The Origins of Random Telegraph Noise in Highly Scaled SiON nMOSFETs

J.P. Campbell[1], J. Qin[1,2], K.P. Cheung[1*], L. Yu[1,3], J.S. Suehle[1], A. Oates[4], K. Sheng[3]
[1]Semiconductor Electronics Division, NIST, Gaithersburg, MD 20899 *kpckpc@ieee.org
[2]Departmemt of Mechanical Engineering, University of Maryland, College Park, MD 20740
[3]Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ
[4]TSMC Ltd., Hsin-Chu, Taiwan 300-77, R.O.C.

ABSTRACT

Random telegraph noise (RTN) has recently become an important issue in advanced circuit performance. It has also recently been used as a tool for gate dielectric defect profiling. In this work, we show that the widely accepted model thought to govern RTN behavior cannot be used to describe our experimental observations. The basis of this model (charge exchange between inversion layer and bulk oxide defects via tunneling) is inconsistent with our RTN observations on advanced SiON nMOSFETs with 1.4 nm physical gate oxide thickness. Alternatively, we show that RTN is *qualitatively consistent with* the capture and emission of inversion charge by interface states. Our results suggest that a large body of the low-frequency noise literature very likely needs to be re-interpreted.

INTRODUCTION

Random telegraph noise (RTN) is a phenomenon in which MOSFET drain current (I_D) exhibits random discrete fluctuations or switching events as a function of time [1-3]. These fluctuations have been shown to be significant in highly scaled devices in which the channel length and width are reduced [3, 4]. There are also indications that the relative RTN amplitude increases as the gate-overdrive is reduced and can become quite large in the sub-threshold regime [5]. RTN fluctuations are particularly worrisome for memory devices such as SRAM and Flash where large RTN has already aroused reliability concerns [4, 6, 7].

It has been experimentally and theoretically shown that extrinsic (defect related) 1/f noise is actually a superposition of RTN events [3]. Thus, the physical origin of both RTN and low-frequency 1/f noise is the likely the same [3]. Over the last half century, numerous papers in the literature model MOSFET RTN and extrinsic 1/f noise as a charge exchange between defects in the gate dielectric and the inversion layer (channel) via a tunneling mechanism [8-10]. Such models are so well established that RTN, as well as 1/f noise, have been used by many groups as a depth profiling tool to measure defects in the gate dielectric [11-15]. Recently, 1/f noise trap profiling has been used in combination with frequency-dependent charge-pumping to suggest that electrical stress can generate defects in high-k dielectric layers of advanced gate stacks [14]. Such work has added fire to a controversy with far-reaching implications on the reliability of advanced high-k/metal gate technology. In this paper, we present experimental evidence which strongly indicates that the prevailing RTN and consequent 1/f models cannot be correct.

In this study we examine RTN behavior of highly-scaled transistors from the sub-threshold to super-threshold regime. We examine the effect of measurement bandwidth on RTN by making repeated measurements on the same device with several different amplifier rise-times. These measurements reveal very large RTN fluctuations which enable highly reliable characteristic time constant extraction. The highly-scaled geometry of the devices used in this study coupled with our extracted time constants provide a very unique set of observations which allow for a clear demonstration that the prevailing RTN model (inversion layer charge tunneling to bulk defects) cannot be correct.

EXPERIMENTAL

We utilize 0.085 μm x 0.055 μm SiON nMOSFET devices with a physical dielectric thickness of 1.4 nm. The RTN measurement apparatus is schematically illustrated in figure 1. The source and gate electrodes are biased using battery-powered variable voltage sources, while the substrate electrode is grounded for all measurements. All measurements were performed at room temperature with the source electrode fixed at -50 mV. The drain current is monitored by using a low-noise current amplifier with selectable bandwidth. The amplifier output is directly captured using a digital storage oscilloscope with a large memory depth (20 x 10^6 Samples). A common limitation for RTN measurements is the conflicting need to measure switching events of various durations with sufficient statistics. With our experimental set-up, we are able to measure a relatively large bandwidth (300 Hz to 30 kHz) while maintaining adequate statistics for the slower switching events.

