Benchmarking and optimization of trench-based multi-gate transistors in a 40 nm non-volatile memory technology

Romeric Gay, Vincenzo Della Marca, Hassen Aziza, Arnaud Regnier, Stephan Niel, Abderrezak Marzaki

Benchmarking and optimization of trench-based multi-gate transistors in a 40 nm non-volatile memory technology

Romeric Gay
STMicroelectronics
13106 Rousset, France
romeric.gay@st.com

Vincenzo Della Marca
Aix-Marseille University
CNRS, IM2NP UMR 7334
F-13397 Marseille, France
vincenzo.della-marca@im2np.fr

Hassen Aziza
Aix-Marseille University
CNRS, IM2NP UMR 7334
F-13397 Marseille, France
hassen.aziza@im2np.fr

Arnaud Regnier
STMicroelectronics
13106 Rousset, France
arnaud.regnier@st.com

Stephan Niel
STMicroelectronics
38926 Crolles, France
stephan.niel@st.com

Abderrezak Marzaki
STMicroelectronics
13106 Rousset, France
abderrezak.marzaki@st.com

***Abstract*— This paper addresses the design and characterization of different architectures of novels high-density multi-gate transistors manufactured in a 40 nm embedded Non-Volatile Memory technology. The proposed multi-gate architectures are based on vertical transistors integrated in deep trenches built alongside the main transistor. Thanks to the built-in trench, the proposed manufacturing process increases the transistor width without impacting its footprint. The electrical behaviour of the different multi-gate transistor architectures is studied and compared based on I-V characteristics. Relevant physical and electrical parameters such as the device footprint, the ON and OFF currents along with the threshold voltage and subthreshold slopes are extracted in order to determine the best candidate among the three studied architectures.**

Keywords—Multi-gate, Triple gate transistor, Dual gate transistor, Trench, Non-volatile-memory.

I. INTRODUCTION

The continuous scaling of physical feature sizes of silicon-based MOSFET transistors is highly desirable in the semiconductor industry in order to increase functionality and reduce the cost of a large variety of integrated circuits and systems. However, continuous scaling has to solve physical limits and challenges imposed by the CMOS manufacturing process [1]. The latter are dictated by the optical lithography which requires new methods such as EUV laser when scaling down below 10 nm [2] and ultimately by the manufacturing cost. Scaling directly impacts the physical parameters of integrated devices such as the gate oxide thickness, the leakage current [3], and the short-channel effects [4]. Therefore, as most of conventional transistor parameters are now reaching fundamental limits, novel transistor architectures are highly desirable to meet cost and performances requirements of high-volume products. This is particularly true for embedded Non-Volatile Memory (e-NVM) technology which has generated interest as a potential solution for several key mass market including Internet of Things (IoT) applications.

In this context, alternatives to classical planar MOSFET have emerged in the last few years such as multiple-gate transistor [5], triple gate transistor [6], gate all around [7], FinFET [8]. Alternatively, and mostly for integrated passive devices, deep trenches are used in various ways, including as insulators, sensors, or capacitors to increase integration density [9].

The aim of this work is to compare the performances of three Multi-Gate Transistors (MGT) architectures manufactured in a 40 nm e-NVM. Multi-Gate Transistors are fabricated directly from a classical MOS transistor where lateral conduction channels are added by integrating deep trench transistors alongside the planar transistor channel. After fabrication, electrical measurements are performed for each MGT transistor configurations to assess their performances versus the equivalent planar transistor considered as the reference architecture.

Section II presents the trench transistor concept, its manufacturing process and the novel MGTs architectures based on trench transistor structure. In Section III, we report an experimental analysis of electrical parameters for different MGTs. The section IV discusses the proposed MGTs architectures. Finally, Section V concludes this paper.

