Design Methodology for Area and Energy Efficient OxRAM-based Non-Volatile Flip-Flop

M. Nataraj¹, A. Levisse¹,², B. Giraudeau², J-P. Noel², P. Meinerzhagen³, J.M. Portal¹, P.-E. Gaillardon¹

¹The University of Utah, ECE Department, Salt Lake City, UT, 84112, USA
²Univ. Grenoble Alpes, F-38000 Grenoble, France; CEA, LETI, MINATEC Campus, F-38054 Grenoble, France
³Aix-Marseille Université, IM2NP, CNRS UMR 7334, F-13453 Marseille, France

Abstract—With the introduction of the Internet of Things (IoT), power consumption became a major design issue in modern system-on-chips. In advanced technologies, leakage power has become a dominant component, especially during sleep periods. Leakage mainly comes from volatile memory elements, e.g., flip-flops that cannot be power-gated in order to retain their states. Non-Volatile Flip-Flop (NVFF) using emerging memory technologies, such as Resistive Random Access Memories (RRAM), are popular solutions to address this issue. In NVFF design, the resistance values of the memory element have a direct impact on the area and energy overhead of the structure. In this paper, we present a design methodology for area and energy efficient RRAM-based NVFF. By characterizing the optimal lower bound of the RRAM resistance ratio required for properly restoring the FF, the store and restore operations can be performed using optimal programming circuit area and energy. Four Transmission-Gate (TG) NVFF topologies implemented in 180nm CMOS technology were analyzed using the proposed methodology. The presented methodology shows that differential NVFF provides minimum restore resistance ratio down to 1.02 considering CMOS and RRAM variability. This enables improvements in terms of store energy (34%) and area overhead (40%) compared to reported state-of-the-art NV-TGFFs design approaches.

Keywords—Non-volatile Flip-Flop; RRAM; Methodology;

I. INTRODUCTION

Modern state-of-the-art power aware Very Large Scale Integration (VLSI) systems such as the ones designed for Internet-of-Things (IoT) applications require high-energy-efficiency and generally tend to have a low activity [1]. A limitation to the energy efficiency comes from volatile elements, such as latches or Flip-Flops (FF) that cannot be power-gated during inactivity period in order to retain their states. The reduction of leakage power consumption during long duration inactivity periods can be achieved by introducing Non-Volatility (NV) in embedded latches and pipeline registers. Introducing Non-Volatile Flip-Flops (NVFFs), allows the system to save its current state before going to long zero-consumption sleep periods. In addition, fine-grain Resistive Random Access Memories (RRAM) integration inside the FF provides fast sleep/wake-up transitions by avoiding access to external out-of-core memories.

Among the various RRAM technologies [2,3,4], Oxide-based Random Access Memories (OxRAM) are arguably among the most promising technologies thanks to their fast switching behavior, simple CMOS co-integration (materials compatible with back-end-of-the-line process) and high integration density. Thereby, OxRAM-based NVFF will be considered in the following.

In order to guarantee fast and robust restore operations, many state-of-the-art NVFF designs use a differential structure comprising of two complementary programmed RRAM to save state of the FF [5,6]. In such topologies, the values of the High Resistance State (HRS) and Low Resistance State (LRS) has direct impact on the performance of the cell. For instance, a small LRS value requires larger programming currents and energy, but at the same time, it leads to a larger resistance ratio that improves the reliability of the restore operation [7].

In this paper, we present a novel design methodology of NVFF design to minimize area and energy overhead while guaranteeing reliable restore operations. By identifying minimal HRS/LRS ratio admissible by the FF circuit, it is possible to reduce the programming energy and therefore, programming circuit area. The methodology has two main steps: 1) Characterization of the lower bound of HRS/LRS that guarantees a reliable restore for a considered topology while accounting CMOS process variations. Key to minimizing this ratio is balancing the load capacitance seen by the FF internal storage nodes. 2) Sizing of programming transistors to achieve the estimated HRS and LRS values. This methodology was applied to four NVFF topologies based on a Transmission-Gate FF (TGFF) architecture and exploiting different store/restore circuits. Simulated in a 180nm CMOS process accounting for process variability, we demonstrate from 38% to 76% of reduction in minimum resistance ratio that, in turns, leads to a 34% reduction in store energy, a 18% reduction in restore energy and a 40% reduction in area overhead compared to reported state-of-the-art NV-TGFFs.

