

# Gate-Last vs. Gate-First Technology for aggressively scaled EOT Logic/RF CMOS

A. Veloso, L.-Å. Ragnarsson, M. J. Cho, K. Devriendt, K. Kellens, F. Sebaai, S. Suhard, S. Brus, Y. Crabbe, T. Schram, E. Röhr, V. Paraschiv, G. Eneman<sup>1</sup>, T. Kauerauf, M. Dehan, S.-H. Hong<sup>2</sup>, S. Yamaguchi<sup>3</sup>, S. Takeoka<sup>4</sup>, Y. Higuchi<sup>4</sup>, H. Tielens, A. Van Ammel, P. Favia, H. Bender, A. Franquet, T. Conard, X. Li<sup>5</sup>, K.-L. Pey<sup>6</sup>, H. Struyf, P. Mertens, P. P. Absil, N. Horiguchi, and T. Hoffmann  
IMEC, <sup>1</sup>also K. U. Leuven and FWO, assignees at IMEC from <sup>2</sup>Samsung, <sup>3</sup>SONY, and <sup>4</sup>Panasonic, Kapeldreef 75, B-3001 Leuven, Belgium; <sup>5</sup>IME, A\*STAR, <sup>6</sup>SUTD, Singapore  
Tel.: +32-16-28 17 28, Fax: +32-16-28 17 06, Email: Anabela.Veloso@imec.be

## Abstract

We report on gate-last technology for improved effective work function tuning with ~200meV higher p-EWF at 7Å EOT, ~2× higher  $f_{\max}$  performance, and further options for channel stress enhancement than with gate-first by taking advantage of the intrinsic stress of metals and gate height dependence. Additional key features: 1) scavenging technique yielding UT-EOT down to ~5Å is demonstrated in gate-last, with high-k deposited first, through the use of an Etch-Stop-Layer with composite nature and similar TDDB reliability to gate-first; 2) controlled alloying for EWF engineering is obtained by careful material selection and tuned metals thicknesses ratio; 3) suppression of abnormal  $L_{\text{gate}}$ - and  $W_{\text{gate}}$ -dependence on  $J_G$ , EOT and NBTI for devices with both high-k and metal deposited last ( $L_{\text{gate}} \geq 35\text{nm}$ ,  $W_{\text{gate}} \geq 80\text{nm}$ ) demonstrates the potential for improved UT-EOT control down to small devices with this scheme.

## Introduction

The recent introduction of high-k/metal gate (MG) into manufacturing has enabled continuance of device scaling and Moore's law [1-3]. Key challenges remain on high-k scalability towards ultra-thin EOT (UT-EOT), with two main approaches being followed for metal gate integration: gate-first (GF) and gate-last (GL=RMG). The latter broadens the range of material options for effective work function (EWF) tuning by allowing low thermal budget after metals deposition, decoupling it from the junction activation (RTA). However, the smallest EOTs reported to date have been demonstrated with GF [4,5]. RMG was shown to considerably enhance the channel stress in short channel devices during dummy gate removal, increasing the benefits from other stress techniques such as embedded-SiGe S/D [2,6]. In this work we explore the impact of metals intrinsic stress vs. gate height ( $H_{\text{gate}}$ ), benchmarking RMG with GF also for EWF(EOT) and RF performance. We show that the scavenging technique previously demonstrated in GF to obtain UT-EOT [5] can be applied in RMG, with no reliability penalty, by exploring the use of an Etch-Stop-Layer (ESL) in high-k first RMG flow (RMG-HKF), and tuning the metals thicknesses ratio for controlled alloying. Finally, we demonstrate the potential for improved UT-EOT at scaled  $L_{\text{gate}}$  and  $W_{\text{gate}}$  by depositing both high-k and metal last (RMG-HKL).

