Dynamic off-state TDDB of ultra short channel HKMG nFETS and its implications on CMOS logic reliability

S. Kupke, S. Knebel, S. Rahman, S. Slesazeck and T. Mikolajick
NaMLab gGmbH, TU Dresden
Dresden, Germany
phone: (+49) –(351)- 212499036, e-mail address: steve.kupke@namlab.com

T. Mikolajick
Institute of Semiconductors and Microsystems, TU Dresden
Dresden, Germany

R. Agaiby and M. Trentzsch
GLOBALFOUNDRIES Module One LLC & Co. KG
Dresden, Germany

***Abstract*— The impact of DC off-state and AC gate + off-state stress on the time dependent dielectric breakdown (TDDB) of ultra-short channel high-k/metal gate (HKMG) nMOSFETs was investigated. Under high DC off-state (drain) bias, the dielectric wear out was found to be caused by hot hole injection at the drain side. The breakdown time scaled with the gate length in accordance with a higher impact ionization rate by an increased subthreshold leakage current for shorter channels. At identical bias, drain-only stress results in a less severe degradation in comparison to gate-only stress. However, the combination of alternating gate and off-state stress results in a lower lifetime compared to DC and AC gate-only stress. The AC gate + off-state pattern exhibits a similar degradation behavior as bipolar AC stress, attributed to continuous charge trapping and detrapping in the gate oxide. The TDDB failure distribution did not obey the Poisson area scaling assuming randomly generated defects. The spatial asymmetric breakdown localized at the drain edge could be described applying a more general negative binominal yield model.**

Keywords- BTI, TDDB, off-state, reliability, CMOS, logic circuits.

I. INTRODUCTION

Before being able to benefit from MOSFET scaling, several physical limitations must be surmounted by continuous device engineering. Short channel effects (SCE) and a reduced durability have become an increasing issue [1]. Therefore, it is of crucial importance to qualify the existing and upcoming technologies in the best possible manner. For a proper reliability assessment and meaningful end of lifetime estimation several requirements need to be met. On the one side, mandatory adjustments on the acceleration models are necessary to better describe the underlying physics, especially for continuously decreasing equivalent gate oxide thicknesses. On the other side, the measurement methods need to be improved to account for e.g. fast relaxation processes under bias temperature instability (BTI) stress [2]. In addition, the reliability assessment should as closely as possible reproduce the actual operation condition of the CMOS device. The gate dielectric integrity is normally tested by means of high DC or AC gate-only stress measurements. In case of HKMG MOSFETs, alternating gate stress was reported to better account for the device switching behavior by minimizing charging effects under DC stress [3]. Despite this, the stress pattern can be further improved considering CMOS logic operation.

Taking a CMOS inverter for example (Fig. 1a), the supply voltage constantly switches between the gate and drain terminal. This results in an alternating BTI and off-state stress pattern as shown in Fig. 1b.

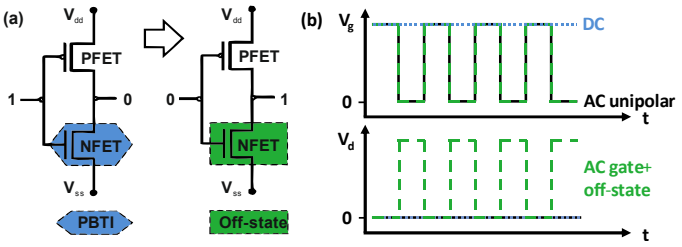

Figure 1. (a) Schematic CMOS inverter logic states. The nFET stress mode changes from PBTI to off-state if the input signal switches. (b) Schematic TDDB stress pattern for DC, unipolar AC and AC gate + off-state stress.

Under off-state conditions ($V_d = V_{dd}$, $V_{g,s} = 0$ V), a non-negligible subthreshold leakage $I_{ds,leak}$ can flow since I_{on}/I_{off} ratios are as low as 10^4 for high performance HKMG devices. This current can have a significant influence on the dielectric integrity. It was also shown, that the gate length has been

This work was financially supported by the European Fund for Regional Development and the Free State of Saxony (COOL MEMORY project).

scaled to a critical size, where the PBTI threshold voltage degradation/relaxation can be well influenced by the applied drain potential [4]. So far, the impact of the high source-to-drain leakage and CMOS realistic alternating gate-drain stress on the gate oxide dielectric breakdown for ultra-short channel devices has not been investigated.

