Impact of Work Function Metal Stacks on the Performance and Reliability of multi-V_{th} RMG CMOS technology

J. Franco*, H. Arimura, S. Brus, E. Dentoni Litta, K. Croes, N. Horiguchi, B. Kaczer
imec, Kapeldreef 75, 3001 Leuven, Belgium; *Jacopo.Franco@imec.be

***Abstract*—Multi-V_{th} CMOS device technologies have become standard for System-on-Chip designs. In Replacement Gate technologies, distinct device V_{th}'s are achieved by deploying different work function metal stacks, and thus concerns exist about the possible chemical interaction of different gate metals with the underlying dielectrics potentially affecting the device performance and reliability. We present a comprehensive study, comprising both electrical measurements and simulations, carried out on a planar transistor platform with state-of-the-art gate stacks. Two different metal stacks are deployed to fabricate low-V_{th} and ultra-high V_{th} pMOS and nMOS device flavors. The study provides fundamental insights on the impact of TiAl-based gate metal on EOT, gate leakage, interface quality, carrier mobility, short channel performance, PBTI and NBTI reliability.**

***Keywords*—RMG, HKMG, NBTI, PBTI, Multi-V_{th}.**

I. Introduction

The offering of multiple V_{th} device flavors has become standard in recent CMOS technology nodes, due to the enhanced flexibility in terms of power/performance trade-offs for System-on-Chip design. In Replacement Gate (RMG) technologies, and in particular in lowly-doped architectures such as finFETs and nanosheets, the distinct device V_{th}'s (up to six different ones [1]) are achieved by deploying different work function metal (WFM) stacks. Despite the reduced thermal budget applied after metal deposition in a RMG flow as compared to traditional Gate-First integrations, concerns exist about the possible chemical interaction of different WFM's with the dielectric stack, with potential implications on the device performance and reliability.

Contrasting accounts have been reported in literature on the reliability of low V_{th} (LVT) and high V_{th} (HVT) devices: in [2] the BTI-induced V_{th} shift was observed to scale with the gate stress overdrive in different device flavors, while in [3] HVT devices were shown to require a larger than expected stress V_g to reach a 50mV V_{th} shift in a ramped voltage stress (RVS) test. We have recently presented a preliminary investigation of this issue by means of BTI simulations [4], and experimentally on MOS capacitors fabricated in low thermal budget flows for Sequential 3D applications [5].

In this paper, we present a new comprehensive study, comprising both electrical measurements and simulations, carried out on a planar transistor platform with state-of-the-art RMG gate stacks. Two different WFM stacks (TiN-only for *p*-WF vs. TiN/TiAl/TiN for *n*-WF) are deployed to fabricate LVT and ultra-HVT (UHVT) pMOS and nMOS device flavors (**Fig. 1**). The study provides fundamental insights on the impact of the TiAl-based *n*-WFM stack on EOT, gate leakage, interface quality, carrier mobility, short channel performance, PBTI and NBTI reliability.

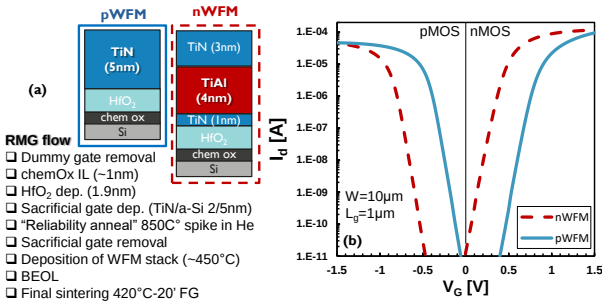

Figure 1: (a) Sketch of the RMG fabrication flow and of the two WFM stacks studied. The *p*-WFM stack is used for pMOS LVT and nMOS UHVT devices; the *n*-WFM for nMOS LVT and pMOS UHVT devices. (b) I_d - V_g curves of the two nMOS and two pMOS device flavors, illustrating the V_{th}-tuning by WFM stack.

