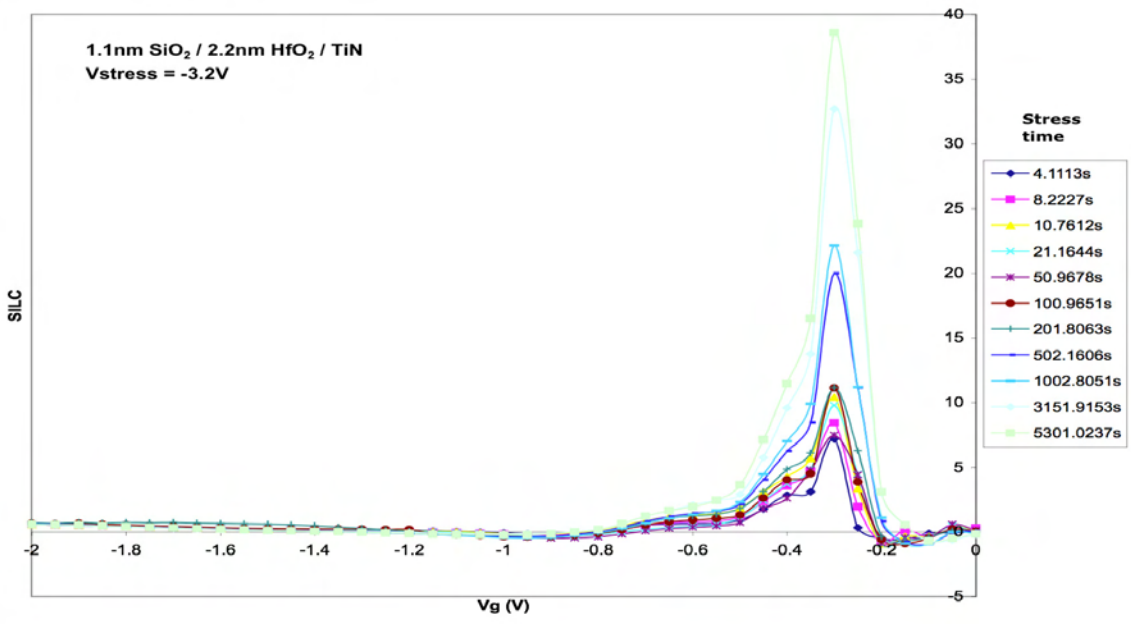

Figure 74: Time evolution of SILC curves for a 2.2nm sample at Vg = -3.2V

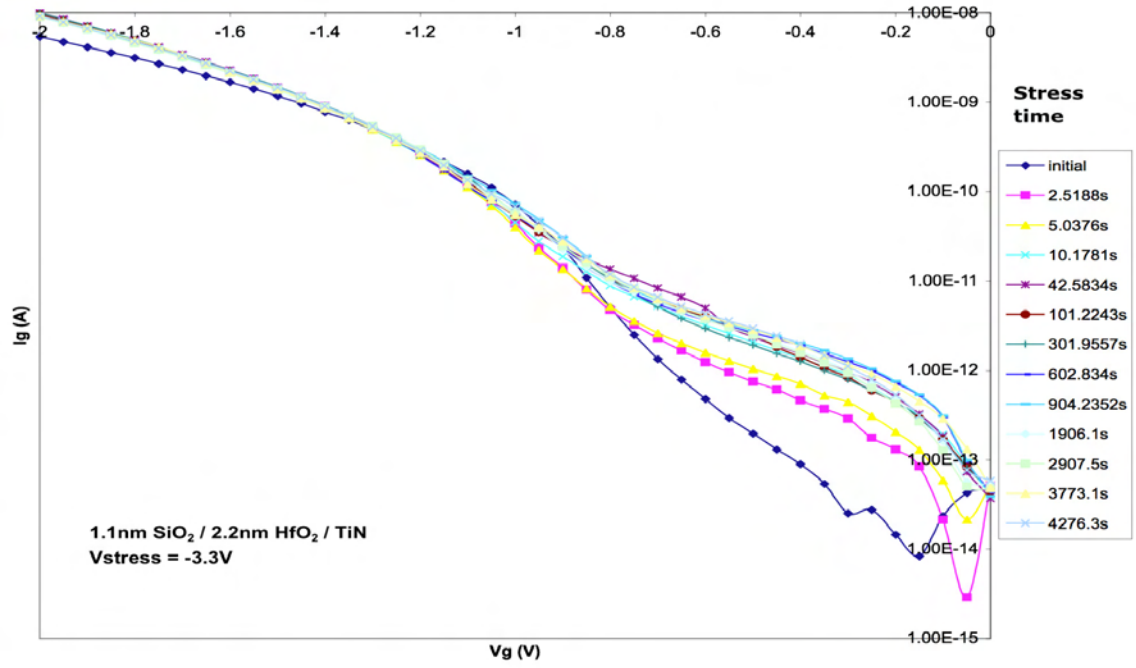

Figure 75: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3.3V

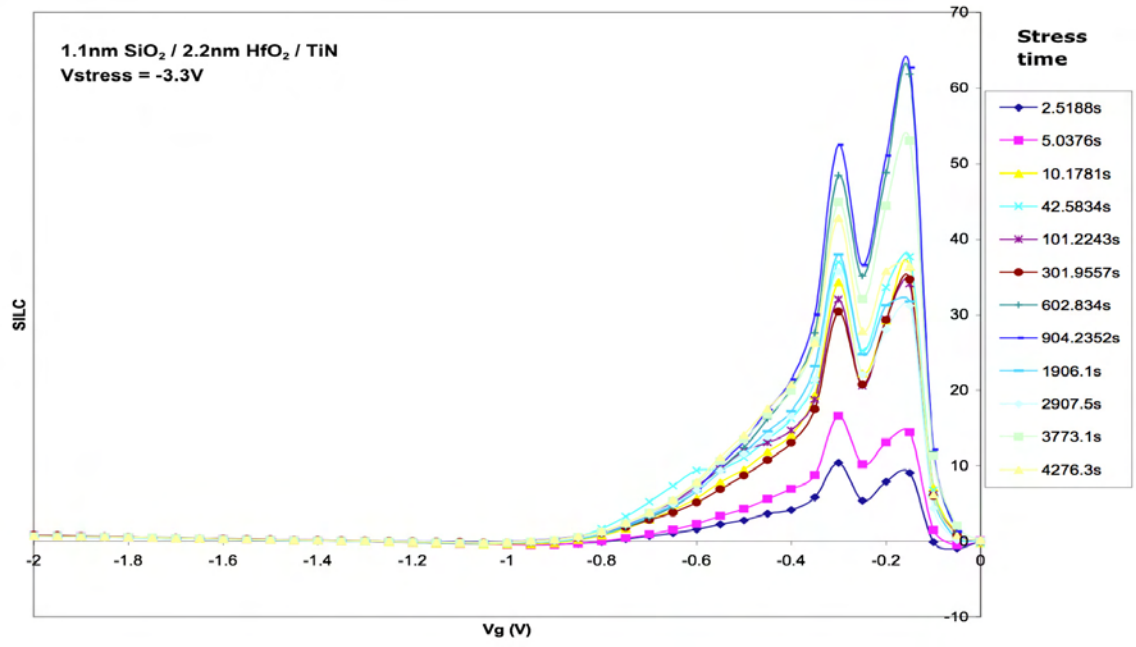

Figure 76: Time evolution of SILC curves for a 2.2nm sample at Vg = -3.3V

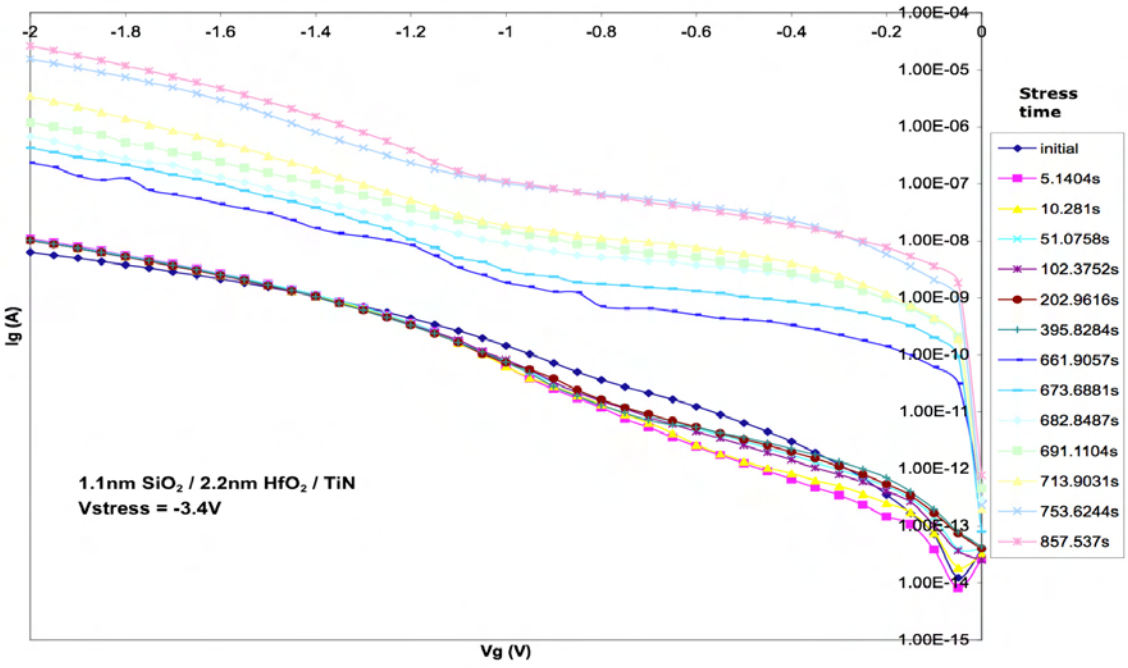

Figure 77: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3.4V

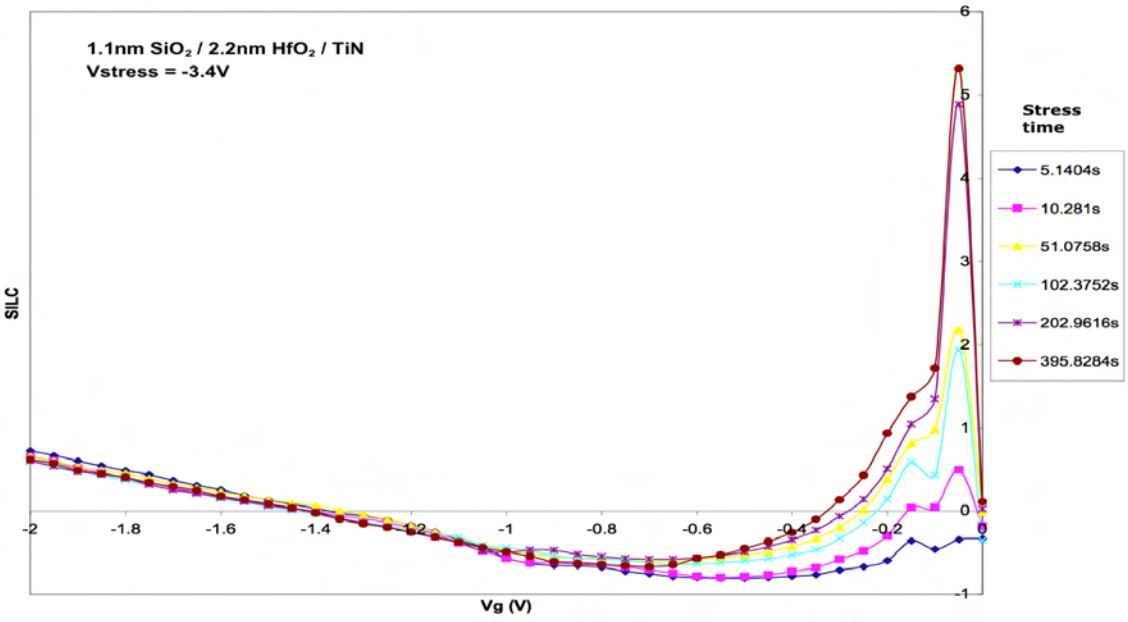

Figure 78: Time evolution of SILC curves for a 2.2nm sample at Vg = -3.4V

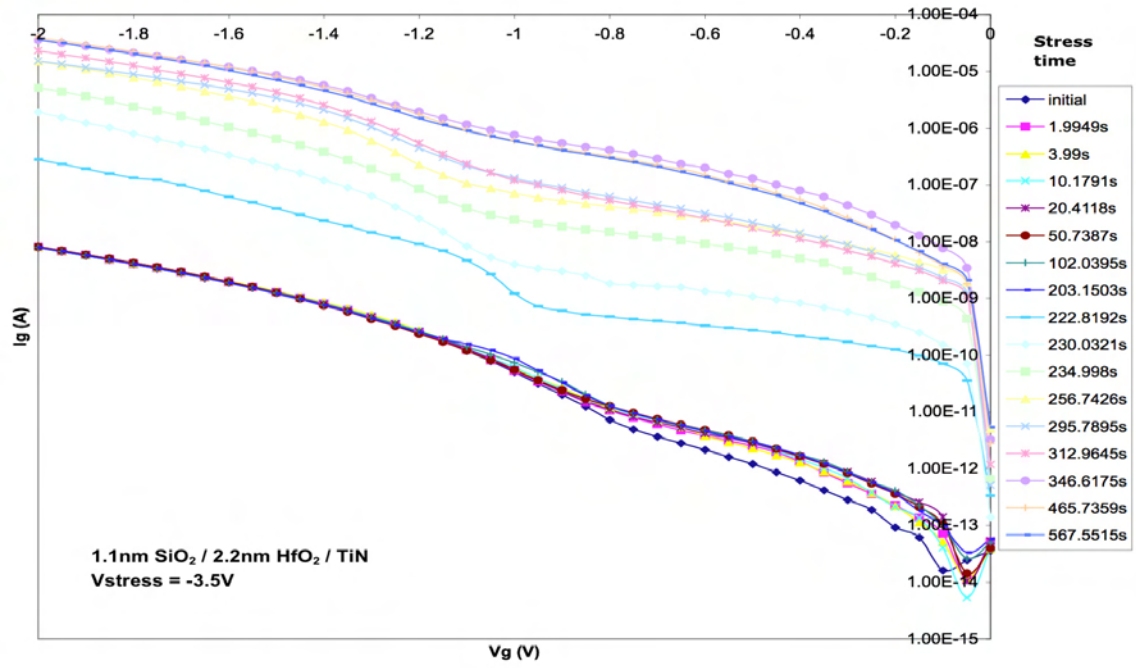

Figure 79: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3.5V

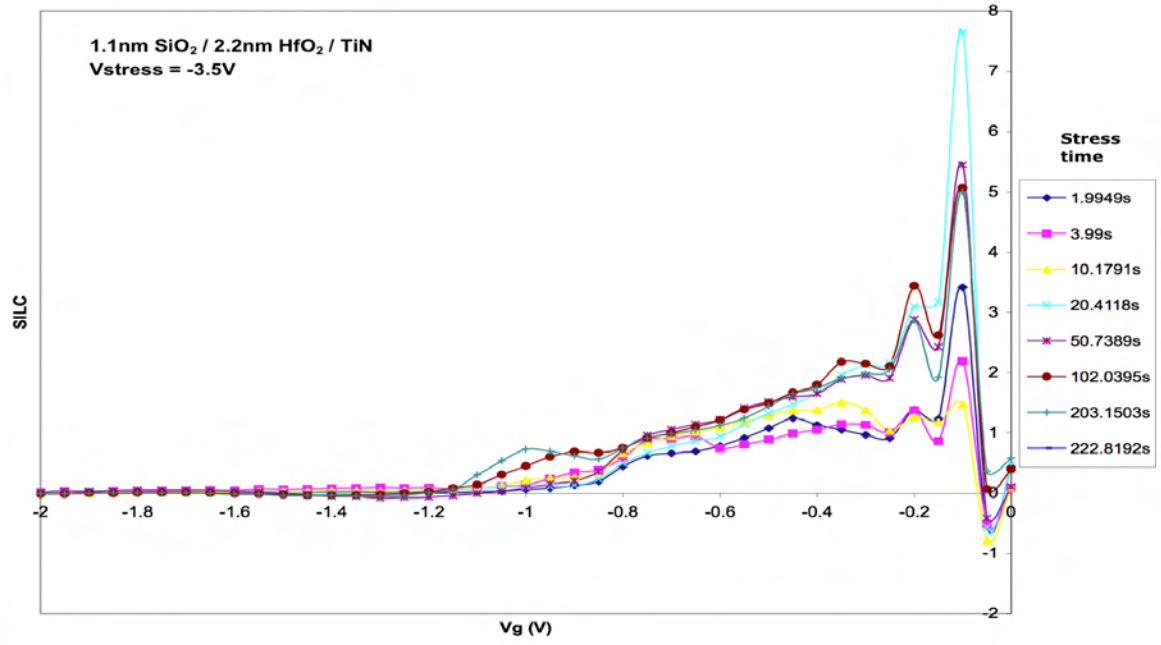

Figure 80: Time evolution of SILC curves for a 2.2nm sample at Vg = -3.5V

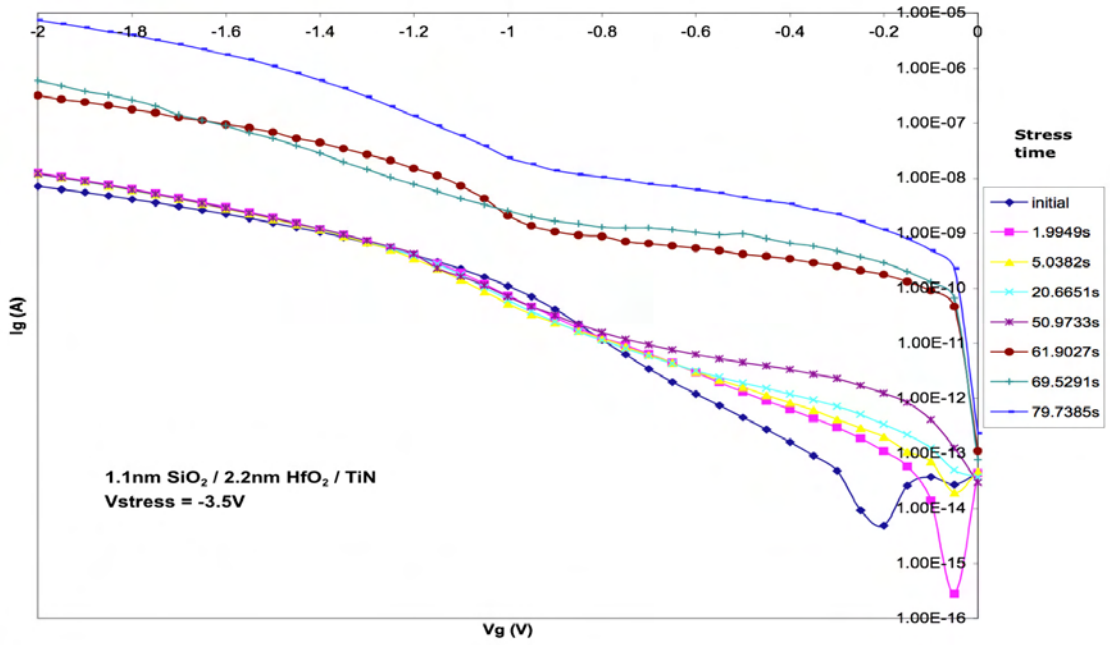

Figure 81: Time evolution of Ig – Vg curves for a 2.2nm sample at Vg = -3.5V