## Device fabrication

The process flow used for device fabrication is schematically illustrated in Fig.1 with options for RMG-HKF or RMG-HKL. Standard processing is used till salicide and CESL deposition. An oxide hard-mask is used for dummy poly gates patterning to prevent silicided gates and HDD implant species diffusing into poly, which would detrimentally impact the poly wet removal step later on.  $H_{\text{gate}}$  is set in 1<sup>st</sup> order by ILD0 CMP [7], with the ILD0 stack optimized to minimize ILD0-loss during dummy poly and dummy dielectric (RMG-HKL) removal steps. In HKL, oxidation in  $O_2$  is used for IL-SiO<sub>2</sub> formation prior to high-k dielectric deposition. Final  $H_{\text{gate}}$  is set by metal-CMP following the WF- + fill(W)-metal depositions. This is a critical step due to the high intrinsic stress of W films (~1GPa as-deposited, exponentially increasing with reduced film thickness) and W grains size, filling gates with  $H_{\text{gate}} \leq 60\text{nm}$  and  $L_{\text{gate}} \geq 35\text{nm}$ .

## Results and discussion

Fig.2 shows an EWF(EOT) overview for GF and RMG stacks, with a clear benefit of the latter towards more p-type EWF values. RMG devices also outperform GF in  $f_{\max}$  (Fig.3) thanks to their lower gate resistance. Several factors play a role in setting the final  $H_{\text{gate}}$ , namely process control [CMPs and conformality of metal depositions vs. gates aspect-ratio ( $AR = L_{\text{gate}} \cdot H_{\text{gate}} / W_{\text{gate}} \cdot H_{\text{gate}}$ )]. An additional factor is stress. The metals intrinsic stress can be used as another strain-enhancement  $H_{\text{gate}}$ -dependent technique (Fig.4), with flexibility to change the films stress by a post-deposition thermal treatment and/or combining CMP with wet/dry etch vs. CMP-only for removing the excess metal from the wafers.

HRTEM images in Fig.5 show that the scavenging technique developed in GF to obtain UT-EOT can also be applied in RMG-HKF, with similar TDDB reliability (Fig.6) [8]. For this, the

presence of an optimized ESL in the gate stack on top of the high-k dielectric is critical. Fig.7 shows a ~3.8Å increase in EOT when skipping the ESL resulting in ~25% performance loss. Lower  $D_{\text{it}}$  indicates ESL's protective role of the underneath high-k dielectric, namely during dummy poly gate removal. Further insight into its nature is provided by physical analysis (EDS in Fig.7c) confirming it to be a composite layer. This is in agreement with simple thermodynamic calculations of the reaction of metal deposited on top of high-k dielectric (e.g., 2nm TiN + Si-cap) with Si during poly deposition and RTA, followed by surface oxidation after dummy poly gate removal (e.g., TiN → TiN + TiSi<sub>x</sub>/TiON). The composite nature also makes ESL a 1<sup>st</sup> order barrier against diffusion of species through it into the high-k, such as Al from a TiAl layer deposited on it. Fig.8a shows that stacks with similar deposited layer thicknesses but different processing behave quite differently in terms of the amount of Al diffusing into and through the ESL (= TiN + TiSi<sub>x</sub>/TiON) vs. TiN from a TiAl layer on top. The considerably smaller  $J_G$  with ESL means less Al diffusing through it. The amount increases for higher thermal budgets. Diffusion is also retarded via oxidation (Fig.8b, with TiN + O<sub>2</sub> → TiN + TiON). Keeping diffusion through the thin layers in the gate stack under control is critical for achieving both low- $V_{\text{TS}}$  (e.g., with Al diffusing into TiN for a more n-type EWF) and CMOS (e.g., by preventing Al diffusing into PMOS WF-metal). Metal layers selection is key. Fig.9 shows TOF-SIMS of two metals screened to compare their resistance to Al diffusion with increased thermal budget. Fig.10 shows a strong correlation of EWF and  $J_G$  on the nominal thicknesses ratio of TiAl/TiN deposited on HfO<sub>2</sub>: a thicker TiAl (for similar TiN underneath) leads to an increase of the amount of Al diffusing into the TiN, shifting EWF towards more n-type values with no leakage penalty. However, beyond a certain ratio, Al starts diffusing through the TiN into the HfO<sub>2</sub> resulting in a sharp leakage increase with no EWF benefits. This ratio control is impacted by the conformality of TiAl and TiN depositions and the gates AR. Fig.11a shows that with good WF-metal thickness vs.  $L_{\text{gate}}$  control, tight low- $V_{\text{T}}$  distributions can be obtained down to small devices. MG thickness variations can lead to  $\Delta V_{\text{T}}$ , as illustrated in Fig.11b for NMOS devices which got reduced TiAl/TiN ratio for  $L_{\text{gate}} < 55\text{-}60\text{nm}$  (as verified by TEM).