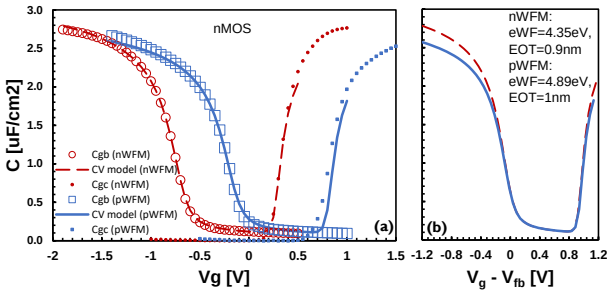

Figure 2: (a) Split *C*-*V* curves (symbols) measured on nMOS devices with the two different WFM's, used to calibrate quantum-corrected electrostatics models of the stacks (lines) for EOT and eWF extraction. (b) Calibrated *C*-*V* models [6] replotted vs. (V_g - V_{fb}) to highlight the C_{ox} increase (i.e., EOT thinning) induced by the *n*-WFM. (note: similar EOT and eWF values were obtained from pMOS *C*-*V*'s, not shown).

II. Impact of WFM on MOS properties and device performance

To estimate the EOT and effective WF (eWF) of the two different gate stacks, the split *C*-*V* curves measured on the LVT and UHVT devices were modeled with a Schrödinger-Poisson solver [6] (**Fig. 2**). An eWF modulation of ~0.5V was estimated, in line with the V_{th} difference of the two device flavors (cf. Fig. 1). The *n*-WFM (used in the LVT nMOS and UHVT pMOS) was found to induce a ~1Å thinner EOT (note: due to the non-ideal permittivity of chemical oxides [7], this could result from a physical thickness reduction up to ~1.5-2Å), likely due to oxygen-scavenging mechanism [8] inducing a reduction of the SiO_2 interfacial layer (IL) thickness despite the limited thermal budget applied during/after the *n*-WFM deposition (max T=450°C).

I_g-V_g curves revealed indeed a larger gate leakage for the devices with *n*-WFM (**Fig. 3**): in nMOS devices, a ~10× increase is consistent with a ~2Å reduced physical IL thickness. Interestingly, an even larger *n*-WFM-induced

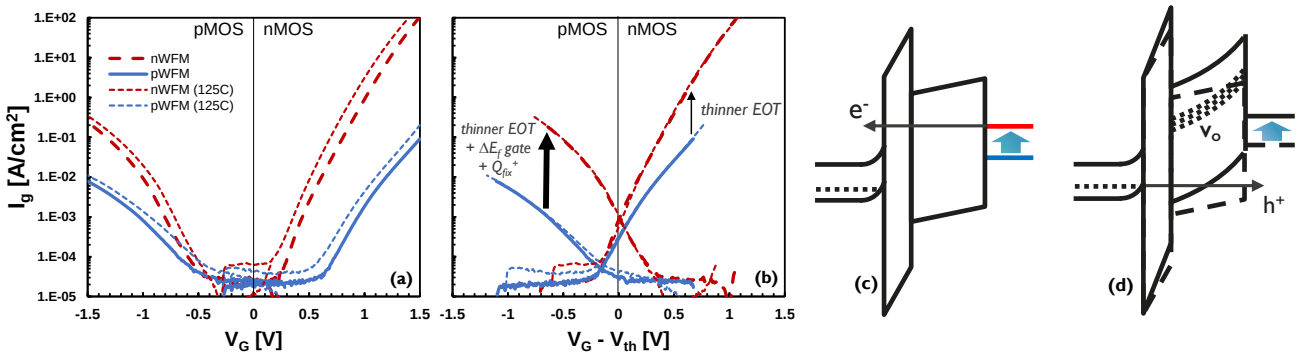

Figure 3: (a) $I_g - V_g$ curves measured on the two nMOS and two pMOS device flavours at 25°C and 125°C. At a given V_g, a larger gate leakage is observed with *n*-WFM in both nMOS and pMOS. (b) I_g curves replotted vs. (V_g - V_{th}). No temperature-dependence is observed, which suggests trap-assisted tunnelling is negligible, while (elastic) direct tunnelling dominates. The enhanced gate leakage with *n*-WFM in nMOS is consistent with the reduced EOT (cf. Fig. 2b), while a larger penalty observed in pMOS suggests additional mechanisms are involved [5,9], namely (c) enhanced gate electron injection with nWFM due to a reduced injection barrier and, likely, (d) enhanced valence band hole tunnelling due to the presence of positive fixed charge across the dielectric stack induced by the *n*-WFM through O-scavenging [8,13,14].

