Gate Overlap Junction Engineering for Off-State Leakage Control in Vertical n-MOS Transistors

Joris Viano*†, Pascal Masson†, Jordan Innocenti*, Vincenzo Della-Marca‡, Christian Rivero*, Julien Delalleau*, Jean-Marc Bourgogne*, Joel Metz*, Brice Arrazat*, Arnaud Regnier*
Email : joris.viano@st.com
*STMicroelectronics, 13106 Rousset, France
†Polytech'Lab, University of Nice Sophia Antipolis, 06100 Nice, France
‡Institute of Materials, Microelectronics and Nanosciences of Provence, Aix-Marseille University, 13451 Marseille, France

***Abstract*—A vertical MOS transistor with front-side access is proposed as an alternative to planar high-voltage MOS devices for charge-storage nonvolatile embedded memories, enabling significant footprint reduction at the array periphery. An anomalous off-state leakage is identified through TCAD simulations and temperature-dependent measurements as gate-induced drain leakage driven by band-to-band tunneling at the gate–drain overlap. Mitigation strategies based on shallow trench isolation engineering and local oxide spacers are evaluated to reduce the electric field locally, and a surface poly-spacer scheme is introduced to laterally shift the junction overlap away from the vertical gate. The resulting trade-offs between on-state conduction, off-state leakage, and reach-through isolation are quantified, providing integration guidelines for the adoption of scaled vertical MOS devices in high-voltage applications.**

Keywords— Band-to-band tunneling, memories, gate-induced drain leakage, high-voltage transistor, off-state leakage, TCAD, vertical MOS transistor

I. Introduction

Vertical transistors (VMOS) were first introduced in the early 1970s by T. John Rodgers and James D. Meindl [1]–[2]. Nowadays, these devices are widely adopted in power MOS technologies [3]. However, the continuous push toward higher levels of integration and footprint reduction is now promoting the use of vertical devices in a broader range of applications. As previously presented in [4], we explore a vertical MOS architecture (Fig. 1) specifically designed to replace the current planar high-voltage transistors on embedded charge storage non-volatile memory (eNVM) technology. This device reuses process operations already available, thereby minimizing its cost. For a given gate–drain distance, the proposed architecture is expected to provide a length reduction in the range of 78–86%. These transistors, located at the NVM array periphery are used as control-gate switches, i.e., to distribute program and erase voltages to the selected memory cells while ensuring isolation of unselected lines from the high-voltage domain. Isolation criteria impose a breakdown voltage (BV) exceeding 13 V with the lowest off-state leakage current (I_{OFF}). As described later, the reverse-biased drain junction characteristic of VMOS devices shows an unexpected early leakage, limiting its use. This study addresses the origin of this leakage and the possible approaches to mitigate it. Device process flow and morphology are detailed in the section II. Thereafter, the off-state leakage is described, and its origin is investigated by temperature measurements and simulations. Finally, two architectures are proposed to remove early leakage by using local spacers.

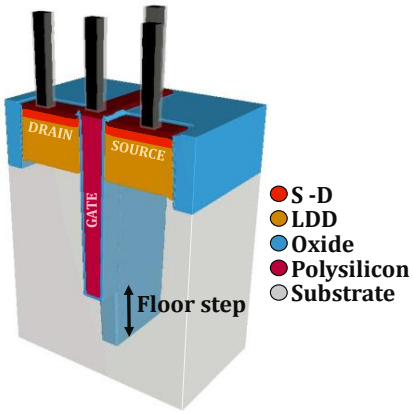

Figure 1 : VMOS 3D simulation using Sentaurus Process Explorer (SPX)

II. Process Flow and Device Morphology

The layout of the device and the in-line scanning electron microscopy (SEM) top-view image are shown in Fig. 2(a) and Fig. 2(b) respectively.

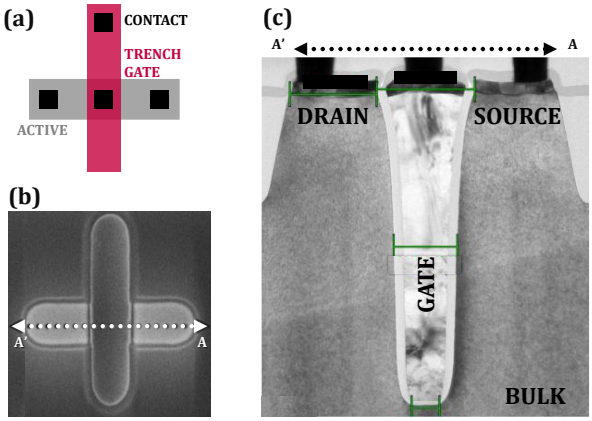

Figure 2 : (a) VMOS layout and (b) in-line SEM top-view. (c) TEM cross-section A-A' along device length.