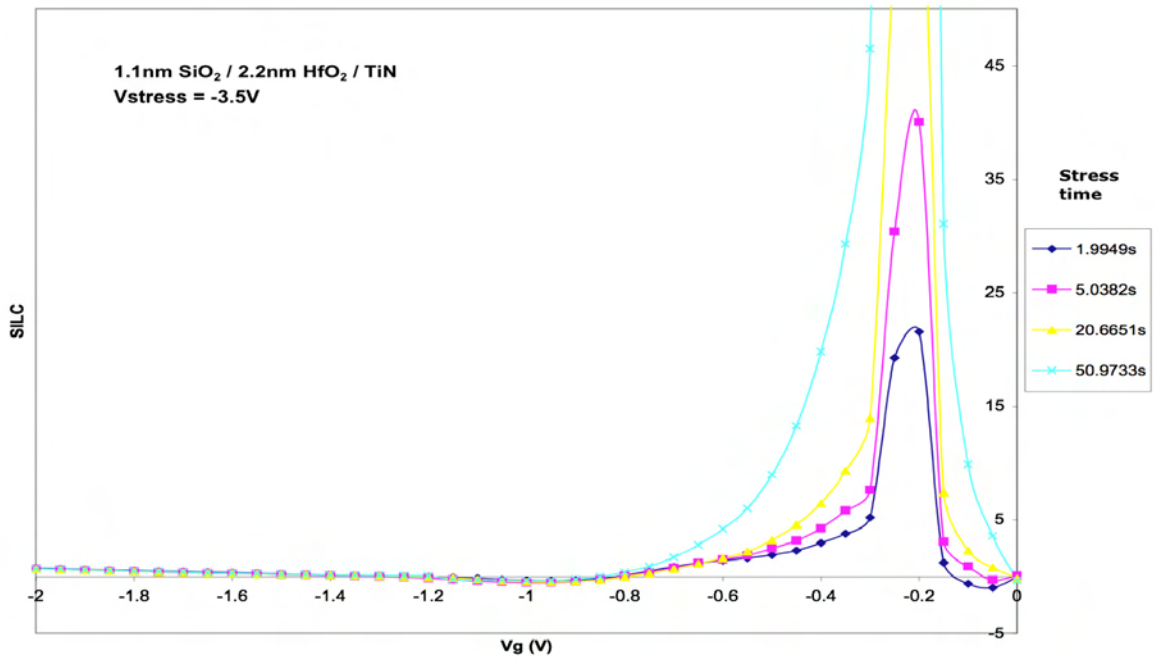

Figure 82: Time evolution of SILC curves for a 2.2nm sample at Vg = -3.5V

It's clear from the I_g-V_g graphs that the leakage current tends to increase with stress time. Once it reaches breakdown, there is a significant increase in the I_g curve. In cases with

soft and hard breakdown, an increase from the pre-breakdown family of curves to the soft breakdown curves is seen, with another increase to the hard breakdown point (see Figs. 61, 63, 77, 79, 81). The pre-breakdown gate leakage curves are a bit noisy in the region between 0 and -.6V due to the very low currents that are near the bottom of our instrument's detection range. This extra noise is responsible for the erratic spikes seen in the SILC-V_g graphs. There is a characteristic "two-bend" area in all of the leakage curves appearing in the region between -.8V and -1.2V, where the gate leakage increases as the