TU Delft

Delft University of Technology

Stress Aware Quiescent Current Test Optimization

Shrivastava, S.; Gunnes, J.; Gebregiorgis, A.; Hamdioui, S.

DOI
10.1109/ITC58126.2025.00016

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 2025 IEEE International Test Conference (ITC)

Citation (APA)
Shrivastava, S., Gunnes, J., Gebregiorgis, A., & Hamdioui, S. (2025). Stress Aware Quiescent Current Test Optimization. In L. O'Conner (Ed.), *Proceedings of the 2025 IEEE International Test Conference (ITC)* (pp. 102-110). (Proceedings - International Test Conference). IEEE. https://doi.org/10.1109/ITC58126.2025.00016

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository
as part of the Taverne amendment.

More information about this copyright law amendment
can be found at https://www.openaccess.nl.

Otherwise as indicated in the copyright section:
the publisher is the copyright holder of this work and the
author uses the Dutch legislation to make this work public.

Stress Aware Quiescent Current Test Optimization

Shubhendu Shrivastava
BL-AAES PL-IVN
NXP Semiconductors
Nijmegen, the Netherlands
shubhendu.shrivastava@nxp.com

Jo Gunnes
BL-AAES PL-IVN
NXP Semiconductors
Nijmegen, the Netherlands
jo.gunnes@nxp.com

Anteneh Gebregiorgis
dept. of Quantum & Computer Engineering, TU Delft
Delft, the Netherlands
A.B.Gebregiorgis@tudelft.nl

Said Hamdioui
dept. of Quantum & Computer Engineering, TU Delft
Delft, the Netherlands
S.Hamdioui@tudelft.nl

2025 IEEE International Test Conference (ITC) | 979-8-3315-7041-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ITC58126.2025.00016

***Abstract*—The test escapes due to latent gate oxide (GOx) shorts have been challenging the relentless pursuit of zero defects, despite of voltage stress testing executed to screen such defects. This scenario underscores a prevailing uncertainty in semiconductor testing, "Are we stressing enough?". Moreover, the increasing complexity of digital circuits, coupled with stringent test time requirements, makes 100% fault coverage an unrealistic target.**

This paper presents a solution to optimize voltage stress methodology, and quantify and maximize stress coverage for Integrated Circuits (ICs). The proposed solution involves three key methods. The first method, Critical Thickness Model (CTM), addresses the question "Are we stressing enough?" by determining the minimum stress period of n and p type MOSFET with gate oxide (GOx) thickness in the sub-3nm range. The second method, Stress Coverage Quantification Algorithm (SCQA), assesses actual defect coverage by calculating the percentage of transistors stressed. The third method, Coverage Maximization Algorithm (CMA), aims to reduce customer returns due to GOx shorts by minimizing test escapes. Finally, the paper explores the possibility of Stress Aware ATPG and discusses the trade-offs between under-stressing and over-stressing. The application of CTM resulted in stress time reduction by a factor of 10^3, thereby reducing test cost and improving yield. Furthermore, SCQA reveals that the ATPG reported coverage is overestimated and differs with SCQA by 6.2%. CMA selected patterns resulted into 2.88% higher coverage, reducing voltage stress test escapes by 10%, improving the quality of voltage stress test.

***Index Terms*—latent gate oxide shorts, voltage stress, stress coverage, stress time, defect activation, test cost, yield loss**

I. Introduction

Moore's Law and Dennard's Scaling Law, introduced in 1965 and 1974, respectively, revolutionized the semiconductor industry with scaling trends [1] [2]. Along with increasing transistor count and shrinking sizes, the challenge of screening production defects in highly dense integrated circuits (ICs) is increasing [3]. Concurrently, advancements in the automotive industry is driving a demand for more reliable ICs, especially for higher levels of automation [4]. A significant challenge to achieving higher automation level in the vehicles is the activation of latent GOx defects when devices are deployed in the field [4]. To screen such defects, a widely used technique is the Burn-in test, which has proven effective in identifying defective devices [5]. However, Burn-in has notable drawbacks: it is time-consuming, requires significant infrastructure, can damage products, and is inefficient in terms of time-to-market, often consuming up to 80% of the test time [5].

To address these limitations, the voltage stress (Vstress) methodology was introduced [6]. In this method, a Device Under Test is subjected to a higher voltage level outside its operational range to assess its behavior in extreme conditions [7]. Burn-in and Vstress, both accelerate the early failure phase of the Bathtub curve, but they differ in their execution and efficiency. While Burn-in uses high temperatures over extended period, Vstress involves subjecting the device under test (DUT) to high voltages, with exponentially quicker activation and screening time [8]. Figure 1 demonstrates the

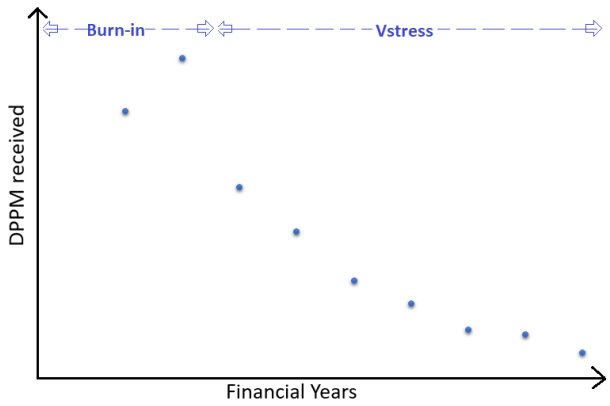

Fig. 1. Vstress v/s Burn-in [9]

impact of Burn-in and Vstress on automotive ICs in terms of GOx returns over nine years. Burn-in resulted in comparatively higher customer returns as compared to Vstress. However, despite these improvements, GOx failures still contribute to customer returns, indicating that the Vstress methodology has reached a saturation point in its efficacy [10]. Figure 2 shows the distribution of gate oxide customer returns over 15 financial quarters, where Vstress was employed. The trend-line indicates that the effectiveness of this methodology is nearing its limit approaching saturation, with only small improvements (shown as the deflection from the saturated trend line) attributed to advancements in the fabrication process and other DfT techniques. This saturation suggests the need for further optimization of the voltage stress methodology to improve defect screening efficiency.