The body well is realized using a depthwise profile, a dry etch is then performed to define the vertical gate. Because trench definition is carried out through both active and shallow trench isolation (STI) areas, gate etching process is performed in two steps. First, by using oxide-selective step to remove STI and secondly, by using silicon-selective etch. This sequence leads to a shallower trench inside active regions and induces a pronounced variation in trench depth along the device width as visible in Fig. 1 at the trench floor. A 3D process simulation of the resulting morphology using Sentaurus Process Explorer (SPX) is shown in Fig. 1. The gate oxide is grown by thermal oxidation, and trench silicon orientation leads to an anisotropic oxide thickness between the trench sidwalls and bottom. The trench is filled with in-situ doped polysilicon followed by a polishing step. Unlike the VMOS used in power applications, here the source and the drain regions (S–D) are

located at the substrate surface to maintain compatibility with front-side device access and operation. They are defined by implanting the lightly doped drain (LDD) at relatively high energy followed by surface S–D implants. A transmission electron microscopy (TEM) cross-section of the VMOS at the process end is shown Fig. 2(c).

III. Early Off-state Leakage Investigation

In operating mode, a bias of 13 V is applied to the drains of the unselected transistors in the off-state, where they should exhibit I_{OFF} as low as possible. To verify this behavior, we performed a drain-voltage sweep in the off-state with the gate voltage, V_G set to 0 V. The measurements, highlighted an unexpected early leakage current as shown in Fig. 3(a). This behavior is defined by an uncontrolled leakage that appears at 7 V and finally catching up with a classical avalanche breakdown at higher field. This leakage current is measured independently of the trench depth or the MOS width (ranging from 10 µm to 0.22 µm). The gate current is negligible, ruling out the hypothesis of gate leakage through the trench oxide. Potential residues, e.g., from polysilicon depositions and etching or oxide/nitride steps used for eNVM and planar MOS processes could degrade the device. This hypothesis was also ruled out by measuring short-loop wafers manufactured with limited process steps required for the vertical MOS transistor only. Leakage onset is not uniform and varies between the different dies and wafers tested. Moreover, Fig. 3(b) reports that repeatedly performing the measurement, shifts the early leakage onset, revealing a clear high-field breakdown behavior. Heating the samples causes the early leakage to reappear, confirming that this phenomenon is transient and the initial offset is not permanent. This behavior, known as junction walk-out, has been observed in planar structures [5] and associated to charge trapping within the gate oxide, decreasing the local field.

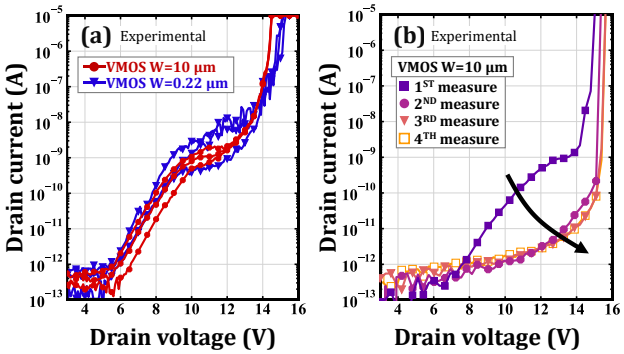

Figure 3 : (a) ID-VD of multiple n-VMOS with different widths (W). (b) Junction walk-out observation when repeating the measure.

Two-dimensional technology computer-aided design (TCAD) numerical simulations in Fig. 4(a) revealed a favorable band-to-band (BTB) tunneling region located at the gate-drain overlap. As described in the band diagram in Fig. 4(b), the surface band bending exceeds the bandgap energy, leading the valence-band energy states to overlap with the conduction-band states. This condition leads to significant tunneling current generation [6]. This mechanism, induced by a lateral electric field, is commonly known as gate-induced drain leakage (GIDL) [7] and is mostly described for planar transistors.

