

Energy/Reliability Trade-Offs in Low-Voltage ReRAM-Based Non-Volatile Flip-Flop Design

Ibrahim Kazi, Pascal Meinerzhagen, *Member, IEEE*, Pierre-Emmanuel Gaillardon, *Member, IEEE*, Davide Sacchetto, *Member, IEEE*, Yusuf Leblebici, *Fellow, IEEE*, Andreas Burg, *Member, IEEE*, and Giovanni De Micheli, *Fellow, IEEE*

Abstract—The total power budget of Ultra-Low Power (ULP) VLSI Systems-on-Chip (SoCs) is often dominated by the leakage power of embedded memories as well as status registers. On the one hand, supply voltage scaling down to the near-threshold (near- V_T) or even to the subthreshold (sub- V_T) domain is a commonly used, efficient technique to reduce both leakage power and active energy dissipation. On the other hand, emerging CMOS-compatible device technologies such as Resistive Memories (ReRAMs) enable non-volatile, on-chip data storage and zero-leakage sleep periods. For the first time, we present and compare ReRAM-based Non-Volatile Flip-Flop (NVFF) topologies which are optimized for low-voltage operation (including near- V_T and sub- V_T operation). Three low-voltage NVFF circuit topologies are proposed and evaluated in terms of energy dissipation and reliability. Using topologies with two complementary programmed ReRAM devices, Monte Carlo simulations accounting for parametric variations confirm reliable data restore operation from the ReRAM devices at a sub- V_T voltage as low as 400 mV. A topology using a single ReRAM device exhibits lower write energy, but requires a near- V_T voltage for robust read. Energy characterization is performed at nominal, near- V_T , and sub- V_T supply voltages. The minimum energy point is reached for near- V_T read operation with a total read+write energy of 735 fJ.

Index Terms—Flip-flops, low-power electronics, nonvolatile memory.

I. INTRODUCTION

ULTRA-LOW POWER (ULP) VLSI systems such as wireless sensor nodes [1] and biomedical implants [2], running for many days or even for several years on a single battery charge, have extremely low power budgets. Embedded memo-

ries as well as status and pipeline registers consume a dominant share of the total power and area of such systems [3], while for the digital signal processing core the share is often small. This power dominance of memories is further exacerbated for systems with only short active computational periods and long sleep periods requiring data and program state retention. In such cases, the leakage power of embedded memories and registers accounts for almost the totality of the VLSI system's power consumption.

Supply voltage scaling to the near-threshold (near- V_T) regime or even down to the subthreshold (sub- V_T) regime is an efficient technique to reduce leakage power consumption (as well as active energy dissipation), at the cost of a speed degradation and increased sensitivity to process parameter variations [4].

While near- V_T and sub- V_T operation enables extremely low leakage power, emerging device technologies allowing the integration of non-volatile memory devices on top of CMOS chips bear the potential of zero-leakage sleep states [5]. Among many technological options, Oxide Memories (OxRAMs) [6] are a promising candidate for next generation non-volatile memory applications. Compared to traditional Flash memories, OxRAMs have better scalability and faster programming time [6]. While a lot of research effort targets OxRAM-based stand-alone memories, this work focuses on the seamless integration of OxRAM devices into CMOS flip-flops for use as non-volatile, distributed storage elements. Previous works on non-volatile flip-flops were based on the “memristor” [5], [7], on bipolar OxRAM [8], and Magnetic Tunneling Junction (MTJ) devices [9]–[11]. All these works consider circuit operation at a high supply voltage, normally corresponding to the CMOS technology's nominal voltage.

In this work, we combine the advantages of an emerging non-volatile memory device technology (namely, OxRAM stacks) with the assets of low-voltage (near- V_T and sub- V_T) circuit operation by specifically optimizing the NVFF topologies for these operating regimes, thereby enabling future VLSI systems with ultra-low active energy dissipation in addition to non-volatile data storage with zero leakage. This article extends our previous work [12] by introducing two new NVFF topologies and by presenting a detailed comparative analysis of all topologies.