voltage is swept below ~ -1V. By constructing an approximate band diagram, it becomes clear that this is the region where the conduction band of the gate electrode lines up with the top of the potential barrier in the conduction band of the Si substrate. This suggests that the increase along the curve is due to more electrons tunneling from the electrode to the substrate, rather than an increase in hole tunneling through the gate.

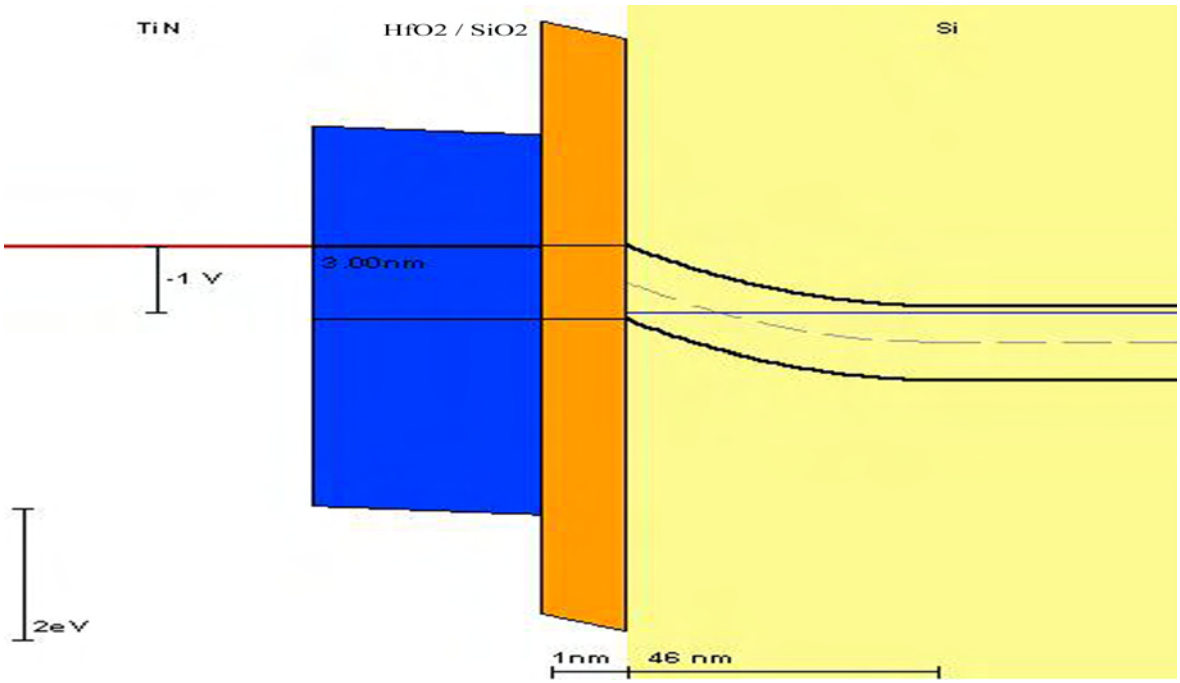

Figure 83: Band diagram for 3nm gate stack at Vg = -1V

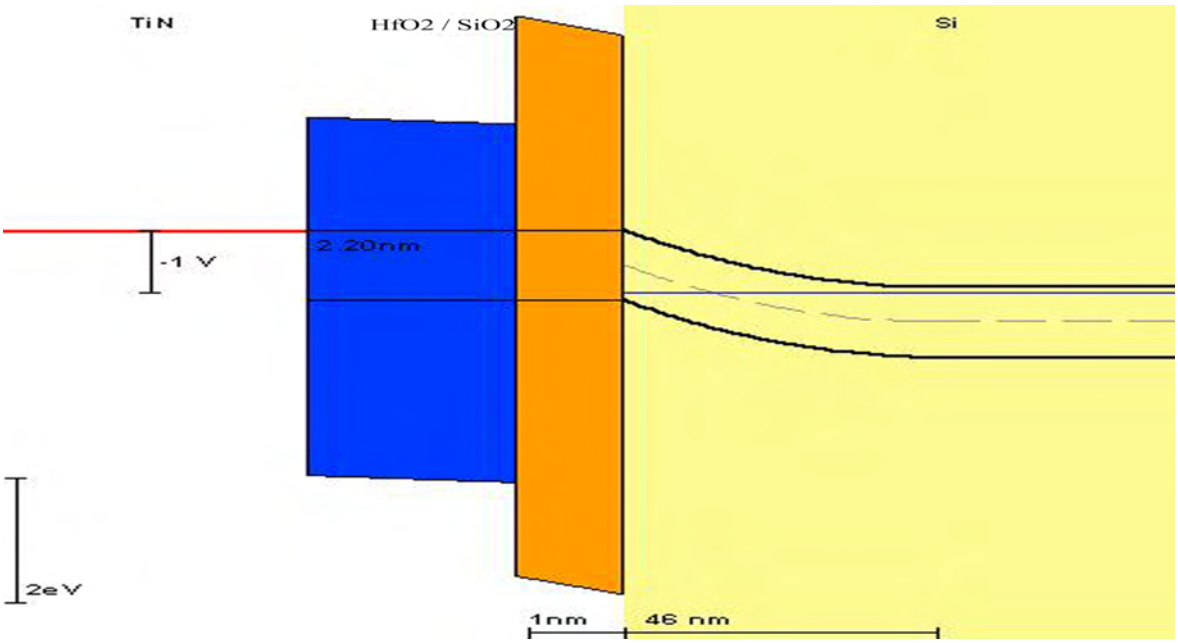

Figure 84: Band diagram for 2.2nm gate stack at Vg = -1V

The increase along the curve does not happen exactly at -1V in each case due to effects from the trapped charges along the interfaces, but the band diagrams support the idea that this region is where electrons at the electrode start to have a higher probability of tunneling through the gate.