In this publication, we report on the dynamic off-state influence on the gate oxide reliability of 28 nm HKMG nMOS transistors. The article is structured as followed: In section III, the breakdown characteristics and cause of wear out under sole DC off-state stress are presented. This is continued by CMOS realistic TDDB measurements under alternating gate and off-state bias. Results for different stress voltages and frequencies are discussed and a modified area scaling law is derived to describe the non-uniform breakdown characteristic.

II. EXPERIMENTAL

The breakdown measurements were done on 28 nm state of the art nMOS HKMG transistors with a thin SiON interlayer and a hafnium based high-k dielectric [5]. The electrical measurements were performed using a Keithley 4200-SCS analyzer and an arbitrary waveform generator HP 8110 A with two synchronized output channels. For the AC gate + off-state stress, the voltage was periodically switched between the two terminals. The duty cycle was slightly adjusted to 49.7 % with a delay of 50.2 % for the drain output to avoid any hot carrier degradation by simultaneous high gate and drain bias. The stress time was then corrected for the actual experienced gate stress. The DC/AC stress condition was constantly interrupted (300 ms delay) and the gate leakage current was determined at either ($V_g = 1.1$ V, $V_d = 0$ V) or ($V_g = 0$ V, $V_d = V_{stress}$) while all other terminals were grounded. The breakdown criteria was set to $I_{BD} = 10^{-6}$ A.

III. DC OFF-STATE DEGRADATION

In order to analyze the drain bias impact on the gate oxide reliability, DC off-state TDDB measurements were conducted. Figure 2 shows the corresponding Weibull distribution under high DC drain stress. In comparison, much higher voltages had to be applied to obtain similar breakdown times as for gate-only stress, which is due to the smaller stressed area of gate-to-drain overlap [6]. The current at the gate and drain after breakdown were found to coincide, confirming that the percolation path in the dielectric stack was truly created at the drain edge. Similar behavior was recently verified using charge pumping measurements by Varghese et al. [7]. The negative threshold voltage shift at high drain biases (inset Fig. 2) indicated a positive charge injection into the gate oxide. The injected holes can stem from drain avalanche hot carriers (DAHC) [8] generated by the subthreshold leakage current and from surface band to band tunneling (BTBT) at the drain edge for $q\Phi_B \geq E_g(\mathrm{Si}) \approx 1.12$ V [9]. The possible hot carrier contribution was further studied by means of gate length dependent current and breakdown measurements. As

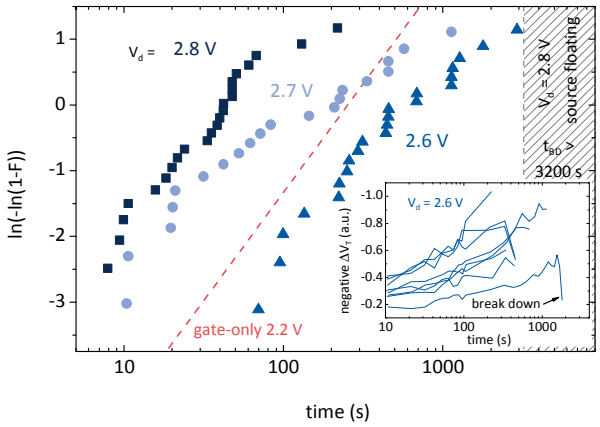

Figure 2. Failure distribution of the dielectric breakdown for off-state stress (symbols) vs. gate-only stress (dashed line). For measurements with a floating source (no channel current), the breakdown occurred beyond 3200 s as indicated by the dashed area. The inset shows the corresponding negative threshold voltage shift under off-state stress.

depicted in Fig. 3, the body and gate current stayed virtually constant from 2 µm down to 0.026 µm gate length. However, $I_{ds,leak}$ strongly increased for shorter gate length L with $I_{ds,leak} \propto \exp(-1/L)$ and exceeded the body current, which is a measure of the BTBT component, below 0.042 µm.