leakage penalty was observed in pMOS devices, suggesting contributions from additional mechanisms: a low gate WF is expected to enhance gate electron injection due to a reduced tunneling barrier height (**Fig. 3c**), while the possible presence of fixed positive charge across the dielectric stack, induced by the O-scavenging mechanism [8,13,14], can enhance the Si valence band hole tunneling towards the gate (**Fig. 3d**) [5,9]. Notice that the lack of temperature dependence of the gate leakage (**Fig. 3b**) suggests trap-assisted tunneling is negligible at operating voltages as compared to direct tunneling through the scaled dielectric stack.

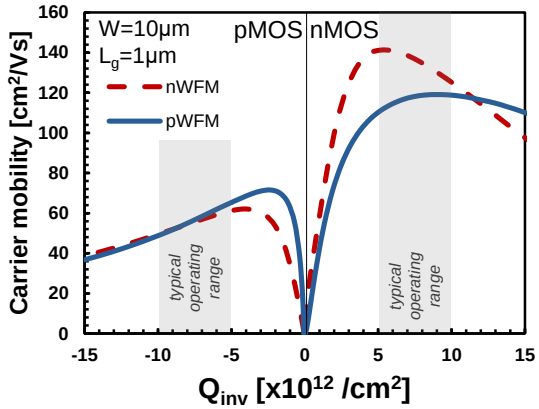

Figure 4: (a) Carrier mobility estimated with the split-CV technique in the two nMOS and two pMOS device flavors. *n*-WFM induces an enhancement of the electron mobility and a reduction of the hole mobility. (note: no stressors were deployed in the transistor test vehicle, accounting for the relatively modest mobility values at ultra-thin EOT).

A clear dependence of the carrier mobility on the WFM was also observed (**Fig. 4**): best hole mobility was obtained with a *p*-WFM, while best electron mobility was obtained with *n*-WFM. I.e., the LVT flavors of both pMOS and nMOS offer better intrinsic performance. This observation was confirmed also by the I_{ON}-I_{OFF} figure of merit of short channel devices (**Fig. 5**). Interface state density estimations based on the MOS parallel gate conductance peak in depletion revealed a smaller D_{it} on the E_v side and a larger D_{it} on the E_c side when the *p*-WFM was deployed (**Fig. 6**). This is likely related to the efficiency of the P_b defect passivation during the final Forming Gas anneal depending on the built-in potential of the stack [10] and could be possibly solved with more effective high-pressure hydrogen anneals [11]. According to our transport compact model experimentally validated in [12], the measured asymmetry in the D_{it} profile can account for a ~10% discrepancy in carrier mobilities, i.e., for only a part of the observed effect. However, this is likely due to the simple D_{it} characterization technique adopted being sensitive at room temperature only to a small energy range near Si mid-gap; the actual D_{it} asymmetry at band edges might be larger.

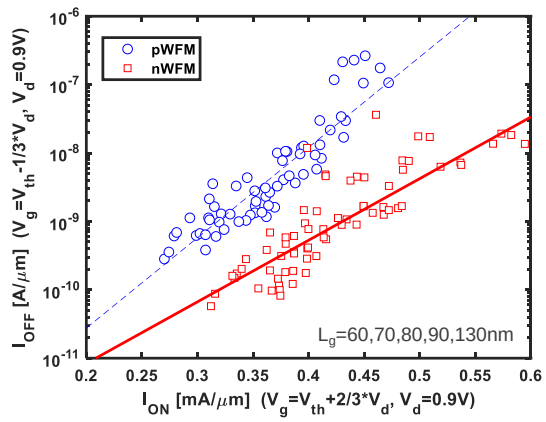

Figure 5: I_{ON}-I_{OFF} plot of short channel nMOS with the two different WFM's, benchmarked at constant offset from V_{th} to account for the different eWF. Enhanced drive current is observed with *n*-WFM, consistent with the observation of enhanced mobility in long channels (cf. Fig. 4).

