Modeling of HKMG Stack Process Impact on Gate Leakage, SILC and PBTI

Dimple Kochar, Tarun Samadder, Subhadeep Mukhopadhyay and Souvik Mahapatra
Department of Electrical Engineering, Indian Institute of Technology Bombay (IIT Bombay), Mumbai 400076, India
Phone: +91-222-572-0408, Email: souvik@ee.iitb.ac.in

***Abstract*—Gate stack process (pre-clean, IL, IL/HK interface, HK, post-HK Nitridation) impact on gate leakage (I_{G0}), SILC ($\Delta I_G/I_{G0}$) and PBTI trap generation at IL/HK interface ($\Delta N_{IT\text{-}HK}$) is analyzed. IL and HK thickness (T_{IL} and T_{HK}), channel/IL and IL/HK energy-barrier offsets (ϕ_{BIL} and ϕ_{BHK}) impact on I_{G0} and $\Delta I_G/I_{G0}$ response from generated bulk traps inside IL ($\Delta N_{OT\text{-}IL}$) and HK ($\Delta N_{OT\text{-}HK}$) are quantified. $\Delta N_{OT\text{-}IL}$, $\Delta N_{OT\text{-}HK}$ and $\Delta N_{IT\text{-}HK}$ time kinetics is simulated by Reaction-Diffusion-Drift (RDD) framework. Model is validated using measurements from differently processed HKMG stacks.**

Introduction: Gate stack thickness scaling is desirable to control electrostatics in FinFETs and GAA-NSFETs and enable channel length scaling. The primary impediment is I_{G0}, which increases by ~10X/2Å for IL and ~10X/1Å for HK EOT scaling [1]. PBTI and SILC in NMOS, NBTI (SILC is absent) in PMOS and TDDB in NMOS and PMOS are other concerns [2]-[5], these effects get worse at scaled EOT. Past I_{G0} modeling hasn't considered impact of ϕ_{BIL} and ϕ_{BHK} changes due to different processes [1], [6]. The correlation of $\Delta N_{OT\text{-}IL}$ and $\Delta N_{OT\text{-}HK}$ to $\Delta I_G/I_{G0}$ for changes in T_{IL}, T_{HK}, ϕ_{BIL} and ϕ_{BHK} is never modeled. Trap generation kinetics for SILC/TDDB and PBTI were empirically modeled in [7], [8]. PBTI is linked to $\Delta N_{IT\text{-}HK}$ (dominating component) and RD model is used for its time kinetics, although process impact is not studied [9]. Recently RD and RDD models are used for NBTI interface and bulk trap generation in PMOS [10]-[12]. This work aims to model direct tunneling leakage, SILC and PBTI in NMOS with deeply scaled (down to EOT~7Å) HKMG stacks.

Scope of work: Impact of changes in T_{IL}, T_{HK}, ϕ_{BIL} and ϕ_{BHK} on I_{G0} and $\Delta I_G/I_{G0}$ is modeled and verified with measured data from NMOSFETs having differently processed HKMG stacks (details of device processes and measurements in [4]). The time kinetics of SILC extracted $\Delta N_{OT\text{-}IL}$, $\Delta N_{OT\text{-}HK}$ and DCIV measured $\Delta N_{IT\text{-}HK}$ is modeled by RDD framework. Impact on generated traps due to scaling of T_{IL} and T_{HK}, post-HK Nitridation (PHKN) and various pre-IL and pre-HK interfacial treatments (IFT) is analyzed.

Device details: HKMG stacks with Chem-Ox (D1, D2) and RTP (D3-D7) IL with different T_{IL} and T_{HK}, different pre-IL and pre-HK IFT, without and with PHKN are used, Table-I. Pre-IL and T_{IL} impact ϕ_{BIL}, pre-HK, T_{HK} and PHKN impact ϕ_{BHK}. Scaling T_{IL} (lower IL bandgap) reduces ϕ_{BIL}, but scaling T_{HK} and PHKN (lower HK bandgap) increases ϕ_{BHK}. These quantities influence I_{G0} and $\Delta I_G/I_{G0}$ response from generated $\Delta N_{OT\text{-}IL}$ and $\Delta N_{OT\text{-}HK}$.

