Prospects for energy-efficient edge computing with integrated HfO 2-based ferroelectric devices

Ian O'Connor, Mayeul Cantan, Cedric Marchand, Bertrand Vilquin, Stefan Slesazeck, Evelyn T Breyer, Halid Mulaosmanovic, Thomas Mikolajick, Bastien Giraud, Jean-Philippe Noel, et al.

Prospects for energy-efficient edge computing with integrated HfO_2-based ferroelectric devices

Ian O'Connor, Mayeul Cantan,
Cédric Marchand, Bertrand Vilquin
Lyon Institute of Nanotechnology
University of Lyon – Ecole Centrale de Lyon – CNRS
Ecully, France
ian.oconnor@ec-lyon.fr

Stefan Slesazeck[1], Evelyn T. Breyer[1],
Halid Mulaosmanovic[1], Thomas Mikolajick[1,2]
[1]NaMLab GmbH, [2]Chair of Nanoel. Materials, TU Dresden
Dresden, Germany
stefan.slesazeck@namlab.com

Bastien Giraud, Jean-Philippe Noël
Univ. Grenoble Alpes, CEA, LETI, MINATEC Campus,
Grenoble, France
bastien.giraud@cea.fr

Adrian Ionescu, Igor Stolichnov
Nanolab
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
adrian.ionescu@epfl.ch

***Abstract*— Edge computing requires highly energy efficient microprocessor units with embedded non-volatile memories to process data at IoT sensor nodes. Ferroelectric non-volatile memory devices are fast, low power and high endurance, and could greatly enhance energy-efficiency and allow flexibility for finer grain logic and memory. This paper will describe the basics of ferroelectric devices for both hysteretic (non-volatile memory) and negative capacitance (steep slope switch) devices, and then project how these can be used in low-power logic cell architectures and fine-grain logic-in-memory (LiM) circuits.**

Keywords—ferroelectric devices, non-volatile memory, steep slope switch, low-power logic, logic-in-memory

I. INTRODUCTION

Data size and functionality requirements for computing are increasing, according to the expectation that hardware performance will continue to improve, irrespective of the actual implementation. This is particularly true for emerging distributed computing paradigms for the Internet of Things, such as Edge Computing and Fog Computing, which are placing extraordinarily stringent constraints on computing hardware performance. Such paradigms are necessary to guarantee low-latency, secure and contextualized computation on inhomogeneous sensory data, as close as possible to the data source. This usually implies that energy sources are limited and consequently, that hardware energy efficiency must be maximized. Indeed, optimal usage of the constrained resources such as memory, bandwidth, processor, and most importantly power of IoT devices is necessary for sustainable and long-life IoT deployments.

Low-power microcontroller units (MCU) with embedded non-volatile memory (NVM) are typically to be found at the heart of today's IoT devices, hierarchically placed between the sensor nodes and the host microprocessor and given the task of pre-computing the data to reduce heavy loading of the host processor. In IoT applications, most of the power is consumed while the MCU is inactive, so NVM is used to realize a "normally-off" MCU, thus drastically cutting power consumption. The contents of the CPU are stored in the NVM and subsequently CPU power is shut down to zero power consumption sleep mode before the data are restored during MCU wake-up period.

While a centralized data transfer from the CPU to the NVM and vice-versa is a straightforward approach, more efficient approaches are highly desirable. In distributed memory concepts, NVM elements can be embedded in an advanced CMOS platform and distributed close to the logic circuits to store contents locally and essentially make logic circuits non-volatile (e.g. NV-registers, NV-SRAM, NV Code memory). This allows inactive logic circuits to be shut down, thus optimizing power savings, minimizing memory cycling and increasing reliability. The distributed memory concept is also in line with more advanced "fine grain" logic-in-memory (LiM) concepts with tighter integration of logic and memory and aiming to reduce latency and energy cost during data transfer. However, storing and re-storing CPU content costs energy, which imposes constraints on the NVM characteristics to be used. At present, the standard NVM used in MCUs is eFlash because it is high density, manufacturable and low cost. However, it suffers from low write speed, high power requirements, low endurance and vulnerability to radiation.

