Nonvolatile Memory and Computing Using Emerging Ferroelectric Transistors

Xueqing Li, Longqiang Lai
The Department of Electronic Engineering
Tsinghua University
Beijing, China
xueqingli@tsinghua.edu.cn

Abstract—Ferroelectric FETs (FeFETs) are emerging as a promising nano device candidate for the next-generation energy-efficient embedded nonvolatile memory (NVM). This promise comes from not only the CMOS-scaling compatibility, but also the compact fusion of logic and non-volatility in a single device that provides opportunities for efficient memory access and in-memory computing. This talk investigates circuit opportunities that harness these intriguing FeFET device features, providing insights into new computation paradigms beyond existing solutions.

Keywords—Ferroelectric FET (FeFET); negative capacitance FET (NCFET); nonvolatile memory; emerging devices; beyond-CMOS; in-memory computing.

I. BACKGROUND AND MOTIVATION

With increasing number of edge devices due to the booming of Internet-of-Things (IoT) and sensors, how to power the ubiquitous computing is indeed a big design constraint [1]. While the battery is indeed improving with better electric-chemistry understanding, the gap between the existing battery expectation and available on-the-shelf products is increasing. For most portable devices, the limited battery life and sometimes the safety problems have raised lots of inconvenience and even life threats.

Effective approaches to lowering the power consumption have been observed in various aspects and levels, ranging from devices, circuits, architectures, algorithms, and systems [2]. Some efforts can lead to a better drop-in replacement for an existing design block, for example, designing a better inverter with simply smaller transistors with smaller capacitance, or with transistors that can operate at a lower supply voltage [3]. More importantly, the effectiveness of some efforts may strongly depend on the progress of efforts in other aspects and levels. It has been increasingly demanding co-design and co-optimization, as illustrated in Fig. 1. This will be further demonstrated in this talk.

While conventional low-power digital computing and memory design approach using the CMOS Boolean solutions has led to orders of magnitude power improvement, the challenge of further scaling the CMOS technology has made this approach much more opaque than before [4]. Even if this CMOS scaling can continue till beyond 1nm with accurate modeling, low-parasitics contacts, sufficiently-low fabrication costs, small variation and good yield, there are fundamental bottlenecks that the existing CMOS computer solution can not break theoretically in physics.

The first well-known one is the CMOS OFF-state leakage current limited by the >60mV/decade room-temperature sub-threshold slope (SS) [3][5][6]. For large-scale integrated circuits, such leakage can cause significant amount of static power consumption by both logic and memory (e.g. SRAM), even with CMOS device tuning (e.g. threshold voltage engineering), circuit innovations (such as proper transistor sizing and new circuit topology creation) and architecture optimizations (such as power gating, dynamic voltage and frequency scaling, pipelining, parallelism, etc.)

The second bottleneck can be more related to the “memory wall” of the conventional von Neumann computer architectures, in which the memory data access can be costly in both time and energy [7]. This is essentially caused by the separation of computing logic and the storage memory elements which finally causes long-distance data movement. With the emergence of new computing architectures like new neural networks that support in-memory computing or near-memory computing, this bottleneck has a higher chance to be relieved but it still highly depends on how much long-distance data access can be eliminated. For example, recent high-performance machine-learning-based neural networks still highly rely on high-bandwidth memories (HBM) [8].

The bottlenecks above do not indicate lower importance of other potential barriers towards higher power efficiency, but these being highlighted are fundamental limitations that do not seem to have a good solution if we stick to the current device and architectures. Meanwhile, beyond-CMOS solutions provide significantly extra design space to mitigate the two abovementioned bottlenecks and also promising results, especially in some specific application scenarios, as to be presented later in this talk.
Regarding the first bottleneck, mitigation by beyond-CMOS solutions can be obtained with steep-slope Boolean transistors that can switch more abruptly with lower applied gate voltage. The steep slope characteristics ensure lower leakage current while providing the same amount of ON-state current for dynamic performance. Possible steep-slope transistors can include negative capacitance FET (NCFET) [9][10], tunneling FET (TFET) [3][5], etc. Meanwhile, emerging nonvolatile memory (NVM) devices could be adopted to reduce and even fully eliminate the static leakage current of both idle CMOS digital logic gates and CMOS SRAM as these NVM can sustain the stored data even if the power supply is shut off [1][11]-[17].