Our detailed, simulation-based analyses show that the state of the art of NVFF design is extended on the following fronts: 1) the NVFF topologies using two complementary programmed ReRAM devices reliably recover the saved data on wake-up with a sub- V_T supply voltage and a standard deviation of up to 5% of the nominal value of the ReRAM resistance; 2) in addition, all proposed NVFF circuits are able to reliably recover data

Manuscript received February 14, 2014; revised April 23, 2014; accepted May 26, 2014. Date of publication July 16, 2014; date of current version October 24, 2014. This work was supported in part by the ERC senior grant NanoSys ERC-2009-AdG-246810 and the Swiss National Science Foundation under the projects number PP002-119057 and number 200021-146600. The work of P. Meinerzhagen was supported by an Intel Ph.D. fellowship. This paper was recommended by Associate Editor Corinto Fernando.

I. Kazi, P.-E. Gaillardon, and G. De Micheli are with the Integrated Systems Laboratory (LSI), EPFL, Lausanne, VD, 1015 Switzerland (e-mail: pierre-emmanuel.gaillardon@epfl.ch).

P. Meinerzhagen was with the Telecommunications Circuits Laboratory (TCL), EPFL, Lausanne, VD, 1015 Switzerland. He is now with Intel Labs, Intel Corporation, Hillsboro, OR 97124 USA (e-mail: pascal.meinerzhagen@epfl.ch).

D. Sacchetto is with the Integrated Systems Laboratory (LSI), EPFL, Lausanne, VD, 1015 Switzerland. He is also with the Microelectronic Systems Laboratory (LSM), EPFL, Lausanne, VD, 1015 Switzerland.

Y. Leblebici is with the Microelectronic Systems Laboratory (LSM), EPFL, Lausanne, VD, 1015 Switzerland.

A. Burg is with Telecommunications Circuits Laboratory (TCL), EPFL, Lausanne, VD, 1015 Switzerland.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSI.2014.2334891

with a High Resistance Value (HRV) to Low Resistance Value (LRV) ratio of less than two, even though a near- V_T voltage is required for the NVFF topology based on a single ReRAM device to do so; and 3) thanks to circuit optimizations allowing for aggressive voltage scaling (down to the sub- V_T domain for most circuits), the read energy has been drastically reduced to 5.4% of the total read+write energy, whereas for operation at nominal voltage, read and write would have similar energy costs.

The remainder of this paper is organized as follows. Section II introduces the manufactured ReRAM stacks and elaborates on their switching behavior. Next, Section III introduces three fundamentally different NVFF topologies based on one and two ReRAM devices and expatiates on their circuit-level optimizations for reliable low-voltage operation. Section IV presents detailed circuit simulation results of all NVFF topologies and compares their read robustness as well as their read and write energy at nominal, near- V_T , and sub- V_T supply voltages, and Section V concludes this paper.

II. RESISTIVE MEMORY: MANUFACTURING PROCESS AND SWITCHING CHARACTERISTICS

Among many ReRAM candidates, OxRAMs base their working principle on the change in resistance of an oxide layer. Different physical mechanisms can be identified in the switching of ReRAMs [6]. In the following, we focus only on the Bipolar Resistive Switching (BRS) [13], related to the O_2 vacancy redistribution in transition metal oxide layers upon application of a voltage across the oxide. Oxide-based memories are promising for circuit applications thanks to their fast write time [14] and high endurance up to 10^{11} cycles [15].

We realized memory stack prototypes of Al/TiO₂/Al from bulk-Si wafers passivated by a 100-nm thick Al₂O₃ layer. 70-nm thick Bottom Electrode (BE) lines were patterned by lift-off and e-beam evaporation. Then, a 50-nm thick TiO₂ layer was deposited by Atomic Layer Deposition (ALD) at 200°C. Finally, vertical Top Electrode (TE) lines were defined with a second lift-off step together with contact areas used for electrical characterization. Each fabricated device occupies an area of $1.5 \mu\text{m}^2$. A Scanning Electron Microscopy (SEM) image of the fabricated devices is shown in Fig. 1. Note that a crossbar architecture was chosen only to ease the device demonstration and characterization. In the following circuit proposals, however, we consider single ReRAM devices embedded in CMOS NVFF topologies. As opposed to stand-alone memories, high integration density is not the primary objective of the present work. In fact, such OxRAM devices are eventually manufactured on top of standard, mature CMOS chips, with relaxed design rules.