To address the limitation of voltage stress methodology and

Regular Paper

2378-2250/25/$31.00 ©2025 IEEE
DOI 10.1109/ITC58126.2025.00016

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

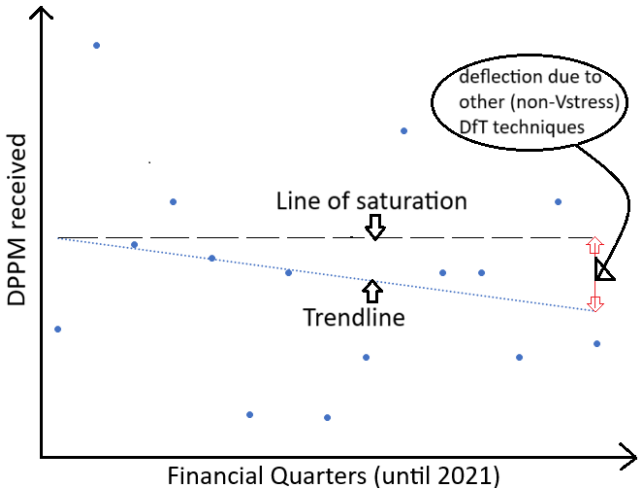

Fig. 2. Saturation in Vstress Efficacy [10]

improve the screening efficiency, this paper presents a structural method to stress that establishes first a relation between stress voltage and minimum stress time that can be used to minimize the stress test time and therefore minimizing stress test cost, following which it present two algorithms which eventually deliver ATPG patterns, that have the maximum transistor level coverage, to execute voltage stress methodology at the tester, minimizing stress test escapes (by 10% for the testchip in experiment).

The forthcoming part of the paper is organized as follows: Section II presents the problem statement; Section III discusses about State of the Art solutions and their shortcomings; Section IV presents novel Critical Thickness Model; Section V provides insights about Stress Coverage Quantification Algorithm; Section VI discusses Coverage Maximization Algorithm; Section VII presents the results and the observations followed by conclusion of this work in Section VIII.

II. Problem Statement

A. Uncertainties regarding Defect Activation

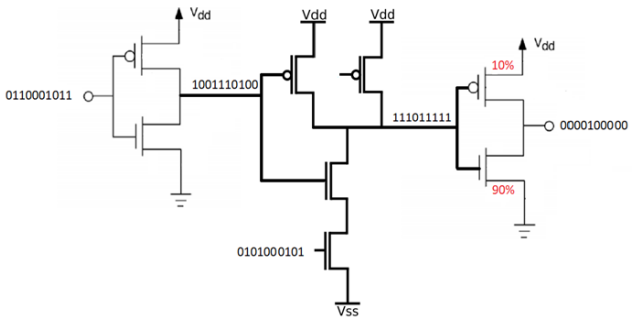

Fig. 3. Schematic of Inverter-NAND-Inverter Circuit

The most widely accepted methodology to screen out the latent GOx shorts is to subject the product to Vstress-Idd$_q$ signature [11] [12]. A Vstress-Idd$_q$ signature is the methodology to accelerate the ageing by applying the extremely high voltage potentials to the IC aiming at eating up the Infant Mortality regime of the Bathtub Curve before delivering the IC into market [13]. This signature is based on Pseudo Stuck-At (PSA) fault model, and hence is preferred in industrial applications as it comes with the coverage of Stuck-At faults [13]. However, it does not say if the fault model is optimal for pattern stressing. Consider the sample circuit in figure 3. The stress coverage of the GOx of the pull-up in the third CMOS stage illustrates how the Pseudo Stuck-At fault model may not provide adequate stress coverage, leading to potential test escapes due to insufficient activation of latent defects and keeping the question open, “Are we stressing enough?”. And on the other hand, there is a probability that the useful lifetime of the pull-down in the same CMOS stage is being eaten up by the same pattern. Conclusively, not only the fault model is not optimal to execute voltage stress methodology as not all the transistors are stressed but also among the stressed transistors there are some which got stressed for 10% of the time and some got stressed for 90%, leading us to an uncertainty, how much stress efforts are optimal for screening latent defects.

B. Meagre Stress Coverage of the Screening Test

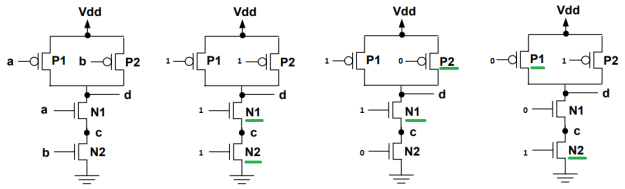

Fig. 4. (a) NAND schematic with (b) exhaustive PSA pattern {1,1} (c) exhaustive PSA pattern {1,0} (d) exhaustive PSA pattern {0,1} on its inputs a,b underlining the stressed MOSFETs [14]

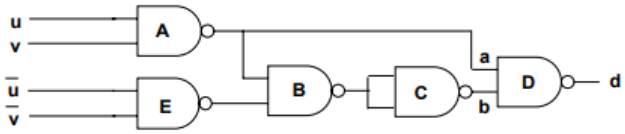

Fig. 5. A NAND based circuit [14]

The exhaustive Pseudo Stuck-At pattern set of a 2-input NAND gate as presented in figure 4 is {(1,1) (1,0) (0,1) }. However, there is no way to stimulate the combination in figure 4(d) featuring 0 and 1 respectively on a and b inputs of the 2-input NAND gate D as shown in figure 5, categorizing the applied PSA pattern as complete but not exhaustive. The complete PSA pattern set therefore fails at stressing MOSFET P1 of NAND gate D in figure 5. This observation denotes that there can be significantly more number of MOSFETs in the circuitry experiencing 0% stress when subjected to Vstress-Idd$_q$ signature, resulting into compromised stress coverage, which is not reported out by ATPG.

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

C. Unstructured way of pattern selection to execute Vstress-Idd$_q$ signature

Based on complete Pseudo Stuck At fault model, of which the shortcomings at the transistor level have been discussed in II.B, ATPG generates n number of patterns for the gate level design. Out of those n patterns, a few are used arbitrarily to execute Vstress-Idd$_q$ signature on an IC. However, there exists no structured solution to figure out which patterns shall be sent to ATE to execute the Vstress-Idd$_q$ signature to cover maximum transistors and to ensure minimum stress test escapes.

Recapitulating, the Vstress-Idd$_q$ signature needs to be optimized due to three prominent problems:

- Improper defect activation (due to Pseudo Stuck-At fault model, voltage potential or low activation time)
- Unknown stress test coverage at transistor level
- Unstructured way of pattern selection

III. THE STATE-OF-THE-ART

In 1988, Lee et al. presented an extrinsic defect-related breakdown model called Thinning Model to model the GOx defects in a MOSFET [15]. These extrinsic defects can lead to a higher local electric field, higher intrinsic defect generation, or higher current density [15]. Thinning model states that the presence of an extrinsic defect will deplete the oxide thickness X_{ox} by an amount of ΔX_{ox}. A term "severity of defect" was introduced in [15] which is measured as $\frac{\Delta X_{ox}}{X_{ox}}$. This model states that a GOx short reduces the effective thickness, and by employing the concept of the 1/E model, calculates the time to breakdown [15].