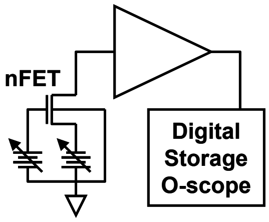

Fig 1: Schematic diagram of the experimental set-up used in this study.

RESULTS AND DISCUSSION

Figure 2(a) illustrates representative RTN $\Delta I_D/I_D$ fluctuations as a function of time for the sub-threshold case when the amplifier bandwidth is greatest (30 kHz). The observed $\Delta I_D/I_D$ is quite large and represents a convenient "test case" to critically examine the RTN phenomenon. The most commonly used parameters to describe RTN behavior are the characteristic times spent in each of the two current states [1, 2]. With reference to figure 2(a), the low current state occurs when an electron has been captured by a defect (which restricts current), and the high current state occurs when the electron has been emitted and the defect state is empty (no current restriction). Consequently, the characteristic time spent in the high current state corresponds to the capture time ($\tau_{capture}$), and the characteristic time spent in the low current state corresponds to the emission time ($\tau_{emission}$). We extract both $\tau_{capture}$ and $\tau_{emission}$ from each RTN measurement by fitting the time distribution to an exponential of the form: $A \exp[-t/\tau]$ [16]. The $\tau_{capture}$ distribution corresponding to figure 2(a) is shown in figure 2(b). The exponential fit illustrated in figure 2(b) is quite good and is representative of all our time constant extractions. We note that many researchers extract $\tau_{capture}$ and $\tau_{emission}$ by simply dividing the total time spent in a given state by the number of switching events [7]. In theory, this averaging approach is

978-1-4244-2195-4/08/$25.00 ©2008 IEEE

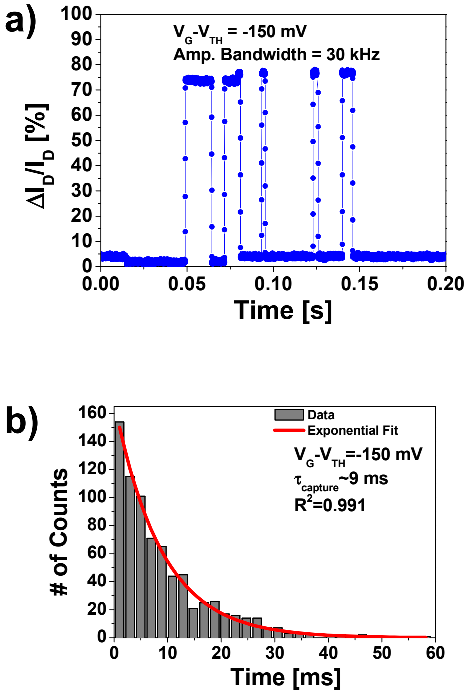

Fig 2: Representative RTN drain current (I_D) fluctuations as a function of time (a) for -150 mV gate overdrive. Note the very large percentage fluctuation. The corresponding representative distribution of time spent in the high state is illustrated in (b). The characteristic time constant ($\tau_{capture}$ in this case) is extracted by fitting the time distribution to an exponential. This plot (b) is representative of the good fits we obtained for our measurements. The amplifier bandwidth for this measurement is 30 kHz.

incorrect. However, we empirically note a relatively good correlation between the "averaged" and exponential time constants (not shown). This somewhat surprising correlation is due, in large part, to the fact that most of the *observable* RTN fluctuations in this work have characteristic times longer than 1 ms, and we rarely observe long duration events. Thus, the impact of sampling bias (omitting the very high and low ends of the distribution) on pure averaging is very minor.