II. TRENCH-BASED MGT

A. Trench transistor manufacturing process

The trench transistor uses a specific process steps already available in the targeted eSTM™ Non-Volatile Memory manufacturing process [10]. Trenches fabrication includes different actions. Firstly, the definition of active and oxide Shallow Trench Isolation (STI) regions presented in Fig. 1. (a) is realized. Then, a photolithography step is carried out and the trench etching is performed (Fig. 1. (b)). Trench oxide is grown at the bottom and along sidewalls of the trench which is then

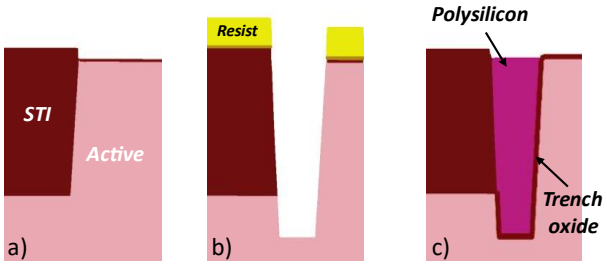


Fig. 1. Trench patterning detailed with (a) STI/Active definition, (b) trench etching and (c) trench filling.

filled with an in-situ doped polycrystalline silicon (Fig. 1c) to build the trench transistor gate. Finally, Chemical and Mechanical Polishing (CMP) is performed to remove the polysilicon excess.

B. MGT transistor architectures

In order to design the different MGT architectures, we started from a planar gate transistor denoted Single Gate Transistor (SGT) (Fig. 2. (a)). The SGT is integrated in the NVM cell environment (as already mentioned), so it uses the process flow of the NVM environment. No additional steps or masks are required to fabricate the device. Therefore, the SGT is processed from the Flash floating gate stack presented in Fig. 2. (a). It consists of a tunnel oxide, a first polysilicon layer (poly1), an oxide/nitride/oxide (ONO) inter-poly dielectric and a last polysilicon layer (poly2). Note that for this device, both poly1/poly2 are shorted, and the contact is performed through the poly2/ONO stack. Finally, source and drain are created with n+ implantations.

For the MGT, trench transistors are integrated at both sides of the SGT architecture represented in Fig. 2. (b). This architecture is referred to as Triple Gate Transistor (TGT).

The resulting TGT has three conduction channels, and therefore it can be considered as three MOS transistors connected in parallel. All the conduction channels are localized between the drain and the source. One is controlled by the planar gate (PG) (poly1+tunnel oxide) and its conduction path is represented by the blue arrow in Fig. 2 (c). The second and third channels are controlled by the gate trench (TG). The current path of one single TG channel is represented by the orange arrow in Fig. 2 (c). The second TG current path is not represented (hidden by the active region). Note that it is possible to integrate only one trench transistor into the SGT architecture. In this case, the final device owns two gates and so is referred to as Dual Gate Transistor (DGT).

C. MGT samples design

In this section, we describe the design of the studied devices. The first layout presented in Fig. 3. is the SGT structure which is the reference architecture, with only one planar gate width (W_P) of 200 nm. The second architecture presented in Fig. 3. is the DGT structure which corresponds to the SGT with one integrated trench transistor. The DGT planar width is reduced to 180 nm because of vertical trench integration. One can notice that the final DGT width depends on the PG plus the vertical

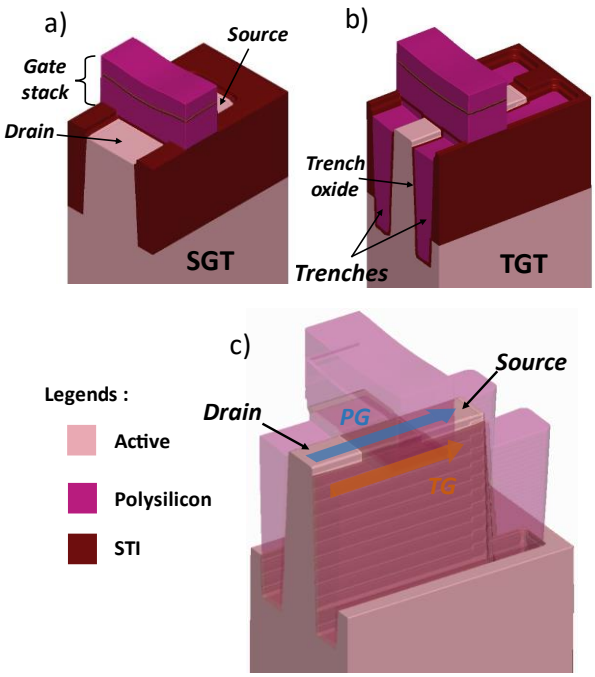


Fig. 2. Trenches integration. (a) Initial MOSFET SGT and (b) Triple Gate Transistor. (c) Conduction mode of TGT.