In 1999, following a similar direction as that in [15], but on a different path, H. Katto proposed an approach that also fundamentally focused on probing the effective oxide thickness, but through voltage-to-breakdown (V_{bd}) measurements [16]. H. Katto proposed a model that relies on the assumption that a device having a certain V_{bd} would have a short time to breakdown τ_0 when subjected to the stress potentials of V_{bd} [16]. Here, τ_0 represents the applied stress time and is constant for all devices. The V_{bd} measurements were taken using an idealistic approach, where the breakdown voltages of devices with different thicknesses were recorded when subjected to varying electric fields. Then the V_{bd} was selected according to the acceptable values of electric field across the gate oxide, and under the aforementioned assumptions, modifying the E model, minimum stress time was calculated in [16].

Further in 2018, Wim Dobbelaere et al., adding to the direction presented in [15], defined a parameter of activation for each latent defect, which can be calculated as :

$Activation = \int_0^{\text{Test Time}} \frac{1}{\text{MTF}}\, dt$ [17]

where, if the activation parameter of a latent defect scores 1 or more, the latent defect should be considered as activated. Here, the parameter MTF represents the function for modeling the breakdown over time of the latent defect, notably based on the famous 1/E model of the TDDB mechanism, which is equal to the τ_{bd} value calculated in [15] [17].

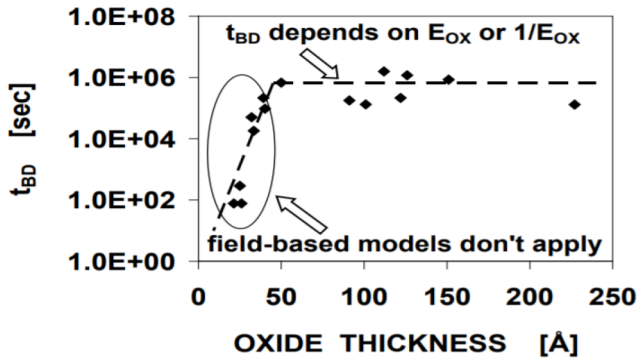

Fig. 6. Compiled TDDB data for different oxide thickness [18]

Putting in a nutshell, the three contributions in the direction of probing the GOx defects, whether using breakdown voltage measurements or time to breakdown, were fundamentally dependent on the field-driven empirical breakdown models E and $1/E$.

Field-driven TDDB models exhibit a key property: for a given gate oxide area and test temperature, TDDB occurrence is independent of oxide thickness [18]. Figure 6 shows a set of TDDB data compiled by McPherson referenced from various sources for constant of 8MV/cm electric field across GOx and 125 deg C test temperature. In the range of GOx thickness from 5nm to 25nm, the field-driven models hold true. However, for oxides thinner than 5nm, a significant deviation is observed. The breakdown time diverges, suggesting that factors beyond the electric field are influencing oxide breakdown [18]. This challenges the applicability of field-driven models for sub-5nm oxides. Sune et al. in [19] have demonstrated that oxide breakdown in ultra-thin gate oxides is energy-dependent rather than thickness-dependent. The widely accepted Power-Law Model, based on the Anode Hydrogen Release (AHR) physical breakdown model, typically applies at high voltage potentials [20] [21]. However, in ultra-thin SiON oxides, Nicollian [18] showed that AHR is valid for trap generation and breakdown at low stress potentials ($< 3V$) and not high stress potentials due to the release of hydrogen species (H+ and H0) from the anode caused by vibrational excitation from tunneling electrons. Referring figure 7, V Model, based on physical mechanisms of Bulk Trap Generation, similar to Power-Law Model, does not strongly depend on thickness for oxides thinner than 3nm [22].

In conclusion, neither Electric Field driven nor Energy-Driven dielectric conduction models effectively estimate breakdown voltage or time to breakdown for ultra-thin gate oxides. The problem lies in the energy of the tunneling carriers, not the oxide thickness, which causes defect generation leading to breakdown. Thus, a new model, with a stronger dependence on oxide thickness or energy-related physical parameters, is required for these ultra-thin oxides.

The work presented in further sections was carried out in

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

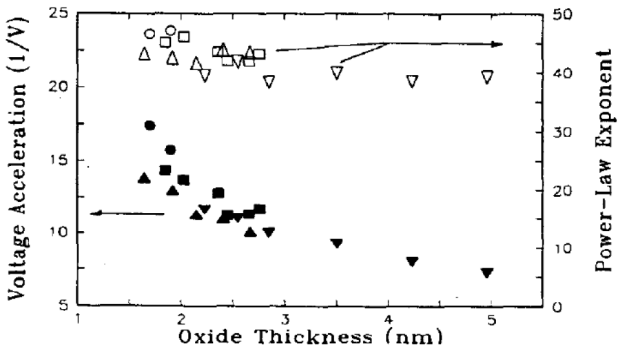

Fig. 7. V-Model & Power Law Model v/s oxide thickness [22]

2020-2021, and we are glad to see that in the meantime, a research has been conducted by S. Eggersglüß and A. Glowatz in 2024 that proposed an approach based on Cell-aware Transistor State Stress Model (TSSM) [31]. There, Siemens, with the access to ATPG algorithm, reports out the transistor level coverage for both Elevated Voltage Stress (EVS) and Dynamic Voltage Stress (DVS). The approach uses a TSSM view of a cell for analysis and then executes Idd_q and Toggle patterns to check the stress coverage [31]. Section VII discusses about the comparison of this approach with one of the outcomes of this paper, i.e., Stress Coverage Quantification Algotithm.

IV. Critical Thickness Model

The Critical Thickness Model adopts the approach from Dobbelaere et al. [24] for probing gate oxide thickness to ensure proper defect activation, discarding the 1/E model for time-to-breakdown estimation. Instead, it advocates using the Direct Tunneling Model for time-to-breakdown, aligning with the Thinning Model. Since the Direct Tunneling current's tunneling length is 3-4nm, it inevitably flows from anode to cathode of SiON dielectric-based automotive products considered in the paper [23]. Given the low energy required for tunneling electrons, aligned with the product's low stress potential ($< 3V$), the Direct Tunneling current was selected as a basis for Critical Thickness Model being highly sensitive to oxide thickness, making it crucial for probing defect activation. The model is also compatible with various gate and dielectric materials, including high-k gate stacks, provided the equivalent oxide thickness is used when talking about the oxide thickness.