Figures 3(a) and (b) illustrate $\Delta I_D/I_D$ as a function of time for gate voltages in the sub-threshold (V_G-V_{TH} = -150 mV) and the super-threshold (V_G-V_{TH} = +25 mV) regimes, respectively. It is clear from figure 3 that the frequency and duration of the RTN fluctuations are dependent on gate overdrive. The gate overdrive dependence is also illustrated by examining the corresponding characteristic time constants. Figures 4(a) and (b) illustrate the extracted $\tau_{capture}$ and $\tau_{emission}$ as a function of gate overdrive for several different amplifier bandwidths (300 Hz, 3 kHz, and 30 kHz). We note that the extracted time constants are strongly dependent on the gate overdrive. We also note that the amplifier bandwidth plays a relatively small role in the extracted $\tau_{capture}$ and $\tau_{emission}$ values with the bandwidth limitation resulting in an error of only a factor of two. One would expect, for the case of very short $\tau_{capture}$ and/or $\tau_{emission}$, the error will be greater. However, as illustrated in figures 4(a) and (b), such cases only happen at high gate-overdrive where the RTN signal becomes too small to detect. Thus, for these measurements, our 30 kHz bandwidth is clearly sufficient. The extracted $\tau_{capture}$ and $\tau_{emission}$ values of figure 4, coupled with the ultra-thin gate dielectric in our devices, provide a critical data set to examine the origin of RTN behavior.

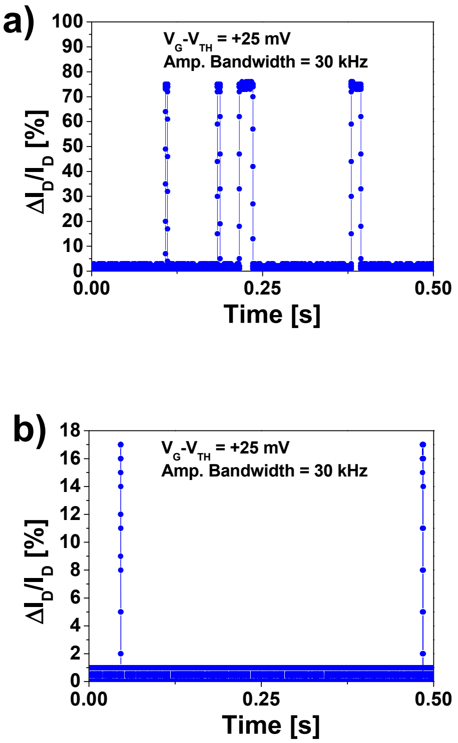

Fig 3: (a) and (b) illustrate the RTN fluctuations as a function of time for -150 mV and +25 mV gate overdrive, respectively. Note the large change in spectrum duration and frequency with gate overdrive. The amplifier bandwidth for this measurement is 30 kHz.

The next few paragraphs compare and contrast the expected RTN behavior (assuming the most common model of inversion layer tunneling to bulk dielectric defects [8-10]) with our observations. We clearly show that this commonly accepted RTN scenario cannot be correct. Figure 5 schematically illustrates the expected time constant relations (assuming the commonly accepted RTN model) for different defect locations. Depending on the distance between the inversion layer and the bulk defect, the de-trapping process can proceed in one of two ways: (1) the electron can tunnel back to the inversion layer, or (2) the electron can tunnel out of the dielectric to the gate electrode. If the bulk defect is located close to the interface, scenario (1) would be favored, while scenario (2) would be favored if the trap is located closer to the gate electrode. In the middle region of the dielectric, de-trapping involves a combination of both scenarios.

For bulk defects dominated by scenario (1):

The trapping and de-trapping time constants should be approximately equivalent. This is because tunneling can only happen at energy levels within a narrow range of the Fermi level. Consequently, there should be roughly equal densities of occupied and unoccupied states in the channel which can participate in the tunneling process. This situation dictates that the probability of tunneling into the defect equals the probability of tunneling out of the defect (Fermi's golden rule). Thus, in this case, the observed RTN should have an equal distribution of high and low levels. This distribution ($\tau_{capture}$ = $\tau_{emission}$) is never experimentally observed (Figures 4(a) and 4(b)).