Therefore, a new, more robust NVM with higher speeds, lower power and high endurance is required to replace eFlash in (normally-off) MCUs to reduce the energy spent during storage and retrieval of CPU content and to allow flexible LiM designs with further improved energy efficiency. A number of NVM candidates with high speed/low power characteristics have emerged (STTRAM, ReRAM, and FeRAM) and several working prototypes have been demonstrated but there is no clear winner at present. FeRAM has the highest endurance of all candidates, low energy per bit and power consumption which could make it a good candidate to replace Flash in embedded applications. However, current embedded FeRAM devices with perovskite materials have serious problems with regard to memory cell scaling, compatibility with Si processing, manufacturability and cost that inhibit development as a mainstream NVM solution. New FE materials to overcome the shortcomings of present day FeRAM are needed.

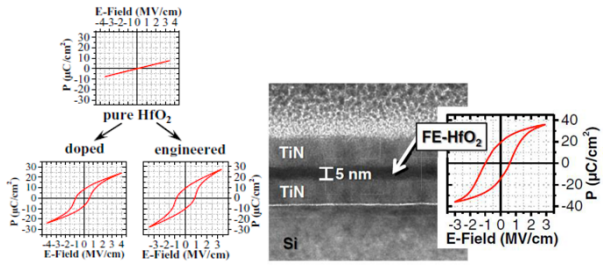
In this paper, we explore the use of new FE HfO_2-based materials to develop a competitive and versatile FeRAM technology for NVM solutions. The structure of the paper is as follows: in section II, we describe device operation of hysteretic and steep-slope devices based on HfO_2 ferroelectric FETs (FeFETs). Section III gives examples of the use of such devices to build non-volatile flip-flops and reconfigurable logic gates. Finally, section IV discusses projections to new computing architectures based on such devices.

II. MATERIALS AND DEVICE OPERATION

A. HfO_2 as a ferroelectric material

High-k HfO_2 is key for modern nanoelectronics since it is compatible with silicon technology and, thanks to its high

dielectric permittivity, has allowed downscaling without the prohibitive leakage currents associated with the traditional gate oxide. Used as a high-k gate oxide, HfO_2 is amorphous. In crystalline form, HfO_2 is generally centrosymmetric with a high temperature tetragonal phase and room temperature monoclinic phase. The tetragonal phase can be realized by doping HfO_2 with certain dopants and actually has the best high-k performance. However, under certain conditions of stoichiometry, doping and/or strain the polymorphism can be extended to a non-centrosymmetric orthorhombic phase in doped HfO_2 [1], and this has led to the recent discovery [2] of ferroelectricity in HfO_2 (Fig. 1). This material may thus impact embedded memory by allowing scaling of FeRAM cells [3] to increase storage capacity.



(a) (b)

Fig. 1. (a) Ferroelectric HfO_2 may be obtained by doping and by strain engineering (b) 5nm HZO capacitor

In terms of silicon compatibility, ferroelectric HfO_2 can be integrated into transistor gates for 1T FeFET memory cells [4] with non-destructive read, integrated in the front end of line (FEOL) with CMOS. This opens the possibility of realizing new, fine-grained LiM designs [5] to enable the merging of logic and memory and to significantly improve energy efficiency of processing and storage units.

Finally, ferroelectric HfO_2 integrated in the gate of transistor could also produce negative capacitance FETs (NCFETs) [6] functioning as low power steep slope switches to further boost low power/high performance operation of LiM circuits.

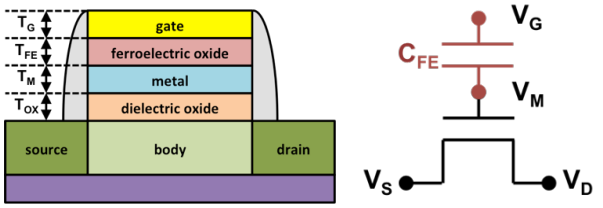
In summary, the functionality and versatility of ferroelectric HfO_2 could have a significant impact on embedded NVM solutions, tightly integrated with logic to increase the energy efficiency of computation.