A. Assumptions

The Critical Thickness Model is based on these assumptions:

- GOx of 130nm node technology and smaller experience similar types of shorts, caused by protruding poly/semiconductor, crystalline defects in the substrate, surface roughness at weak spots, and not foreign particles contamination. This is supported by NXP Quality Department observations, which cite advancements in fabrication and cleanroom processes that minimize particle contamination.
- Intrinsic defect density is uniform across the wafer, depending on silicon wafer quality, cleanliness, and oxidation annealing conditions.
- Tunneling electrons create a trapezoidal barrier. For thin oxides ($t_{ox} < 10nm$), electrons transition from Direct Tunneling to Fowler-Nordheim Tunneling (FNT) when stress voltage is high enough, independent of oxide thickness. This is supported by analysis in [24] [25].
- Image force-induced potential barrier lowering is neglected, as it can yield inaccurate results for thin oxides [26]
- As CTM relies on the Philips MOS Model 11 and Thinning Model, the assumptions of these models form the foundation of Critical Thickness Model.

B. The Model

The Critical Thickness Model (CTM) describes gate oxide breakdown due to intrinsic TDDB caused by extrinsic critical latent gate oxide shirts. While technological advances have reduced the number of gate oxide shorts, they still persist [27]. However, not all shorts need to be screened; only the critical latent gate oxide shorts, which have the potential to cause device failure during field operation or testing. The criticality depends on the product's desired lifetime (e.g. a minimum of 10 years for automotive products). CTM defines the concept of critical thickness $t_{ox.critical}$, which helps distinguish between critical and non-critical shorts. As shown in figure 8, different

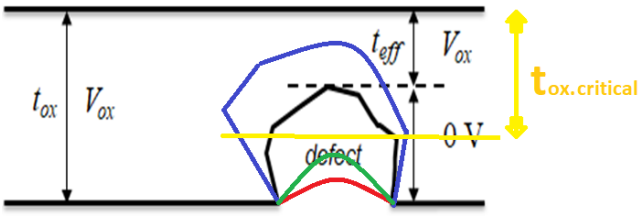

Fig. 8. Critical Thickness Model

gate oxide shorts vary in size and deplete the oxide thickness to an effective thickness, $t_{eff} = t_{ox} - t_{defect}$. If the latent short exists with the thickness of $t_{ox.critical}$, it is categorized as critical and must be screened out. This model assumes that each MOSFET has a latent gate oxide short that depletes the thickness to $t_{ox.critical}$. The critical thickness depends on application and device physics, varying for pMOS and nMOS at a specific oxide thickness and process node. The calculation of $t_{ox.critical}$ requires understanding of sub-3nm dielectric breakdown mechanisms. CTM is applicable only for gate oxides with thicknesses in the sub-3nm range. As discussed in previous section, Direct Tunneling (DT) is the dominant failure mechanism in such thin oxides. Tunneling

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

electrons can lead to intrinsic defects like electron traps, interface states, and trap creation processes. The applied stress potential ($< 3V$) triggers Direct Tunneling, invoking trap creation. Once defect density reaches a critical level, it causes catastrophic breakdown [23] [25]. The tunneling electrons, with energies over 2.3 eV, cause breakdown as the current is inversely proportional to oxide thickness. The breakdown time $t_{bd} = Q_{bd}/J_{DT}$ [22], is used to estimate the critical thickness. CTM relates the time to breakdown of a good GOx to one with critical defects, providing a ratio between original and minimum desired operational lifetimes at nominal voltage and temperature.

$$\frac{t_{bd.life}}{t_{bd.crit}} = \frac{Q_{bd.life} * J_{DT.crit}}{J_{DT.life} * Q_{bd.crit}} \tag{1}$$

where,
$t_{bd.life}$ is expected lifetime of a good GOx
$Q_{bd.life}$ is charge to breakdown for a good GOx
$J_{DT.life}$ is Direct Tunneling current through a good GOx
$t_{bd.crit}$ is minimum desired lifetime of the product
$Q_{bd.crit}$ is charge to breakdown for a critically defected GOx
$J_{DT.crit}$ is Direct Tunneling current through a critically defected GOx and

$$Q_{bd} = q * N_{bd}/P_g \tag{2}$$

where, q is charge of the tunneling carrier,
N_{bd} is critical intrinsic defect density,
P_g is intrinsic defect generation rate

Stathis and DiMaria, in their experiment on a 3nm gate oxide with 2V of gate voltage, showed that the intrinsic defect generation rate (Pg) is independent of oxide thickness and exponentially depends on gate voltage [25]. Their experiment also established that the critical intrinsic defect density at breakdown is linearly dependent on oxide thickness until 3nm, after which it becomes independent of thickness. This result implies that charge to breakdown (Qbd) is unaffected by oxide thickness and strongly dependent on gate voltage. Consequently, the charge to breakdown for both good and critically defected gate oxides is equal when operating at nominal voltage and temperature. Substituting this piece of information in equation 1, we get :

$$\frac{t_{bd.life}}{t_{bd.crit}} = \frac{J_{DT.crit}}{J_{DT.life}} \tag{3}$$

Using the Philips MOS Model 11 [28], the equation 3 can be further expanded as shown in the following equation :

$$\frac{t_{bd.life}}{t_{bd.crit}} = \frac{t^2_{ox.life}}{t^2_{ox.crit}} * exp(-B * t_{ox.life}(\frac{(\oint_b - V_{ox.life} * q)^{3/2} - \oint_b^{3/2}}{V_{ox.life} * \oint_b^{3/2}}) * exp(B * t_{ox.life}(\frac{(\oint_b - V_{ox.life} * q)^{3/2} - \oint_b^{3/2}}{V_{ox.life} * \oint_b^{3/2}}) \tag{4}$$

where,
$t_{ox.life}$ is thickness of a good GOx
$t_{ox.crit}$ is thickness of critically defected GOx
$V_{ox.life}$ is nominal voltage across the GOx
β is $\frac{-4\sqrt{2m_{ox}}\oint_b^{3/2}}{3hq}$