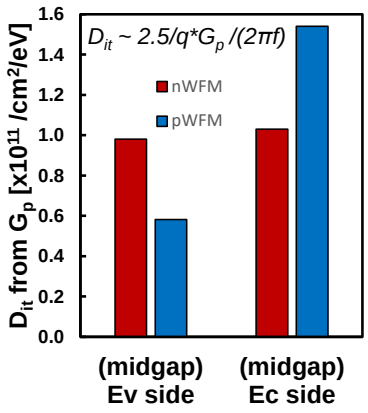

Figure 6: Interface density estimations based on the measured peak of the MOS gate parallel conductance in depletion (corrected for gate leakage). As compared to the *p*-WFM ref., the *n*-WFM yields a worse/better interface passivation on the E_v/E_c sides, respectively.

III. IMPACT OF WFM ON BTI

LVT device flavors consistently showed also smaller BTI-induced ΔV_{th} at a given stress V_{ov} (**Fig. 7**): the *p*-WFM exposed the presence of an additional nMOS PBTI tail at low stress V_{ov} (**Fig. 7a**), while the *n*-WFM was observed to enhance the pMOS NBTI at any stress V_{ov} (**Fig. 7b**). These two observations, apparently complementary, have likely distinct root causes. The PBTI penalty with *p*-WFM can be ascribed to the presence of a second defect band in HfO_2 at deeper energy levels as compared to the main PBTI electron trap band. The initial occupancy of these deep traps is determined by the gate Fermi level (E_f) [4]: a low gate WF maintains these traps in equilibrium, while with a high gate WF they become available for charge exchange with the channel (**Fig. 8**). NBTI is virtually insensitive to (deep) HK traps [4], while it is extremely sensitive to near-interface

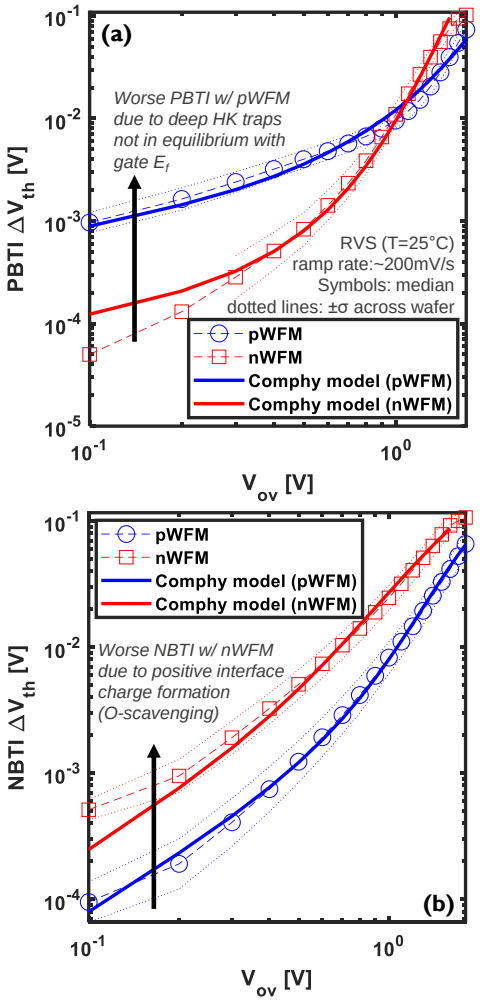

Figure 7: (a) nMOS PBTI and (b) pMOS NBTI ramped voltage stress data. The BTI-induced ΔV_{th} is plotted vs. stress overdrive to account for the initial V_{th}'s of the different device flavors. Best PBTI reliability is observed with n-WFM, while best NBTI reliability is observed with p-WFM (i.e., the LVT flavors show always best intrinsic reliability; this observation was reproduced also at long stress times and elevated T, not shown). The symbols and dashed lines represent the median ΔV_{th} data measured on multiple devices across wafer; the dotted lines represent ±1σ variation across wafer (note the measurement error bar of ~0.06mV – related to the current measurement resolution and the device subthreshold slope – is smaller than the device-to-device variation across wafer, and thus negligible); the solid lines represent calculations of the unified Comphy BTI model, see Fig. 8 and Table I.