B. Ferroelectric transistor (FeFET) operation

From a structural point of view, a ferroelectric transistor (FeFET) is simply an extension of a regular bulk or FDSOI MOSFET with an additional layer of ferroelectric material inside the gate stack (Fig. 2) [8]. This leads to a functionality based on the inclusion of a ferroelectric capacitance located between the external gate and the "internal" gate, which actually controls the state of the FET channel. The potential simplicity of the process has led to speculation that every transistor in a standard CMOS process could be transformed into a non-volatile memory or steep-slope device.

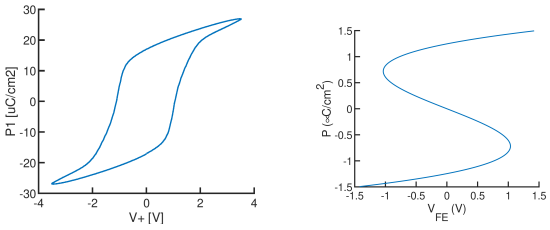
FeFETs operate in two different modes: a non-volatile mode, which requires hysteretic operation, and a steep switching mode, which can be hysteretic or non-hysteretic. The ratio between the ferroelectric capacitance and the dielectric capacitance determines the FeFET operation mode.

1) Non-volatile mode: this mode leverages the hysteretic polarization vs. voltage characteristic of the ferroelectric material (P versus V_{FE}) as already shown in Fig. 1. When this material is placed in the gate stack, it forms a capacitance in series with the gate of a transistor. From the point of view of the overall device, the I_{DS}-V_{GS} transfer characteristic then also becomes hysteretic (Fig. 3(a)). For V_{GS}=0V and a centered ferroelectric hysteresis, the FeFET demonstrates bistable states corresponding to positive and negative polarization retention in the ferroelectric layer. Hence the FeFET channel resistance is either in a high-resistance state (HRS) for low I_{DS}, or a low-resistance state (LRS) for high I_{DS}. For an n-type FeFET, HRS is achieved for P<0 and LRS for P>0, while the conditions are opposite for a p-type FeFET. The ferroelectric capacitance must be small with respect to the dielectric capacitance as the series capacitance leads to a reduced hysteretic window of P vs. V_{GS}. This usually means that the ferroelectric layer has to be relatively thick in order to achieve non-volatile FeFET operation.



(a) (b)

Fig. 2. (a) Ferroelectric HfO_2 device gate stack (b) equivalent circuit schematic



(a) (b)

Fig. 3. Polarization versus voltage characteristics (a) of a non-volatile FeFET (b) of an NCFET

2) Steep switching mode: this mode relies on a direct non-hysteretic transition between positive polarization for negative voltage and negative polarization for positive voltage, originating from the "S-shaped" P vs. V_{GS} curve described by Landau-Ginzburg-Devonshire theory, without considering domain formation. This particular region of the curve displays *negative* differential capacitance (C=δQ/δV), which would be physically unstable in a standalone device, but can be associated in series with the dielectric capacitance in a stable way if the overall capacitance is positive. This condition typically requires a ferroelectric capacitance that is larger than the dielectric capacitance, implying that the ferroelectric layer is relatively thin. On the other hand, the higher ferroelectric capacitance means a weaker negative capacitance effect. Therefore, both capacitances have to remain comparable. The voltage division across the ferroelectric and dielectric capacitances gives

$$\frac{V_M}{V_G} = \frac{C_{FE}}{C_{FE} + C_{MOS}}$$

where V_M and V_G represent respectively the intermediate and gate potentials as shown in Fig. 2(b), and C_{FE} and C_{MOS} represent the ferroelectric and dielectric capacitances

respectively. Usual voltage division would mean that $V_M<V_G$; but with a negative ferroelectric capacitance value, $V_M>V_G$ and is actually *amplified*. As V_M is the surface potential controlling the FET channel, the drain current I_{DS} is enhanced and from the point of view of the gate voltage V_G, the transistor can achieve "steep switching" with a subthreshold slope below the Boltzmann limit of 60mV/dec. However the use of an internal metal as shown in Fig. 2(a) can impair the stabilization of the negative capacitance state due to leakage currents and domain formation [10][11]. Therefore, experimental devices should avoid using such an internal metal layer. Such devices (commonly termed "negative capacitance FETs" or NCFETs) have clear advantages for low-power logic operation, and the intrinsic compatibility with CMOS coupled with the potential for non-volatile devices makes the case for a ferroelectric technology platform all the more convincing.