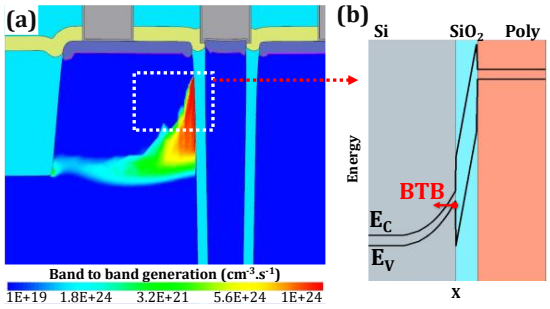

Figure 4 : (a) 2D TCAD simulation of BTB generation field on VMOS and (b) surface band diagram.

Although the different BTB models (*NonLocalPath* [8], *Schenk* [9], and *Hurkx* [10]) available in Sentaurus Device allow us to capture the onset of current generation, the triggering potential appears to be higher than in the measurements. The GIDL current is described by the following BTB theory equation [11]:

$$I_{GIDL} = \xi_S \cdot A \cdot \exp\left(\frac{-B}{\xi_S}\right), \quad (1)$$

where ξ_S is the oxide electric field, A is a pre-exponential factor and B is expressed in [12] following a simplification of Kane's theory [6] as:

$$B = \frac{4 \cdot (2m_r)^{1/2} \cdot E_g(T)^{3/2}}{3q\hbar}, \quad (2)$$

where m_r is the reduced effective mass of the electron, $E_g(T)$ the silicon bandgap, q the electronic charge and $\hbar$ the reduced Planck constant. ξ_S is expressed in [13] as:

$$\xi_S = \frac{V_{DG} - \psi_S}{\frac{\varepsilon_{Si}}{\varepsilon_{SiO_2}} T_{ox}}, \quad (3)$$

with T_{ox} the gate oxide thickness, ε the different dielectric constants, and ψ_S the surface potential defined as:

$$\psi_S = V_{DG} + \frac{T_{ox}^2}{\varepsilon_{ox}^2} q N_D \varepsilon_{Si} - \sqrt{\left(V_{DG} + \frac{T_{ox}^2}{\varepsilon_{ox}^2} q N_D \varepsilon_{Si}\right)^2 - V_{DG}^2}, \quad (4)$$

with V_{DG} the potential difference between the drain and the gate and N_D the donor concentration at gate oxide interface. From Eq. (1) , the BTB current is primarily controlled by the electric field at the $Si - SiO_2$ interface, while its temperature dependence slightly arises through the bandgap $E_g(T)$. Knowing that, we performed electrical characterizations in a wide range of temperature to assess whether a cumulative trap-assisted two-step tunneling (TATT) current overlaps the direct BTB characteristics as described in [14] or if TCAD underestimates the early leakage. Measurements were performed until $V_D = 11.5\ V$ to ensure leakage repeatability as the walk-out phenomenon occurs for voltages above 12 V. Fig. 5(a) shows the reverse-bias sweeps for several temperatures. From this characteristic, we extract the activation energy using an Arrhenius-type analysis to assess the temperature dependence of the underlying leakage mechanisms. At $V_D = 6\ V$ (Fig. 5(b)), the extracted activation energy E_{A1} is close to

0.63 eV , which is consistent with a classical SRH generation–recombination mechanism. For $V_D > 6\,V$ when the other leakage contribution arise, activation energy becomes smaller, reaching values below 0.2 eV at $V_D = 11.5\,V$ Fig. 5(c). This confirms that the leakage is not a trap-assisted mechanism. Fitting TCAD BTB models using fitted parameters rather than implementing a cumulative current contribution through interface traps is coherent. The comparison between TCAD simulations and measurements for a VMOS with a W=10µm is presented in Fig. 6.

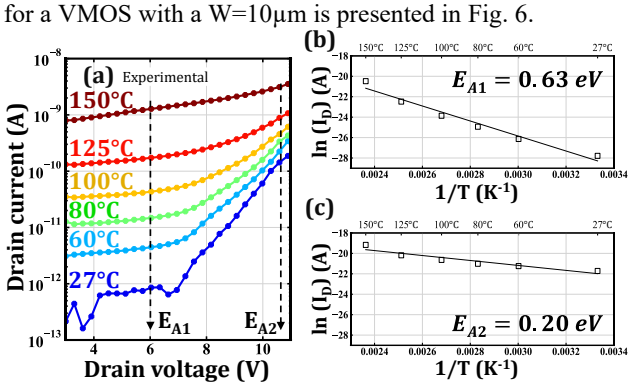

Figure 5 : (a) Early leakage measurement at different temperature. Activation energy extracted at (b) VD=6V and at (c) VD=11.5V.