For bulk defects dominated by scenario (2):

The de-trapping probability should be much higher than the trapping probability. This is because the defect's proximity to the

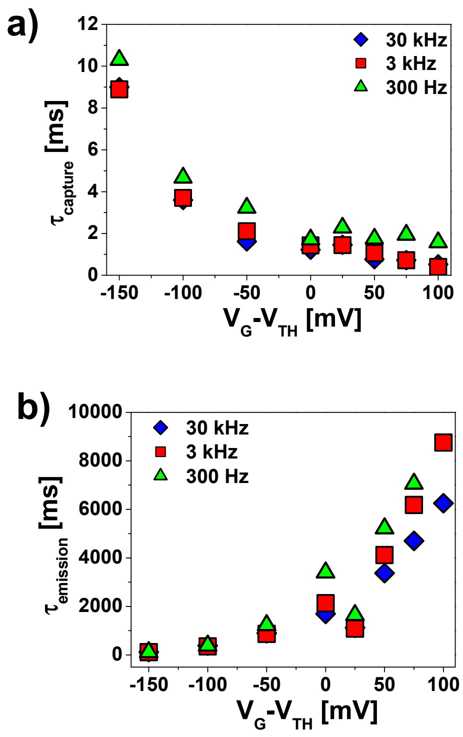

Fig 4: Extracted capture (a) and emission (b) time constant as a function of gate overdrive for several amplifier bandwidths. Notice that the relative values of the extracted time constants are much larger than those predicted by the commonly accepted RTN model.

gate electrode presents a much smaller tunneling barrier for electrons to tunnel out to the gate. Consequently, the defect should remain empty most of the time. Thus, in this case, the observed RTN should spend most of the time in the high current state. This situation ($\tau_{capture}$ >> $\tau_{emission}$) is also never experimentally observed (Figures 4(a) and 4(b)).

For bulk defects where both scenarios (1) and (2) contribute:

One would still expect the de-trapping probability to be greater than the trapping probability. This is because the trapped electron now has two de-trapping pathways (back to the inversion layer or on to the gate electrode). Thus, in this case, the observed RTN should also spend more time in the high current state than the low current state. This situation ($\tau_{capture}$ > $\tau_{emission}$) is also never experimentally observed (Figures 4(a) and 4(b)).

With reference to the above tunneling scenarios, if RTN is due to tunneling processes to and from bulk dielectric defects, $\tau_{emission}$ must always be less than or equal to $\tau_{capture}$ (for any defect location). It is also clear that the maximum $\tau_{emission}$ must be associated with the defects near the midpoint of the gate dielectric layer. We can estimate this maximum $\tau_{emission}$ by assuming that $\tau_{capture}$ = $\tau_{emission}$ at this point. $\tau_{capture}$ for bulk defects at the midpoint can be calculated using the tunneling front model [17]. Figure 6 illustrates the tunneling front calculations for an SiO_2 dielectric at various dielectric fields [18]. (While tunneling in our SiON dielectric is slightly different than the modeled SiO_2 dielectric, the error introduced is minimal.) The capture time for defects located at the midpoint of the dielectric (1.4 nm / 2 = 0.7 nm) is $\leq 10^{-9}$ s (see figure 6). In this calculation, we use a characteristic tunneling time constant, τ_0, of 6.6 x 10^{-14} s [19]. Note that many researchers choose $\tau_0 = 10^{-10}$ s for the tunneling front calculation [10]. This larger τ_0 value is actually based on speculation [20] while the 6.6 x 10^{-14} s was measured experimentally and supported by theoretical calculation [19]. Even if we allow for this ~3 orders of magnitude error in τ_0, the capture time for defects located 0.7 nm into the dielectric changes to $\leq 10^{-6}$ s. Therefore, in our RTN measurements we should expect our $\tau_{emission}$ values to be less than 10^{-6} s under all possible conditions.