In GF devices, an  $L_{\text{gate}}$  dependence on the IL-SiO<sub>2</sub> thickness was observed by HRTEM (Fig.12), with no differences seen for HfO<sub>2</sub> and MG layers. IL-SiO<sub>2</sub> is thicker at gate edges and for smaller  $L_{\text{gate}}$ , and the difference increases for thicker EOTs. This is verified electrically (Fig.13), using a simple fitting model based on the TEM observations. Interestingly, a less steep leakage increase is observed with increasing  $L_{\text{gate}}$  than with the decreasing MG thickness responsible for smaller EOTs [5]. Oxygen diffusion under the gate edges during high thermal steps is thought to be the root cause. In RMG-HKL, the IL-SiO<sub>2</sub> and high-k dielectric are introduced into the devices only after all FEOL high temperature steps, thus more easily avoiding potential sources of oxidation. As a result, less IL-SiO<sub>2</sub> thickness variation (smaller  $\Delta\text{EOT}$ ) is expected, which is electrically confirmed in Fig.14 by flatter  $J_G(L_{\text{gate}})$  and  $J_G(W_{\text{gate}})$  curves, and in Fig.15 by a less steep slope for the NBTI lifetime increase with scaled  $L_{\text{gate}}$ . This makes it potentially more attractive to obtain UT-EOT down to small  $L_{\text{gate}}$  and  $W_{\text{gate}}$  with RMG-HKL.

## Conclusions

Improved EWF tuning, ~2× higher  $f_{\max}$ , and additional options for channel stress enhancement with gate-last were reported. UT-EOT was demonstrated for RMG-HKF by IL-SiO<sub>2</sub> scavenging, using an ESL, with no TDDB reliability impact, and the thicknesses ratio of selected metals tuned for controlled alloying. RMG-HKL potential for improved UT-EOT control was demonstrated by suppression of the abnormal  $L_{\text{gate}}$ - and  $W_{\text{gate}}$ -dependence on  $J_G$ , EOT and NBTI.

## References

- [1] K. Mistry *et al.*, IEDM Tech. Dig. 2007, 247; [2] S. Natarajan *et al.*, IEDM Tech. Dig. 2008, 941; [3] C.-H. Jan *et al.*, IEDM Tech. Dig. 2010, 604; [4] T. Ando *et al.*, IEDM Tech. Dig. 2009, 423; [5] L.-Å. Ragnarsson *et al.*, IEDM Tech. Dig. 2009, 663; [6] J. Wang *et al.*, VLSI Tech. Dig. 2007, 46; [7] K. Devriendt *et al.*, ICPT Tech. Dig. 2010, 130; [8] T. Kauerauf *et al.*, IRPS2011 (accepted).

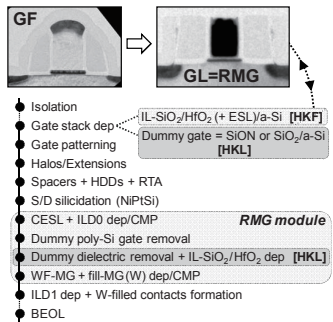


**Fig.1** – Schematics of process flow used for fabrication of high-k/metal gate-last (GL=RMG) devices. TEM images on top highlight the differences with gate-first (GF) MOSFETs.

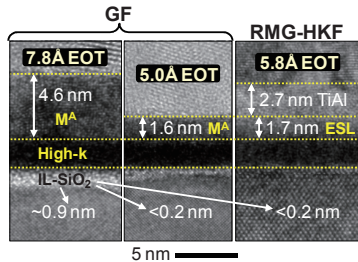


**Fig.5** – HRTEM show reduced IL-SiO<sub>2</sub> thickness for GF stacks with thinner metal electrode (M<sup>A</sup> + Si-cap) => smaller EOT. The same applies to RMG-HKF with a thin ESL introduced before poly removal and WF-metal deposition.