The relative change of I_g from it's initial value (SILC) curves turned out to be less useful in understanding the breakdown of the devices. Part of the problem was due to the large amount of noise seen in the I_g-V_g curves between 0 and -.6V. This noise shows up as very sharp peaks that could skew one side of any meaningful peaks at a slightly higher voltage bias. Another problem encountered was the fact that the PMOS devices did not always give a consistent initial I_g-V_g curve. The I_g-V_g curves for two devices without any stress applied are shown in Figs. 85 & 86. There is quite a bit of variance in the gate leakage current at the smaller gate biases. As shown in Fig. 86, the leakage current nearly doubling in the region from 0 to -.4V bias from one sweep to the next, and then

falls back down on the next voltage sweep. Because of these discrepancies, any of the calculated SILC data between 0 and -.6V is questionable.

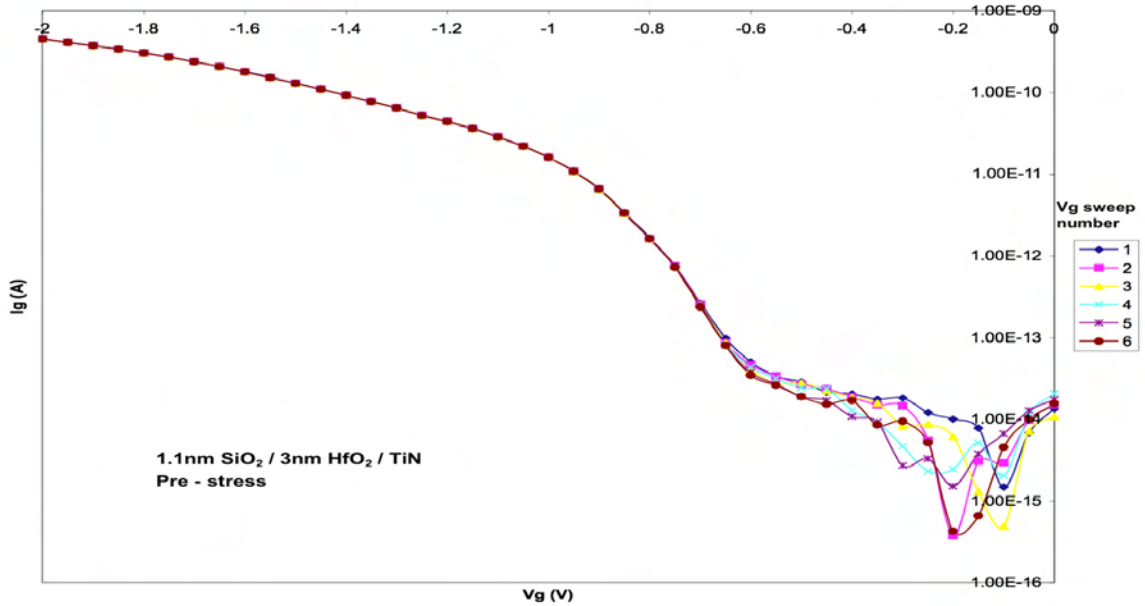

Figure 85: Pre-stress Ig – Vg curves for a 3nm sample

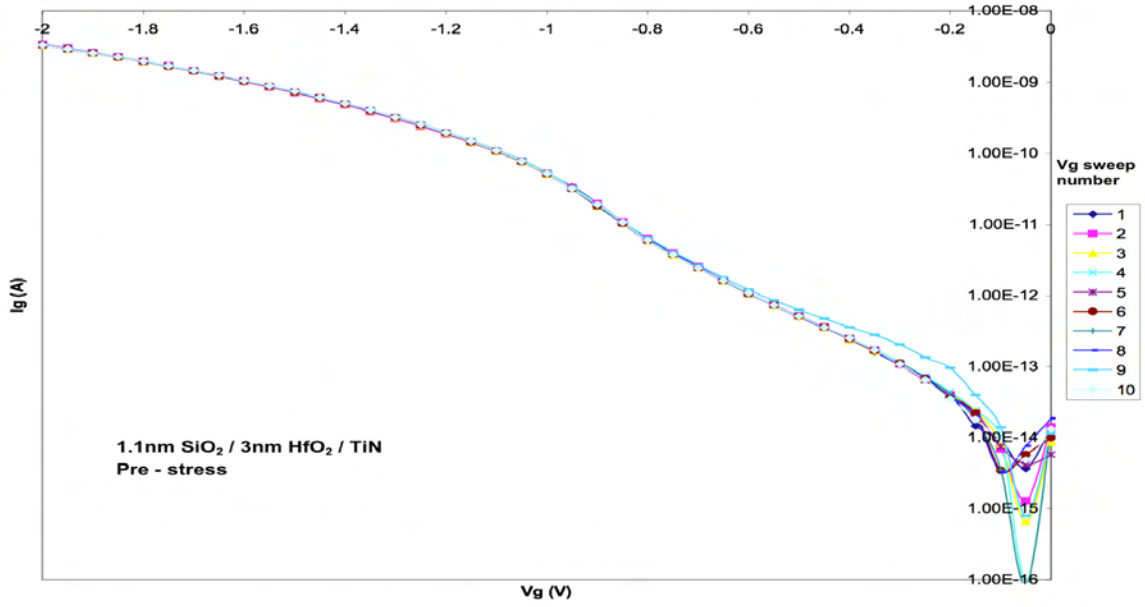

Figure 86: Pre-stress Ig – Vg curves for a 3nm sample

Some of the SILC curves did show time-progressive peaks at a certain gate bias, yet other showed none or even a negative peak, indicating that leakage at that gate bias decreased from it's initial value with stress. Because of these difficulties, the SILC data did not yield much useful information about the breakdown of the devices, other than a general trend for the leakage current to increase along most of the voltage bias range. By taking the SILC data at any apparent peak, or as near as could be found, a correlation with the trap density and the relative change of the trap density (ΔN_{it}) at high and low frequencies was sought out. However, this did not yield any apparent connection or reliance on stress voltage for either sample (see examples in Fig. 87 - 90).