TG channel widths. The third architecture presented in Fig. 3. is the TGT1 structure which is composed of the SGT planar transistor and two vertical trench transistors. In this case the TGT1 width includes the widths of the two TG and the W_P of the PG which is defined by the space between the trenches (160 nm). Finally, the fourth architecture is the TGT2 and corresponds to the TGT1 architecture with a smaller W_P=110 nm. For each architecture presented in Fig. 3, the equivalent electrical schematic is provided to highlight the number of transistors as well as the area. The SGT and the DGT have a similar footprint, 0.304 µm² and 0.338 µm² respectively. Thus, the integration of one trench has a slight impact on the area with respect to the SGT transistor. Regarding the TGT1 structure, the integration of a double trench results in an area increase. The scaling of W_P in the TGT2 enables a gain of 10% in terms of occupied surface.

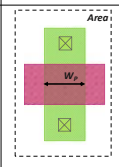
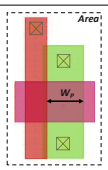
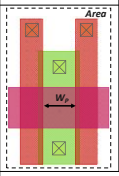
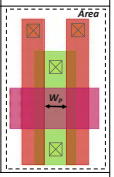
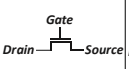
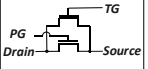
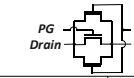
	SGT	DGT	TGT1	TGT2
Layout				
Electrical Schematic				
W_P (nm)	200	180	160	110
Area (µm²)	0.304	0.328	0.412	0.374

Fig. 3. Layout, electrical schematic, planar width (W_P) and area of each multi-gate architectures.

III. Experimental Results

The aim of this section is to study the performances of the three MGT architectures. This evaluation is conducted by analyzing drain current (I_D) versus gate voltage (V_G) electrical characteristics. Performance parameters extracted from the experimental results will be compared and discussed. The SGT being considered as the reference transistor. For the multi-gate transistors, TG and PG are shorted to study the equivalent transistor behavior.

A. MGT transistors benchmark

Fig. 4 shows I_D-V_G electrical characteristics in linear and saturated mode with drain voltages (V_D) of 0.1 V and 2 V respectively for: (a) SGT, (b) DGT, (c) TGT1 and (d) TGT2. A summary of the most relevant parameters in saturated mode is shown in Table I. The ON current parameter (I_{ON}) is extracted for V_G = 4 V and the OFF current parameter (I_{OFF}) for V_G = 0 V. The threshold voltage (V_t) is measured at the maximum of the transconductance, while the subthreshold slope (S_s) is extracted on two decades of drain current characteristics from 0.1 nA to 10 nA.

In Fig. 4, the standard SGT (Fig. 4. (a)) and the DGT (Fig. 4. (b)) characteristics are reported. The DGT conduction current is higher than the SGT as we can see on linear curves. This difference is due to the presence of the vertical trench transistor which leads to an additional conduction channel for the DGT architecture. This conduction current increase is also observed in Fig. 4. (c)-(d) for both triple gate transistor configurations TGT1 and TGT 2. It is more pronounced because each TGT architecture benefits from a triple channel conduction. These observations are confirmed by the I_{ON} current parameter extracted in saturated mode presented in Table I. Indeed, we observe that the TGTs structures provide an I_{ON} driving current three times higher than the SGT I_{ON} and two times higher than the DGT one. One can notice that the driving current of the TGT1 is close to the TGT2 with a smaller area.

Regarding the I_{OFF} parameter in Table I which represents the leakage current, the SGT and DGT values are equivalent (I_{OFF} = 3E-14 A). Thus, the integration of one trench transistor does not impact the device leakage current. However, we observe an increase of one decade for the TGT1 architecture and two decades for the TGT2 architecture. These variations are due to the two trenches integration on the planar gate structure, inducing the hump effect. This effect is discussed in the next paragraph. Note that the leakage current increases when the planar width is decreased.

In Fig. 4. (a)-(d) for drain current characteristics in logarithmic scale, we can notice a variation of subthreshold slopes (S_s parameter in Table I) for the different architectures. We observe that S_s decreases with the number of integrated trenches. Indeed, the S_s value of the DGT device is 95 mV/dec, lower than the SGT one which is equal to 107 mV/dec. Besides, both Ss values of the TGT architectures are lower than the DGT.