defects and fixed charge [5]. The observed NBTI penalty with *n*-WFM is likely related to the presence of positive fixed charge at the Si/SiO_2 interface enhancing the oxide electric field (as suggested also by the pMOS gate leakage trends, see Fig. 3d), formed due to the *n*-WFM-induced oxygen scavenging (**Fig. 8, inset**), which contributes to the reduction of the eWF [13,14].

We implemented these two scenarios in our BTI modelling framework Comphy [15]. The model includes: 1) hole traps in the SiO_2 IL (most likely hydroxyl-*E'* defects forming due to interface strain [7]), with energy levels distributed below the Si valence band, and with density exponentially decaying away from the channel interface [16]; 2) electron traps in HfO_2 (most likely oxygen vacancies), with energy levels distributed above the Si conduction band; 3) deep traps in HfO_2 [15], with energy levels distributed within the Si bandgap, and thus overlapping with the typical range of gate E_f required for V_{th} tuning. Both HfO_2 defect bands are assumed to have uniform density across the dielectric thickness. As reported in Table I, most of the model parameters are assumed to be identical for the *p*-WFM and the *n*-WFM gate stacks, except the IL thickness (thinner for the *n*-WFM gate stack, in line with the experimental observation, see Fig. 2), the density of shallow electron traps in HfO_2 (as oxygen scavenging by *n*-WFM is expected to enhanced the density of oxygen vacancies [13,14]), and the gate E_f. Determining the latter appropriately for the *n*-WFM gate stack, given that the effective WF shift as determined by the *C-V* curves includes the contribution from positive interface charge induced by oxygen scavenging, is not trivial. We rely on the assumption that the enhanced pMOS NBTI with *n*-WFM is solely due to the oxygen scavenging-induced positive fixed interface charge [5]. Based on this assumption, as NBTI is insensitive to the gate E_f [4], we determined through Comphy simulations the positive fixed charge density required to quantitatively justify the observed NBTI enhancement (Q_{fix} ~8×10^{20}/cm^2). Next, the gate E_f shift for the *n*-WFM stack w.r.t. the TiN *p*-WFM reference was obtained from the experimental eWF shift (see Fig. 2) after accounting for the contribution from Q_{fix} (note: our model suggests that only ~60% of the eWF modulation comes from a real change of the gate E_f, with positive charge formation accounting for the remaining ~40%). With these assumptions, an excellent agreement between the calibrated Comphy model simulations and the measured BTI shifts (RVS) was achieved for pMOS and nMOS with *p*-WFM or *n*-WFM (**Fig. 7a-b**, lines vs. symbols), thus confirming the validity of the insights about the contribution of deep HfO_2 defects to PBTI in (U)HVT nMOS, and the impact of *n*-WFM-induced oxygen scavenging on NBTI in (U)HVT pMOS due to the formation of fixed positive interface charge.

Finally, it is worth noticing that comparing BTI in different device flavors only at large degradation levels, as typically done in industrial environments by considering the voltage required in a RVS test to reach an end-of-life V_{th} shift criterion (e.g. 50mV [2,3]), might result in incorrect benchmarking. In particular, with such criterion the (U)HVT nMOS device with *p*-WFM would incorrectly come out as more reliable w.r.t. to the LVT nMOS with *n*-WFM (i.e., a larger V_{ov} is necessary to reach a 50mV shift, cf. Fig. 7a). On the contrary, the reliability of the (U)HVT nMOS is severely limited by the additional degradation related to the deep HfO_2 defect band, which is clearly visible only at small degradation level, as at high stress voltages the degradation related to the shallow HfO_2 defect band takes over (and thus the *n*-WFM gate stack

appears to be worse due to its larger density of oxygen vacancies, see Table I). By accounting for the very short stress time duration of the RVS (typically a few seconds, i.e., dramatically short than the 10 year operating lifetime for which a 50mV failure criterion is defined), a better practice would consist in adopting a very stringent failure criterion when benchmarking gate stacks based on such quick reliability test (e.g., V_{ov} necessary to reach a 2mV shift). Notice how such approach would consistently result in a correct reliability ranking of the gate stack flavors here studied (Fig. 7).