In our ReRAM devices, the switching operations are enable after cycling the voltage across the device, as opposite to the traditional methodology, which involves larger voltages and current compliance to electrically form the devices. After 50 burn-in cycles, the resistive switching behavior stabilizes to the behavior shown in Fig. 1. Consistent BRS with a High Resistance State (HRS) and a Low Resistance State (LRS) is achieved. The SET and RESET threshold voltages range from -2 V to $+2 \text{ V}$. Moreover, the switching operation is limited by a low current compliance of $10 \mu\text{A}$, allowing the use of small (close to minimum size) programming transistors, which leads to a compact design. As reported in literature, there exist oxide

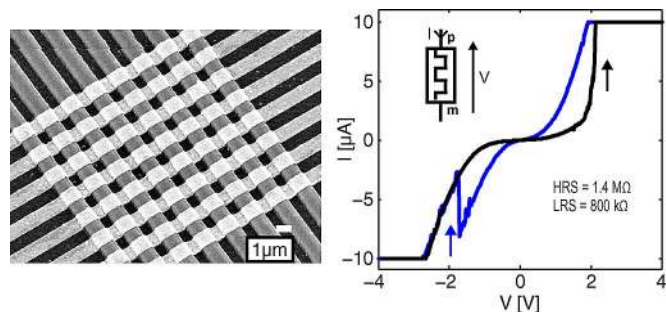


Fig. 1. $1.5 \mu\text{m}^2$ Al/TiO₂/Al ReRAM stack crossbar SEM image and single node switching characteristic under $10 \mu\text{A}$ current compliance. The crossbar structure is solely used for device characterization, while all NVFF topologies contain only one or two separate ReRAM stacks.

stacks with higher compliance current, lower LRS, and a much higher HRS/LRS ratio, compared to our considered device. However, in this study, we chose a ReRAM device with low current compliance and high LRS on purpose, in order to minimize the size of the programming transistors and the write/read energies, while accepting a low HRS/LRS ratio. Note as well that this HRS/LRS ratio, smaller than two, can be considered a worst case ratio with respect to the impact of variability compared to other devices, but it does neither significantly decrease the endurance, nor increase the variability [16]. In the following, we show that under this worst case assumption, it is still possible to obtain robust NVFF operation even at low voltages with proper circuit design. Higher HRS/LRS ratios would tremendously facilitate the circuit design and would improve the read robustness at low voltages even further.

III. NON-VOLATILE FLIP-FLOP TOPOLOGIES AND OPERATION

This section explains the design and the operating principle of the various herein proposed ReRAM-based non-volatile flip-flop topologies. In all the designs, a conventional master/slave flip-flop based on tri-state inverters is modified to add the possibility of non-volatile data storage, through additional read and write sub-circuits to/from the ReRAM device(s), as shown in the block diagram in Fig. 2(a). The various NVFF topologies differ from each other mainly in their method of recalling the saved state from the ReRAM device(s) to the CMOS slave latch, and in the number and the position of the ReRAM devices inside the circuit. The writing process to the ReRAM device(s) is almost identical for all the circuits.

We first introduce a conventional master-slave flip-flop in CMOS technology with minor modifications to accommodate two complementary programmed ReRAM devices for non-volatility. Unfortunately, circuit simulations show that this baseline NVFF design works reliably only at high (close to nominal) supply voltages. Therefore, we then introduce three different NVFF topologies which are optimized for robust operation at low voltages (near- V_T and sub- V_T supply voltages). Two of these optimized NVFF topologies are similar to the baseline design in that they also use two complementary programmed ReRAM devices. However, circuit optimizations lead to robust operation at near- V_T and even at sub- V_T supply voltages. The third topology requires only a single ReRAM device and exhibits lower energy consumption than all previous topologies, which comes at the cost of reduced robustness.