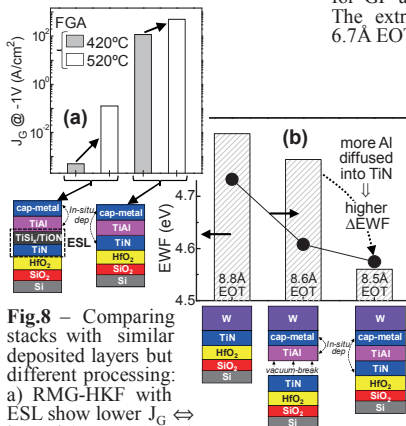


**Fig.8** – Comparing stacks with similar deposited layers but different processing: a) RMG-HKF with ESL show lower  $J_G$  => less Al (none at 420°C) diffused into HfO<sub>2</sub>, with ESL than for TiAl/TiN/HfO<sub>2</sub>; b) lower  $\Delta EWF$  => less Al diffused into TiN from TiAl, if TiN is oxidized.

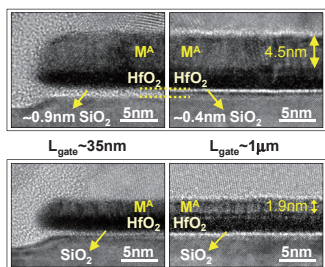


**Fig.12** – HRTEM images show an  $L_{gate}$  dependence on the IL-SiO<sub>2</sub> thickness in GF devices: thicker for smaller  $L_{gate}$  with the difference increased for larger EOTs. Oxygen diffusion under the gate edges is thought to be the root cause.

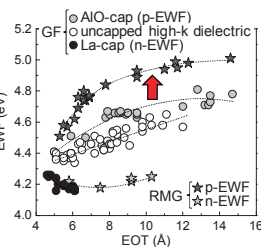


**Fig.2** – EWF vs. EOT overview for GF and RMG gate stacks. For PMOS, a clear improvement towards a more p-type EWF is obtained with RMG.

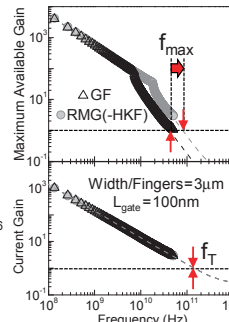


**Fig.3** – RMG outperforms GF in  $f_{max}$  thanks to its lower gate resistance, even if both type of devices exhibit similar  $f_T$  due to comparable drive currents.

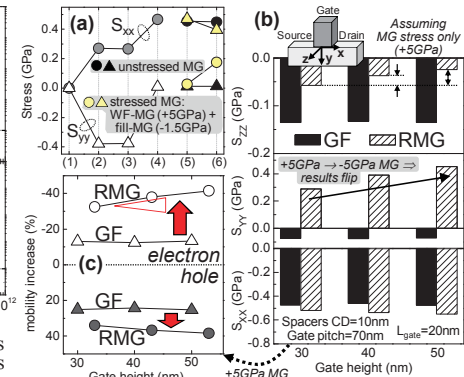


**Fig.4** – a) Simulated stress build-up through RMG flow after: 1) dummy gate patterning; 2) CESL and ILD0 dep; 3) ILD0 CMP; 4) dummy gate removal; and 5+6) WF- + fill-MG dep. Impact of MG intrinsic stress vs. gate height, RMG vs. GF, explored in b), c).

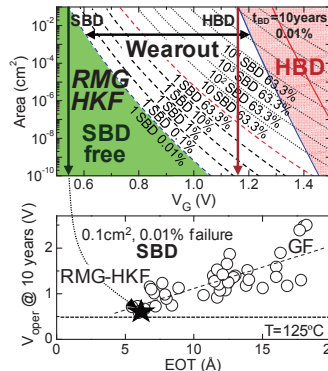


**Fig.6** – Similar TDDDB reliability for GF and RMG-HKF at UT-EOT (bottom). The extrapolated 10 years lifetime map for 6.7Å EOT RMG-HKF is shown on top.