Regarding the second bottleneck, the introduction of unique computing or storage primitives provide completely new opportunities that can reshape the design space. For example, the integration of Boolean logic and nonvolatile memory (NVM) storage within each ferroelectric FETs for digital applications, and the nonlinear switching behavior of resistive memory (ReRAM) and metal-insulator-transition devices for neuromorphic and coupled-oscillator complex problem solvers, respectively [18].

This talk will use FeFET as an example to highlight the opportunities that can be enabled by emerging device-circuit co-design [12]-[16]. It is believed that FeFETs are promising because of their CMOS compatibility, the capability of being designed to be a steep-slope device or a nonvolatile memory, and also the memory-logic integration with each single transistor which enables unique in-memory computing flexibilities. While most efforts in this talk will cover the summary of the FeFET NVM and nonvolatile logic designs that fit well with existing computer architectures, it is expected that FeFETs can also be explored for more sophisticated architectures, including neural networks and array-style in-memory computers.

In the rest of this talk, Section II will briefly review the basics of FeFET devices, with the focus on highlighting the difference from a conventional MOSFET. Section III will summarize some recent FeFET-based memory designs. Section IV will review recent FeFET-based nonvolatile logic designs, specifically, nonvolatile flip-flops. Their application scenario will also be introduced as well. Section V discusses the future work and Section VI concludes this presentation.

II. FeFET BASICS AND ITS OPPORTUNITIES

A. Device Structure and General Operating Theories

A conceptual FeFET device is illustrated in Fig. 2(a), with its equivalent simplified model in Fig. 2(b), and typical I-V characteristics in Fig. 2(c) [19][20]. An FeFET is essentially a MOSFET with an extra ferroelectric gate insulator, such as doped hafnium dioxide, making it compatible with the existing commercial CMOS process. The adoption of the ferroelectric material in this structure could achieve the steep switching behavior with a sub-threshold swing below 60mV/decade so that the transistor could be used to build lower-power logic gates [19]. It is achieved by making use of the voltage booting function of the negative capacitance of the ferroelectric material to increase the internal MOSFET gate voltage. It was also predicted in theory and confirmed by recent experiments that, by increasing the ferroelectric layer thickness ($T_{FE}$), when the negative ferroelectric capacitance is smaller than the positive MOSFET gate capacitance, hysteresis appears and may exhibit distinct ON and OFF states with zero gate-source voltage ($V_{GS}$) based on the direction of the ferroelectric material polarization, as shown in Fig. 2(c) [19]. For conventional logic gates, hysteresis should be strictly controlled or minimized to comply with the logic operation. On the contrary, it is intriguing to use the hysteresis for low-power NVM applications.

In this talk, unless otherwise pointed out, we focus on using the hysteretic FeFETs for memory applications.
Fig. 2(d) can show over four orders of difference in magnitude, leading to low-cost sensing schemes to distinguish the state difference [13][20]. This can be superior to most existing FeRAM, STT-RAM, ReRAM, and PCRAM devices. The sharp transitioning between different states also helps to maintain a larger noise margin. These advantages come from the unique FeFET features: (i) the settling-down transition behavior in the energy landscape as a passive amplification for $V_{\text{MOS}}$, and (ii) the gain of the internal MOSFET from $V_{\text{MOS}}$ to sensed current $I_{DS}$. For FeRAM, no such intrinsic gain is provided and sensing is more complex and sensitive to bit-line parasitics.

- **Tunable Low-Voltage Operation.** With proper MOSFET work function engineering and ferroelectric material design that matches the MOSFET properties, e.g. the gate capacitance, it is possible to locate the FeFET I-V hysteresis window around zero $V_{GS}$ [22]. By tuning $T_{FE}$, the hysteresis width could also be optimized to work under a proper supply voltage.

- **Logic-Memory Integration.** The FeFET has integrated the NVM storage and the logic transistor operating as a memory state amplifying reader. Such integration not only provides the opportunity to design a simplified low-power sensing scheme, but also opens up new space for future memory-oriented computing [13][14].