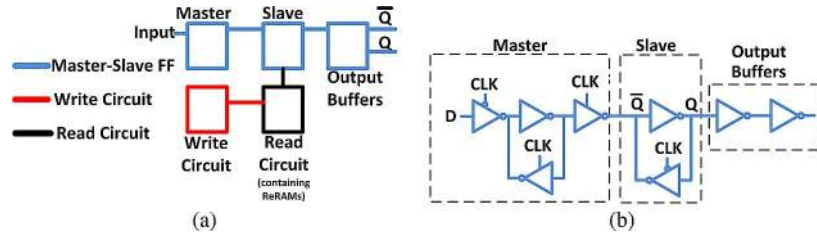


Fig. 2. (a) Conceptual block diagram of a non-volatile flip-flop; and (b) conventional master-slave flip-flop.

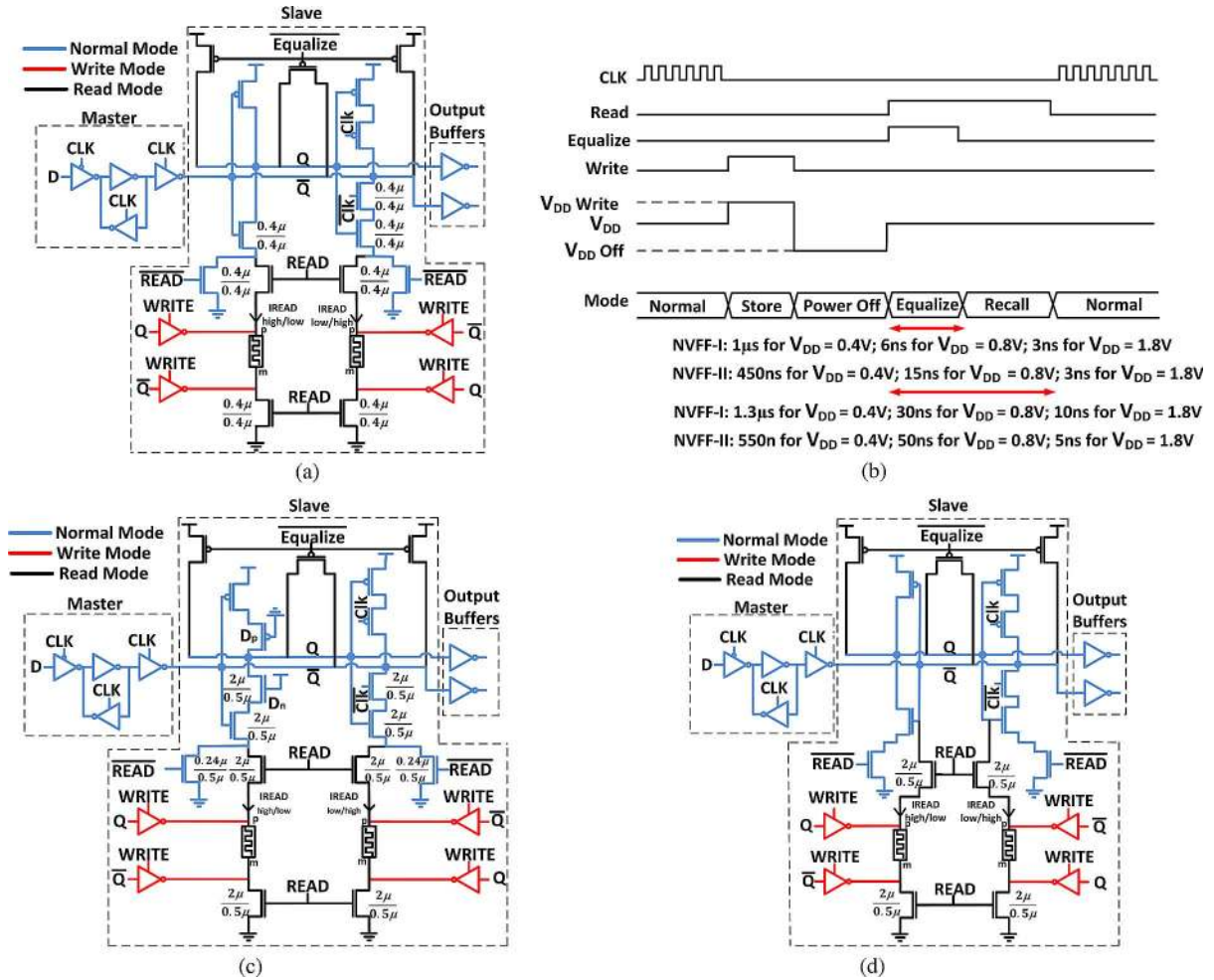


Fig. 3. (a) Baseline non-volatile flip-flop design (NVFF-0); (b) control signals sequence for ReRAM read and write operations for the complementary ReRAM-based NVFF circuits; (c) sub- V_T optimized NVFF structure, based on NVFF-0 topology (NVFF-I); and (d) sub- V_T optimized NVFF structure, inspired from [18] (NVFF-II).

A. Baseline Non-Volatile Flip-Flop (NVFF-0)

In order to enhance a conventional CMOS flip-flop structure, such as the one shown in Fig. 2(b), with non-volatile data storage, two ReRAM devices are inserted in the current sink of the cross-coupled inverter pair in the slave latch. The resulting baseline circuit, referred to as NVFF-0, is shown in Fig. 3(a). A similar circuit topology has been used in [17], which, however, used magnetic memory devices. In this baseline NVFF, the two ReRAM devices are always used in a complementary fashion, i.e., one device is programmed to the HRS, while the other one is programmed to the LRS. Dedicated non-volatile programming (or ReRAM write) circuits are highlighted in red color in

Fig. 3(a), while dedicated restore on wake-up (or ReRAM read) circuits are shown in black color.