IV. Conclusions

We have compared the performance and reliability of two pMOS and two nMOS gate stacks, with *p*-WFM and *n*-WFM deployed to fabricate LVT and (U)HVT device flavors. LVT flavors (i.e., pMOS w/ *p*-WFM and nMOS w/ *n*-WFM) showed best performance, in terms of carrier mobility and drive current, and best BTI reliability. TiAl-based *n*-WFM was found to induce oxygen-scavenging from the dielectric stack, despite the limited thermal budget involved in the RMG process flow after metal deposition, yielding a ~1Å EOT reduction and a larger density of oxygen vacancies in HfO_2 and fixed positive interface charge. The latter was identified as the cause of enhanced NBTI in (U)HVT pMOS. Furthermore, the PBTI of (U)HVT nMOS w/ *p*-WFM was found to be enhanced due to the additional contribution of deep HfO_2 defects, which are instead kept in equilibrium by the gate Fermi level in the presence of a *n*-WFM (nMOS LVT). Finally, we have presented a unified calibrated Comphy model, capable of reproducing the PBTI and NBTI trends in nMOS and pMOS with different gate WFM's, and provided guidelines for correct multi-V_{th} gate stack BTI benchmarking based on accelerated RVS test.

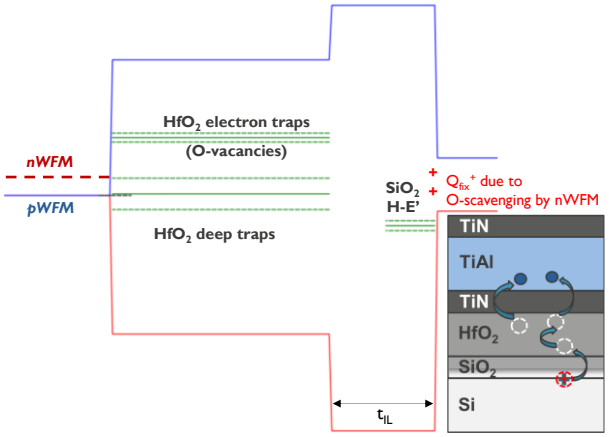

Figure 8: Unified Comphy BTI model, including hole traps in the SiO_2 IL, shallow electron traps and deep traps in HfO_2. The initial occupancy of the latter is determined by the gate E_f. The model includes also positive interface charge, reduced IL thickness and enhanced HfO_2 electron trap density induced by (inset) oxygen scavenging mechanism by the n-WFM [8,13,14].

Table I - Comphy BTI model parameters

	w/ p-WFM	w/ n-WFM
SiO_2 hole traps (hydroxyl-E' [16])		
N_{t0} [$\times 10^{20}$/cm^3]	3.8	
μ E_t [eV]	-0.87	
σ E_t [eV]	0.105	
$N_t=N_{t0}$*exp(-x/x_0), x_0 [nm]	0.5	
HfO_2 electron traps (O-vacancies [5])		
N_t [$\times 10^{20}$/cm^3]	4.7	32
μ E_t [eV]	0.79	
σ E_t [eV]	0.134	
HfO_2 deep traps [15]		
N_t [$\times 10^{20}$ /cm^3]	1.2	
μ E_t [eV]	-0.2	
σ E_t [eV]	0.33	
Other model parameters		
Gate E_f [eV]	-0.172	0.065
Positive interface charge [$\times 10^{12}$/cm^2]	0	8
SiO_2 IL thickness [nm]	1.05	0.95
SiO_2 IL relative dielectric constant	5	

**All energies referred to Si mid-gap*

***NMP parameters (S and R) pinned to typical values from [15-16], as they are not critical for reproducing short stress time degradation (RVS data)*

References

[1] B. Sell *et al.*, "22FFL: A high performance and ultra low power FinFET technology for mobile and RF applications", in *IEEE International Electron Devices Meeting (IEDM)*, pp. 29.4.1-4, December 2017.