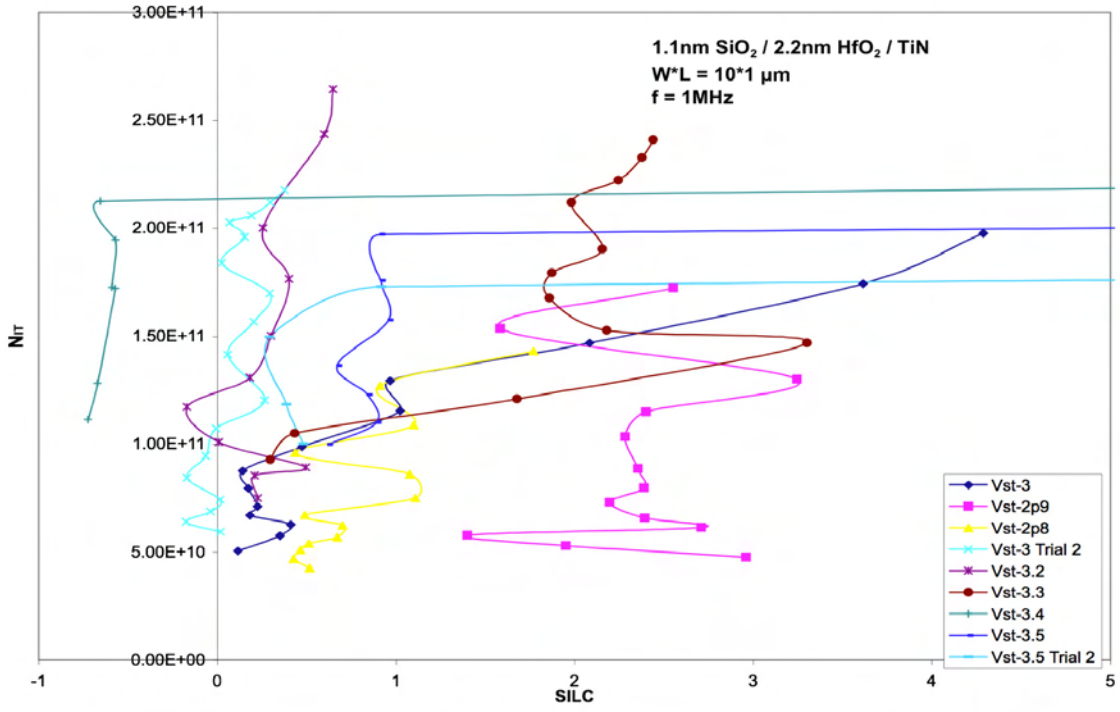

Figure 87: 1MHz Nit – SILC for 2.2nm samples

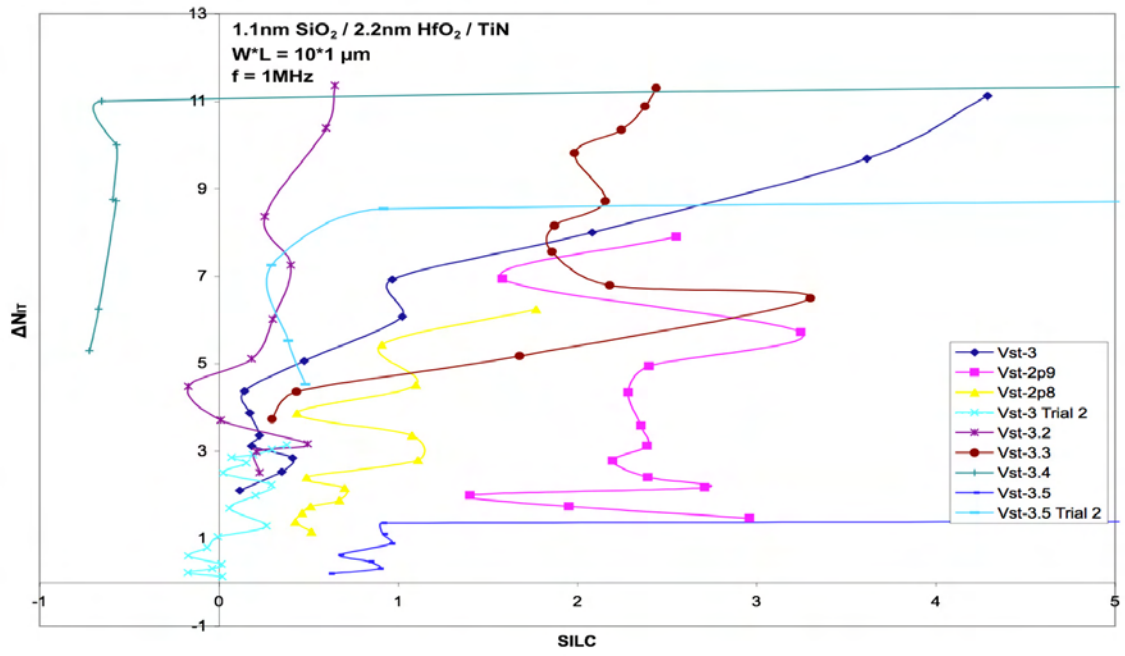

Figure 88: 1MHz relative ΔNit – SILC for 2.2nm samples

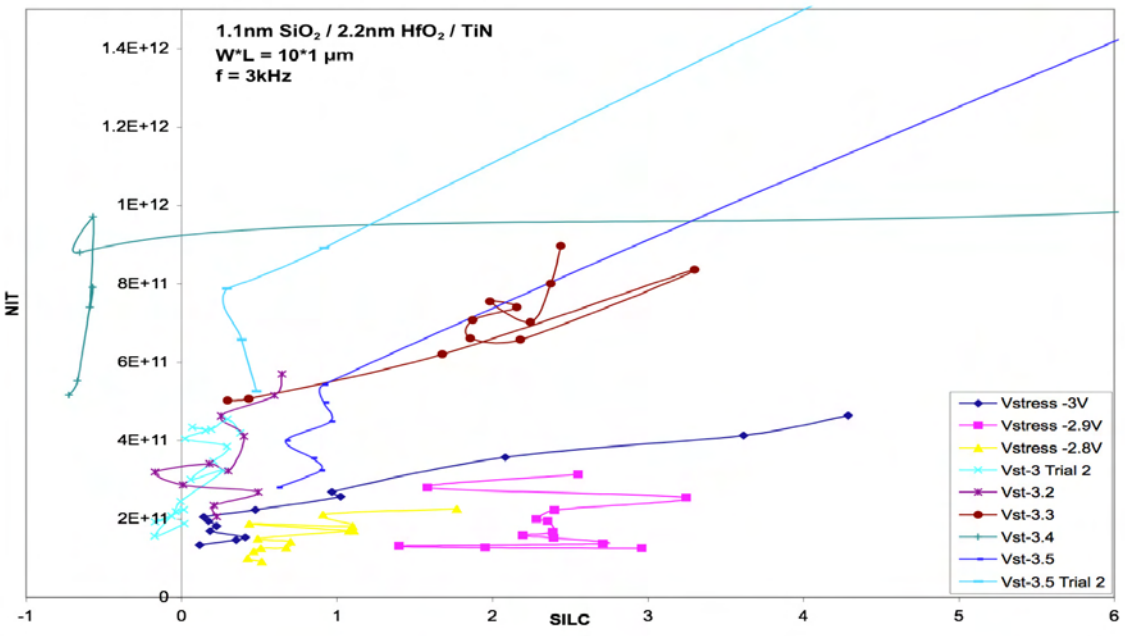

Figure 89: 3kHz Nit – SILC for 2.2nm samples

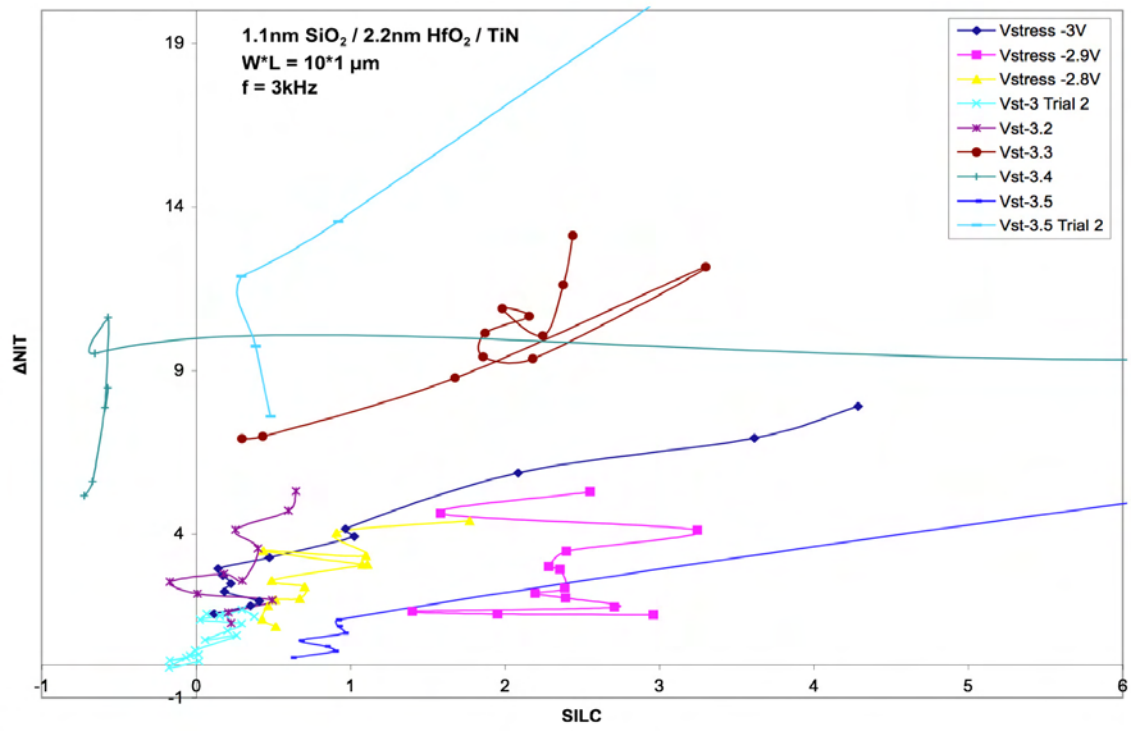

Figure 90: 3kHz relative ΔNit – SILC for 2.2nm samples

Due to the procedure of the SILC measurements, it is possible to measure the current on all four terminals; thus, the carrier separation technique could be applied to the data. The carrier separation technique showed a number of things consistently throughout both samples. First, prior to breakdown, the source/drain current dominated the bulk current, indicating that more holes are being lost to leakage than electrons being injected from the gate electrode. This is consistent with what was seen during the carrier separation of the CVS measurements. Second, after breakdown, the bulk current dominated the source drain current, indicating that once a breakdown path is established, more electrons flow from the gate electrode to the substrate than holes from the substrate to the gate. However, during the voltage sweep, it is seen that after breakdown the source/drain current has a dramatic increase in the region between -.8V and -1.2V. As

was shown earlier, this is the region where electrons become more likely to tunnel from the gate electrode to the substrate. Once the source/drain current becomes comparable to the bulk current, the bulk current, which signifies the number of electrons tunneling from the electrode to the substrate, falls off a small amount and no longer follows as closely to the total gate leakage current (Figs. 91 – 95). Since the source/drain current is actually a measurement of the number of holes which leave one side of the device and don't make it to the other, this drastic increase in source/drain current when coupled with the decrease in the bulk current indicates that there is a larger number of holes and electrons recombining. The increase in hole/electron recombination is because the electrons at the gate electrode are no longer constrained to the breakdown pathway(s) as their most likely way to reach the substrate; this means that a larger number of electrons may tunnel through the gate at multiple spatial points, as opposed to the few points available via the breakdown path(s). This increases the likelihood that a hole and electron will come close enough to recombine.

