Extended Abstracts of the 2005 International Conference on Solid State Devices and Materials, Kobe, 2005, pp.444-445

H-4-1

Feasibility analysis of direct tunneling through medium-κ dielectrics for embedded RAM applications

B. Govoreanu, R. Degraeve, T. Kauerauf[(1)], W. Magnus, D. Wellekens, G. Groeseneken[(1)], J. Van Houdt

IMEC, Kapeldreef 75, B-3001 Leuven, Belgium
[(1)]also with Katholieke Universiteit Leuven, ESAT Dept., Belgium

Phone: +32-16-281337, Fax: +32-16-281844, E-mail: govorean@imec.be

Abstract – *We propose medium-κ dielectrics for direct tunneling floating gate memory devices, targeting embedded RAM applications. We found that SiON offers best performance if voltage reduction overrules refresh time, while Hf-silicates would be preferred if the refresh time is more critical. Our analysis is based on a direct tunneling current model, a Response Surface Methodology and experimental data on small MOSFET's. The impact of dielectric degradation during cycling is studied for scalability towards the 32 nm node.*

Introduction

Floating gate (FG) MOS structures with ultra-thin SiO_2 tunnel dielectric (TD) [1] have potential for embedded RAM applications. In this work, we propose the use of medium-κ materials (κ < ~10) as TD's in a FG direct tunneling (DT) memory device for embedded RAM applications. A systematic analysis of the intrinsic performance of the FG device is carried out, using in-house developed tools. We show that medium-κ TD's improve the performance of the DT memory. SiON is best suited for low-voltage/high-speed operation, whereas HfSiON combines very long refresh times with a low EOT. Furthermore, based on our extensive measurement database, we infer reliability-imposed scalability limits of the medium-κ DT-RAM concept to be beyond the 32 nm technology node.

Device Principle

The band diagrams in **Fig. 1** show the device operating principle. The erased state is associated with the neutral FG, whereas the programmed state corresponds to excess electrons stored on the FG, after a positive programming pulse was applied. In order to have slow FG discharge of the programmed state, the FG potential variation should remain below the FG/substrate workfunction difference. This constraint defines the optimal range for the threshold voltage shift, ΔV_T, between the two logic states.

Models

The Si/TD/FG structure is treated quantum-mechanically and the potentials of the stacked gate structure (**Fig. 1**) are determined self-consistently, using a FG capacitor model. The DT current through the TD is calculated using our recently developed model [2], based on the estimation of the lifetimes of the quasibound states in the inverted MOS channel. The model is used to fit with excellent accuracy experimental I_G-V_G curves (**Fig. 2**), from which the relevant material parameters have been consistently extracted, for **(a)** an In-Situ Steam Generated (ISSG) oxide with Decoupled Plasma Nitridation at 10 mTorr, for 30 s (here referred to as SiON) and **(b)** a Hf-silicate, deposited by MOCVD, with a 1 min, 800 C NH_3 PDA, following an IMEC clean [3], and with a 10 s poly activation anneal at 1000 C (HfSiON). A set of 4 factors, namely the control gate (CG) voltage (V_{CG}), TD thickness (EOT_{id}), coupling factor (α_G), and substrate doping (N_s), expected to be most important for the memory performance assessment of the FG structure are considered simultaneously, by using a Response Surface Methodology (RSM) [4]. We have derived RSM models for the main RAM performance factors, including the V_T-shift ($\mathbf{\Delta V_T}$), the programming speed ($\boldsymbol{\tau_P}$), the refresh time ($\boldsymbol{\tau_R}$) and the read disturb time ($\boldsymbol{\tau_D}$), for structures with different TD's, focusing on SiON and HfSiON, but also including conventional SiO_2 (as reference) and HfO_2. In all cases, the model quality regression factor exceeded 0.98, making them genuine design charts, able to account for the entire parameters space at once.