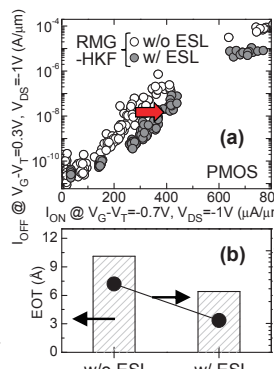


**Fig.7** – a-b) Higher performance and lower EOT for RMG-HKF devices built with a thin ESL. Lower  $D_{IT}$  indicates its protective role of the underneath high-k dielectric, namely during dummy poly gate removal. c) EDS after device fabrication confirm ESL composite nature as illustrated in d).

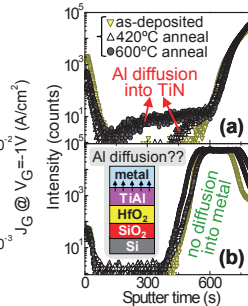


**Fig.9** – Diffusion-barrier metal screening with TOF-SIMS: a) Al diffusion into TiN already occurs at 420°C; b) no diffusion into top metal layer seen here, even after 5min 600°C anneal.

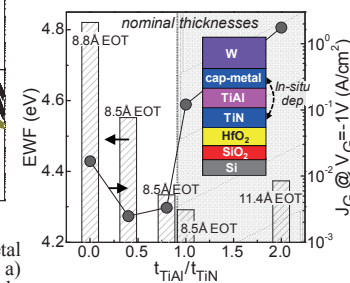


**Fig.10** – The amount of Al diffusing into TiN is controlled by the thicknesses ratio TiAl/TiN. Higher  $\Delta EWF$  w/o  $J_G$  penalty => optimized ratio.

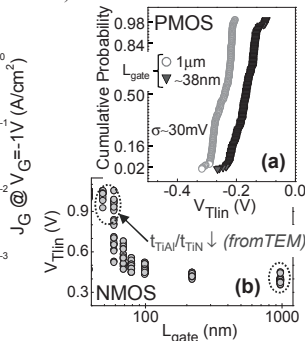


**Fig.11** – a) Good  $V_T(L_{gate})$  control with  $L_{gate}$ -independent WF-metal thickness.  $\Delta V_{TMG}$  can cause  $\Delta V_{T3}$  as shown in (b) where smaller devices got reduced TiAl/TiN ratio in MG.

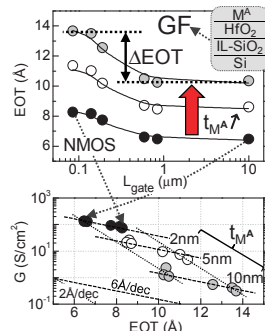


**Fig.13** – ~2-3Å thicker EOT for short channel GF devices. Data fitted by a simple model assuming thicker IL-SiO<sub>2</sub> at the edges (top). Less steep leakage ↑ with  $L_{gate}$  ↑ than with  $t_{MA}$  ↑ (bottom).

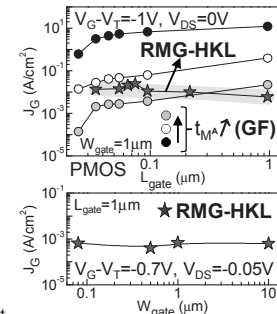


**Fig.14** – Flatter  $J_G(L_{gate})$  for RMG-HKL than for GF devices, suggesting smaller  $\Delta EOT(L_{gate})$  and also no significant  $\Delta J_G(W_{gate})$ .

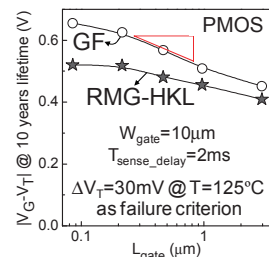


**Fig.15** – Improved NBTI lifetime for smaller GF devices is derived from their thicker EOT. A less steep slope for RMG-HKL corroborates the smaller  $\Delta EOT(L_{gate})$  for these devices.