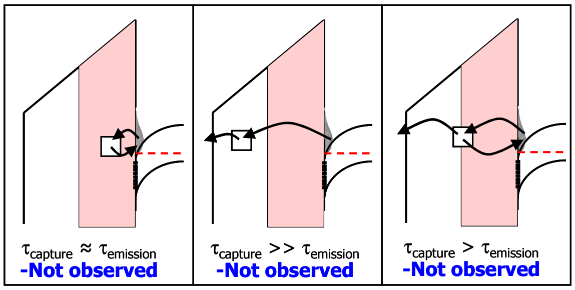
$\tau_{capture} \approx \tau_{emission}$
-Not observed
$\tau_{capture}$ >> $\tau_{emission}$
-Not observed
$\tau_{capture}$ > $\tau_{emission}$
-Not observed

Fig 5: Schematic diagram illustrating the possible RTN reaction pathways (assuming the common RTN model [8-10]) through the dielectric for different defect locations.

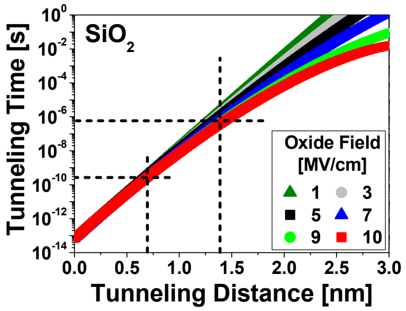

Fig 6: The calculated tunneling times [18] and distances for electrons tunneling through SiO_2. The calculations are realized using a tunneling front model [17] with SiO_2 effective mass = 0.5 m_0 and τ_0 = 6.6 x 10^{-14} s.

Both of these characteristic tunneling times are much faster than our observed RTN results (figures 4(a) and (b)). If the inversion layer tunneling to bulk defect model was correct, the 10 µs (30 kHz) minimum rise time of our current amplifier would prevent us from observing any RTN from our sample.

As further evidence that the inversion layer to bulk defect tunneling model is incorrect, we note that the observed RTN dependence on gate-overdrive is only possible if there is a continuous distribution of traps in space as well as in energy in the dielectric layer. This assumption conflicts greatly with the typical understanding of gate dielectric defects. *The body of evidence discussed thus far collectively compels us to conclude that the commonly accepted model for RTN behavior and the subsequent analysis must be incorrect.*

On the other hand, our experimentally observed $\tau_{capture}$=10^{-3}s to 10^{-2} s and $\tau_{emission}$=10^{-1}s to 10^{+1}s trends are *qualitatively consistent* with an RTN mechanism involving the capture and emission of inversion layer charge by interface states. The observation that the drain current spends much more time in the low state than high state indicates that the defects are much more likely to be filled rather than empty. This is what one would expect for interface states in inversion.

The capture rate for interface states should be proportional to the inversion charge density. At weak inversion, the inversion charge

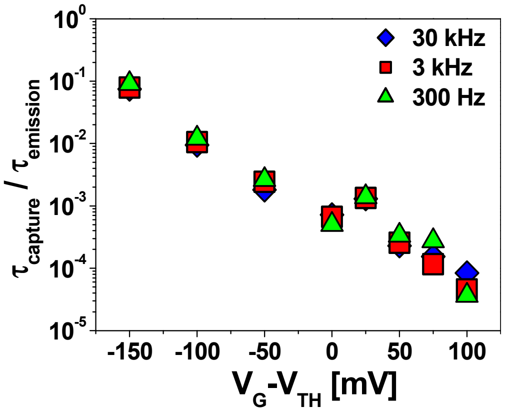

Fig 7: $\tau_{capture}/\tau_{emission}$ as a function of gate overdrive. This gate overdrive dependence clearly illustrates that $\tau_{emission}$ increases much faster than $\tau_{capture}$ with increasing gate overdrive. This scenario would be expected in the case of an RTN mechanism dominated by interface state charge capture and emission.

density is low, which dictates longer capture times. At strong inversion, the inversion charge density is high, dictating very short capture times. This gate-overdrive dependence is in agreement with our observations in figure 4(a). One would also expect that the emission rate for interface states should decrease with gate-overdrive. As inversion charge density increases, available empty states lie at higher energy and the emission probability decreases (longer $\tau_{emission}$). This gate-overdrive dependence is in agreement with our observations in figure 4(b). This correspondence is further illustrated by realizing that increasing gate-overdrive should result in an exponential decrease in $\tau_{capture}$ and the simultaneous exponential increase in $\tau_{emission}$. The ratio between the two characteristic times should remain an exponential function of gate-overdrive. This is indeed the case as shown in figure 7.