Results

A. The intrinsic performance

We discuss the design charts for SiON, some of which are shown in **Fig. 3**. For programming, the CG voltage and the TD thickness are the most important factors (**Fig. 3,a**), whereas N_s was the least sensitive from the 4-factor set, in an interval from 2.10^{17} to 8.10^{17} cm^{-3}. Programming times of 10 ns and below can be achieved for a V_T-shift of 0.5 V, for any point to the left of the dashed line. A 2.5 V CG pulse allows for 10 ns programming for a SiON of 1.42 nm EOT, whereas programming at 2 V (i.e., only $2V_{DD}$, in a 45 nm technology node) would be possible for an EOT of 1.32 nm. The coupling factor (**Fig.3,b**) is only of secondary importance for programming, as revealed from the ΔV_T chart corresponding to τ_P = 10 ns. Its optimal value depends on the programming voltage, and is typically less than 0.7 (dashed line). This is the balance between the competing effects of a higher amount of the charge/cycle that has to be transferred onto the FG to achieve a given V_T shift, for an increasing α_G and a smaller voltage drop over the TD, which reduces the electron injection from the inverted channel, for a decreasing α_G. The coupling factor is however important for improving the refresh performance, and a higher value allows for a thinner TD, at identical refresh times (**Fig. 3,c**). Assuming S/D overlaps similar to contemporary MOS devices, refresh times of 64 ms or longer are not compatible with TD's enabling 10 ns low-voltage programming. The design window for 200 μs refresh time at 2.5 V programming ("ABC" in **Fig. 3,c**) is very narrow and requires a very high coupling factor, of more than 0.85, due to the S/D extension-dominated FG discharge (inset, **Fig. 2**,**a**). Removing the S/D overlaps by a technology workaround causes the 200 μs refresh time to be no more an issue and significantly enlarges the 64 ms design window ("ABCD", **Fig. 3,d**) down to coupling factors of 0.7. Read disturb has also been investigated and read-out times of up to 1 μs are not causing more than 20 % charge loss, for read pulses as high as 1.2 V. Erasing at 10 ns is not an issue when the S/D overlap is present, while erasing in less than 100 ns is feasible when removing it. All the considered materials were subject to a similar treatment. **Fig. 4** summarizes the main performance results for conventional and medium-κ TD's. SiON is the best candidate for low-voltage embedded DT-RAM, offering about 0.7 V decrease of the programming pulse, as compared to SiO_2 at similar speed and allowing for a $2V_{DD}$ CG pulse at 10 ns/0.5 V-shift and with a 200 μs refresh at 1.3 nm EOT. 64 ms refresh is possible for α_G's higher than ~0.75, for a 1.4 nm EOT. HfSiON requires higher programming pulses, due to the reduced DT current. However, it has clearly superior refresh times at lower EOT's, due to a built-in κ-asymmetry, making the tunnel barrier less sensitive to the applied bias (Reverse-VARIOT effect) [5], thus reducing the FG discharge. SiO_2/HfO_2 TD's behave similar to HfSiON, however their use is questionable due to Q_{BD} limitations [6].

B. Scalability & Reliability projections

Small-area ($0.25x0.35\mu m^2$) transistors have been subjected to positive CVS and the current through a single generated trap has been extracted, as discussed in our recent work [7]. Single-trap IV curves have been divided into clusters, using a Jarvis-Patrick nearest-neighbor partitioning algorithm [8]. Most of the traps cause a current relatively more important in the low bias range (**Fig. 5,a**, inset), as compared to the feature-size (F) dependent DT current. The average current of a typical IV cluster is fitted using a trap-assisted tunneling (TAT) model (**Fig. 5.a**). This permitted extraction of the parameters of the generated traps, which were subsequently used in assessing the impact of a typical defect on the FG charge retention. For a programmed device, the electrons leak from the FG to the substrate. The contribution of the average current of a typical cluster to the total FG discharge current (at V_{FG}<0) through the TD becomes more important when the device area is scaled down, as shown in **Fig. 5,b** for F down to 32 nm and adversely affects the FG charge loss (**Fig. 6,a**). We noticed that traps situated over the S/D overlap region draw even more current, which is one more argument in favor of a tight control of the FG to S/D overlap. The remaining V_T-window is reduced as compared to the trap-free case. However, there is still a sufficient margin for a tolerated loss of the V_T-window of up to 20 %, even below the 32 nm node (**Fig. 6,b**). **Fig. 7** shows the estimated number of cycles to reach the Q_{BD} (calculated as a 1 μA current increase), for SiON and HfSiON, with F = 45 nm. The degradation in 1 cycle is determined from the time-dependent voltage drop over the TD, assuming ΔV_T = 0.5 V. For the measured samples, SiON shows the best performance, with over 10^{22} cycles. Available SiO_2/HfO_2 samples of ~1 nm EOT reached the Q_{BD} criterion at a 10 ns-compatible V_{CG} after very few cycles and cannot be used as a TD for DT-RAM memory, with the present targets. Further statistical analysis of the trap effects is in progress.