III. FeFET Based Logic and Memory Circuits

In this section, we examine non-volatile memory cells and reconfigurable logic cells based on ferroelectric devices.

A. Non-volatile memory

An example of a non-volatile memory cell is the Black and Das non-volatile flip-flop depicted in Fig. 4. This structure combines the operation of a regular SRAM cell for normal operation, and two FeFETs for non-volatile storage, which can be written to independently. When the circuit is powered, the values stored in the ferroelectric capacitors can be written to the SRAM through the *sense* transistor. If the FeFETs are in different resistance states, the imbalance will determine the rest state the SRAM will return to, as the FeFET with the lowest resistance will pull its branch down. Such an approach reduces the number of write cycles, and allows the SRAM to retain voltage compatibility with traditional structures. Moreover, this device remains operational when HRS and LRS are relatively close.

B. Reconfigurable logic

As well as pure non-volatile memory applications, FeFETs have also been integrated inside logic gates themselves [9][13]. Two types are presented here: a 1-FeFET reconfigurable NAND/NOR logic gate and a 2-FeFET X(N)OR logic gate. In both cases, the ferroelectric material is HfO_2. To achieve sequential logic functionality, one logic input is stored in the polarization state of the FeFET by applying a write pulse to the gate terminal, whereas the second logic input is subsequently applied as readout voltage to the gate terminal (see inputs A and B in Fig. 5(a) and (b)).

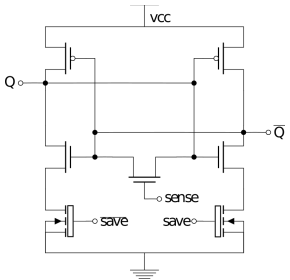


Fig. 4. Non-volatile FeFET based latch

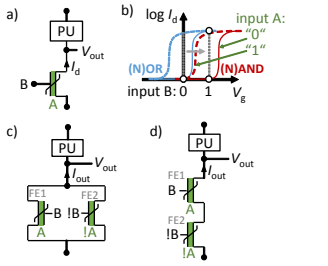


Fig. 5. a) and b) 1-FeFET NAND/NOR schematic and basic functionality c) and d) 2-FeFET X(N)OR logic gates

Input A denotes the internal polarization state of the FeFET and can be either logic 0 (high resistance state – solid line) or 1 (low resistance state – dashed line). The second input – input B – is applied as a readout voltage to the gate terminal. Note that the applied voltage for logic input B is small enough not to switch the polarization state. Therefore, these gates constitute a sequential logic that immediately links the memory and the logic functionality of the FeFET.

Electrical measurement results prove the feasibility of the NAND/NOR logic gate concept for 22nm FD-SOI FeFETs. By applying a back bias voltage to the FeFET, the I_dV_g characteristics of the FeFET can be shifted along the V_g axis. As a result, the logic functionality switches between NAND and NOR behavior and electrical reconfigurability is introduced (see Fig. 5(b) and Fig. 6(a)). As a next step, two FeFETs of NAND or NOR functionality are connected in parallel or in series, respectively. If the inputs of the second FeFET are the logical complements of the inputs of the first FeFET (Fig. 5(c) and (d)), the resulting structure exhibits X(N)OR functionality (Fig. 6b). A direct integration into existing AND/NAND memory arrays is possible [13].