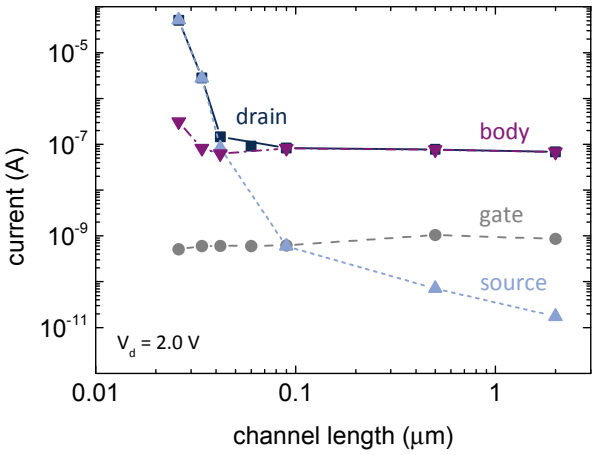

Figure 3. Drain, source, gate and body current as a function of the channel length for a fixed drain voltage, while all other terminals were grounded. The drain current is composed of all three contributions. For ultra-short channel devices the subthreshold leakage becomes dominant.

The time to breakdown for different short channel nFETs below 0.1 µm as shown in Fig. 4 correlated with the terminal current measurements. The average lifetime decreased for smaller gate length. The 0.026 µm devices with the highest source to drain leakage proved to have the lowest dielectric lifetime. At 0.042 µm gate lengths, both the subthreshold leakage and the BTBT current had about the same strength and could equally contribute to the degradation. When the gate length was further increased, the breakdown time did not saturate as one could expect since the drain current now governed by the BTBT current stays constant. However, the

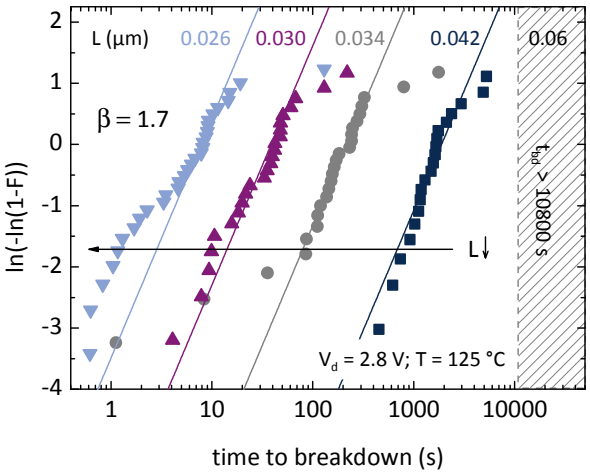

Figure 4. Time to breakdown distribution for different nFET gate lengths at a fixed high drain voltage. The lifetime decreased for short device length and was outside the measurement window for 0.06 µm gate length devices.

breakdown time for 0.06 µm n-channel MOSFETs did further increase and was above the measurement window of 10800 s. Time to breakdown measurements with a floating source junction, eliminating the channel current contribution, revealed a similar behaviour with very long lifetimes as indicated by the dashed area in Fig. 2. The floating source prevents any current contribution from the subthreshold leakage and thus any hole injection by impact ionization.

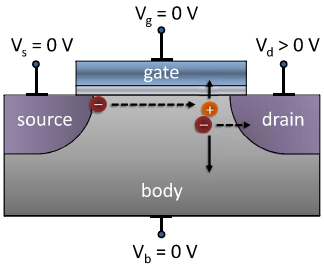

Figure 5. Schematic nFET degradation by hot hole injection under high off-state stress. The high channel current generates hot carriers by impact ionization. Hot electrons are collect by the drain contact, while the hot holes are repelled by the high positive drain potential and are either injected into the gate oxide or collected by the body contact.