ReRAM Read Operation: During system wake-up (power-on), the slave latch would ideally be directly restored, based on the data stored in the ReRAM devices, during ramp-up or connection of the power supply. However, this is impossible due to a number of reasons: 1) the clock and the READ signal are not controlled yet; 2) there might be uncontrolled, residual charges on the internal nodes Q and \bar{Q} ; and 3) different power-gating approaches (mechanical, footer and/or header transistors, driving the supply to ground level) result in different wake-up scenarios. Therefore, the following wake-up sequence is proposed, as shown in Fig. 3(b): 1) turn on the

power supply; 2) at the system level, silence the clock signal to low; 3) enable the READ and the EQUALIZE control signals; and 4) upon de-assertion of EQUALIZE, the slave latch is correctly restored based on the value of the ReRAM devices. Using this wake-up sequence, both internal storage nodes Q and \bar{Q} are first pre-charged and equalized using three dedicated PMOS transistors controlled by $\overline{\text{EQUALIZE}}$. Following this pre-charge phase, the internal nodes Q and \bar{Q} are connected to ground through the ReRAM devices. The complementary resistance states of the two ReRAM devices modulate the discharge currents (the branch with HRS has a lower discharge current w.r.t. the branch with LRS), starting a race condition. As soon as one internal node is discharged to $V_{DD} - V_{T,PMOS}$, the PMOS transistor driven by that node turns *on* and starts to pull up the other internal node. This decides the race, before the feedback of the latch restores full logic levels.