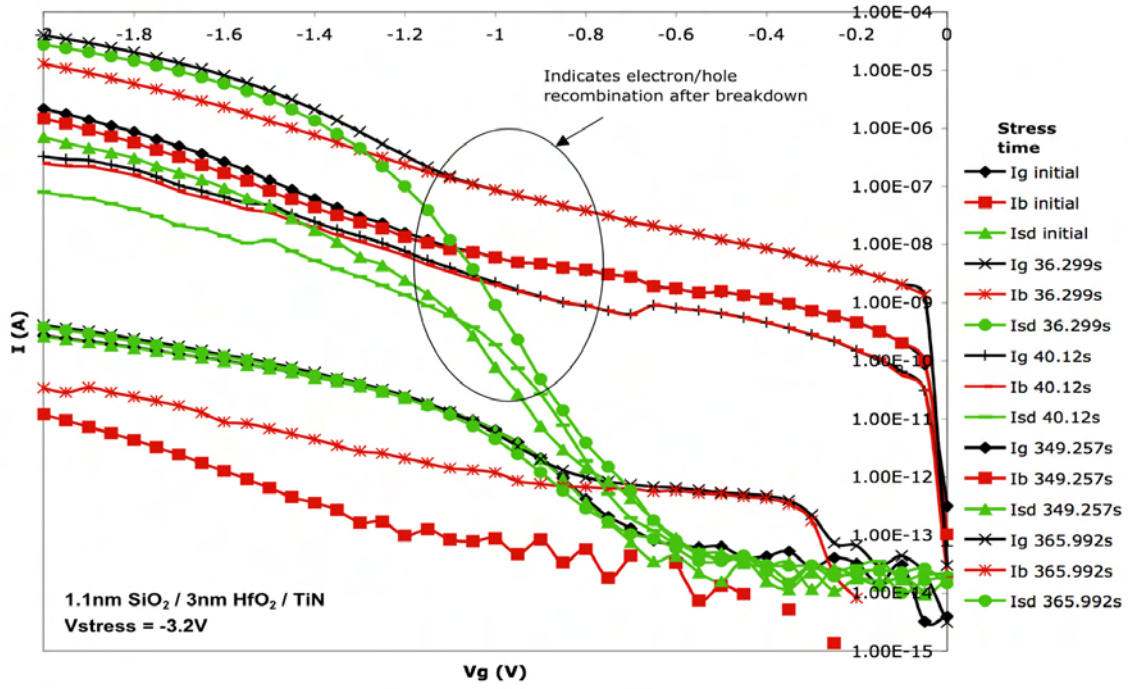

Figure 91: Time evolution of selected carrier separation I – Vg curves for a 3nm sample at Vg = -3.2V

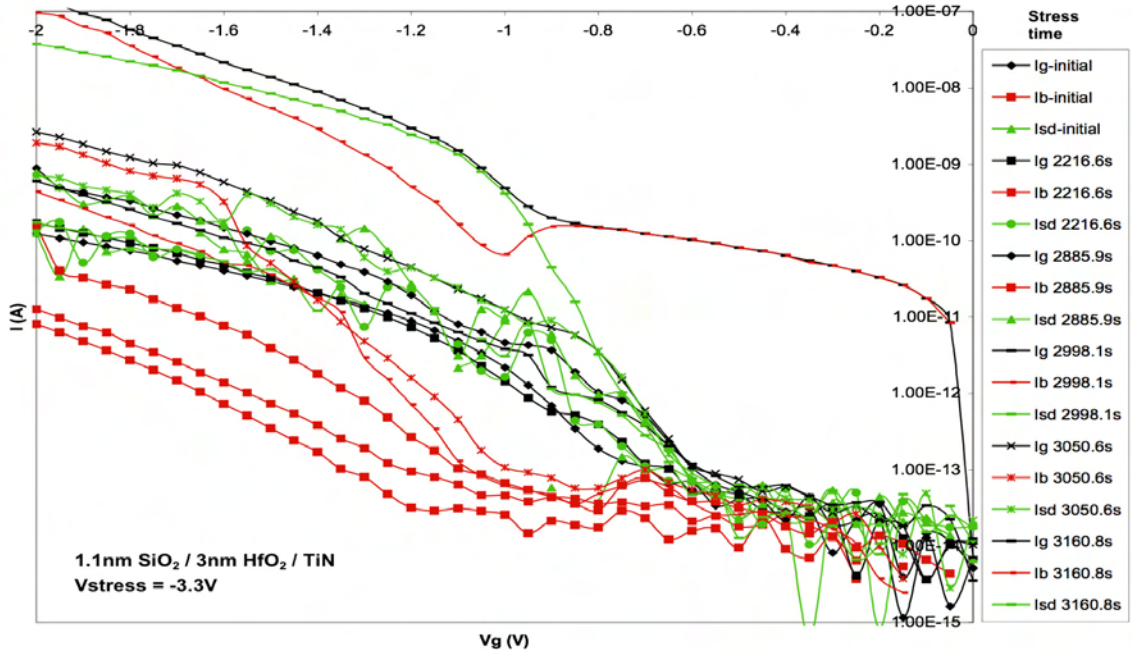

Figure 92: Time evolution of selected carrier separation I – Vg curves for a 3nm sample at Vg = -3.3V

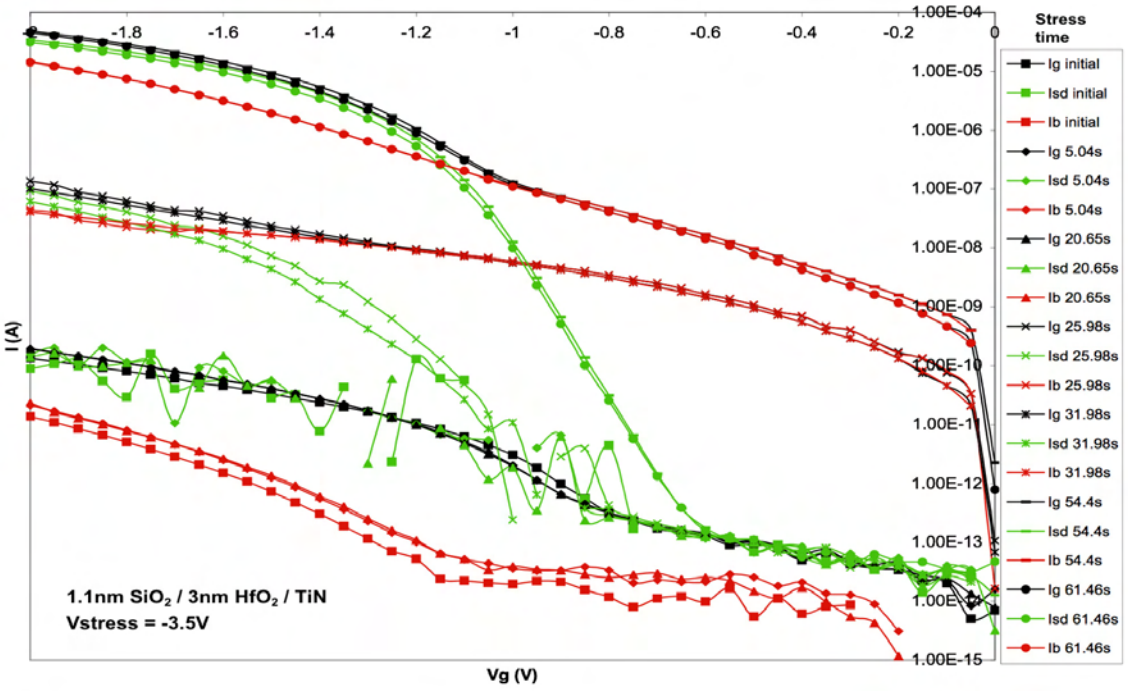

Figure 93: Time evolution of selected carrier separation I – Vg curves for a 3nm sample at Vg = -3.5V

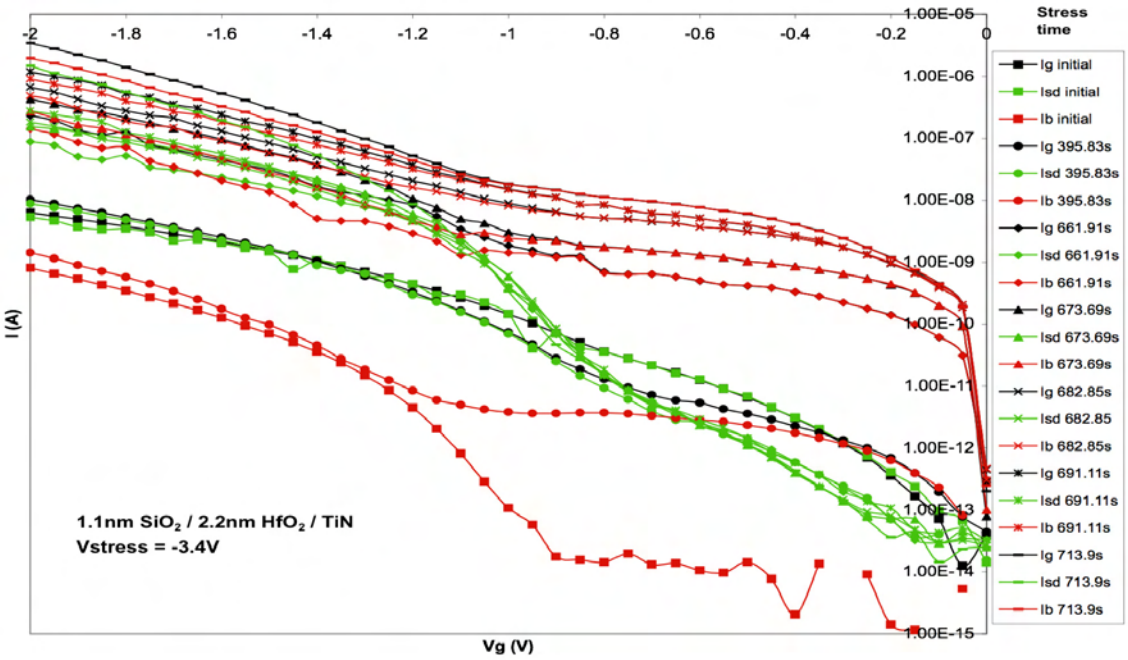

Figure 94: Time evolution of selected carrier separation I – Vg curves for a 2.2nm sample at Vg = -3.4V