This together with the length dependent TDDB measurement rules out that BTBT could be the major contributor to the insulator degradation. From that we deduce, that channel electrons traveling from source to the drain side and related DAHC can be accounted for the main part of the oxide damage. The electrons, generated by impact ionization, are attracted by the positive drain bias, while the holes are repelled and are eventually injected into the gate oxide causing damage (Fig. 5).

IV. AC Gate + Off-state TDDB

The off-state degradation discussed in the last section appeared at high drain voltages, which are generally not applied in high performance MOSFETs and are rather important for I/O and high power devices, where a high off-state voltage has to be tolerated [10]. Nevertheless, the findings are important to generally understand possible degradation mechanisms that can occur under off-state bias and are considered fundamental in order to carry on with the AC stress measurements.

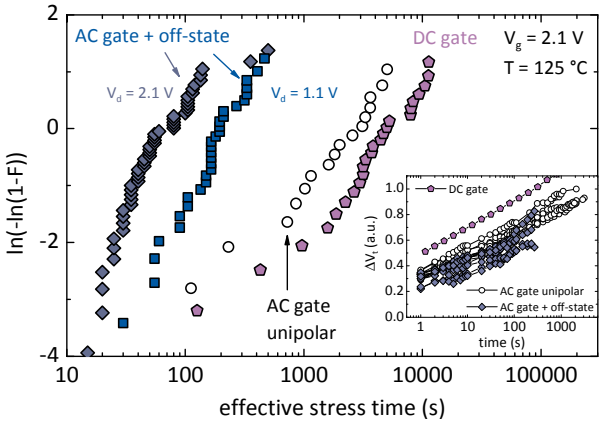

Figure 6. TDDB failure distribution for three different stress patterns: DC, unipolar AC, AC gate + off-state applying the same gate stress voltage. The alternating gate and drain bias resulted in the lowest device lifetime. All AC conditions were applied using f = 10 Hz at 50 % duty cycle. The stress time was corrected for the effective gate stress period. The inset shows the threshold voltage shift for the different stress cases. The lowest V_t degradation for the AC gate + off-state case yields in a reduced shielding effect against the external stress field, resulting in an enhanced wear out.

For CMOS logic circuits, the stress frequently switches between BTI and off-state degradation. Thus, compared to the conventional DC and AC test patterns, an AC gate + off-state pattern is considered to be a reasonable stress case.

Figure 6 compares the extracted failure times for the different stress patterns as introduced in Fig. 1b. Remarkably, the AC gate + off-state stress results in the lowest lifetime of all three cases. The decreased lifetime in comparison to unipolar stress indicates a strong contribution of the off-state bias under AC conditions. In contradiction, the previous DC off-state results indicated that much higher drain voltages than 2.1 V were necessary for the device to fail within a reasonable time frame. From that it is evident, that the physical origin for DC off-state and AC gate + off-state breakdown are different.

Under high gate stress, electrons are trapped into preexisting defects in the high-k dielectric. Under DC conditions, the resulting threshold voltage shift can reach up to 400 mV before breakdown occurs.

The accumulated charge effectively shields the external applied field, which relaxes the stress conditions. The off-state bias acts as if a negative voltage would be applied at the gate close to the drain edge. This accelerates the release of trapped charges, similar to results reported for bipolar AC stress on HKMG devices [3]. Subsequently, when reapplying a positive gate stress, a higher effective stress field is again present, resulting in an enhanced wear out [11]. The enhanced degradation is still existent when the drain bias is lowered down to use conditions (Fig. 6), where side effects from drain

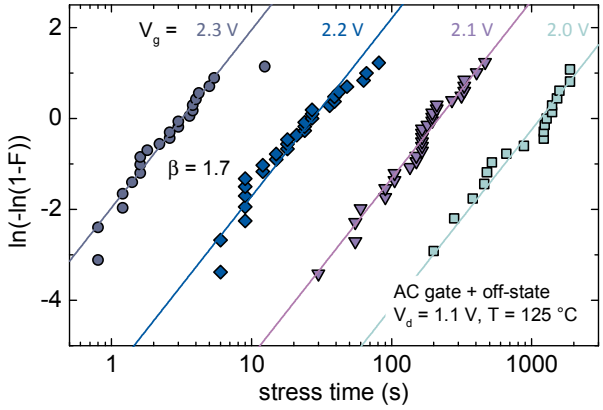

Figure 7. TDDB failure distribution for AC gate + off-state TDDB measurements using $V_d = V_{dd}$ but different gate voltages.