From (4), we calculate the critical thickness that classifies a latent gate oxide short as critical or non-critical. With the value of parameter of critical thickness being known, the following step calls for determining the minimum stress time required to activate the critical latent gate oxide defect under a fixed stressed voltage. CTM establishes a ratio between the time to breakdown of a good gate oxide at operational voltage and temperature, and the time to breakdown of a critically defected gate oxide at stress voltage. Substituting Philips MOS MODEL 11 for Direct Tunneling current in equation 1, we get :

$$\frac{t_{bd.life}}{t_{bd.crit}} = \frac{V^2_{ox.crit} * t^2_{ox.life}}{V^2_{ox.life} * t^2_{ox.crit}} exp(-B * t_{ox.life}(\frac{(\oint_b - V_{ox.life} * q)^{3/2} - \oint_b^{3/2}}{V_{ox.life} * \oint_b^{3/2}}) * exp(B * t_{ox.life}(\frac{(\oint_b - V_{ox.life} * q)^{3/2} - \oint_b^{3/2}}{V_{ox.life} * \oint_b^{3/2}}) * \frac{\sqrt{V_{ox.life} * (q/\oint_b) + 1} - 1}{\sqrt{V_{ox.crit} * (q/\oint_b) + 1} - 1}^2 * \frac{q_{bd.life}}{q_{bd.crit}} \tag{5}$$

Equation (5) provides $t_{bd.crit}$, the breakdown time for the critically defected gate oxide or the minimum stress time to activate the critical latent gate oxide short. The minimum screen time will differ for pMOS and nMOS due to their physical properties. Additionally, Vstress-Idd$_q$ signature is conducted under temperature stress to enhance the impact, and (3) includes the stress temperature parameter in the form of charge to breakdown (Q_{bd}), as Q_{bd} is temperature-dependent [29] [30].

Summarizingly, Critical Thickness Model fundamentally comprises of two equations:

- equation (4) : to find out the critical thickness of the gate oxide
- equation (5) : to deduce the minimum screen time based on the critical thickness

and hence, profiles the latent gate oxide short activation when the device is subjected to Vstress-Idd$_q$ signature in presence of temperature stress. Therefore, only if a device experiences stress for the minimum stress time for a certain stress voltage and stress temperature, it shall be considered as stressed.

V. Stress Coverage Quantification Algorithm

Critical Thickness Model addresses the first concern of improper defect activation. The next step in optimizing the Vstress-Idd$_q$ signature is determining how many devices in

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

the DUT actually were stressed, undrestressed, overstressed, addressing the second concern of meagre fault coverage.

Stress Coverage Quantification Algorithm (SCQA) works on the image of transistor level schematic captured during the capture cycle for each pattern. Following are the I/Os of SCQA:

- Input Configuration File : This is a text file which is Python-readable and prompts user to provide the complete paths to flat netlist, and hierarchical netlist and the capture time based on the minimum stress time calculated from Critical Thickness Model
- Output Text File : This is a text file which is written by the algorithm and contains the information about:
  - Stress coverage of each pattern
  - Detailed information of the devices that got stressed for each pattern
  - Stress coverage of whole pattern set
- Output Database : This database contains complete profile of applied stress on each device by the complete pattern set. The database lets the user know for how much time each transistor has been stressed, how many devices got stressed for 0%, 10%, 20%,..., 100% of the times, and most importantly it categorizes each device as stressed, overstressed or unstressed according to the minimum stress time read in, which is calculated by Critical Thickness Model.

Figure 9 shows the block diagram of Stress Coverage Quantification Algorithm.

VI. Coverage Maximization Algorithm

As not all patterns are executed at ATE due to test time constraints, Coverage Maximization Algorithm (CMA) sorts ATPG-generated patterns based on maximum transistor coverage. Based on the Greedy algorithm (Dijkstra's), CMA maximizes stress coverage by minimizing repeated devices. CMA has two inputs and an output:

- Pattern Set Stress Database: Tracks how often each MOSFET is stressed in the pattern set, generated by SCQA.
- Per Pattern Stress Database: Provides a stress profile for each pattern, also generated by SCQA.
- Maximized Coverage Pattern Set: An output text file containing a pattern set with maximum coverage at transistor level.

Figure 10 shows the block diagram of Coverage Maximization Algorithm

VII. Results and Discussion

Validation of Critical Thickness Model involves determining minimum stress time based on critical thickness and validating it through device-level simulations. These simulations calculate critical thickness and minimum stress time, with TCAD (Sentaurus) used to model device characteristics in the absence of experimental data. Critical thickness and minimum stress time, defined by equations (4) and (5), depend on device parameters. Finite Element Modeling within TCAD

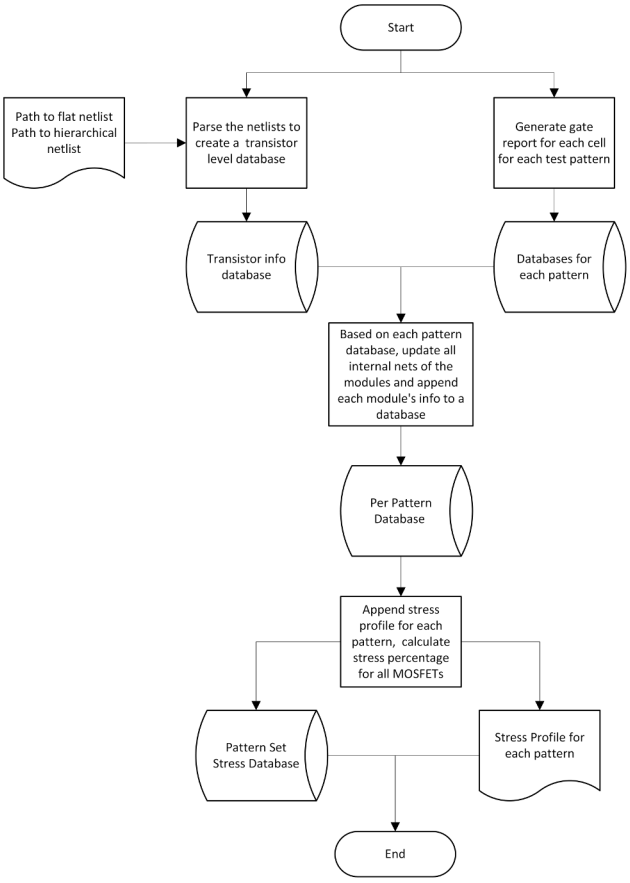

Fig. 9. SCQA flow chart

was used to solve the equations (4) and (5) as TCAD solves fundamental, physical and partial differential equations for discrete geometrics. For long-channel devices, use Brew's model, while the 2D charge sheet model applies for short-channel devices. Critical Thickness Model calculated critical thickness for 3nm thick SiON GOx was reported to be 2.4 nm for a given stress voltage for the application lifetime of 10 years. The ratio of calculated minimum stress time to the in-practice stress time was reported to be in order of 10e-3 seconds, for stressing a transistor (both p and n type), when stressed at a certain stress voltage ($< 3V$).