ReRAM Write Operation: Prior to an active-to-sleep transition, the data stored in the slave latch needs to be written to the non-volatile ReRAM devices. To this end, as shown in Fig. 3(b), the clock is silenced and kept low for the entire duration of the ReRAM write operation, thereby forcing the slave latch to be non-transparent and isolated from the master. As shown in Fig. 3(a), during write, the ReRAM devices are completely disconnected from the slave latch and from the read circuits, so that the voltage drop across their terminals can be set by the write drivers (highlighted in red). The write drivers are controlled by the internal nodes Q and \bar{Q} . In general, the write pulse width depends on the chosen ReRAM technology. Here, a write pulse width of 10 ns is used to successfully program the ReRAM devices. As previously illustrated in Fig. 1, a voltage of +2 V or -2 V is required for successful switching. To be able to use small programming transistors (with a non-negligible voltage drop across their channel) and limit the programming current, the write drivers are supplied with a voltage as high as 2.4 V. This voltage is only slightly above the nominal supply voltage range of the core transistors in the considered 0.18 μm CMOS technology and does neither seriously enhance the risk of oxide breakdown, nor compromise junction reliability, nor considerably accelerate aging.

While the ReRAM write operation requires a high voltage, the ReRAM read operation as well as the normal flip-flop operation can be performed at scaled voltages for better energy efficiency. Several architectural alternatives for the distribution of the high supply voltage are discussed in [12]. Here, we adopt the approach of dynamically rising the supply voltage of all NVFFs during a write operation, as shown in Fig. 3(b).

Normal Operation and Endurance: Note that during normal operation, the presented hybrid CMOS/ReRAM NVFF, and all other NVFF topologies introduced in the following, fully rely on CMOS transistors, which are known to exhibit high endurance. The part of the NVFF circuits containing ReRAM devices, whose endurance is not yet comparable with the one of CMOS transistors, do not switch very frequently (only during active-to-sleep state transitions), which guarantees high overall system endurance. Finally, note that the setup and hold times during normal operation of all herein proposed NVFF topologies are solely determined by the master latch and are not affected by the insertion of the ReRAM devices in the slave latch. The data call window of the presented NVFFs is therefore equal to the one of the conventional CMOS flip-flop shown in Fig. 2(b).

B. NVFF Topologies for Robust Sub- V_T and Near- V_T Operation (NVFF-I and NVFF-II)

For the baseline NVFF circuit (NVFF-0) shown in Fig. 3(a), a correct read operation depends on the modulation of the discharge currents (discharging nodes Q and \bar{Q}) by the complementary ReRAM devices. However, the discharge currents might be altered due to the following reasons: 1) different pull-down networks in the two branches due to the use of either a simple or a tri-state inverter; and 2) mismatch between the transistor pairs (in the inverters and in the dedicated read transistors) and the ReRAMs, caused by local variations. Post-layout circuit simulations show that the read operation of the NVFF-0 circuit works reliably at nominal supply voltage, whereas the Pull-Down Network (PDN) mismatch other than the complementary resistance values leads to read failures at low voltages (see Section IV for detailed simulation results). Therefore, in order to enable the straightforward integration of ReRAM-based NVFFs into ultra-low power VLSI SoCs, i.e., operated in the near- V_T or sub- V_T domain, we introduce in the following two NVFF topologies which are optimized for operation at low voltages.

The two sub- V_T optimized NVFF-I and NVFF-II structures are shown in Fig. 3(d) and Fig. 3(c), respectively, with the dedicated programming (ReRAM write) circuits highlighted in red and the restore (ReRAM read) circuits in black. Both NVFF-I and NVFF-II use the same control signal sequence, shown in Fig. 3(b) during active-to-sleep and sleep-to-active transitions as the baseline circuit NVFF-0. Similarly to NVFF-0, the two ReRAM devices are used in a complementary way, i.e., one device is programmed to the HRS, while the other one is programmed to the LRS. The main difference between these two architectures is the connection of the read circuit to the slave latch of the CMOS flip-flop. Inspired by [18], the read circuit of NVFF-II is connected to the output nodes, Q and \bar{Q} , while, in the NVFF-I topology, it is connected to the source of the NMOS transistors of the slave latch. Hence, during read operation, the PDN of NVFF-II consists of only the portion shown in black, whereas for NVFF-I, the PDN is composed of the circuit shown in black and the NMOS transistors shown in blue (which are part of the inverters forming the slave latch).