depletion and strong impact ionization can be excluded. This opens up the possibility to solely apply a conventional gate voltage acceleration, while keeping $V_d = V_{dd}$ for a reasonable end of lifetime projection. The TDDB failure distribution for different gate voltages is shown in Fig. 7. The shape parameter $\beta = 1.7$ is similar to the one obtained by the length dependent DC TDDB measurements.

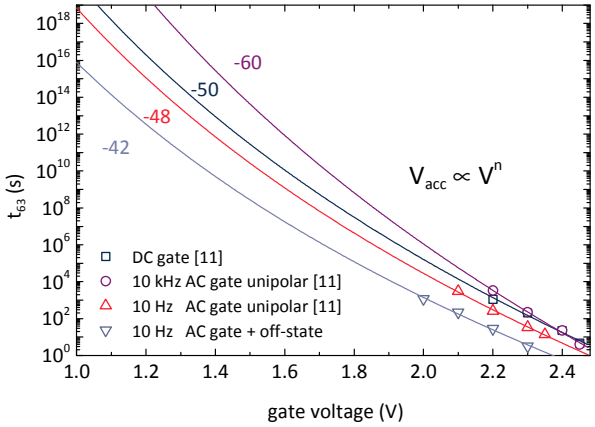

Figure 8. 63% quantile as obtained by TDDB measurements as a function of acceleration voltage. DC and AC unipolar gate stress [11] are compared with the new AC gate + off-state method, which results in the lowest lifetime and smallest slope.

The voltage acceleration can be described by a power law, which is commonly used for ultra-thin gate dielectrics, as shown in Fig. 8. For comparison, data for similar devices and different stress patterns taken from [11] are plotted. The AC gate + off-state has a lower power law exponent $n = -42$, which proves that drain voltage even under use condition well influences the detrapping with consequences for the breakdown time. However, due to the asymmetric breakdown at the drain side, we cannot exclude a divergent behavior from the power law without extensive low voltage TDDB measurements.

CMOS circuits are normally operated at high frequencies, which is also important to be covered by the reliability

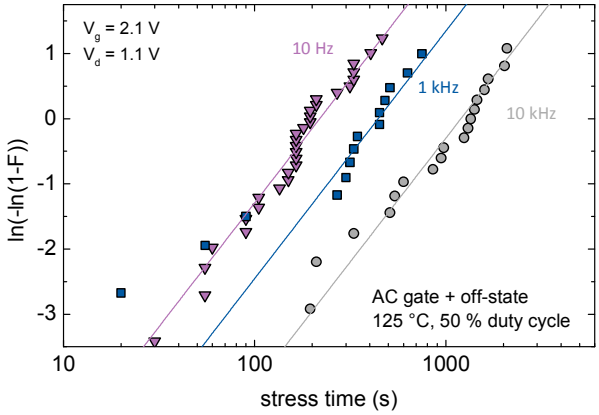

Figure 9. Frequency dependent TDDB failure distribution for AC gate + off-state stress using $V_g = 2.1$ V and $V_d = 1.1$ V at 50 % duty cycle.

assessment. There, the situation can be relaxed applying shorter pulse sequences as shown in Fig. 9.The average time to breakdown increased by almost one order of magnitude when increasing the frequency from 10 Hz to 10 kHz. The high frequency stress exhibited a comparable threshold voltage degradation (not shown) to the 10 Hz case in accordance to [12]. Charge carriers may not be able to follow the alternating field for short pulses and an equilibrium between trapping and detrapping is reached. However, the cause of the lifetime increase is not fully understood.