Leakage and SILC: WKB tunneling probabilities via IL (T_1) and HK (T_2) and supply function in cathode (=substrate in inversion) govern I_{G0}, Fig.1(a). Typical rate of I_{G0} increase for scaling IL, Fig.2, and HK, Fig.3, is enhanced when ϕ_{BIL} and ϕ_{BHK} changes are also considered. Fig.4 compares the measured and modeled I_{G0} across devices. Due to different pre-IL and pre-HK process, ϕ_{BIL} and ϕ_{BHK} do not follow the expected T_{IL} and T_{HK} trends. SILC is due to Trap-Assisted-Tunneling via traps generated at IL and HK bulk during PBTI/TDDB stress, Fig.1(b). Based on trap location (in IL or HK), the probabilities T_1, T_2 and T_3 are calculated by WKB, and energy relaxation is also considered. The maximum SILC response point is inside HK, Fig.5, which is verified over IL of 3-7Å, Fig.6, and HK over 17-24Å, Fig.7. The maximum response point moves deeper into HK for lower IL, Fig.6, more so due to lower ϕ_{BIL}, and towards IL/HK interface for lower HK, Fig.7, more so due to higher ϕ_{BHK}. Fig.8 compares the measured and modeled $\Delta I_G/I_{G0}$ across devices. Identical ϕ_{BIL} and ϕ_{BHK} values are used as Fig.4, $\Delta N_{OT\text{-}HK}$ is mentioned. Magnitude of $\Delta N_{OT\text{-}IL}$ is 1/300 of $\Delta N_{OT\text{-}HK}$, to remain consistent with TDDB bimodal Weibull distribution in IL/HK dual layer stacks [7]. Due to dominant contribution from HK, Figs.5-7, SILC is dependent primarily on $\Delta N_{OT\text{-}HK}$ and insensitive to changes in $\Delta N_{OT\text{-}IL}$, which stays valid across different stress oxide electric field (E_{OX}), Fig.9.

Time kinetics, RDD framework: Fig.10 plots the time evolution of measured and modeled $\Delta I_G/I_{G0}$ at various stress V_G, for different IL devices, the corresponding $\Delta N_{OT\text{-}HK}$ is shown in Fig.11 together with the RDD model lines. Energetic electrons and electric field trigger bond dissociation [13] and release of Hydrogen atoms (H), H induced bond dissociation and eventual diffusion and drift of different molecular (H_2) and ionic (OH^-) species control $\Delta N_{OT\text{-}HK}$ (and $\Delta N_{OT\text{-}IL}$) time kinetics, Fig.12, [(1)+(2) and (3)]. The H_2 to OH^- ratio (~ k_{F2} to k_{F3} ratio) governs the time kinetics power-law slope *n* at long time, Fig.13. Pure H_2 and pure OH^- results in *n*~1/6 and ~1/2, while mixed H_2/OH^- results intermediate values [12]. These devices show *n*~1/3, although reported SILC slopes range from *n*~1/2 to ~1/4 [14]-[16]. $\Delta N_{IT\text{-}HK}$ time kinetics is modeled for different V_G, Fig.14, using pure H_2 diffusion [RD model, (1)+(2)], Fig.12. Modeling of $\Delta N_{IT\text{-}HK}$ and $\Delta N_{OT\text{-}HK}$ time kinetics at different V_G and across all devices is done respectively using only 2 and 3 adjustable model parameters (all data are at iso temperature).

Process impact: Fig.15 plots measured and modeled $\Delta N_{IT\text{-}HK}$ and $\Delta N_{OT\text{-}HK}$ for changes in (a) T_{IL}, (b) PHKN and (c) pre-HK IFT to control moisture before ALD HK process. $\Delta N_{OT\text{-}HK}$ and $\Delta N_{IT\text{-}HK}$ behaves differently for IL reduction (reduces and increases) and PHKN (increases and reduces), but similarly with IFT (increases with higher moisture content). Other processes are also modeled in a similar manner, but not shown here for brevity.