Since it is crucial to modulate the discharge current of the precharged nodes Q and \bar{Q} exclusively by the value of the ReRAM devices to ensure correct read at low voltages, any other type of mismatch in the PDN needs to be avoided. To this end, in case of NVFF-I, two always-*on* transistors (D_n and D_p) are inserted into the simple inverter (used in NVFF-0) to mimic the tri-state inverter in the other pull-down branch. The insertion of dummy transistors can be avoided in the NVFF-II topology, because the ReRAMs are connected to the output nodes rather than integrated in the PDN of the inverters. Moreover, all transistor pairs in the PDNs of both NVFF-I and NVFF-II are upsized to further improve matching (see Fig. 3(d) and Fig. 3(c) for the exact transistor dimensions).

C. NVFF Topology With Single ReRAM Device (NVFF-III)

All NVFF topologies considered so far require two ReRAM devices and a non-trivial wake-up sequence. Unfortunately, the total write energy increases with the number of ReRAM devices used in the NVFF topology. Therefore, this section proposes an

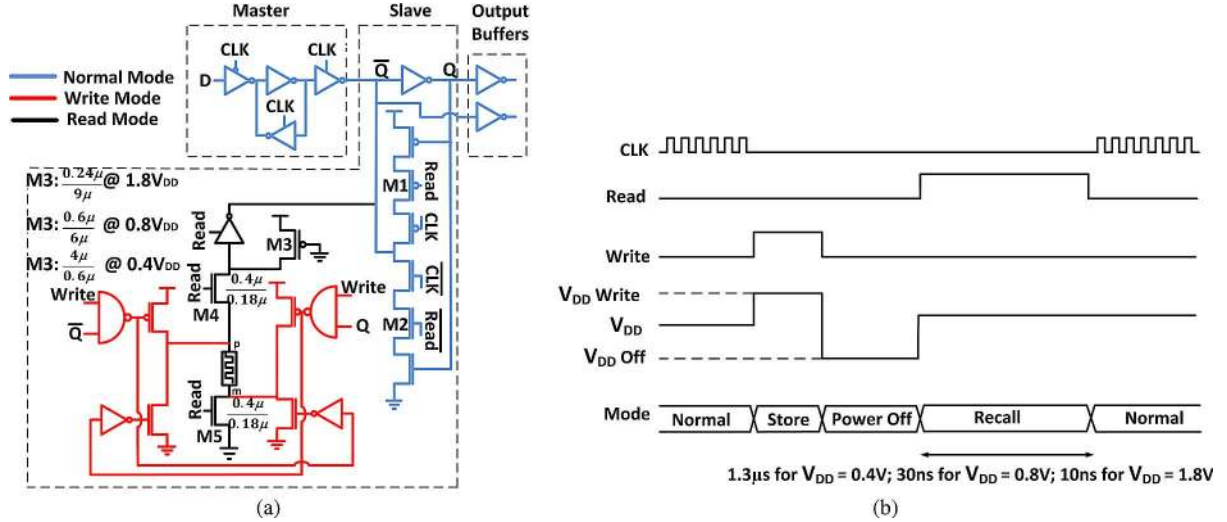


Fig. 4. (a) NVFF architecture with a single ReRAM device (NVFF-III); and (b) the corresponding control sequence for read and write operation.

NVFF topology which is based on a single ReRAM device, that leads to lower write energy, and allows for a simplified wake-up control signal sequence. Our NVFF topology is similar to [5] but uses our custom oxide stack (see Section II) instead of HP’s “memristor” [7]. This single-ReRAM NVFF architecture, referred to as NVFF-III, is depicted in Fig. 4(a).