For gate bias stress, the Poisson area scaling is normally applied to extrapolate to use conditions, which assumes that the defects are randomly generated and distributed within the gate oxide area A, resulting in the well-known equation:

$$\frac{t_{63}^{(2)}}{t_{63}^{(1)}} = A_{acc} = \left(\frac{A_1}{A_2}\right)^{1/\beta}. \qquad (1)$$

This implies that the defect generation is also independent of the stress history with no influence of the momentary trap density on the following new defect position. Figure 10 shows the acceleration factor versus the length (W = const.) or width (L = const.) ratio. The black line ($\alpha = \infty$) represents the traditional Poisson scaling law with the derived shape parameter $\beta = 1.7$ from the breakdown measurements.
The length scaling proved to have a smaller influence on the acceleration factor than the Poisson law, whereas an increase in width stronger impacts the failure time scaling. This is clear indication of a non-uniform defect distribution/generation under alternating gate and drain stress and supports that the breakdown is strongly localized, presumably at the drain side.
Therefore, the acceleration factor A_{acc} cannot be described as function of the area. The width and length scaling have to be separately addressed and the acceleration factor can be written as

$$A_{acc} = L_{acc} \cdot W_{acc}. \qquad (2)$$

By considering the asymmetric breakdown, one can already give an explanation for the diverse behavior. We assume a random gate area and that the breakdown occurs primarily at the drain region as depicted in the inset of Fig. 10. A length scaling (L + ΔL) would not significantly change the failure time, since the effective breakdown area A_{eff} stays almost constant. In contrast, a width scaling severely influences the wear our time due to a direct change in size of the effective breakdown area. Therefore, a potential model needs to account for the change in the ratio between the overall gate area and the effective breakdown area in case of a length scaling.

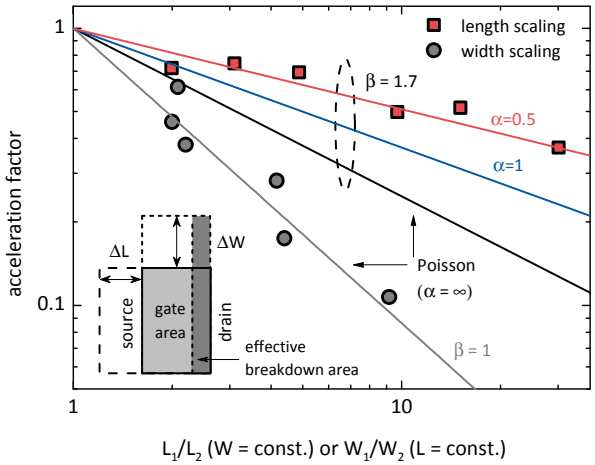

Figure 10. Acceleration factor as a function of the length (red squares) or width (grey circles) ratio. The upper three curves belong to a β = 1.7. The length scaling can be described using the negative binominal yield model with the clustering parameter α = 0.5. For comparison, the Poisson area scaling (α = ∞) with the same β value is plotted. The influence of the width scaling on the acceleration factor is more pronounced and can be well fitted with the Poisson scaling law with β = 1. The schematic dimension scaling (inset) demonstrates the stronger influence of the width on the effective drain side breakdown area.

A more general approach to quantitatively describe the scaling behavior is given by the so called compound Poisson distribution, which expresses the probability to get x fatal defects within an area A for a certain defect density D and defect distribution f(D):

$$Prob\,(X = x) = \int_0^\infty \frac{e^{-AD}(AD)^x}{x!} f(D)dD\,. \qquad (3)$$

It was shown by Stapper et al. [13], that the failure probability can be modeled using a gamma distribution, which results in the following equation

$$Prob\,(X = x) = \frac{\Gamma(x+\alpha)}{x!\,\Gamma(\alpha)} \frac{(A\lambda)^x}{(A\lambda+1)^{x+\alpha}}\,, \qquad (4)$$

where $\alpha = (\sigma\mu)^2$ is the coefficient of the standard deviation and the mean, and $\lambda = D/\alpha$. The parameter α is also called the clustering parameter and a measure of the failure symmetry.
Equation (4) is also known as the negative binominal distribution and is a generalization of the Poisson distribution, which can be derived for the limit $\alpha \to \infty$. Considering that the device fails for a breakdown, the yield is given for X = 0 to be