Conclusion: Not just thickness, the composition and quality of the IL and HK integration processes impact leakage and reliability of ultra-thin HKMG stacks. T_{IL}, T_{HK}, ϕ_{BIL} and ϕ_{BHK} impact I_{G0}, and changes in barriers should be considered for proper estimation of I_{G0} increase at reduced T_{IL} and T_{HK}. The $\Delta I_G/I_{G0}$ response location and magnitude due to given generated trap density is governed by the above parameters. The dominating contribution to SILC is due to $\Delta N_{OT\text{-}HK}$, changes in $\Delta N_{OT\text{-}IL}$ has negligible impact on $\Delta I_G/I_{G0}$ in these stacks. The generic RDD framework (RD being a subset) is able to model the time kinetics of $\Delta N_{OT\text{-}HK}$ ($\Delta N_{OT\text{-}IL}$) and $\Delta N_{IT\text{-}HK}$ and explain their power-law time slope at long time. The process dependence of $\Delta N_{OT\text{-}HK}$ and $\Delta N_{IT\text{-}HK}$ is modeled. This work would help guide co-optimization of HKMG EOT scaling and reliability.

References: [1] Wu, SSE, p. 1164, 2006, [2] Ramey, IRPS, p. 1023, 2009, [3] Cartier, IEDM, 18-4, 2011, [4] Joshi, IRPS, 4C-2, 2013, [5] Wu, TED, p. 4523 & 4535, 2019, [6] Ando, EDL, p. 865, 2011, [7] Nigam, IRPS, p. 523, 2009, [8] Yang, IRPS, 5D-4, 2012, [9] Mukhopadhyay, TED, p. 1474, 2017, [10] Parihar, TED, p. 1699, 2018, [11] Kumar, TED, 2020, (early access), [12] Samadder, TED, 2020, (submitted), [13] Padovani, JAP, p. 155101, 2017 [14] Okada, VLSI, p. 158, 1998, [15] Nigam, IRPS, p. 381, 1999, [16] Nicollian, IEDM, p. 392, 2005

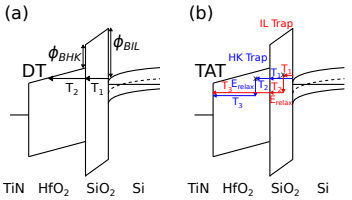

Fig. 1: NMOS band diagram, in inversion, with (a) Direct Tunneling (DT) for DT current (I_{G0}) and (b) Trap Assisted Tunneling (TAT) for SILC (ΔI_G) processes shown

Device	Pre-IL	IL	Pre-HK	HK	Nitridation
D1	Type-A	Chem-Ox 6.5Å	Type-I	18Å	No
D2	Type-A	Chem-Ox 6.5Å	Type-I	23Å	Yes
D3	Type-A	RTP-5Å	Type-IV	23Å	No
D4	Type-B	RTP-3Å	Type-IV	23Å	No
D5	Type-B	RTP-3Å	Type-III	23Å	No
D6	Type-C	RTP-3Å	Type-II	18Å	No
D7	Type-C	RTP-3Å	Type-II	23Å	No
D8	Type-C	RTP-3Å	Type-IV	23Å	Yes

Table 1: Process splits for the HKMG stacks used (Gate First HKMG NFETs), details in [4]

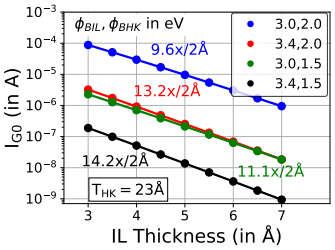

Fig. 2: Simulated I_{G0} vs. IL thickness for different ϕ_{BIL}, ϕ_{BHK} for V_G = 1.3V. Typical rate of I_{G0} increase for scaling IL changes when ϕ_{BIL}, ϕ_{BHK} variations are considered.

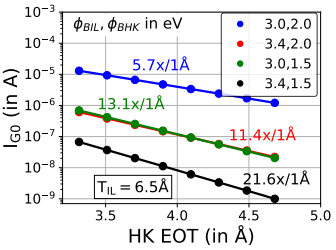

Fig. 3: Simulated I_{G0} vs. HK EOT (k=20) for different ϕ_{BIL}, ϕ_{BHK} for V_G = 1.3V. Typical rate of I_{G0} increase for scaling HK changes when ϕ_{BIL}, ϕ_{BHK} variations are considered.