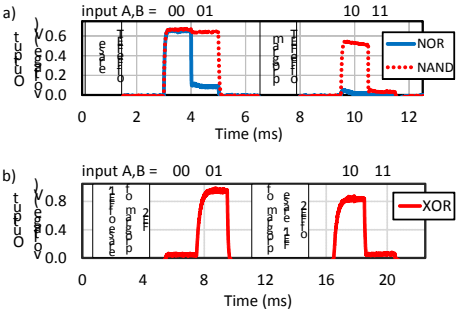


Fig. 6. Transient voltage measurements of (a) the proposed 1-FeFET logic NAND/NOR and (b) the 2-FeFET logic XOR gates

IV. Prospects for Energy-Efficient Computing

Integrated ferroelectric devices allow data storage close to logic circuits to reduce energy cost of data transfer, allow smart gating for "normally-off" computing and open the way to novel energy-efficient computing paradigms such as logic-in-memory (LiM).

A. Normally-off computing

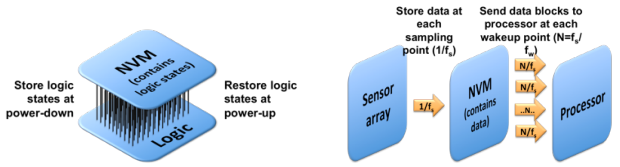
Normally-off (N-Off) computing uses a non-volatile memory array to immediately store the context of logic (data at logic nodes) during power-down, which enables significant reduction of activity duty cycles and energy consumption in IoT nodes due to leakage currents (Fig. 7(a)). An extension of this model also uses dedicated non-volatile memory close to data sources (e.g. sensors) to store and accumulate acquired data without expensive wakeup cycles for IoT node processors (Fig. 7(b)). The priority criteria are:

- For store (at power-down): 1. Low write power, 2. Low write time, 3. High-density
- For restore (at power-up): 1. Low read power, 2. Low read time 3. High-density
- For accumulation (off): 1. Low write power, 2. Low write time, 3. Low read power, 4. Low read time, 5. High density, 6. High endurance, 7. Low leakage

B. Logic-in-memory (LiM)

LiM represents a logically enhanced memory array that can be programmed to realize arithmetic and logic operations in an endurance-aware way. Several concepts and terms exist to identify means of associating logic with memory, such as LiM (logic in memory), IMC (in-memory computing) and PiM (processing in memory). In order to define clearly the

contribution of HfO_2 FeFETs in this spectrum, we firstly make explicit our perception of these terms.



(a) (b)

Fig. 7. Normally-off computing to: (a) store/restore logic states (b) accumulate data at sensor nodes

1) PiM - Processing in Memory: move PROCESSING functionality INto modified MEMORY. Here, memory is enhanced with a lightweight processor (complete with control unit and instruction register / decoder) located close to memory macros. The main processor offloads complete portions of tasks onto the lightweight processor which has direct and fast access to local data as they are physically situated inside the memory, and is capable of carrying out a set of simple but frequently used operations.

2) IMC - In Memory Computing: INside modified MEMORY add elementary COMPUTING functionality. It can be useful to enhance memory to enable it to do some computation locally. In this case, instead of loading operands and storing results from and to the memory, the processor can ask the enhanced memory to carry out elementary operations on operands before sending the result for subsequent (more complex) calculations. This alleviates load on processor-memory data communication but requires a richer processor-memory control communication.

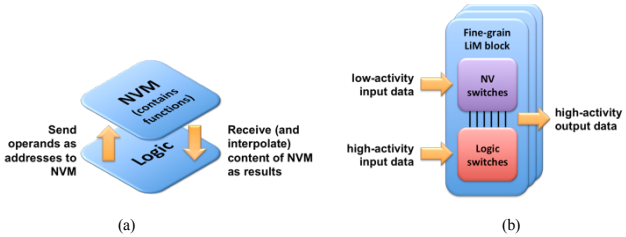
3) LiM - Logic in Memory: do LOGIC operations IN existing MEMORY resources. *Coarse-grain LiM* requires a full (non-volatile) memory matrix that can be integrated sufficiently close to the processor to be used with the same latency as a L1 cache. The memory is used to contain known results of a frequently used operation (multiplication is a simple example). A key characteristic is that no additional logic is required inside the memory plane, although some additional interpretation of results (e.g. interpolation, normalisation) may be necessary in the logic plane. *Fine-grain LiM* is different and more prospective. No clear use model has yet emerged for this, but the key characteristic is that a (non-volatile) memory element (or small network of elements) can be connected to a small logic switch network (in different planes) to achieve novel functions. A non-volatile flip-flop is an example using the level of intimacy between logic and memory required for fine-grain LiM [15].