$$Prob\,(X = 0) = \left(1 + \frac{AD}{\alpha}\right)^{-\alpha}, \qquad (5)$$

which is called the negative binominal yield model. This yield model is already well known and employed to describe the manufacturing yields of VLSI circuits with failures due to particle clustering [14]. The Weibull distribution describing the breakdown statistics can be then equated to the yield model, resulting in

$$e^{-(t/t_{63})^\beta} = \left(1 + \frac{AD}{\alpha}\right)^{-\alpha}, \qquad (6)$$

where the defect density D = D(t) is now considered to be time dependent with D(t=0) = 0. It was shown by Hong et al. [15], that the negative binominal yield model leads to the following area scaling law:

$$A_{acc} = \frac{\alpha^{(1+1/\beta)}}{\Gamma(1+1/\beta)} \int_0^\infty \frac{\left[ln\left(\frac{A_1}{A_2}u+1\right)\right]^{1/\beta}}{(u+1)^{\alpha+1}} du\,, \qquad (7)$$

with

$$u = \frac{A_2}{A_1}\left[e^{\frac{1}{\alpha}(t/t_{63})^{1/\beta}} - 1\right]. \qquad (8)$$

Equation (7) has no analytical solution but can be nummerically calculated. Based on a non-symmetrical defect distribution it only holds true for different device gate length, where the breakdown mainly occurs at the gate-drain region. The negative binominal area scaling result to be less dependent on the length ratio than the Poisson scaling law and can well describe the experimental values for α = 0.5 and β = 1.7. Fitting the Poisson law to the data would have resulted in β ≈ 4, which is unreasonable high for ultra-thin gate dielectrics.

The obtained acceleration factor for different widths exhibited a steeper dependence than the Poisson and binominal yield model. Applying a width scaling, the ratio between the gate area and the effective breakdown area stays constant (inset Fig. 10), which is not taken into account by the negative binominal model. Therefore, eq. (6) cannot reproduce the experimentally obtained data with the given β = 1.7. However, in a first approximation the width scaling can be modeled by considering the effective breakdown area A_{eff} and the remaining gate area ($A-A_{eff}$) and their corresponding defect distribution to be independent. This approach is equivalent to the dual layer percolation model assuming different defect generation rates for the interfacial and high-k layers [16]. Assuming different but constant defect generation rates within both areas, the width scaling is then describable with the standard Poisson scaling law. The shape parameter becomes a composite function of both statistical processes, with a likely major influence of the effective breakdown area scaling. Using eq. (1), the width dependence was found to be well described

by β = 1, which results in a linear correlation between the breakdown time and width:

$$W_{acc} = \frac{t_{63}^{(2)}}{t_{63}^{(1)}} = \frac{W_1}{W_2}. \quad (9)$$

The separate addressing of the acceleration laws for length and width showed to have one limitation. In principle, both laws are only applicable for single transistor scaling. The factors cannot be used to directly scale to a target area A_{target} by simply multiplying W_{target} x L_{target}. As a way to resolve this drawback one can use a certain average length considering the technology node ($L_{target} = L_{avg}$) and apply eq. (7) to obtain the length scaling factor. Then, one needs to sum up the width of all transistors n within the target area ($W_{target} = \sum_{i=1}^{n} W_i$) to calculate W_{acc}. The overall acceleration factor is then given by eq. (2).

The AC gate + off-state stress pattern has revealed some unique feature in comparison to the standard DC and AC TDDB assessments. The newly introduced method can also make an important contribution to a better understanding of circuit reliability, e.g. ring oscillators [17]. There the high frequency together with the continuous switching between gate and off-state stress results in a circuit degradation behavior which cannot always be easily described by the sum of all individual transistors assuming standard DC/AC degradation.