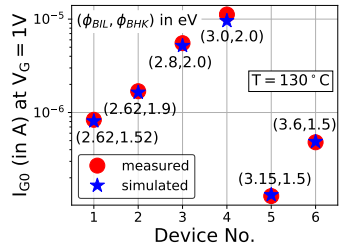

Fig. 4: I_{G0} modeling of measured data for D1-D6, and their ϕ_{BIL} and ϕ_{BHK}, which do not follow the expected T_{IL} and T_{HK} trends due to different pre-IL and pre-HK processes

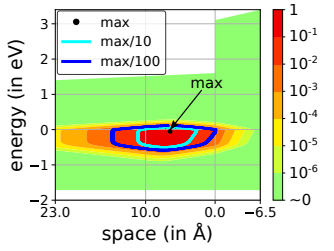

Fig. 5: SILC contour plot (normalised) for an HKMG stack, T_{IL} = 6.5Å, T_{HK} = 23Å, V_{sense} = 1V, T = 130°C. The maximum SILC response point is inside HK

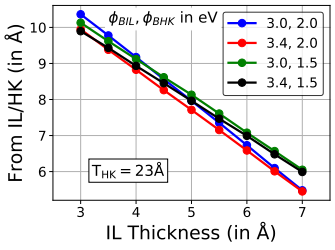

Fig. 6: Peak SILC response location vs. IL thickness for different ϕ_{BIL}, ϕ_{BHK}. The maximum response point moves deeper into HK for lower IL, more so due to lower ϕ_{BIL}

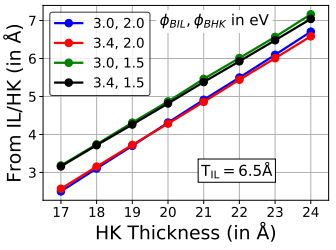

Fig. 7: Peak SILC response location vs. HK thickness for different ϕ_{BIL}, ϕ_{BHK}. The maximum response point moves towards IL/HK interface for lower HK, more so due to higher ϕ_{BHK}

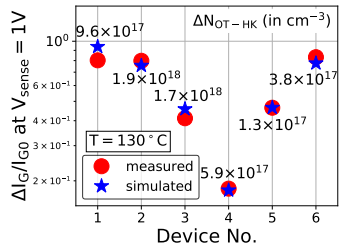

Fig. 8: SILC (at fixed V_G=1.6V) modeling of measured data for D1-D6, with their ΔN_{OT-HK} using IL:HK = 1:300 trap generation rate ratio

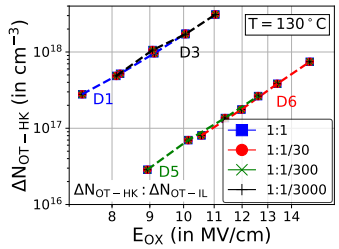

Fig. 9: ΔN_{OT-HK} vs. E_{OX} for D1, D3, D5 and D6, varying in IL or HK thickness, needed to model SILC, for different trap generation rates. SILC is dependent primarily on ΔN_{OT-HK} and insensitive to changes in ΔN_{OT-IL}

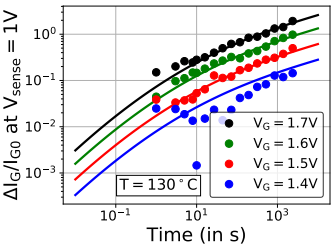

Fig. 10(a): Measured SILC time kinetics and modeling (different V_G) for D6 with T_{IL} = 3Å, T_{HK} = 18Å. Symbols: Data, Lines: TAT Simulation

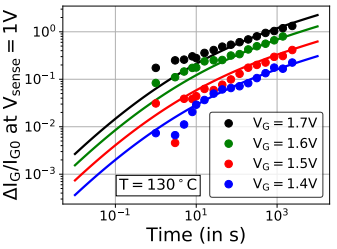

Fig. 10(b): Measured SILC time kinetics and modeling (different V_G) for D2 with T_{IL} = 6.5Å, T_{HK} = 23Å. Symbols: Data, Lines: TAT Simulation