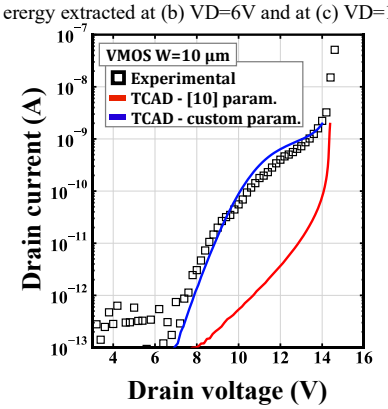

Figure 6 : TCAD BTB simulation using Hurkx model with [10] and custom fitting parameters using Sentaurus release W-2024.09.

IV. Off-State Leakage Control

We have shown that the early leakage origin observed in the VMOS is GIDL. In this section, we describe two architectural solutions designed to minimize the effect of the gate potential on the drain junction.

A. Oxide spacer

Whereas, in planar MOS devices, it is possible to minimize the gate overlap by adjusting the LDD implants and spacers, here the vertical gate here overlaps the entire source and drain regions. Equation (4) shows that the gate-induced band bending can be modulated by increasing T_{ox}. By etching the trench through the STI layer rather than in the active region, we can easily create a local oxide spacer at the overlap region witout additional cost. The channel gate oxide below the STI remains unchanged and the resulting trench depth is increased and becomes uniform along device width. BTB simulations (Fig. 7(a)) were performed, and they confirmed a reduction in band bending (Fig. 7(b)), thereby reducing the tunneling mechanism. Electrical characterizations presented in Fig. 8(a) confirm the initial shift of BTB onset after impact ionization becomes dominant and show a clear hard-breakdown voltage. Despite overcoming the off-state leakage issue, another limitation hinders its use. As mentioned above, the STI layer restricts the influence of the gate potential. In the on-state, the inversion layer will not form in this region. As a result, a resistive area is formed between the LDD implant and the channel, limiting the effective drive current. A solution using a deeper phosphorus implant has been implemented, and the corresponding electrical characterizations are shown in Fig. 8(b).

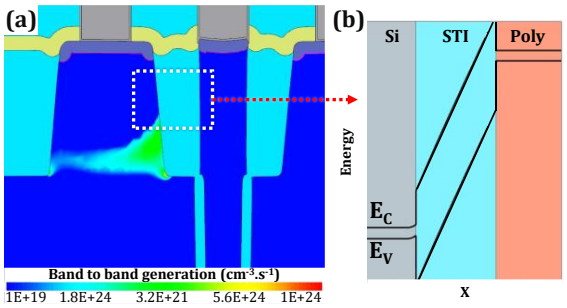

Figure 7 : (a) TCAD simulation of BTB generation field on VMOS using STI spacer and (b) surface band diagram showing lower band bending.

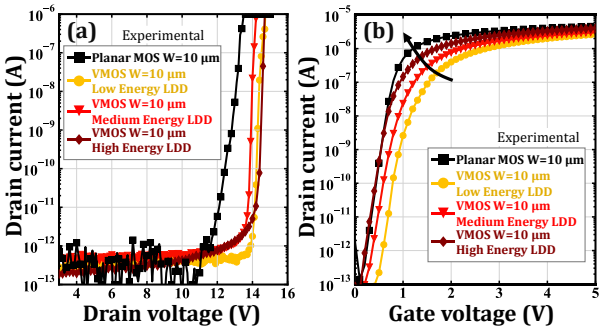

Figure 8 : (a) ID-VD and (b) ID-VG characteristics of n-VMOS using STI spacer with different LDD ernergy

On the one hand, increasing the LDD implant energy makes it possible to identify an operating point for a trench-gate architecture that closely matches the specifications of the planar MOS reference. On the other hand, this process reduces the parasitic resistance between neighboring active regions. Drain-to-drain leakage below STI isolation, also called reach-through leakage, has been monitored on dedicated structures (Fig. 9(a)). As shown in Fig. 9(b), the reach-through current increases as the LDD energy increases. A perspective for this architecture is to minimize the STI thickness on both sides of the gate trench. The challenge is to recover the channel access without degrading the reach-through isolation, while avoiding reintroducing GIDL. Keeping the same TCAD parameter set and applying it on VMOS with different STI thicknesses (Fig. 10(a)) allows us to predict a maximal STI shrink up to 45%, before a non-negligible level of GIDL current reappears (Fig. 10(b)). Co-integration of this device could be enhanced by two methods. The first is to use a deeper local STI as a lateral