V. CONCLUSION

In summary, the TDDB behavior of 28 nm HKMG nFETs considering realistic CMOS operation was studied by employing different stress patterns. Under DC off-state stress, the dielectric failure was caused by hot carrier injection originating from impact ionization of channel electrons. The time to failure was correlated to the device length, where the source to drain subthreshold current increased for smaller channel length. Higher drain voltages were necessary to obtain similar breakdown times in comparison to gate-only stress. The AC gate + off-state stress pattern was introduced and shows to have similar origin as what was reported for bipolar AC stress. The voltage acceleration was obtained with a lower power law exponent in comparison to AC unipolar gate-only stress. The area scaling did not obey the Poisson distribution due to a primarily localized breakdown at the drain edge. The more general negative binominal distribution, assuming asymmetrical defect generation, was successfully applied to describe the length scaling data.

REFERENCES

[1] T.-C. Chen, "Challenges for silicon technology scaling in the nanoscale era", ESSCIRC, pp. 1-7, 2009.

[2] A. Kerber, K. Maitra, A. Majumdar, M. Hargrove,R. J. Carter and E.A. Cartier, "Characterization of Fast Relaxation During BTI Stress in Conventional and Advanced CMOS Devices With HfO_2/TiNGate Stacks", IEEE TED, vol. 55, pp. 3175 – 3183, 2008.

[3] A. Kerber, A. Vayshenker, D. Lipp, T. Nigam, and E. Cartier, "Impact of charge trapping on the voltage acceleration of tddb in metal gate/high-k n-channel mosfets", IRPS, 2010.

[4] S. Kupke, S. Knebel, G. Roll, S. Slesazeck, R. Agaiby, M. Trentzsch and T. Mikolaijick., "Off-state induced threshold voltage relaxation after PBTI stress", IIRW, 2012.

[5] J. Yuan et al., „Performance elements for 28nm gate length bulk devices with gate first high-k metal gate", IEEE Solid-State and Integrated Circuit Technology (ICSICT), pp. 66-69, 2010.

[6] E. Wu, E. Nowak and W. Lai, "Off-state mode TDDB reliability for ultra-thin gate oxides: New methodology and the impact of oxide thickness scaling", IRPS, 2004.

[7] D. Varghese and A. Muhammad, "Charge Pumping as a Monitor of OFF-State TDDB in Asymmetrically Stressed Transistors", IEEE EDL, vol. 30, no. 9, pp. 972-974, 2008

[8] N. Koike and K. Tatsuuma, "A drain avalanche hot carrier lifetime model for n- and p-channel mosfets", IEEE TDMR, vol.4, no.3, 2004.

[9] D. Varghese et al., "off-state degradation in drain-extended nmos transistors: Interface damage and correlation to dielectric breakdown", IEEE TED, vol. 54, no. 10, pp. 2669-2678, 2007.

[10] Schlunder, S. Aresu, G. Georgakos, W. Kanert, H. Reisinger, K. Hofmann and W. Gustin,"HCI vs. BTI? - Neither one's out", IRPS, 2012

[11] S. Knebel et al., "Influence of frequency dependent time to breakdown on high-k/metal gate reliability", IEEE TED, vol. 60, no. 7, pp. 23, 2013.

[12] K. T. Lee, J. Nam, M. Jin, K. Bae, J. Park, L. Hwang, J. Kim, H. Kim, and J. Park, "Frequency dependent TDDB behaviors and its reliability qualification in 32 nm high-k/metal gate CMOSFETs, IRPS, 2011

[13] C. H. Stapper, "On a composite model to the IC yield problem", IEEE Solid-State Circuits, vol. 10, pp. 537-539, 1975.

[14] C. H. Stapper, "On yield, fault distributions, and clustering of particles", IBM Journal of Research and Development, vol. 30, pp. 326-338, 1986

[15] C. Hong, L. Milor, M. Choi and T. Lin, "Study of Area Scaling Effect on Integrated Circuit Reliability Based on Yield Models", Microelec. Rel., vol. 45, pp. 1305-1310, 2005

[16] T. Nigam, A. Kerber, P. Peumans, "Accurate Model for Time-Dependent Dielectric Breakdown of High-K Metal Gate Stacks", IRPS, 2009

[17] J.-J. Kim et al., "Reliability monitoring ring oscillator structures for isolated/combined NBTI and PBTI measurement in high-k metal gate technologies", IRPS, 2011