In Fig. 4. we can note at the beginning of I_D-V_G characteristics on logarithmic blue curve, a hump effect for the TGT is present (Fig. 4. (c)-(d)) while is less pronounced in the DGT characteristics (Fig. 4. (b)) and absent for the SGT (Fig. 4. (a)). This hump effect is explained by the abrupt transition between the active region and the trench transistor [11]. This latter leads to an overconsumption of oxide at the active area corner, inducing a gate electric field modification at the channel edge. Moreover, this topologic modification can reduce the TG channel doping. Therefore, all these contributions activate the parasitic transistor at the active/trench corner with a reduced threshold voltage. This effect has an impact on the subthreshold region and can explain the I_{OFF} leakage current increase for the TGT architectures [12].

The following conclusions can be extracted from Table I. A higher threshold voltage is measured for the SGT (V_t = 1.42 V) compared to the TGT architectures (V_t = 0.95 V for TGT1 and V_t = 1.06 V for TGT2). This variation can be explained by the different planar and vertical transistors parameters depending on the TGT architecture. Indeed, the PG width is reduced by the trenches integration. For the TG, its width corresponds to the source and drain implantation depth and its length corresponds to the source/drain space. Moreover, the

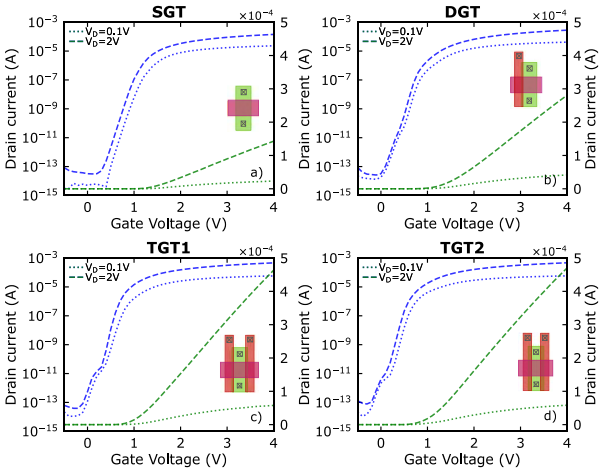


Fig. 4. I-V characteristics in linear and saturated conduction modes in log and linear scales for: (a) SGT, (b) DGT, (c) TGT1 and (d) TGT2.

	I_{ON}(µA)	I_{OFF}(A)	V_t(V)	S_s(mV/dec)
SGT	143	3e-14	1.42	107
DGT	280	3e-14	1.24	95
TGT1	463	6e-13	0.95	79
TGT2	470	4e-12	1.06	83

Table I. Main parameters for the three MGT architectures in saturated mode (V_D = 2V).

trench oxide thickness slightly decreases down toward the trench sidewalls and the TG channel implantation is non-uniform versus the substrate depth. On the other hand, the doping reduction of the active corner can impact both PG and TG channel conduction. Therefore, all these contributions can enable a V_t shift for the TGT transistors compared to the SGT transistor.

B. Discussions

Table II provides a comparison of the three MGT architectures with respect to the planar SGT transistor. Three device parameters are considered: the I_{ON} current, the S_s and the area. Moreover, we consider two figures of merits for the multi-gate architectures: the I_{ON}/I_{OFF} and I_{ON}/Area ratios.

We can see in Table II a driving capability enhancement of about 96% for the DGT architecture and roughly 220% for both TGT architectures with respect to the SGT one. Regarding the S_s parameter, we observe that for the DGT, a 11% decrease is reported and for the TGT1 and TGT2, a 26% and 49% decrease are reported respectively.

Concerning the area, all the MGT structures have a larger footprint than the SGT one. Two trenches integration leads to a 36% increase of the area for the TGT1 architecture and a 23% increase for the TGT2. However, the one trench integration (DGT architecture) results in a lower increase of only 8% compared to the SGT one.

According to the I_{ON}/I_{OFF} figure of merit, the two TGT structures show a degradation of this parameter with a 84% and 92% decrease. However, the DGT architecture shows a strong improvement of the I_{ON}/I_{OFF} ratio by 96% with respect to the SGT architecture.

Finally, the Ion/Area ratio increases versus the number of integrated trenches with an optimal value for the TGT2 architecture.