- **Low-Power Write Operation** [13]-[16][20]. The polarization switching is accomplished by applying a positive or negative voltage across the ferroelectric material. Different from the state change in resistive memory devices like ReRAM and STT-RAM, no static DC current is consumed for FeFET (biased with $V_{DS} = 0V$). Furthermore, when considering the resistive memory device variations of required write pulse duration, even more energy could be saved.

As pointed out above, the ferroelectric material in FeFETs could be the same as that in FeRAM, leading to similar memory features of retention time, endurance, etc. On the other hand, the FeFET memory read operation is non-destructive, which outperforms FeRAM. More importantly, as analyzed above, FeFET NVM is fundamentally superior to FeRAM with better distinguishability and access interface.

### C. Recent Device Fabrication Progress

The initial fabrication of stacking ferroelectric materials into the gate was reported long ago [23]. Recent material and process development makes FeFETs more attractive for logic and memory applications [21]-[38]. Table 1 summarizes some reported results. Notably, several important milestones related to FeFET fabrication and their fundamental understanding have been achieved recently. While ferroelectric materials can be BTO, PZT, PT, BST, and SBT, recent advances mostly come from the doped hafnium (Hf) material solution, which is found to be compatible with the CMOS process and scales down well in a fin structure [24].

### D. Device Modeling

There are a few FeFET models and Landau-Khalatnikov (LK) equation has been used [20][39]-[41]. Most results in this paper uses the calibrated FeFET model in [20] with an embedded 10nm or 65nm FinFET PTM as the baseline MOSFET. The FeFET device design, including capacitance matching, ferroelectric switching mechanism, etc. has been discussed in [20].

### III. FeFET Nonvolatile Memory Arrays

This section reviews some recent designs of FeFET-based nonvolatile memory arrays, starting from the 10-transistor (10T) per cell FeFET-based nonvolatile SRAM (nvSRAM) [15], then the 2-transistor (2T) per cell design [12], and then projected 1-transistor (1T) per cell design. The trade-off is discussed among different designs. Finally, the potential application and future work is discussed.

Evaluation of an array-style memory design should be done considering both the cell design and the peripherals. The drain, source, gate, and body (if there is), should all be properly biased or controlled during the power-off, idle, read, and write modes. Read and write operations have been introduced in the previous section, and at the circuit level, access transistor may be required for desired isolation.

#### A. 10-T nvSRAM Design

The concept of nvSRAM is to back up the conventional SRAM state to an in situ distributed nonvolatile storage cell and to restore the data back to the SRAM when necessary, e.g. power-gating. The reason of not directly using the nonvolatile storage cell is mostly for the purpose of keep some virtues of CMOS SRAM, such as speed and endurance. Varying with different applications, the main design and optimization targets of the backup storage cell can include density, backup and restore energy and latency, as well as other specifications like variation and yield, supply voltage range and number of required voltage levels, etc.

Fig. 4(a) shows the 10-T nvSRAM circuit topology [15]. During the idle state, the restore control voltage $V_{str}$ is grounded, and the FeFET gate voltage $V_{bkp}$ is biased at VDD/2, or some other similar voltage levels to prevent unnecessary FeFET polarization switching activities when the SRAM state changes. If the SRAM supply voltage is sufficiently low, the FeFET gate voltage $V_{bkp}$ can be biased.
at any voltage between the ground and VDD. On the other hand, for a given FeFET, if the SRAM supply voltage is too high, it can be impossible to find a Vbkp biasing that can prevent FeFET polarization switching when the SRAM state changes. When there is a demand of backup, Vrstr stays grounded, and the gate voltage Vbkp goes to VDD (to switch one FeFET to positive polarization) and then ground (to switch the other one to negative polarization), and then back to the idle state Vbkp. After the backup operation accomplishes, the SRAM power supply can be safely turned off and the FeFET polarization remains. When there is a demand of restore (while the SRAM supply is grounded), Vbkp goes to VDD/2 and Vrstr goes to VDD, and then the SRAM supply voltage is gradually increased to VDD. As the FeFET backup cell has a huge difference in pulling down the two internal SRAM nodes to the ground (one floating and the other grounded), the SRAM states can be restored.