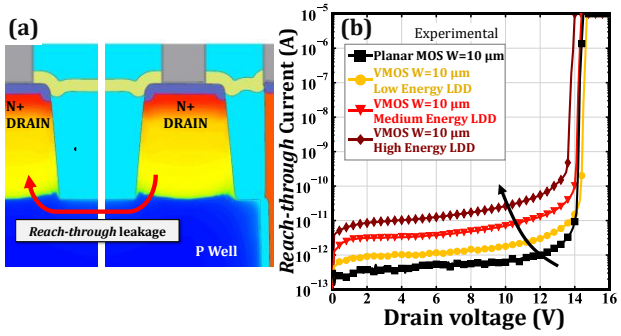

Figure 9 : Reach-through leakage representation using TCAD (a) and its measurements on n-VMOS using different LDD energy (b)

isolation, to reduce reach-through leakage. The second would be to use a dedicated shallower oxide spacer around the trench gate to minimize the channel access resistivity. However, these solutions require additional process steps beyond those already available in the technology platform, and despite the resulting area gain, they significantly lessen the industrial interest of this new architecture. In the last section, we will present an alternative solution that can mitigate the GIDL effect without adding any additional process steps.

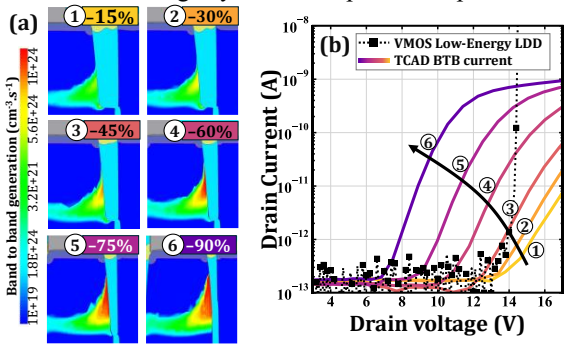

Figure 10 : (a) TCAD simulation of BTB generation field and (b) ID-VD characteristics on n-VMOS depending on the STI thickness shrink.

B. Alternative Architecture Using Poly Spacer

Taking advantage of the eNVM polysilicon stack allows us to use it as an implant-masking layer for a new VMOS architecture. As depicted in Fig. 11(a), the poly spacer protects the gate oxide interface from implantation steps and shifts the BTB region at the silicon surface, far from the weak spot. Note that the resulting junction is now two-dimensional. Controlling the gate–drain overlap through lateral diffusion below the surface polyspacer is closer to that of conventional planar MOS. An initial large-area test structure was evaluated to demonstrate its benefit in mitigating early leakage. Measurements in Fig. 11(b) report the off-state characteristics of different VMOS architectures. No early off-state leakage is observed in the polyspacer device compared with the standard layout, using the same LDD process.

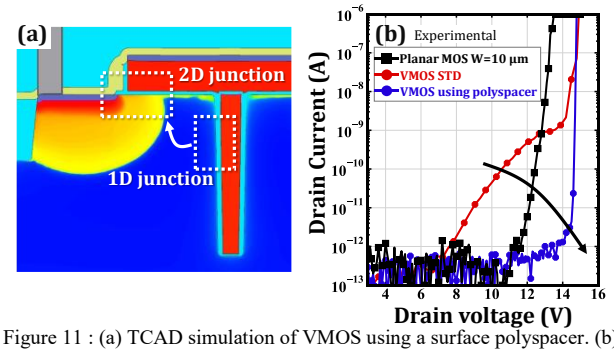

Figure 11 : (a) TCAD simulation of VMOS using a surface polyspacer. (b) Electrical characteristics of the corresponding structure.

V. Conclusion

In this study, we have demonstrated that the main obstacle to integrating vertical MOS transistors into an eNVM technology lies in early off-state leakage due to gate-induced drain leakage at the gate–drain overlap region. By combining measurements on a wide range of temperature with TCAD simulations, this leakage has been linked to band-to-band tunneling. A first solution, using shallow trench isolation to reduce the electric field at the overlap has been presented. This solution restores robust off-state behavior but comes with a drawback: reduced gate control in the STI region, which either creates a resistive access segment or aggravates reach-through leakage as the LDD energy is increased. TCAD exploration of STI spacer shrinking provides a quantitative view of the trade-off and predicts a shrinkage limit of 45%. Finally, we have introduced a surface-spacer-based approach, in which a polysilicon layer is used to shield the gate and move the junction overlap away from the vertical gate.