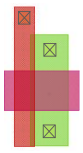
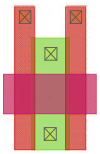
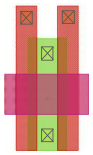
	DGT	TGT1	TGT2
			
I_{ON} (µA)	+ 96 %	+ 224 %	+ 229 %
S_s (mV/dec)	- 11 %	- 26%	- 22 %
Area (µm²)	+ 8 %	+ 36 %	+ 23 %
I_{ON}/I_{OFF}	+ 96 %	- 84 %	- 92 %
I_{ON} / Area (µA/µm²)	+ 81 %	+ 139 %	+ 167 %

Table II. Comparisons of multi-gate architectures performances with respect to SGT reference transistor.

IV. Conclusion

We have investigated the electrical behavior of three different architectures of a novel multi-gate transistor. The MGT is made of distinct conduction channels controlled by independent gates (PG and TG). Different MGT architectures have been manufactured with single and double trench integration (DGT, TGT1 and TGT2). Experimental results demonstrate the functionality of the MGT architectures. Moreover, the performances of the MGT architectures are compared to a regular planar transistor based on I_{ON}, I_{OFF}, S_s, area and two figures of merit which are I_{ON}/I_{OFF} and I_{ON}/Area ratios. Experimental results demonstrate optimal performances for the Dual Gate Transistor architecture (DGT) with improvements of 96% for the driving capability, 11% decrease of Ss, 96% for the I_{ON}/I_{OFF} ratio and finally 81% for the I_{ON}/Area ratio, making it ideal for analog applications requiring high driving capability as well as low power applications due to its high I_{ON}/I_{OFF} ratio.

References

[1] K. Seshan, “Limits and Hurdles to Continued CMOS Scaling,” in Handbook of Thin Film Deposition, 4th ed., pp. 19–41, doi:10.1016/b978-0-12-812311-9.00002-5.

[2] T. Watanabe, "Current status and prospect for EUV lithography," 2017 7th International Conference on Integrated Circuits, Design, and Verification (ICDV), 2017, pp. 2-7, doi: 10.1109/ICDV.2017.8188625.

[3] M. Koh et al., "Limit of gate oxide thickness scaling in MOSFETs due to apparent threshold voltage fluctuation induced by tunnel leakage current," in IEEE Transactions on Electron Devices, vol. 48, no. 2, pp. 259-264, Feb. 2001, doi: 10.1109/16.902724.

[4] A. Chaudhry and M. J. Kumar, "Controlling short-channel effects in deep-submicron SOI MOSFETs for improved reliability: a review," in IEEE Transactions on Device and Materials Reliability, vol. 4, no. 1, pp. 99-109, March 2004, doi: 10.1109/TDMR.2004.824359.

[5] I. Ferain et al., “Multigate transistors as the future of classical metal–oxide–semiconductor field-effect transistors.” Nature, vol. 479, issue 7373, pp. 310-316, Nov. 2011, doi: 10.1038/nature10676.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[6] B. Doyle et al., "Tri-Gate fully-depleted CMOS transistors: fabrication, design and layout," 2003 Symposium on VLSI Technology, pp. 133-134, Aug. 2003, doi: 10.1109/VLSIT.2003.1221121.

[7] S. Kim et al., "Vertically Stacked Gate-All-Around Structured Tunneling-Based Ternary-CMOS," IEEE Transactions on Electron Devices, vol. 67, no. 9, pp. 3889-3893, Sept. 2020, doi: 10.1109/TED.2020.3011384.

[8] S. Natarajan et al., "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 µm2 SRAM cell size," 2014 IEEE International Electron Devices Meeting, 2014, pp. 3.7.1-3.7.3, doi: 10.1109/IEDM.2014.7046976.

[9] S. Banzhaf et al., "Post-trench processing of silicon deep trench capacitors for power electronic applications," 2016 28th International Symposium on Power Semiconductor Devices and ICs (ISPSD), 2016, pp. 399-402, doi: 10.1109/ISPSD.2016.7520862.

[10] S. Niel et al., "Embedded Select in Trench Memory (eSTMTM), best in class 40nm floating gate based cell: a process integration challenge," IEDM, 2018.

[11] S. Schwantes et al., "Characterization of a new hump-free device structure for smart power and embedded memory technologies", Microelectronic Engineering, vol. 81, pp. 132-139, Jul. 2005, doi: 10.1016/j.mee.2005.04.007.

[12] Y. Joly et al., "Impact of hump effect on MOSFET mismatch in the sub-threshold area for low power analog applications," 2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology, 2010, pp. 1817-1819, doi: 10.1109/ICSICT.2010.566768.