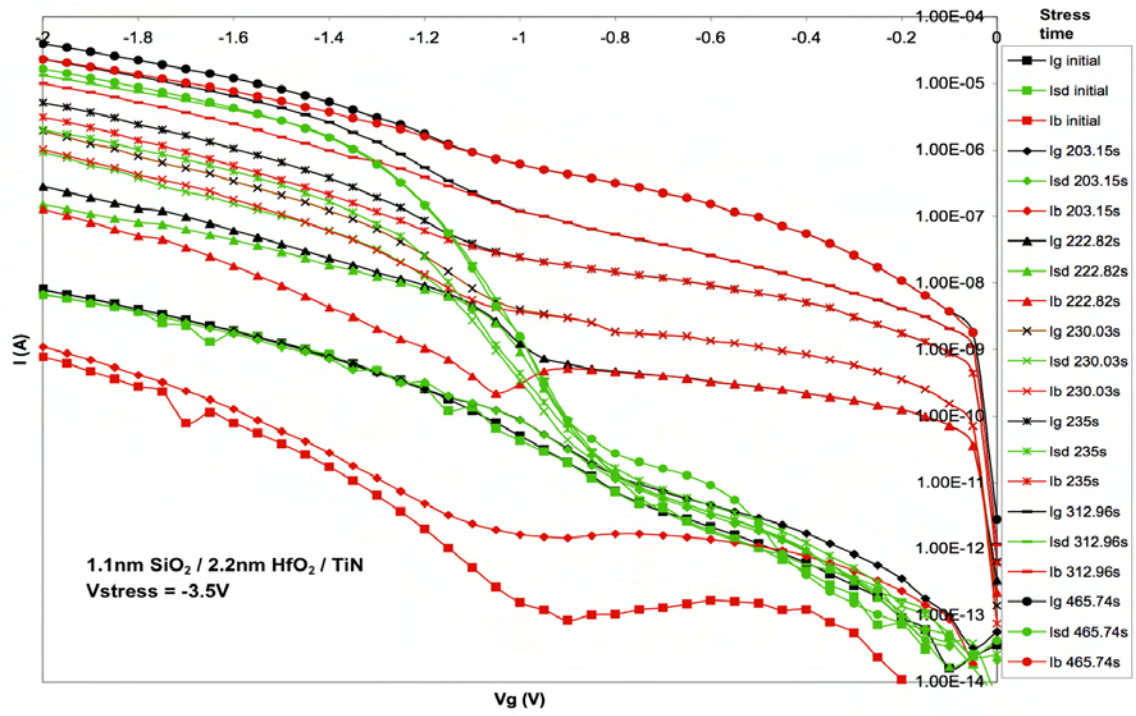

Figure 95: Time evolution of selected carrier separation I – Vg curves for a 2.2nm sample at Vg = -3.5V

A correlation was found between the relative increase of N_{it}, $\Delta N_{it} = \frac{N_{it}(t) - N_{it}(t=0)}{N_{it}(t=0)}$ at high frequency and the leakage current rate of change from the SILC measurements, $I_g(V_g,t) = \frac{I_g(V_g,t) - I_g(V_g,t=0)}{t - (t=0)}$, when taken at a large enough gate bias that it was consistent (Figs. 96 - 99).

