Introduction to FeFET Technology

Sven Beyer[1], Stefan Dünkel[1] and Martin Trentzsch[1]

[1] GLOBALFOUNDRIES Dresden Module One LLC & Co. KG
Wilschdorfer Landstr. 101, 01109 Dresden, Deutschland
Phone: +49 (0) 351 277-4557 E-mail: sven.beyer@globalfoundries.com

Abstract

With the discovery of ferroelectricity in hafnium oxide (HfO_2) based thin films in 2011 and the co-integration of ferroelectric field effect transistors (FeFET) into standard high-k metal gate (HKMG) CMOS platforms 2016/2017 by GLOBALFOUNDRIES (GF), the FeFET has emerged from a theoretical dream to an applicable reality. In the first order this technology will be matured as an extremely cheap, fast, low-power eFLASH replacement that can deliver a cost-reduction roadmap for existing eNVM technology nodes. The number of publications regarding the fundamental understanding of FeFETs, as well as their application, based on GF's maturing technology, is almost growing exponentially. Especially the neuromorphic design community has shifted focus towards this versatile CMOS compatible device with game-changing potential. In this paper, GLOBALFOUNDRIES FeFET technology will be introduced, which is undergoing continuous improvement. Some of the recent applications, discoveries and publications will be discussed.

1. Introduction

The discovery of ferroelectricity in doped hafnium oxide (e.g. Si:HfO_2) [1] has paved the way for the realization of the front-end of line (FEoL) and CMOS compatible hafnium oxide based ferroelectric field-effect transistors (FeFETs) within 28nm HKMG bulk [2] and 22nm HKMG fully depleted silicon-on-insulator (FDSOI) [3] platforms by GLOBALFOUNDRIES (GF). These devices offer fast read/access times, reversible and low-power switching, steady data retention and high cost effectiveness, making them an attractive alternative to classical embedded NOR-Flash (eFLASH). Moreover, FeFETs are characterized by rich time-voltage switching patterns, which can be readily adopted for neuromorphic applications [4], [5]. In addition, like CMOS devices, FeFETs can be dimensioned in transistor width and gate length, and can be flexibly co-integrated and intermixed with standard CMOS FETs [6], offering a true in-memory computing solution [7].

In this paper, the GF state-of-the-art FeFET technology will be introduced and the co-integration with standard CMOS devices will discussed. Then, the device performance, the switching behavior as well as the characteristic switching time-voltage trade-off will be pointed out. Finally, the related implications as well as potential applications for this versatile device will be briefly illustrated.

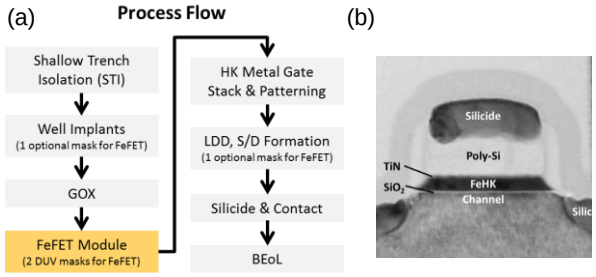

Fig. 1 (a) Schematics of the FeFET technology process flow. (b) TEM cross-section of an embedded FeFET with 9nm thick FeHK HfO_2.

2. FeFET integration

The FeFETs are incorporated into GF 28nm gate first HKMG CMOS platform (28SLP) using a simple dual mask patterning module, as shown in Fig. 1(a). Two additional (optional) masks can be used for flexible and independent device implant targeting. This FeFET module is directly portable into GF 22FDX™ HKMG FDSOI technology with just very minor process adjustments, as demonstrated in [3].

The devices comprise a 0.8nm thick interfacial oxide (SiO_2) followed by an 8 to 10nm thick, ferroelectric doped HfO_2, capped with a TiN metal cap and silicided polysilicon, as shown in Fig. 1(b).

This FeFET integration module is minimally invasive so that the standard CMOS devices are already close to platform specifications, as demonstrated by the universal curves for the standard core devices shown in [8]. Moreover, the co-integration of FeFETs with the 24Mb 0.12µm^2 6T SRAM does not significantly impact its yield, which still exceeds 90% [8].

3. Switching behavior

The FeFETs are characterized by a reversible switching between a low-V_T (LVT) and a high-V_T (HVT) state. This is generally achieved by applying positive and negative rectangular gate pulses [Fig. 2(b,c)], respectively, and the corresponding operations are called Erase (ERS) and Program (PRG). Fig. 1(a) shows the V_T distributions of the two logic states for FeFETs with L = 450nm x W = 450nm across the wafer. They are obtained by using optimized write pulses and yield a mean memory window (MW) of 1.5V and a very good uniformity [9].