Simulation of SCQA was performed at a 130nm test chip. An ATPG framework was used to generate 25 Pseudo Stuck-At test patterns, which were then processed by SCQA. SCQA calculated coverage (transitor level) for these patterns differed from ATPG coverage (gate level) by 6.2%. While at gate level, ATPG reported coverage as 80.86%, SCQA reported it to be 75.82% at the transistor level. The difference of coverages at transistor level and cell level is concerning for automotive applications targeting SAE level 4 and reveals the need of Stress Aware fault model for patterns to execute Vstress-Idd$_q$ signature. SCQA also categorized over 90% of

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

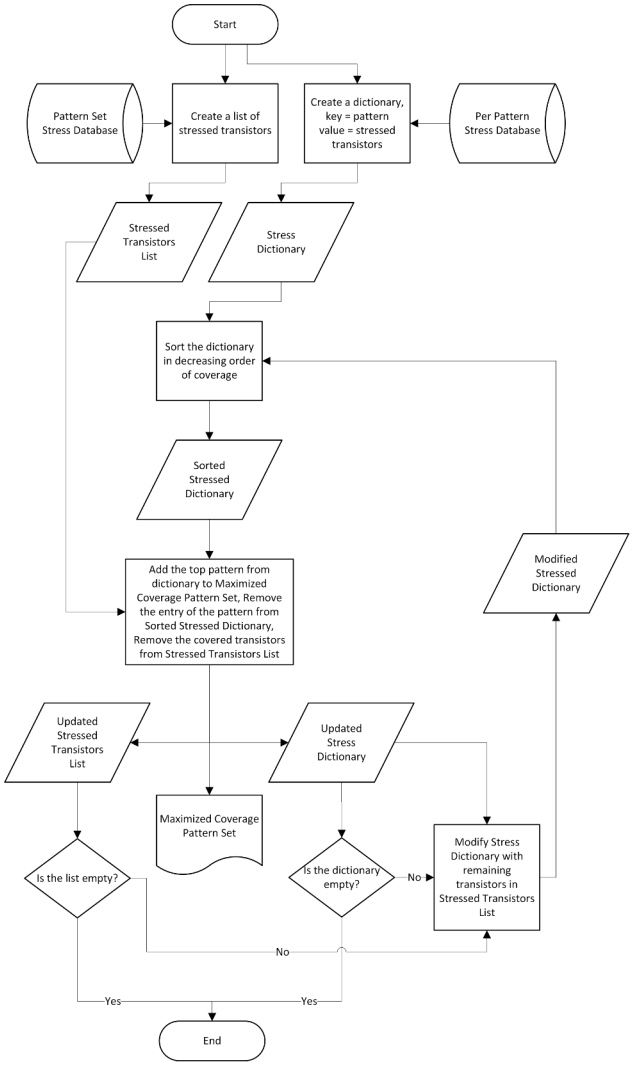

Fig. 10. CMA flow chart

stressed transistors as overstressed (eating useful lifetime and activating non-critical latent GOx shorts). With these results and the observations from the algorithm's way of working, it can be observed that while SCQA and TSSM, both are exhaustive in nature at the gate level using transistor-level simulation to identify specific stress-inducing states, SCQA with the limitation of not having the access to ATPG algorithm, reports out only steady state stress detecting latent GOx shorts coverage but Siemens presented TSSM can report out latent shorts and opens for EVS and DVS [31]. Table I tabulates the differences between SCQA and the algorithm using TSSM.

For the same test chip, Coverage Maximization Algorithm (CMA) was executed to optimize the test pattern set. The top 10 patterns delivered as Maximized Coverage pattern set by CMA showed a stress test coverage (at the transistor level) of 73.96% when compared to 71.08% of (transistor

TABLE I
COMPARISON BETWEEN SCQA AND TSSM BASED APPROACH

Feature	**Stress Coverage Quantification Algorithm (SCQA)**	**Transistor State Stress Model (TSSM)**
Objective	To report EVS coverage at the transistor level	To report EVS and DVS coverage
Inputs	Design netlist, CDS views, ATPG patterns	Design netlist, UDFM, Spice views, ATPG patterns
Target	Latent GOx Shorts	Latent Shorts and Opens
Outputs	Transistor level EVS coverage	Transistor level EVS and DVS coverage
Test Patterns	ATPG generated PSA pattern set	ATPG generated Toggle and PSA pattern set

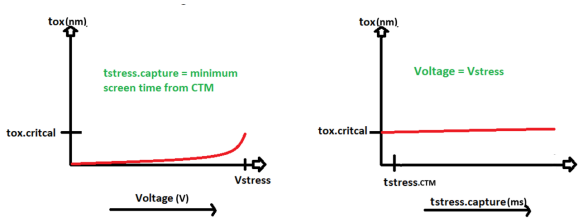

Fig. 11. Time vs Voltage impact on targeting critical thickness

level) coverage of arbitrarily selected patterns. The comparison indicates a 2.88% of coverage enhancement (adding 288,000 more transistors to the coverage for a IC of 10M transistors), when opting for CMA. Loss of this coverage would have resulted in approximately 10% of stress test escapes at the transistor level.

These results bring us to three questions for discussion, "How should we resolve the overstressing and understressing with the help of Critical Thickness Model?", "Like a minimum stress time, is there a need of maximum threshold on stress time ?", "What is the application of Critical Thickness Model?". Let's take them one by one. When a GOx experiences overstress, this invites unwanted activation of non-critical latent defects which if follow irreversible behavior can eventually lead to high gate leakage current. On the flip side, when a GOx experiences understress, the critical latent GOx shorts are missed. From reliability point of view, stressing is understood as an ageing accelerating event which can lead to device degradation, which may eventually cause the breakdown of the device. However, from testing point of view, stressing is an event where time zero non-critical defects become critical (get activated). The impact of stress time on stressing is almost negligible when compared to voltage potential as stress experienced is exponentially dependent on voltage and linearly dependent on stress time. For visual representation, please refer plots in figure 11 which are completely based on in-house calculations performed with a device physicist. Therefore, overstressing and understressing should be understood in terms of stress voltage and not stress time, to ensure that a transistor has experienced the stress