Typically, the restore speed is limited by the supply voltage recovery latency as the supply network usually has large parasitics. And the backup speed can be in the range of nanosecond when a polarization switching activity is needed. Note that this design does not consume static current during the backup and restore phases, leading to significant amount of energy savings when compared with nSRAM designs based on ReRAM and MTJ, as shown in Fig. 4(b). Here a break-even-time (BET) can be used to indicate how the minimum amount of supply shut-down to save sufficient leakage energy to count for the cost of backup and restore energy consumed. Theoretical analysis shows hundreds of times of energy savings per backup and restore operation.

The memory array in Fig. 5(a) could be potentially used for multiply-and-accumulation computing, i.e. “dot production” for two input vectors. The hindrance in using this design includes: (i) it does not support a practical voltage- or current-mode sensing scheme as each output sense line can be connected to multiple read select lines with low resistance if the cross-over FeFET shows up with positive polarization (in this case the sense current would be steered to the read select line instead of purely to the sense amplifier); (ii) wide write voltage range, approximately 2xVDD as both positive and negative voltages are used. In contrast, the 2T cell design based on Fig. 4(a) has no such issues. Therefore, further design optimization for the 2T cell in Fig. 5(a) is required to make it a truly practical memory array that supports both convenient read and write operations.

**B. 2-T NVM Array Design**

When the endurance is not an issue in some applications, using purely the backup cell in the 10-T nSRAM design is feasible. In this case, the two branches of backup storage in Fig. 4(a) can be reduced to only one branches, i.e. keeping M1 and N1 would be sufficient to store a bit. Or, the two branches could be used to store two bits.

In [12], a 2-T per cell NVM array was reported, as shown in Fig. 5(a). In the memory array, each FeFET gate can be accessed through a wordline-controlled access transistor, and the write is accomplished by applying either a positive voltage or a negative voltage to the gate of each FeFET. Fig. 5(b) shows the write performance comparison with FeRAM, which is also based on ferroelectric capacitors using the same ferroelectric material. Evaluation results show that over 10x write energy savings could be achieved.

**C. 1-T NVM Array Design**

Further improving the density of FeFET-based NVM array is beneficial to reduce the overall cost when a large amount of memory is adopted. Based on the abovementioned 2-T per cell designs, further removing the one access transistor in each cell needs extra work of FeFET device-level re-design to ensure that cells not being accessed do not short bitlines and wordlines. A practical FeFET device can be like those reported in [33][35][36]. The required device characteristics are briefly illustrated in Fig. 6(a). This FeFET is turned off and its polarization state is sustained when the gate-source biasing is set to ground. To switch the polarization to positive and negative, the gate source voltage needs to be sufficiently high in positive and negative, respectively. To read the FeFET and tell the $I_{DS}$ difference between the two polarization states, the gate source voltage is biased at a non-zero positive voltage $V_R$, which is high enough to turn on the FeFET with positive polarization, as illustrated in Fig. 6(a).

The 1-T per cell NVM array is similar to the NOR-type FLASH memory array, as illustrated in Fig. 6(b). To read a row, the gate control voltage of the row is set to $V_R$, and the
source bitline SBL is set to GND, and the sense bitline BL is set to VDD. The current flowing through the cell can be sensed either through sensing the voltage change of the precharged sense bitline or through sensing the current flowing at the clamped sense bitline. To write a row, the gate can be grounded with the source and drain bitlines shorted to VDD to switch to the negative polarization or – VDD to switch to the positive polarization.

While the performance is still being evaluated at this moment, the density improvement over prior versions can be guaranteed. As a matter of fact, the energy consumption performance can be good as no static power is consumed, and the read performance can be good due to large ON-state current and ultra-low OFF-state current.

D. Summary and Future Work

Fig. 7 summarizes the FeFET-based NVM performance. Although this summary is a rough evaluation, it can clearly show the advantage towards energy-efficient embedded nonvolatile memory. Future work on variation analysis, endurance improvement, experimental demonstration, and application level evaluation is needed.