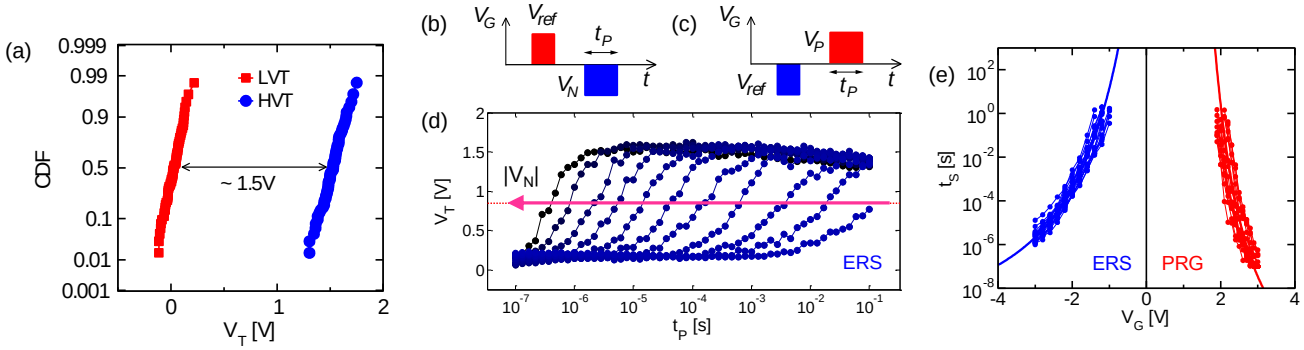

Fig. 2 (a) Distributions of the two saturated logic states across the wafer; Pulse schemes for (b) ERS and (c) PRG operations. (d) ERS switching transition (LVT to HVT) as a function of pulse width t_P for pulse amplitudes V_N from -3 V to -2 V in steps of 200 mV. (e) Switching time t_S as a function of gate voltage amplitude for both ERS and PRG. t_S is extracted as t_P for which half of the full MW is switched at a given V_G. Adapted from [9].

The switching in FeFETs is governed by a pronounced time-voltage trade-off. This is exemplified in Fig. 2(d), which shows the LVT to HVT transition (ERS) as a function of pulse width t_P for different values of pulse amplitude V_N. A linear increase in $|V_N|$ corresponds to an exponential increase in the switching speed, as exemplified in Fig. 2(e). A similar trend is observed for the PRG transition as well. This exponential relationship appears to be universal and holds true regardless of the device size (it is also observable for ultra-scaled FeFETs as shown in [10]), pulse polarity or excitation pattern (i.e. one-shot [9] or accumulative switching [11]).

The t_S-V_G graphical representation in Fig. 2(e) not only is useful to determine the electrical operation points of a FeFET, but it also shows great utility in assessing the device retention as well as disturb-free write/read conditions [9]. In fact, t_S-V_G curves are highly sensitive to fabrication process variations, under which they may shift with respect to the V_G-axis or change the slope etc., clearly indicating the short- and long-term retention trends [9].

4. Beyond memory

Neuromorphic computing

The FeFETs in a multi-domain configuration show highly gradual ERS and PRG transitions, irrespective of the applied pulsing scheme (i.e. increasing pulse amplitude, increasing pulse width or identical pulses) [12]. Accordingly, a large number of intermediate conduction states are available, making the FeFET very appealing for artificial synapses capable of plasticity in spiking neural networks or for analog weights in deep neural networks [4].

Furthermore, FeFETs can undergo a partial or complete switching to the opposite state under a train of pulses, each of which is insufficient to induce any appreciable effect [11], [12]. This accumulative phenomenon is particularly pronounced in small-area FeFETs, which typically display a sharp, almost digital switching [11]. The FeFET abruptly transitions to the other state in an all-or-nothing manner, i.e. only when a critical number of pulses has been accumulated. This behavior is exploited to implement FeFET-based leaky integrate-and-fire neurons, which are compact capacitor-less solutions and promise to largely outperform classical CMOS neurons in terms of power and area consumption.

Logic-in-memory

Since FeFETs are transistors, they can intrinsically be adopted in logic gates. By combining their logic and storage properties, as well as the easy co-integration with standard logic CMOS devices, FeFETs appear to be ideal candidates for normally-off logic gates. Indeed, not only simple gates such as AND, OR [13] and XNOR [14] based on FeFETs have been demonstrated, but also multiplexers with integrated look-up table, 1-bit half and full adders have been successfully implemented [7], with a sensibly lower transistor count and energy consumption with respect to conventional CMOS counterparts.

5. Conclusions

HfO_2 based FeFET is a promising low-power and cost-effective eFLASH replacement due to its very simple and non-invasive integration, fast write/read and low-power operation. In this paper, we have discussed its fabrication, main time-voltage switching properties and potential applications in neuromorphic and logic-in-memory devices.

References

[1] T. S. Böscke, et al., Appl. Phys. Lett., **99** (2011), 102903.
[2] M. Trentzsch, et al., IEEE IEDM, (2016), p. 10.8.1.
[3] S. Dünkel, et al., IEEE IEDM, (2017), p. 19.7.1.
[4] H. Mulaosmanovic, et al., IEEE VLSIT, (2017), p. T176.
[5] H. Mulaosmanovic, et al., Nanoscale, **10** (2018) 21755.
[6] S. Beyer, et al., IEEE NVMTS, (2018), p. 28.
[7] E. T. Breyer, et al., IEEE J. Electron Dev. Soc., **8** (2020), 748.
[8] S. Beyer, et al., IEEE IMW, (2020), p. 1-4.
[9] H. Mulaosmanovic, et al., IEEE Trans. Electron Devices, **67** (2020), 3466.
[10] H. Mulaosmanovic, et al., ACS Appl. Mater. Interfaces, **9** (2017), 3792.
[11] H. Mulaosmanovic, et al., ACS Appl. Mater. Interfaces, **10** (2018), 23997.
[12] H. Mulaosmanovic, et al., IEEE EDTM, (2020), p. 1-4.
[13] E. T. Breyer, et al., IEEE IEDM, (2017), p. 28.5.1.
[14] E. T. Breyer, et al., IEEE ISCAS, (2018), p. 1-5.