Conclusion

We demonstrated that using medium-κ TD's brings significant improvement in the performance of the DT FG memory. Ultrathin SiON allows for low-voltage (down to 2 V) 10 ns operation, with a 200 μs refresh for about 1.3 nm EOT, or 2.5 V/10 ns/64 ms for about 1.4 nm. HfSiON offers very long retention times, combined with a thin EOT. The Q_{BD}–based endurance estimation considerably exceeds 10^{12} cycles, and the structure shows scalability at least down to the 32 nm node, provided that the FG to S/D overlaps are well controlled, ideally removed.

References:

[1] K.Tsunoda, et al: VLSI Tech. Dig., pp. 152-153, 2004.
[2] B.Govoreanu, et al: IEEE Trans. El. Dev., **51**(5): 664-673, 2005.
[3] M.Heyns,et al: IBM J. Res. Dev., **43**(3): 339-350, 1999.
[4] R.Myers, et al: Technometrics, **31**(2): 137-157, 1989.
[5] B.Govoreanu, Ph.D. dissertation, KU Leuven, 2004.
[6] R. Degraeve, private communication.
[7] R.Degraeve, et al: IRPS 2005, in press.
[8] R.Jarvis & E.Patrick: IEEE Trans. Comp., **C-22**: 1025-1034, 1973.

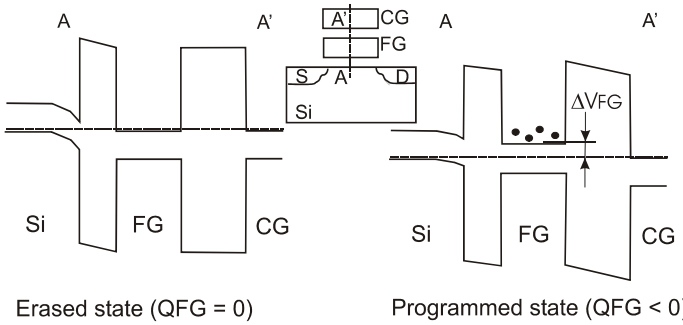

Fig. 1: Band diagrams, showing the memory device operating principle. The erased state corresponds to the neutral FG, while the programmed state is for excess electrons on the FG. The optimal window for the V_T-shift is determined by the FG/substrate workfunction difference (ϕ_{ms}).

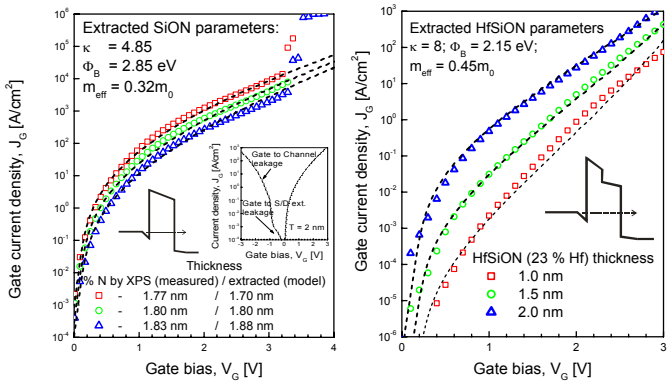

Fig. 2: Measured and calculated [2] direct tunneling currents through (a) SiON and (b) HfSiON-23% samples. All electrical and physical characterization data are in excellent agreement and unique sets of material parameters have been used for all calculations. The inset in (a) shows the dominant current for negative polarity (S/D overlap or channel).