The remainder of this paper is organized as follows. Section II provides a general background. Section III introduces the NV-TGFF topologies under study. Section IV presents the design methodology and applies it to our NV-TGFFs. Experimental results are given in Section V, while Section VI concludes the paper.

II. BACKGROUND

This section provides a brief background related on the technology and design principles behind OxRAM-based NVFF.

A. OxRAM Technology

Among the many candidate technologies, Oxide-based Random Access Memories (OxRAMs) appears as a highly
promising technology due to their fast switching and Back-End-of-Line (BeOL) friendly materials [8]. OxRAM operation is related to the creation and destruction of an oxygen vacancies-based Conductive Filament (CF) formed in a transition metal oxide stack [9]. By applying voltages across the top (TE) and bottom (BE) electrodes of the OxRAM, its resistance state can be modified. After an initial forming operation, i.e., creation of the first CF, the OxRAM can be switched between a High Resistance State (HRS) and respectively Low Resistance State (LRS) by the set and reset operations. The reset operation is achieved by applying a negative voltage across TE-BE. This results in the rupture of the CF, switching the OxRAM to a HRS. The achieved HRS value depends on the reset voltage and the duration of reset pulse [10]. During a set operation, a positive voltage is applied across TE-BE. This results in the widening of the CF, switching the OxRAM to a LRS. The achieved LRS value depends on the injected current during the switching [10]. The dependence between the sizing of the programming circuit and the performance of OxRAMs will be used in this paper to optimize the energy and area of storage elements.

In this work, an OxRAM model considering the field effect on the creation and destruction of oxygen vacancies is used to achieve electrical simulations. In this model, presented in [11] the achieved resistance is directly linked to a CF radius. Voltage differences between TE and BE leads to forming, set and reset operations. While the forming operation is a set operation considering longer time and higher voltages, we considered it as already performed in this work.

B. Non-Volatile Flip-Flop

Thanks to their zero-leakage power consumption during sleep periods, Non-Volatile Flip-Flops (NVFF) are appealing solutions to improve the energy efficiency of the Internet of Things objects. Fig. 1 shows the schematic diagram of a volatile positive edge triggered Master-Slave Transmission Gate Flip-Flop (TGFF) from [12]. This FF will serve as a baseline structure in the rest of the paper. The TGFF state is represented by the logic outputs of the slave latch, named \( \overline{Q} \) and \( Q \) which are buffered out to \( \text{out}Q \) and \( \text{out}Q \).

Such structure can be made non-volatile by adding complementary programmed ReRAM in the structure. A NVFF has three modes of operation namely normal mode, store mode and restore mode. During normal mode, restore and store signals are inactive and the NVFF operates as a normal D-FF. Fig.2 presents several NVFF slaves. Store operation is initiated right before the system goes into a sleep phase. During store operation, the current state of the FF is programmed in two complementary OxRAMs (One is set while the other is reset [6,7,13,14]). Restore operation is initiated when the system wakes-up from a sleep phase. It is performed in two steps: First, \( Q \) and \( \overline{Q} \) are precharged and equalized to \( V_{DD} \). Then, the nodes discharge through the complementary programmed OxRAMs. Due to the resistance values difference in the 2 branches, one node will discharge faster than the other setting up a race condition to turn on one of the \( p \)-type transistors of the cross-coupled inverters, thereby restoring the state.

III. CONSIDERED NVFF TOPOLOGIES

In this section, we introduce four different NV-TGFFs implementations. While the master stages are directly borrowed from a regular TGFF, the proposed topologies differ from their slave latches implementation. The considered slave configurations are shown in Fig. 2.

The first configuration (Config-1) was taken from [13,14]. This topology stores the content of the internal nodes \( Q \) and \( \overline{Q} \) into the OxRAM by using 4 tri-state inverters, i.e., 8 transistors per memory element, directly connected to the electrodes (TE – BE), of the OxRAMs. A restore is performed by (i) connecting the OxRAMs to the internal nodes \( Q \) and \( \overline{Q} \) through \( n \)-type transistors, (ii) meta-stabilizing the latch through a pre-charge and equalize circuit and (iii) letting the latch lock in a given state depending on the resistance stored in the OxRAMs memories. The store and restore paths are independent of each other which minimizes the sizes of \( n \)-type transistors used to connect the OxRAMs in the differential pull-down paths. The second topology (config-2) is a variant of config-1 that was also proposed in [13,14]. In this topology, only the restore circuitry is different. Instead of loading the internal nodes \( Q \) and \( \overline{Q} \), the OxRAMs are inserted in the pull-down network of the cross-coupled inverters. During a restore, the stored resistance of the memories will unbalance the inverters locking the latch in a given state. Compared to config-1, this restore approach leads to no capacitance addition to the nodes \( Q \) and \( \overline{Q} \). The third topology (config-3)