IV. FeFET NONVOLATILE COMPUTING LOGIC

This section reviews some recent designs of FeFET-based nonvolatile latches and flip-flops [13][14][16]. The trade-off is discussed among different designs.

A. Application Scenarios and Key Specifications

Power gating has been widely used, by which the power supply of the idle and leaky digital computing blocks could be fully turned off to reduce the static power consumption. This is illustrated in Fig. 8. This can be more meaningful as the scale of modern processors is increasing with more transistors integrated. Meanwhile, the state of flip-flops in pipelines, state machines, and register files should be backed up during the power shut-down period, and be restored when the power supply is recovered. Fig. 9 illustrates a conceptual nonvolatile flip-flop (nvDFF) and a few recent FeFET-based nvDFFs which can sustain the flip-flop state during power-off periods [13][14][16]. With the development of IoT and energy harvesting techniques, power supply disturbance can be frequent and such nvDFFs are essentially critical to keep the progress with such nonvolatile computing methodology.

Therefore, critical specifications usually include:

- **Area Overhead.** This includes the backup and restore controller, backup and restore circuitry, routing, etc. If a separate supply voltage is used, extra area is needed.

- **Backup and Restore Energy Overhead.** Using more energy than that used to sustain idle leaky circuits is meaningless. Thus reducing this energy overhead can make sure that even if the power supply is shut down for a short period of time, it is still likely to save the overall energy consumption. Break-even time (BET), which has been used for nvSRAM evaluations, has been widely used for nvDFFs as well.

- **Backup and Restore Energy Time.** This is useful when the processor needs prompt wakeup response.

- **Normal Mode Energy-Latency Overhead.** This indicates whether energy-latency performance of the normal-operation mode is negatively affected.

Fig. 8. Concepts of power gating to mitigate static leakage power.

Fig. 9. nvDFFs [13][14][16]: (a) Concept with in situ NVM as the state backup storage; (b) nvDFF1; (c) nvDFF2; (d) nvDFF3.

B. FeFET-Based nvDFFs for Different Optimization Goals

Fig. 9(b-d) shows the circuit scheme of three energy-efficient nvDFF designs with different features [13][14][16]: ultra-low normal-mode overhead for the on-demand nvDFF1, low-normal-mode overhead low area overhead on-demand nvDFF2, and ultra-low area for the intrinsic nvDFF3.

For nvDFF1 in [13], the backup operation is triggered when the backup control signal Bkp is enabled to be high, which leads to VDD or – VDD FeFET biasing for necessary polarization switching. The restore operation is similar to nvSRAM in that the initial pull-down branches with on/off-state FeFETs determine the final settled state during the supply ramp-up period. nvDFF2 in [16] eliminates access transistors and prevents unnecessary polarization switching during the normal mode operation by properly limiting the supply voltage safely within the hysteresis window.

For nvDFF3 in [14], the concept is that by embedding FeFETs into the latch, all DFF state change can finally lead to FeFET polarization change if the clock cycle is
sufficiently long. While more polarization switching activities cause more normal-mode latency and energy consumption, this design eliminates external controls, and needs only two extra transistors to make the DFF nonvolatile.

These achievements originate from the various device features (see Section II) and the circuit techniques that harness them. Table II summarizes the nDFF comparisons which clearly show the advantages of the FeFET solution.

<table>
<thead>
<tr>
<th>Table II: Comparison Between Recent nDFFs (Data From [16])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device Technology</td>
</tr>
<tr>
<td>Area overheads</td>
</tr>
<tr>
<td>Normal-mode EDP overhead</td>
</tr>
<tr>
<td>Backup and restore energy</td>
</tr>
<tr>
<td>Backup and restore speed</td>
</tr>
</tbody>
</table>

V. SUMMARY

FeFETs have been proved to be promising with recent device and circuit progress in future emerging applications. Further device, circuit and application co-design and co-optimization will bring even more opportunities.

ACKNOWLEDGMENT

This work was supported in part by NSFC under grants 61720106013 and 61674094 and in part by the Beijing Innovation Center for Future Chip.

References