For FeFET-based coarse-grain LiM (Fig. 8(a)), the machine can be programmed using the state of NV devices to realize complex logical and arithmetic functions in an endurance-aware way, which can then be read using word lines to address function content. The priority criteria in this scenario are:

- For programming offline: 1. Low write power, 2. Low write time, 3. High-endurance
- For operation online: 1. Low read time, 2. Low read power, 3. High-density.

For FeFET-based fine-grain LiM (Fig. 8(b)), device-level association of logic and memory can be used for long-term variable assignments in mathematical accelerator functions such as function coefficients (e.g. filters) and table content. Both arithmetic and logic functions can also be associated with sequencing. The priority criteria in this scenario are:

- For NV switches: 1. High endurance, 2. High density, 3. Low write time, 4. Low write power
- For logic switches: 1. Low leakage, 2. High speed, 3. High density, 4. Low voltage



(a) (b)

Fig. 8. LiM with non-volatile memory: (a) coarse-grain (b) fine-grain

V. Conclusion

Thanks to its proven compatibility with silicon, ferroelectric HfO_2 can be integrated in FET gate stacks to form non-volatile memory cells and low-power logic. In this paper, we have described the underlying physics and circuit operation and projected how normally-off computing and logic-in-memory can benefit. Future work will focus on experimental demonstration of large-scale HfO_2 FeFET memory arrays integrated with advanced CMOS platforms.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780302 "3εFerro".

References

[1] Böschke et al. Ferroelectricity in hafnium oxide thin films, Appl. Phys. Lett. 99, 102903 (2011)

[2] J. Müller, et al., "Ferroelectricity in simple binary ZrO2 and HfO2", NanoLett. 12, 4318 (12)

[3] J. Müller et al., "Ferroelectric Hafnium Oxide Based Materials and Devices: Assessment of Current Status and Future Prospects", ECS Journal of Solid State Science and Technology, 4 (5) N30-N35 (15)

[4] M. Trentzsch et al., "A 28 nm HKMG super low power embedded NVM technology based of ferroelectric FETs" IEDM 2016, 294-297

[5] X. Yin et al., "Exploiting Ferroelectric FETs for Low-Power Non-Volatile Logic-in-Memory Circuits" ICCAD'16, November 07–10, 2016, Austin, TX, USA

[6] S. Salahuddin, S. Datta, NanoLett. 8 , 405 (08)

[7] M. Hofmann et al., "Direct observation of negative capacitance in polycrystalline ferroelectric HfO2" Adv. Funct. Mater. 2016

[8] A. Aziz et al., "Computing with Ferroelectric FETs: Devices, Models, Systems, and Applications", DATE 2018

[9] M. Hoffmann et al., "Direct observation of negative capacitance in polycrystalline ferroelectric HfO2" Adv. Funct. Mater. 2016

[10] A. I. Khan et al. IEEE Trans. Electr. Dev. 63 (11), 4416-4422 (2016)

[11] M. Hoffmann et al. Nanoscale, 2018,10, 10891-10899 (DOI: 10.1039/C8NR02752H).

[12] E. T. Breyer, H. Mulaosmanovic, T. Mikolajick and S. Slesazeck, "Reconfigurable NAND/NOR logic gates in 28 nm HKMG and 22 nm FD-SOI FeFET technology" IEDM 2017, pp. 28.5.1-28.5.4

[13] E. T. Breyer, H. Mulaosmanovic, S. Slesazeck, and T. Mikolajick, "Demonstration of versatile nonvolatile logic gates in 28nm HKMG FeFET technology" ISCAS 2018, pp. 1-5

[14] M. Trentzsch et al., A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs IEDM 2016, p. 294

[15] N. Jovanovic, et al. "Design Considerations for Reliable OxRAM-based Non-Volatile Flip-Flops in 28nm FDSOI Technology". International Symposium on Circuits and Systems (ISCAS), IEEE, 2016.