[2] W. Liu *et al.*, "Cap Layer and Multi-Work-Function Tuning Impact on TDDB / BTI in SOI FinFET Devices", in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 2A4.1-5, May 2018.

[3] B. Linder *et al.*, "Process Optimizations for NBTI/PBTI for Future Replacement Metal Gate Technologies", in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 4B1.1-5, April 2016.

[4] J. Franco *et al.*, "On the Impact of the Gate Work-Function Metal on the Charge Trapping Component of NBTI and PBTI", in IEEE Trans. on Device and Materials and Reliability 19(2), pp. 268-274, 2019.

[5] J. Franco *et al.*, "Low thermal budget PBTI and NBTI reliability solutions for multi-V_{th} CMOS RMG stacks based on atomic oxygen and hydrogen treatments", in *Proc.* International Electron Devices Meeting (IEDM), pp. 30.4.1-4, 2022.

[6] J. R. Hauser and K. Ahmed, "Characterization of ultra-thin oxides using electrical C-V and I-V measurements", AIP Conference Proceedings 449(1), pp. 235-239 (1998);

[7] J. Franco *et al.*, "Low-temperature atomic and molecular hydrogen anneals for enhanced chemical SiO_2 IL quality in low thermal budget RMG stacks", in *Proc.* International Electron Devices Meeting (IEDM), pp. 31.4.1-4, 2021.

[8] G. Bersuker *et al.*, "Origin of the Flatband-Voltage Roll-Off Phenomenon in Metal/High- Gate Stacks", in IEEE Trans. on Electron Devices 57(9), pp. 2047-2056, 2010.

[9] H. Arimura *et al.*, "Gate Stack Technology for Advanced Logic and Memory Periphery Devices", as discussed at the 53rd IEEE Semiconductor Interface Specialists Conference (SISC), 2022.

[10] Lars-Åke Ragnarsson and Per Lundgren, "Electrical characterization of P_b centers in (100)Si–SiO_2 structures: The influence of surface potential on passivation during post metallization anneal", in Journal of Applied Physics 88, pp. 938-942 (2000);

[11] H. Arimura *et al.*, "Si-passivated Ge nMOS gate stack with low Dit and dipole-induced superior PBTI reliability using 3D-compatible ALD caps and high-pressure anneal", in *Proc.* International Electron Devices Meeting (IEDM), pp. 33.4.1-4, 2016.

[12] Z. Wu *et al.*, "Investigation of the Impact of Hot-Carrier-Induced Interface State Generation on Carrier Mobility in nMOSFET", in IEEE Trans. on Electron Devices 68(7), pp. 3246-3253, 2021.

[13] T. Ando *et al.*, "Simple Gate Metal Anneal (SIGMA) stack for FinFET Replacement Metal Gate toward 14nm and beyond", in Proc. Symp. on VLSI Technology, pp. 1-2 , 2014.

[14] J. Rozen *et al.*, "Ultra-scaled Conformal Scavenging Electrode with Superior Tunability for Short-channel RMG FinFET Workfunction and all-ALD 3D-compatible ReRAM", in *Proc.* International Electron Devices Meeting (IEDM), pp. 36.4.1-4, 2019.

[15] G. Rzepa *et al.*, "Comphy—A compact-physics framework for unified modeling of BTI", in *Microelectronics Reliability*, Vol. 85, pp. 49-65, 2018.

[16] J. Franco *et al.*, "Low Temperature Atomic Hydrogen Treatment for Superior NBTI Reliability—Demonstration and Modeling across SiO_2 IL Thicknesses from 1.8 to 0.6 nm for I/O and Core Logic", in Proc. Symp. on VLSI Technology, pp. 1-2 , 2021.