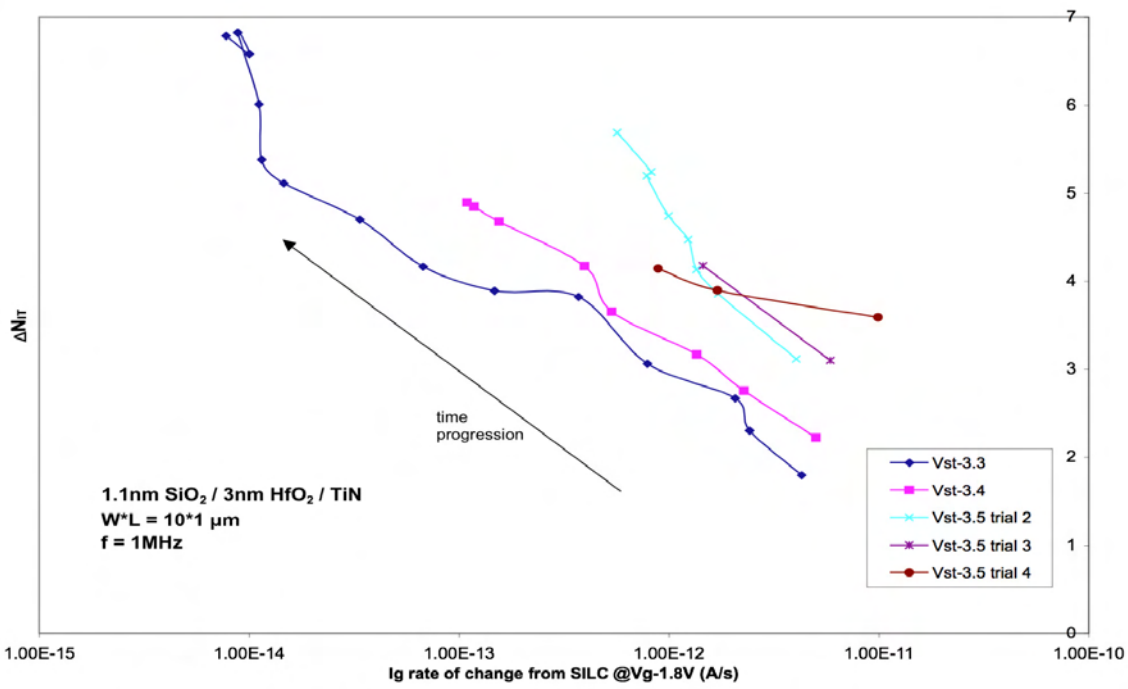

Figure 96: 1MHz relative ΔNit – Ig rate of change for 3nm samples

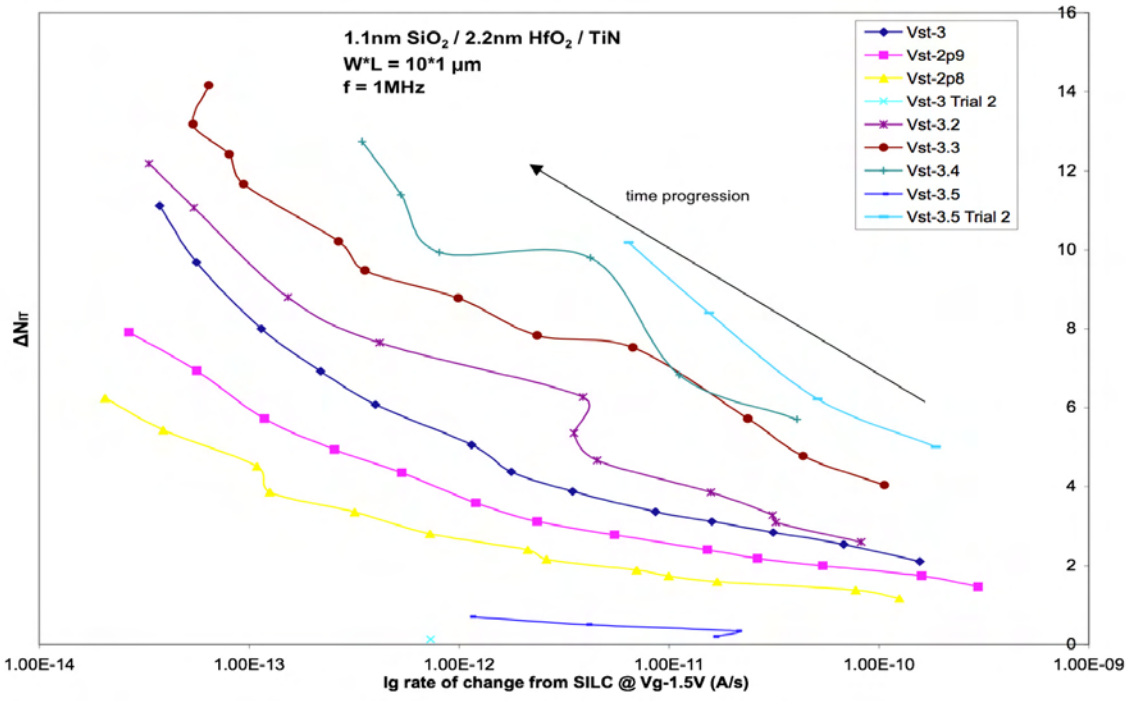

Figure 97: 1MHz relative ΔNit – Ig rate of change for 2.2nm samples

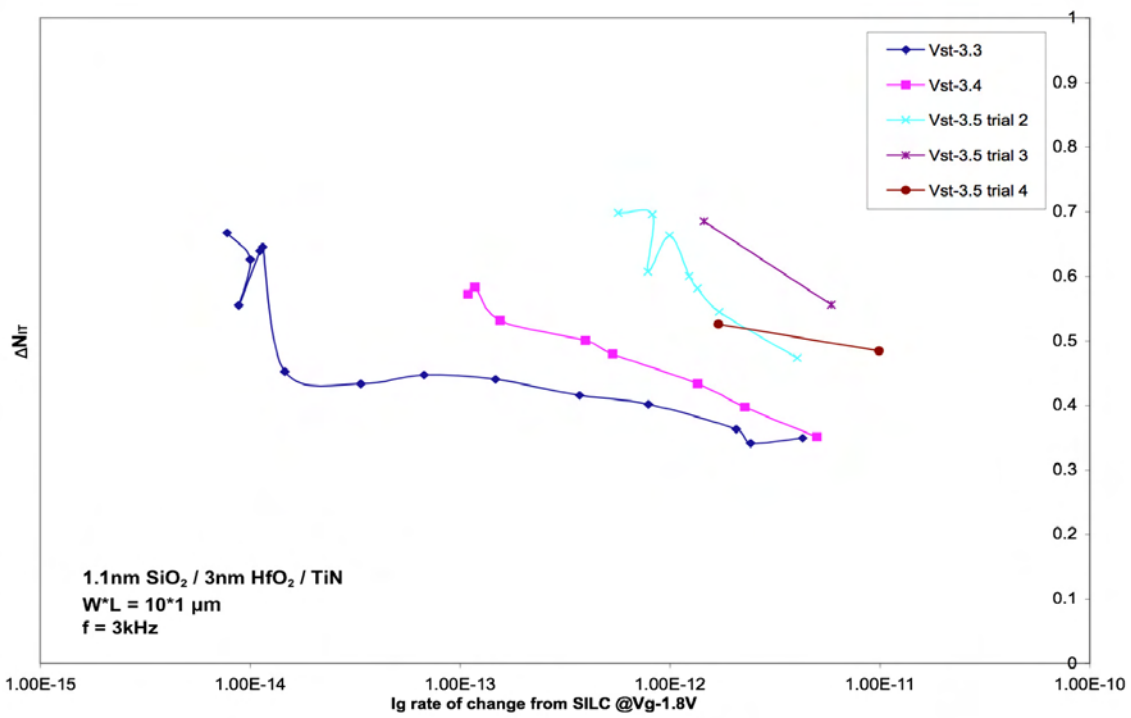

Figure 98: 3kHz relative ΔNit – Ig rate of change for 3nm samples

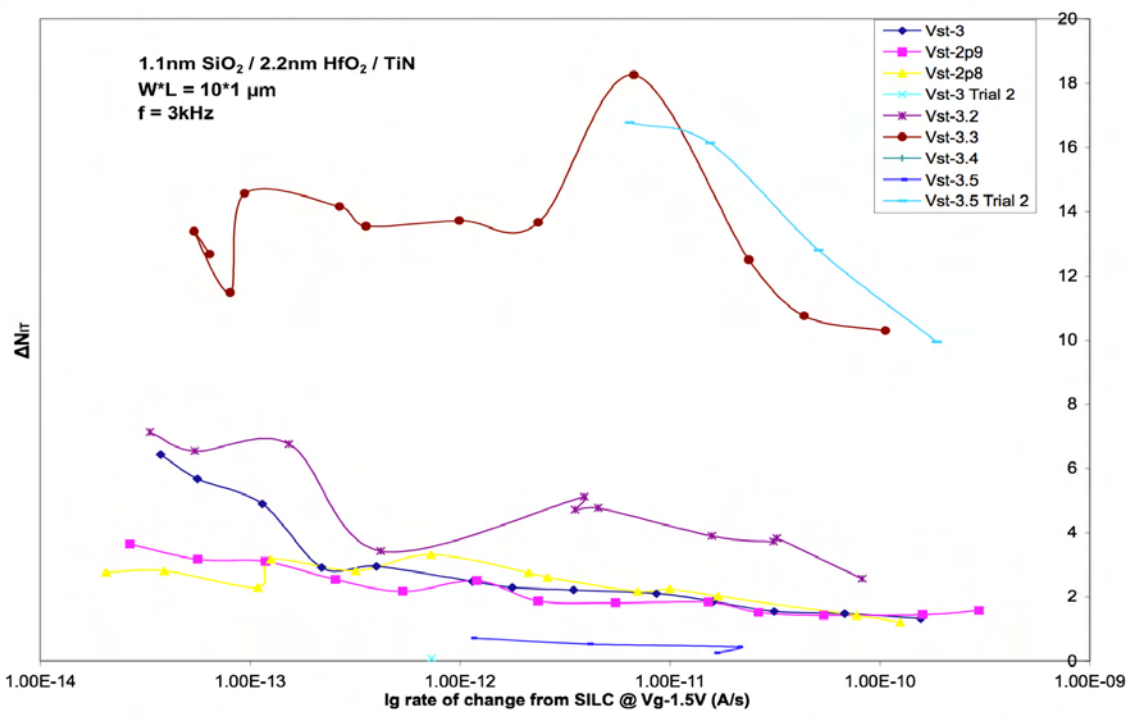

Figure 99: 3kHz relative ΔNit – Ig rate of change for 2.2nm samples

This correlation shows us that although the relative change of trap density tends to be larger for higher stress voltages, it does not seem to affect the leakage current rate of change. The high frequency CP data shows a stronger correlation, but this may be due to the fact that the low frequency data was at such low currents that noise skewed the data. The trend follows a power law, since the data is straight in a log – log graph, with the leakage current growth rate decreasing as the relative change in trap density increases.

CHAPTER IV

CONCLUSION

The physical mechanisms of breakdown in the high-κ dielectric gate stack for PMOS devices was investigated via constant stress voltage (CVS), fixed-amplitude, fixed-frequency charge pumping (CP), stress induced leakage current (SILC), and carrier separation techniques. It was found that the majority of traps activated during stress were located nearer to the SiO_2/Si interface. Higher stress voltages showed a tendency to bring the devices to breakdown in a shorter amount of time, as was expected. The greatest increase in the number of traps generated happened immediately after the application of a stress voltage, corroborating the theories which predict the majority of trap generation during the first few moments of stress. Tunneling through the gate stack appears to be the major mechanism of creating a breakdown pathway from the substrate to the gate electrode, although this needs to be verified through further experimentation. No apparent correlation was found between the relative change in trap density and the relative change in gate leakage current. One trend was found between the relative change of the trap density and the gate leakage rate of change. The extrapolated trendline power, representing the generation rate, showed a stress voltage dependence and a lower rate for traps nearer the HfO_2/SiO_2 interface for samples with a thicker HfO_2 layer.

Further experimentation should be done to investigate the reason for the large initial increase in trap density during the first few moments of stress. Stressing a large number of devices for short amounts of time at varying gate biases and taking CP measurements should yield useful results. The actual cause of breakdown needs to be further investigated. It has been assumed that tunneling is the process causing the generation of traps, but it may be that the electric fields seen near the interfaces are causing the trap activation, and then charge carriers use the traps to facilitate tunneling through the gate stack. This could be investigated by using hot carrier injector structures to separate the effects of the electric field and tunneling. Verification of the spatial placement of defects throughout the gate stack needs to be done to clarify hoe deep fixed-frequency CP measurements can probe. It has been shown that laser tomography can be used as a technique to visualize electrically active defects within a semiconductor.[8] Application of this method may further explain the spatial distribution of traps. Some of these experiments are already being conducted throughout the scientific community, and should help to answer questions remaining from this study.

REFERENCES

[1] D. Heh, C.D. Young, G.A. Brown, P.Y. Hung, A. Diebold, E.M. Vogel, J.B. Bernstein, and G. Bersuker, Spatial Distributions of Trapping Centers in HfO2/SiO2 Gate Stack. IEEE Transactions on Electron Devices, Vol. 54, No. 6, June 2007, pp. 1338-1345.

[2] B. Streetman and S. Banerjee, Solid State Electronic Devices: Fifth Edition. Prentice Hall, 2000, pp. 283.

[3] E. Vogel and D. Heh, Defects in High-k Gate Stacks: Nano-electronic Semiconductor Devices. Springer, 2006, pp. 85-96.

[4] C.D. Young, D. Heh, S.V. Nadkarni, R. Choi, J.J. Peterson, J. Barnett, B.H. Lee, and G. Bersuker, Electron Trap Generation in High-k Gate Stacks by Constant Voltage Stress. IEEE Transactions on Device and Materials Reliability, Vol. 6, No. 2, June 2006, pp. 123-131.

[5] W.Y. Loh, B.J. Cho, M.S. Joo, M.F. Li, D. Chan, S. Matthew, and D.L. Kwong, Charge Trapping and Breakdown Mechanism in HfAlO/TaN Gate Stack Analyzed Using Carrier Separation. IEEE Transactions on Device and Materials Reliability, Vol. 4, No. 4, December 2004, pp. 696-701.

[6] S. Chiang, M.C. Chen, W.S. Liao, J.M. You, M.F. Lu, Y.S. Hsieh, W.M. Lin, S. Huang-Lu, W.T. Shiau, S.C. Chien, and T. Wang, Channel Soft Breakdown Enhanced Excess Low-Frequency Noise in Ultra-Thin Gate Oxide PD Analog SOI Devices. IEEE 43rd Annual International Reliability Physics Symposium, April 2005, pp. 698-699.

[7] http://coen.boisestate.edu/bknowlton/Research/dualBandDiagramProgram.htm.

[8] O.V. Astafiev, V.P. Kalinushkin, and V.A. Yuryev, A New Optical Technique for Characterization of Technological Semiconductor Wafers. Materials Science and Engineering B, Vol. 34, No. 2-3, November 1995, pp. 124-131.

VITA

Brian Gregory Yust was born in El Paso, Texas, on March 5, 1982, the son of Gregory Yust and Linda Yust. After completing his work at J.M. Hanks High School in El Paso, Texas, in 2000, he entered Texas A&M University. He received the degree of Bachelor of Arts in physics from Texas A&M in August 2005. In August 2006, he entered the Graduate College of Texas State University-San Marcos in pursuit of his Master of Science degree. During his graduate studies, he was employed as a laboratory instructor and teaching assistant at Texas State and as a NanoScholar intern at SEMATECH in Austin, Texas.

Permanent Address: 3900 B Wrightwood

Austin, Texas 78722

This thesis was typed by Brian Yust.