One might argue on the basis of high frequency charge pumping measurements that interface state charge capture should occur much faster than our observed time constants [18]. However, charge pumping requires that the device be swept from deep accumulation to deep inversion. The situation is quite different in RTN measurements where the inversion charge densities are substantially less and the system is much closer to steady-state. This explanation for the observed $\tau_{capture}/\tau_{emission}$ gate overdrive dependence is much more natural than the one required for the tunneling to bulk dielectric defects model. Thus, this serves as another indication that the observed RTN is consistent with an interface state capture and emission process.

CONCLUSION

In this study we have examined the origin of RTN fluctuations by extracting the characteristic capture and emission time constants ($\tau_{capture}$ and $\tau_{emission}$) as a function of gate overdrive. A comparison of our extracted $\tau_{capture}$ and $\tau_{emission}$ values with the expected inversion layer to bulk defect tunneling times yields very large discrepancies. However, our $\tau_{capture}$ and $\tau_{emission}$ observations are qualitatively consistent with an RTN process involving the capture and emission of inversion layer charge by interface states. These observations very strongly call into question much of the RTN and consequent 1/f analysis which profiles bulk dielectric defects. Thus, it is quite possible that current understanding of bulk dielectric defect generation and subsequent reliability (especially in high-k bi-layer dielectrics [13-15]) is incorrect.

ACKNOWLEDGMENTS

The authors would like to acknowledge fruitful conversations with C.A. Richter and T. Grasser. The authors acknowledge funding from the NIST Office of Microelectronic Programs. J.P.C also acknowledges funding support by the National Research Council.

REFERENCES

[1] S. Machlup, "Noise in Semiconductors - Spectrum of a 2-Parameter Random Signal" *J Appl Phys,* **25**, pp. 341-343 (1954)

[2] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, "Discrete Resistance Switching in Submicrometer Silicon Inversion-Layers - Individual Interface Traps and Low-Frequency (1-F Questionable) Noise" *Phys Rev Lett,* **52**, pp. 228-231 (1984)

[3] M. J. Uren, D. J. Day, and M. J. Kirton, "1/F and Random Telegraph Noise in Silicon Metal-Oxide-Semiconductor Field-Effect Transistors" *Appl Phys Lett,* **47**, pp. 1195-1197 (1985)

[4] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya, "Random Telegraph Signal in Flash Memory: Its Impact on Scaling of Multilevel Flash Memory Beyond the 90-nm Node" *IEEE J. Solid-St Circ,* **42**, pp. 1362-1369 (2007)

[5] A. Ohata, A. Toriumi, M. Iwase, and K. Natori, "Observation of Random Telegraph Signals - Anomalous Nature of Defects at the Si/SiO$_2$ Interface" *J Appl Phys,* **68**, pp. 200-204 (1990)

[6] A. S. Spinelli, C. M. Compagnoni, R. Gusmeroli, M. Ghidotti, and A. Visconti, "Investigation of the Random Telegraph Noise Instability in Scaled Flash Memory Arrays" *Jpn J Appl Phys,* **47**, pp. 2598-2601 (2008)

[7] N. Tega, H. Miki, M. Yamaoka, H. Kume, M. Toshiyuki, T. Ishida, M. Yuki, R. Yamada, and T. Kazuyoshi, "Impact of Threshold Voltage Fluctuation Due to Random Telegraph Noise on Scaled-Down SRAM" *IEEE Int. Rel. Phys. Symp.* pp. 541-546 (2008)

[8] G. Ghibaudo, O. Roux, C. Nguyenduc, F. Balestra, and J. Brini, "Improved Analysis of Low-Frequency Noise in Field-Effect MOS-Transistors" *Phys Status Solidi A,* **124**, pp. 571-581 (1991)