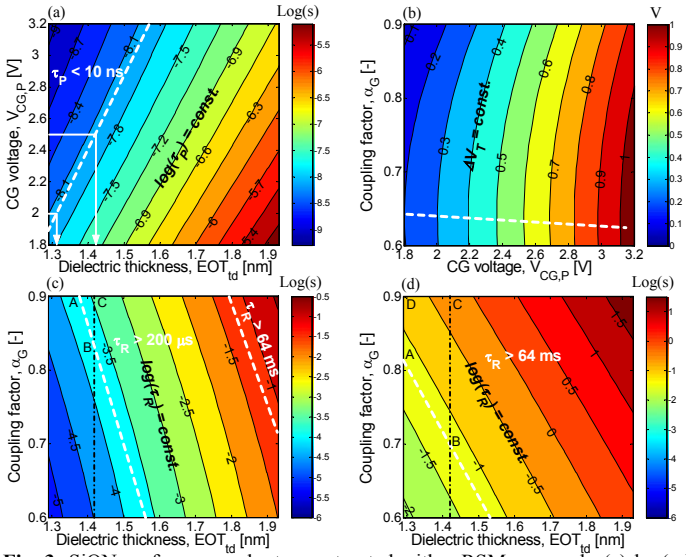

Fig. 3: SiON performance charts constructed with a RSM approach: (a)-log(τ_P) design window vs. programming voltage (V_{CG}) and TD thickness (EOT_{td}), for a coupling factor α_G = 0.75. (b)-the V_T shift vs. α_G and V_{CG} for EOT_{td} = 1.4 nm. (c)-log(τ_R) vs. α_G and EOT_{td} with S/D overlap. (d) Same as (c), but without the S/D overlap. For each chart, the substrate doping has been set at 5.10^{17} cm^{-3}.

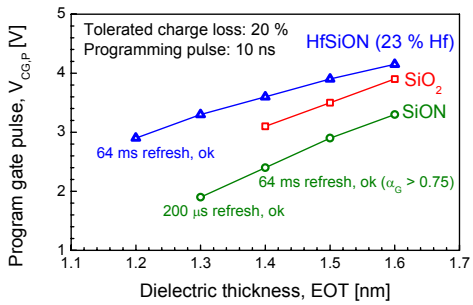

Fig. 4: Programming performance summary, for SiO_2, SiON and HfSiON tunnel dielectrics, for a 10 ns pulse and 0.5 V-V_T-window. The refresh criterion is defined at 20 % charge loss.

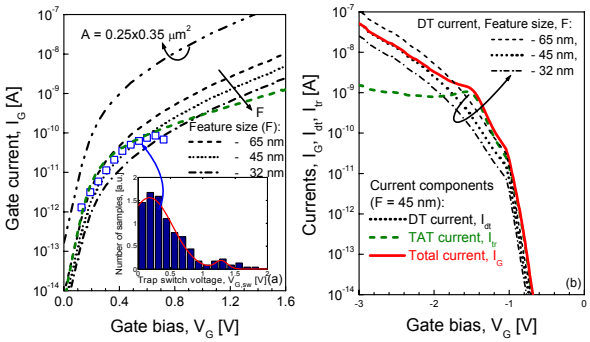

Fig. 5: (a) TD currents through a larger and scaled area cells (black lines), scaling with the feature size-dependent FG area. The symbols show the average current of a representative cluster of measured single-trap currents [7], the dashed green curve is the fitted TAT current using a 1D neutral trap model. The inset shows that most of the traps are relevant at low biases, corresponding to trap depths in a range of 2.4 eV–2.6 eV. (b) Contribution of the current due to one typical defect to the total FG discharge current. All experimental data are taken on 1.44 nm EOT SiON samples.

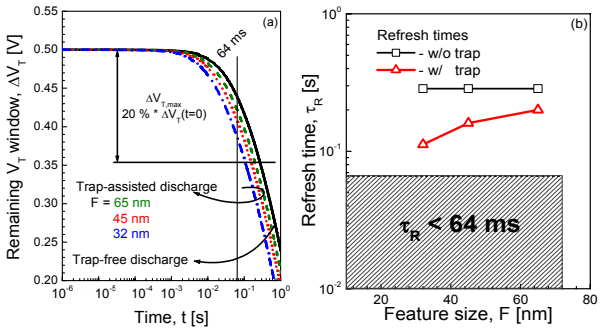

Fig. 6: (a) Impact of a typical defect on the FG charge loss, using the FG leakage currents shown in Fig. 5,b. (b) Impact of the trap on the refresh rate of an F-dependent cell size. For F = 32 nm, the refresh rate is still larger than 100 ms.

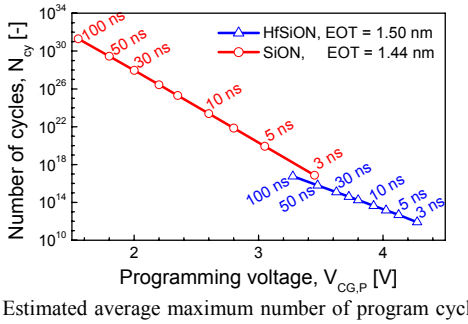

Fig. 7: Estimated average maximum number of program cycles as a function of the program voltage. Failure criterion was Q_{BD} determined at a 1 μA current increase and assuming a trap generation rate of 0.4 [7]. Since the degradation is dominated by the initial part of the programming pulse, a larger V_T shift will need a longer programming time, but produce similar degradation.