![Figure 2: Schematic diagram of the considered NV-TGFF slaves, Config-1 (a) and 2 (b) are from [12] and [13]. Config-3 (c) and 4 (d) are variants of (a) and (b). Command signals are generated using (f) while (e) is the output buffer including an additional TG to balance the internal nodes capacitance. Store circuitry is represented in green while restore circuitry is in blue.](image-url)
uses a similar restore concept like config-1, i.e., the restore is performed using a direct connection of the nodes $Q$ and $\overline{Q}$ to the OxRAMs but uses transmission gates (n-type and p-type transistors) instead of a simple n-type transistor. The use of TG provides a wider voltage dynamic range during restore operation than n-type only. Compared to the previous topologies, the store operation does not rely anymore on additional tri-state inverters. Instead, we let the internal nodes drive one electrode of the OxRAMs, by activating the restore signal, while the other electrode is driven by only two additional tri-state inverters. The last topology (config-4) uses a similar restore approach like config-2, i.e., the OxRAMs unbalance the cross-coupled inverters by loading their pull-down networks. The store circuit is similar to config-3.

The restore mechanisms used by the different topologies rely on unbalancing the loads either at nodes $Q$ and $\overline{Q}$ or within the pull-down network of the inverters. As a result, and as identified in [13,14], the capacitances of the slave latch must be properly matched to maximize the restore reliability. Slave inverters of NV-TGFF are symmetrical compared to basic TGFF, while, in the output of NV-TGFF an extra TG was added to balance the capacitances (as shown Fig. 2-e). These design changes ensured a balanced load for nodes $Q$ and $\overline{Q}$.

IV. PROPOSED DESIGN METHODOLOGY FOR STORE ENERGY AND AREA REDUCTION

In order to reduce the programming energy and the circuit area, we propose to, identify by electrical characterization the optimal lower bound of the OxRAM resistance ratio required for properly restoring the FF state even under CMOS fabrication variability. Once the lower bound is identified, the programming circuit can be sized appropriately. Avoiding a worst-case overdesign is beneficial to operate with minimal energy and area. Several iterations are done between the 2 steps in order to take in account the effect of the store circuit sizing in the restore operation. Thereby, added capacitance and serial resistances in the restore path due to the store circuit are taken in account.

A. Step 1: Minimal HRS to LRS Ratio Characterization

As the restore operation relies on the discharge of nodes $Q$ and $\overline{Q}$ through the OxRAM resistances, characterization of the minimum HRS/LRS ratio is achieved in two steps. First, the capacitances are balanced such that the discharge time during restore operation will be affected only by the OxRAM resistance values. Knowing that HRS depends only on time and voltage [10], we assume that for a given system clock frequency and programming voltage, the HRS value can be considered fixed. Then, we evaluate the maximum tolerable LRS value providing a successful restore by, first sweeping the LRS value until the restore operation fails, as illustrated by the example waveforms in Fig. 5. Then, from the LRS value causing a restore fail, we perform Monte Carlo simulations accounting for CMOS variability and we decrease the LRS value, until the 10,000 statistical runs provide a successful restore. This way, the minimum HRS/LRS ratio enabling successful restore independently of the CMOS variability is determined.

B. Step 2: Sizing of Store/Restore Circuit Transistors

Once the restore conditions are identified, the store programming transistors can be sized to achieve the target HRS/LRS ratio and account for OxRAM variability. A smaller target HRS/LRS ratio requires smaller programming current and time [10]. This lead to smaller writing circuitry, faster programming operations and therefore a lower programming energy. To compensate the OxRAMs intrinsic variability [10], a 10% of variability was considered in the programmed OxRAMs state. Thereby, the store circuit is oversized to ensure successful restore of the programmed OxRAM resistance states. Fig. 4 illustrates how the OxRAM variability is accounted in step 2. The maximum LRS represents the Gaussian upper tail, while the minimum HRS is the Gaussian lower tail. Thereby, the set programming current is increased in order to ensure a sufficiently high HRS/LRS ratio leading to a successful restore. Note that any margin can be accounted by our methodology. 10% OxRAM resistance variability is used as an illustration.