[9] K. K. Hung, P. K. Ko, C. M. Hu, and Y. C. Cheng, "A Unified Model for the Flicker Noise in Metal Oxide-Semiconductor Field-Effect Transistors" *IEEE Trans. Electron Dev,* **37**, pp. 654-665 (1990)

[10] R. Jayaraman and C. G. Sodini, "A 1/F Noise Technique to Extract the Oxide Trap Density Near the Conduction-Band Edge of Silicon" *IEEE Trans. Electron Dev,* **36**, pp. 1773-1782 (1989)

[11] Z. Celik-Butler, P. Vasina, and N. Vibhavie Amarasinghe, "A Method for Locating the Position of Oxide Traps Responsible for Random Telegraph Signals in Submicron MOSFETs" *IEEE Tran. on Electron Dev.,* **47**, pp. 646-648 (2000)

[12] J. Lee and G. Bosman, "Defect Spectroscopy Using 1/f Noise of Gate Leakage Current in Ultrathin Oxide MOSFETs" *Solid State Electron,* **47**, pp. 1973-1981 (2003)

[13] P. Srinivasan, E. Simoen, R. Singanamalla, H. Y. Yu, C. Claeys, and D. Misra, "Gate Electrode Effects on Low-Frequency (1/f) Noise in p-MOSFETs with High-Kappa Dielectrics" *Solid State Electron,* **50**, pp. 992-998 (2006)

[14] H. D. Xiong, H. Dawei, Y. Shuo, Z. Xiaoxiao, M. Gurfinkel, G. Bersuker, D. E. Ioannou, C. A. Richter, K. P. Cheung, and J. S. Suehle, "Stress-Induced Defect Generation in HfO$_2$/SiO$_2$ Stacks Observed by Using Charge Pumping and Low Frequency Noise Measurements" *IEEE Int. Rel. Phys. Symp.* pp. 319-323 (2008)

[15] H. D. Xiong, D. Heh, M. Gurfinkel, Q. Li, Y. Shapira, C. Richter, G. Bersuker, R. Choi, and J. S. Suehle, "Characterization of Electrically Active Defects in High-k Gate Dielectrics by Using Low Frequency Noise and Charge Pumping Measurements" *Microelectron Eng,* **84**, pp. 2230-2234 (2007)

[16] M. J. Kirton and M. J. Uren, "Noise in Solid-State Microstructures - A New Perspective on Individual Defects, Interface States and Low-Frequency (1/F) Noise" *Adv Phys,* **38**, pp. 367-468 (1989)

[17] T. R. Oldham, A. J. Lelis, and F. B. Mclean, "Spatial Dependence of Trapped Holes Determined from Tunneling Analysis and Measured Annealing" *IEEE Trans. Nucl. Sci.,* **33**, pp. 1203-1209 (1986)

[18] Y. Wang, V. Lee, and K. P. Cheung, "Frequency Dependent Charge-Pumping, How Deep Does It Probe?" *IEEE Int. Electron Dev. Meeting* pp. 763-766 (2006)

[19] I. Lundstrom and C. Svensson, "Tunneling to Traps in Insulators" *J Appl Phys,* **43**, pp. 5045-5047 (1972)

[20] S. Christensson, I. Lundstrom, and C. Svensson, "Low Frequency Noise in MOS Transistors I: Theory" *Solid State Electron,* **11**, p. 797 (1968)

Questions and Answers

Q: Elastic tunneling was in your motivation, what changes to low-frequency noise/random telegraph noise would you expect with inelastic tunneling? Would inelastic tunneling explain your results?

A. We cannot rule out the possibility that *inelastic* tunneling could play a major role in our observations. The whole point of this paper is that we can unambiguously rule out the presence of *elastic* tunneling. (This point invalidates much of the recent low-frequency noise defect profiling work.) However, the exact physics governing our observations are difficult to unambiguously determine. Our observations are qualitatively consistent with interface state capture and emission. Although, an inelastic tunneling process or even an interface state mitigated tunneling process could also be dominant.