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

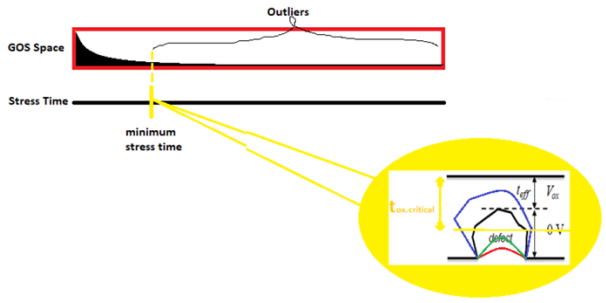

Fig. 12. Latent Gate Oxide Shorts' Trend when stressed

voltage potential for at least CTM provided minimum stress time, guaranteeing the critical latent GOx short screening. However, if there is a limitation of maximum stress voltage, in that case CTM can be used in ensuring the impact of Vstress-Idd$_q$ Signature. This brings us to the aforementioned second question. Indeed the additional impact of stress is low when stressed for higher time than CTM calculated time, however, risk of unintentional activation of non-critical gate oxide shorts is also not zero. Figure 12 shows that although within CTM calculated minimum stress time, all the critical latent GOx shorts are screened out, stressing beyond will still show some failures. However, such failures are yield loss and unintentionally activated non-critical defects, that would never manifest as a defect under normal operational conditions during the lifetime and hence depicted as outliers. Screening out such outliers will only contribute to test cost and yield loss. Moving on to the last question, the in-practice stress time calculation method is entirely based on the maximum allowed stress period calculated from aging acceleration perspective in such a way that at certain voltage Vstress, for a time t_{stress}, only a certain percentage of the lifetime is consumed. Therefore, stress time for each pattern $t_{pattern.stress}$ as of now is calculated as:

$$t_{pattern.stress} = \frac{t_{stress}}{n}$$

where, n is number of ATPG generated patterns, to be used to execute Vstress-Idd$_q$ signature. Although the method has been asserted as good enough going by the falling trends of customer returns (not zero yet, refer section II), the very primitive flaw with this method of deducing stress time is that it does not say anything about whether the allowed aging of the transistor targets critical latent GOx shorts or causes yield loss. To answer such questions, Critical Thickness Model shall be employed. The model renders the information about how much time shall a transistor be stressed to get rid of critical latent GOx shorts. With that particular information in use, $t_{pattern.stress}$ becomes equal to CTM calculated minimum stress time of pMOS (as it is higher for pMOS than nMOS), minimizing both test cost and yield loss while ensuring screening of critical latent gate oxide shorts.

VIII. Conclusion

This paper talks about three key challenges in stress testing of digital ICs, namely improper defect activation (due to Pseudo Stuck-At fault model, voltage potential or low activation time), unknown stress test coverage at transistor level, and unstructured way of pattern selection. Addressing them, this paper focuses on minimizing stress time to activate latent gate oxide shorts, identifying unstressed transistors, and enhancing transistor level stress test coverage at ATE. The contributions of this work include the first attempt to model gate oxide breakdown from a Design-for-Test perspective (screening devices in infant mortality period of the Bathtub curve while limiting test cost), the first effort to quantify stress quality at the transistor level of teh entire design and the first attempt to sort the patterns to execuet stress based on transistor profile. To develop Critical Thickness Model, an extensive review of empirical and physical TDDB models was conducted, along with a study of the applicability of reliability assessment models in the Infant Mortality regime of the Bathtub curve. It was established that existing defect activation models fail to capture the impact of acceleration parameters, which are crucial for effective testing for the gate oxides thinner than 3nm - 4nm. This led to the development of Critical Thickness Model, which incorporates oxide thickness-dependent transport, defect generation, and breakdown mechanisms. The Critical Thickness Model enabled the creation of Stress Coverage Quantification Algorithm (SCQA) to quantify stress test coverage at the transistor level. The SCQA revealed significant gaps in gate level ATPG coverage, motivating the development of a Coverage Maximization Algorithm to ensure maximum stress test coverage within minimal time, reducing test escapes by approximately 10% (adding 288,000 transistors more to coverage for an IC of 10 M transistors).

This paper also presents critical discussions to redefine the understanding of overstressing and understressing in both reliability and defect activation contexts. The thesis provides a foundation for further optimization of the static Vstress-Idd$_q$ signature, with the potential for higher stress test coverage and reduced test escapes (lower test cost and yield loss). On the basis of the results reported by SCQA, a potential future work could be to explore how to develop a fault model for Stress-Aware Pattern Generation as Pseudo Stuck-At fault model does not ensure optimal stressing (also concluded in II-B). A problem that might be faced to develop such model for Stress-Aware pattern generation is that it will consume time which will be outside the acceptable limits of practicality, like Leakage patterns. To overcome this challenge, the idea is to realize the concept of Fault Coning at the transistor level, where the circuit can be distributed in several small cones according to the propagating effect of the real defects and testing on several cones can be performed in parallel to reduce the time consumption. Here, the parallel patterns are expected to ensure activating as many MOSFETs as possible for the stress time calculated by Critical Thickness Model. Another idea is to explore the possibility of realizing Toggle

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.

fault model at transistor level where idea is primarily neither to have compares on Primary Outputs or Scan Outputs nor to have a higher gate level coverage but an approach to target activation ($V_{GS} \neq 0$) of maximum transistors, facility to measure quiescent current during capture mode, followed by a report out from the tool about how many transistors could the pattern set actually activate.

REFERENCES

[1] R.R Schaller, "Moore's law: past, present and future," IEEE Spectrum, vol. 34, pp. 52–59, April 1997.

[2] M. Bohr, A 30 Year Retrospective on Dennard's MOSFET Scaling Paper, vol. 12. IEEE Solid-State Circuits Society Newsletter, Winter 2007, pp.11–13.

[3] M. S. Gaur, V. Laxmi, M. Zwolinski, M. Kumar, N. Gupta and Ashish, "Network-on-chip: Current issues and challenges," 2015 19th International Symposium on VLSI Design and Test, Ahmedabad, India, 2015, pp. 1-3

[4] E. Sperling, Unknowns driving up the cost of auto IC reliability, Feb. 2022, [online] Available: https://semiengineering.com/unknowns-driving-up-the-cost-of-auto-reliability/

[5] M. P. -L. Ooi, Z. A. Kassim and S. N. Demidenko, "Shortening Burn-In Test: Application of HVST and Weibull Statistical Analysis," in IEEE Transactions on Instrumentation and Measurement, vol. 56, no. 3, pp. 990-999, June 2007.