V. NV-TGFF SIZING AND PERFORMANCE ESTIMATION

In this section, the methodology introduced in the previous section is applied to the NV-TGFFs under consideration. Then, the performances of the optimized topologies are evaluated and discussed.

A. Restore HRS/LRS Ratio Characterization

The results of the methodology on the HRS/LRS restore ratio are presented in Fig. 5. The minimum restore ratio are presented for each topologies at 1.8V and 1V. As a reference a non-optimized and unbalanced NV-TGFF using a 1.8× restore ratio is considered as in [13,14]. Config-3 provides the best successful restore ratio (1.02×, a 76% improvement compared to the reference). The use of TGs in the restore path decreases the serial resistance compared to the n-type transistor used in the other topologies and enables a ratio improvement compared to the 38% achieved by config-2. At 1V, the restore ratio of config-2 and more specifically config-4 are degraded due the number of transistors in the restore path, increasing the overall serial resistance and CMOS variability effects.

Figure 3: HRS/LRS ratio characterization by sweeping LRS value with a fixed HRS value and validating the restore operation each case. Here config-2 is simulated: LRS values are swept from 30kΩ to 60kΩ.

Figure 4: Identification of the minimum ratio and sizing of the store circuit accounting of the OxRAM variability.

Figure 5: HRS/LRS ratio characterization by sweeping LRS value with a fixed HRS value and validating the restore operation each case. Here config-2 is simulated: LRS values are swept from 30kΩ to 60kΩ.

Figure 6: Identification of the minimum ratio and sizing of the store circuit accounting of the OxRAM variability.
B. Area and Performance Overhead Comparison

The programming area overheads for each NV-TGFF topologies are presented in Fig. 6 and compared to the reference NV-TGFF. The area estimation considers an oversizing of the programming circuit to account a 10% variability on the memories as discussed in Section IV-b. The topologies optimized with the proposed technique show an improvement of 40% compared to the reference. Among the optimized NV-TGFF, config-3, which provides the smallest HRS/LRS ratio, leads to the smallest area overhead. Fig. 7 shows the effect of the different slave topologies on the FF performance in normal mode, i.e., on its D-Q delay. We see that the different proposed approaches do not lead to significant variations in speed.

C. Energy Overhead Comparison Results

The restore and store energies for the different topologies taking into account the minimum HRS/LRS restore ratio are presented in Fig. 9 and Fig. 10 respectively. A topology with a lower HRS/LRS ratio leads to large store energy reduction, counterbalanced by a slight increase in restore energy. Since the HRS is assumed constant, a lower ratio corresponds to a higher LRS. A high LRS leads to a longer restore time (slower discharge before locking the latch) where both pull-down networks are on and draining current. Overall, our proposed methodology leads to a cycle (store and restore) energy improvement of 40%, while slightly increasing the static leakage current (from 350pA to an average of 550pA for proposed NV-TGFF topologies).

D. Discussions

Among the presented and benchmarked topologies, the use of transmission-gates in the restore path (as in config-3) leads to a smaller admissible HRS/LRS ratio compared to the other restore approaches. Using our sizing methodology, a lower restore ratio leads to the smallest area overhead and store energy. In addition, when small HRS/LRS ratios are used, only small programming currents are required and therefore, it is possible to directly use the cross-coupled inverters of the slave latch as store circuits.

VI. CONCLUSION

In this paper, we introduced a novel design methodology for area and energy efficient OxRAM-based NVFF. This methodology consists in sizing the store circuit depending on the performances of the restore circuit. We showed that for differential NV-TGFF, it is possible to successfully restore HRS/LRS ratios down to 1.02 while accounting for CMOS and OxRAM variability. We then applied our methodology to 4 different NV-TGFF topologies and showed that it enables up to 40% of area improvement and 38% of store energy reduction. For differential NV-TGFF topologies, this methodology shows that energy consumption and area overhead can be strongly improved by co-optimizing store and restore operations.

VII. REFERENCES