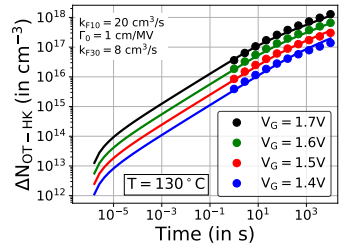

Fig. 11(a): ΔN_{OT-HK} (from SILC in fig. 10(a), trap generation rate ratio IL:HK=1:300) time kinetics and RDD modeling (different V_G) for D6 with T_{IL} = 3Å, T_{HK} = 18Å. Symbols: TAT simulation, Lines: RDD model

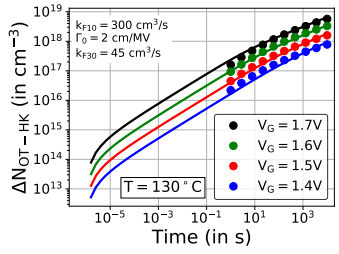

Fig. 11(b): ΔN_{OT-HK} (from SILC in fig. 10(b), trap generation rate ratio IL:HK=1:300) time kinetics and RDD modeling (different V_G) for D2 with T_{IL} = 6.5Å, T_{HK} = 23Å. Symbols: TAT simulation, Lines: RDD model

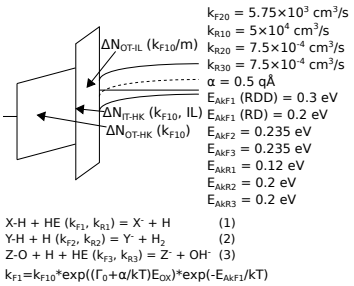

Fig. 12: Trap generation at different locations in the oxide is shown along with RDD chemical equations. Release of H, H induced bond dissociation and eventual diffusion and drift of molecular (H_2) and ionic (OH^-) species control ΔN_{OT} time kinetics

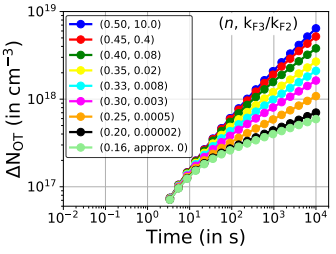

Fig. 13: Simulated stress time kinetics of ΔN_{OT} for different k_{F3} and slope change. The H_2 to OH^- ratio ($\sim k_{F2}$ to k_{F3} ratio) governs the time kinetics power-law slope n at long time. Pure H_2 and pure OH^- results in $n \sim 0.16$ and ~ 0.50, respectively [12]

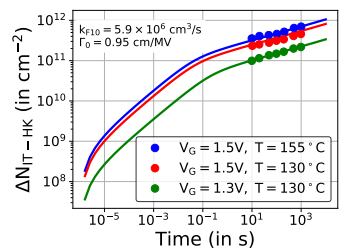

Fig. 14: DCIV-ΔN_{IT-HK} time kinetics and modeling (different V_G) for D7, using pure H_2 diffusion i.e. RD model ((1)+(2) from Fig. 12). Symbols: DCIV, Lines: RD model

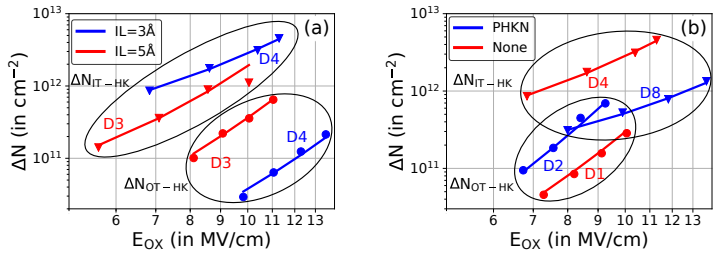

Fig. 15: E_{OX} dependence of ΔN_{IT-HK} and ΔN_{OT-HK} and modeling for variation in (a) IL thickness, (b) PHKN and (c) Pre-HK IFT. ΔN_{OT-HK} and ΔN_{IT-HK} behave differently for IL reduction (reduces and increases) and PHKN (increases and reduces), but similarly with IFT (increases with higher moisture content). Symbols: DCIV or TAT simulation, Lines: RD or RDD model