Acknowledgment

The authors gratefully acknowledge the GQR Q Lab, RCCAL and IM2NP teams for their help with electrical characterization. The authors also thank the process and integration engineers, including J. Fortin, P. Devoge and J. Babic, for their support in device fabrication.

References

[1] T. J. Rodgers and J. D. Meindl, "VMOS: high speed TTL compatible MOS logic," in *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 239-250, Oct. 1974, doi: 10.1109/JSSC.1974.1050509.

[2] T. Rodgers, W. Hiltpold, G. Zimmer, G. Marr and J. Trotter, "VMOS ROM," *1976 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, Philadelphia, PA, USA, 1976, pp. 60-61, doi: 10.1109/ISSCC.1976.1155539.

[3] R. K. Williams, M. N. Darwish, R. A. Blanchard, R. Siemieniec, P. Rutter and Y. Kawaguchi, "The Trench Power MOSFET: Part I—History, Technology, and Prospects," in *IEEE Transactions on Electron Devices*, vol. 64, no. 3, pp. 674-691, March 2017, doi: 10.1109/TED.2017.2653239.

[4] J. Locati, V. Della Marca, C. Rivero, P. Fornara, A. Regnier, S. Niel, K. Coulié, AC stress reliability study on a novel vertical MOS transistor for non-volatile memory technology, Microelectronics Reliability, Volume 114, 2020, 113810, ISSN 0026-2714, https://doi.org/10.1016/j.microrel.2020.113810.

[5] C. R. Lo et al., "Study of the Walk-Out Effect of Junction Breakdown Instability of the High-Voltage Depletion-Mode N-Channel MOSFET for NAND Flash Peripheral Device and an Efficient Layout Solution," 2020 IEEE International Reliability Physics Symposium (IRPS), Dallas, TX, USA, 2020, pp. 1-6, doi: 10.1109/IRPS45951.2020.9129216.

[6] E.O. Kane, Zener tunneling in semiconductors, Journal of Physics and Chemistry of Solids,Volume 12, Issue 2, 1960, Pages 181-188, ISSN 0022-3697, https://doi.org/10.1016/0022-3697(60)90035-4.

[7] T. Y. Chan, J. Chen, P. K. Ko and C. Hu, "The impact of gate-induced drain leakage current on MOSFET scaling," *1987 International Electron Devices Meeting*, Washington, DC, USA, 1987, pp. 718-721, doi: 10.1109/IEDM.1987.191531.

[8] Sentaurus Device User Guide, Version W-2024.09, Synopsys, Inc., Mountain View, CA, USA, September 2024, pp. 558-566.

[9] A. Schenk, Rigorous theory and simplified model of the band-to-band tunneling in silicon, Solid-State Electronics, Volume 36, Issue 1, 1993, Pages 19-34, ISSN 0038-1101, https://doi.org/10.1016/0038-1101(93)90065-X.

[10] G. A. M. Hurkx, D. B. M. Klaassen and M. P. G. Knuvers, "A new recombination model for device simulation including tunneling," in *IEEE Transactions on Electron Devices*, vol. 39, no. 2, pp. 331-338, Feb. 1992, doi: 10.1109/16.121690.

[11] J. Chen, T.Y. Chan, I.C. Chen, P.K. Ko, IEEE Electron Device Lett. 8 (1987) 515–517.

[12] L. Huang, P.T. Lai, J.P. Xu, Y.C. Cheng, Microelectron. Rel. 38 (1998) 1425–1431.

[13] L Lopez, P Masson, D Née, R Bouchakour, Temperature and drain voltage dependence of gate-induced drain leakage, Microelectronic Engineering, Volume 72, Issues 1–4, 2004, Pages 101-105, ISSN 0167-9317, https://doi.org/10.1016/j.mee.2003.12.024.

[14] G. -B. Lee, C. -K. Kim, M. -S. Yoo, J. Hur and Y. -K. Choi, "Effect of OFF-State Stress on Gate-Induced Drain Leakage by Interface Traps in Buried-Gate FETs," in IEEE Transactions on Electron Devices, vol. 66, no. 12, pp. 5126-5132, Dec. 2019, doi: 10.1109/TED.2019.2947603.