[6] J. Yim et al., "A Prevenient Voltage Stress Test Method for High Density Memory," 4th IEEE International Symposium on Electronic Design, Test and Applications (delta 2008), Hong Kong, China, 2008, pp. 516-520, doi: 10.1109/DELTA.2008.93. keywords: Stress;Electronic equipment testing;Degradation;Life estimation;Circuit testing;Breakdown voltage;Voltage control;Random access memory;Temperature;Packaging;voltage stress test;acceleration factor;burn-in test;junction temperature;reliability;constant voltage stress;voltage ramp stress/

[7] M. F. Zakaria, Z. A. Kassim, M. P. . -L. Ooi and S. Demidenko, "Reducing burn-in time through high-voltage stress test and Weibull statistical analysis," in IEEE Design & Test of Computers, vol. 23, no. 2, pp. 88-98, March-April 2006.

[8] R. Kawahara, O. Nakayama and T. Kurasawa, "The effectiveness of IDDQ and high voltage stress for burn-in elimination [CMOS production]," Digest of Papers 1996 IEEE International Workshop on IDDQ Testing, Washington, DC, USA, 1996, pp. 9-13

[9] F. Huisman, Quality Engineer, BL-AA. PL-IVN, NXP Semiconductors

[10] H. Gerritsen, Quality Manager, BL-AA. PL-IVN, NXP Semiconductors.

[11] D.Fernandez, Voltage Margining using the TPS 62130, https://www.ti.com/lit/an/slva489/slva489.pdf

[12] R.M. Kho, A.J. Moonen, V.M. Girault, J. Bisschop, E.H.T. Olthof, S. Nath, Z.N. Liang, Determination of the stress level for voltage screen of integrated circuits, Microelectronics Reliability, Volume 50, Issues 9–11, 2010, Pages 1210-1214, ISSN 0026-2714

[13] T. Kurasawa, R. Kawahara and O. Nakayama, "The Effectiveness of IDDQ and High Voltage Stress for Burn-in Elimination," in IDDQ Testing, IEEE International Workshop on, Washington, DC, 1996, pp. 9

[14] S. T. Zachariah and S. Chakravarty, "A comparative study of pseudo stuck-at and leakage fault model," Proceedings Twelfth International Conference on VLSI Design. (Cat. No.PR00013), 1999, pp. 91-94, doi: 10.1109/ICVD.1999.745130.

[15] J. Lee, I. -. Chen and C. Hu, "Statistical modeling of silicon dioxide reliability," 26th Annual Proceedings Reliability Physics Symposium 1988, 1988, pp. 131-138, doi: 10.1109/RELPHY.1988.23440.

[16] H. Katto, "Breakdown voltage distribution and extrinsic TDDB failures of MOS gate oxides," 1999 IEEE International Integrated Reliability Workshop Final Report (Cat. No. 99TH8460), 1999, pp. 85-91, doi: 10.1109/IRWS.1999.830564.

[17] W. Dobbelaere et al., "Applying Vstress and defect activation coverage to produce zero-defect mixed-signal automotive ICs," 2019 IEEE International Test Conference (ITC), Washington, DC, USA, 2019, pp. 1-4

[18] Nicollian, P. E. (2007). Physics of Trap Generation and Electrical Breakdown in Ultra-thin SiO2 and SiON Gate Dielectric Materials. University of Twente.

[19] J. Sune, et al. Phys. Stat. Sol. (a), V. 1 1 I , p.675, 1989.

[20] Wu, Ernest & Vayshenker, A. & Nowak, E.J. & Sune, Jordi & Vollertsen, R.-P & Lai, Weihua & Harmon, Dasia. (2003). Experimental evidence of T/sub BD/ power-law for voltage dependence of oxide breakdown. Electron Devices, IEEE Transactions on. 49. 2244 - 2253. 10.1109/TED.2002.805606.

[21] J. Sune and E. Y. Wu, "Mechanisms of hydrogen release in the breakdown of SiO/sub 2/-based gate oxides," IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest., 2005, pp. 388-391, doi: 10.1109/IEDM.2005.1609359. [57] Temperature dependence of electrical overstress,

[22] E. Y. Wu et al., "Voltage-dependent voltage-acceleration of oxide breakdown for ultra-thin oxides," International Electron Devices Meeting 2000. Technical Digest. IEDM (Cat. No.00CH37138), 2000, pp. 541-544, doi: 10.1109/IEDM.2000.904375.

[23] J. Appl. Phys. 98, 121301 (2005); doi: 10.1063/1.2147714

[24] DiMaria, D. J., & Stasiak, J. W. (1989). Trap creation in silicon dioxide produced by hot electrons. Journal of Applied Physics, 65(6), 2342–2356. doi:10.1063/1.342824

[25] J. H. Stathis and D. J. DiMaria, "Reliability projection for ultra-thin oxides at low voltage," International Electron Devices Meeting 1998. Technical Digest (Cat. No.98CH36217), 1998, pp. 167-170, doi: 10.1109/IEDM.1998.746309.

[26] Juan C. Ranuárez, M.J. Deen, Chih-Hung Chen, A review of gate tunneling current in MOS devices, Microelectronics Reliability, Volume 46, Issue 12, 2006, Pages 1939-1956, ISSN 0026-2714, https://doi.org/10.1016/j.microrel.2005.12.006.

[27] B. Kruseman, Senior Scientist, Design Enablement, NXP Semiconductors

[28] R. Van Langevelde, "Physical Background and parameter extraction for MOS Model 11", Nat.Lab, Unclassified Report 2002/810, April 2003.

[29] J. J. Tzou, C. C. Yao, R. Cheung and H. Chan, "Temperature dependence of charge generation and breakdown in SiO2," in IEEE Electron Device Letters, vol. 7, no. 7, pp. 446-448, July 1986

[30] Chi-Hung Lin, J. Cable and C. S. Woo, "Temperature and electric field characteristics of time-dependent dielectric breakdown for silicon dioxide and reoxidized-nitrided oxides," in IEEE Transactions on Electron Devices, vol. 42, no. 7, pp. 1329-1332, July 1995

[31] S. Eggersglüß and A. Glowatz, "A Cell-aware Transistor State Stress Model and its Application for Quality Measurement," 2024 IEEE International Test Conference (ITC), San Diego, CA, USA, 2024, pp. 41-45, doi: 10.1109/ITC51657.2024.00016

Regular Paper

Authorized licensed use limited to: TU Delft Library. Downloaded on November 04,2025 at 09:04:01 UTC from IEEE Xplore. Restrictions apply.