ReRAM Read Operation: When the VLSI system is in recall mode (after wake-up), the slave latch is disconnected from the master by keeping the clock signal low, as shown in Fig. 4(b). This is done to ensure that the recalled value is not overwritten by any disturbance from the master latch. Moreover, compared to the standard CMOS master/slave flip-flop shown in Fig. 2(b), two transistors (M1 and M2) are added to disable the feedback mechanism in the slave latch during the ReRAM read operation. Therefore, the inverting buffer controlled by the READ signal (also referred to as sense inverter) can easily impose a value to the slave latch during wake-up, according to the state of the ReRAM device. The READ signal is set high so that the ReRAM can be accessed through the transistors M4 and M5. The PMOS transistor M3 acts as a current source and provides a voltage drop across the ReRAM device. M3 needs to be sized carefully. In fact, its sizing depends on the power supply voltage, the voltage transfer curve of the sense inverter (especially the trip-point), and the HRS and LRS of the ReRAM device. The sizing of M3 ensures that two logically distinct input states for the sense inverter can be produced depending on the resistance of the ReRAM device. Once the READ signal goes high, the sense inverter is turned on and forces a logic state onto the internal node \bar{Q} of the slave latch. If the ReRAM is in LRS, a logic “1” is produced at node \bar{Q} , and vice versa.

ReRAM Write Operation: Prior to an active-to-sleep transition, the ReRAM device is programmed to HRS or LRS during a write operation, depending on the logic state of Q and \bar{Q} . Compared to all previously presented NVFF circuits (NVFF-0, NVFF-I, and NVFF-II), the write circuit is slightly modified in order to enable a slight reduction in the write supply. In fact, in Fig. 3(c) and (d), a tri-state inverter is used to connect the ReRAM devices to V_{DD} (and ground) during write. In this case, there is a stack of two transistor between the ReRAM devices and V_{DD} (internal tri-state inverter topology), and a write voltage of 2.4 V is required to account for the voltage drop

across the transistor stack. In contrast, in Fig. 4(a), the write transistor is directly connected to the ReRAM device and is controlled by a NAND gate. Hence, there is only a single transistor between V_{DD} and the ReRAM device, and the write voltage can be reduced to 2.2 V. During normal mode of operation, the read (black) and write (red) circuits are disconnected from the rest of the main flip-flop.

IV. ENERGY AND RELIABILITY ANALYSIS

In this section, the reliability of low-voltage (near- V_T and sub- V_T) operation as well as the energy dissipation of the various, previously introduced NVFFs are characterized. First, the post-layout simulation setup and the methodology to account for parametric variations in MOS transistors and ReRAM devices are described. Then, detailed, comparative simulation results for read robustness and energy dissipation, during read and write, of all NVFF topologies are presented.

A. Methodology and Simulation Setup

The ULP and especially the biomedical VLSI design community often prefers to use mature CMOS technology nodes for 1) high reliability, 2) low leakage currents, and 3) low cost. Therefore, this study adopts a mature 0.18 μm CMOS process. All simulations, run by CADENCE Spectre, assume a Typical-Typical (TT) process corner at 27 °C. A dynamically adjustable power supply is presumed, switching between 2.4 V for write operations, and a lower value (V_{DD}) for read as well as normal operation (flip-flop sampling operation). V_{DD} assumes the technology’s nominal value (1.8 V), a near- V_T value (0.8 V), and a sub- V_T value (0.4 V). Monte Carlo circuit simulations (1000 runs) account for local parametric variations of all MOS transistors, according to statistical distributions provided by the foundry. While advanced statistical models of the ReRAM devices are not available yet, we assume that the HRS and the LRS follow a Gaussian distribution. The measured, nominal value of HRS (1.4 M Ω) and LRS (800 k Ω) is taken as mean value, denoted by $\mu(\text{HRS})$ and $\mu(\text{LRS})$. The values 40 k Ω , 80 k Ω , and 160 k Ω corresponding to 5%, 10% and 20% of $\mu(\text{LRS})$, respectively, are taken for the standard deviation, denoted by $\sigma(\text{HRS})$ and $\sigma(\text{LRS})$. Please note that these assumed HRS and LRS distributions correspond well

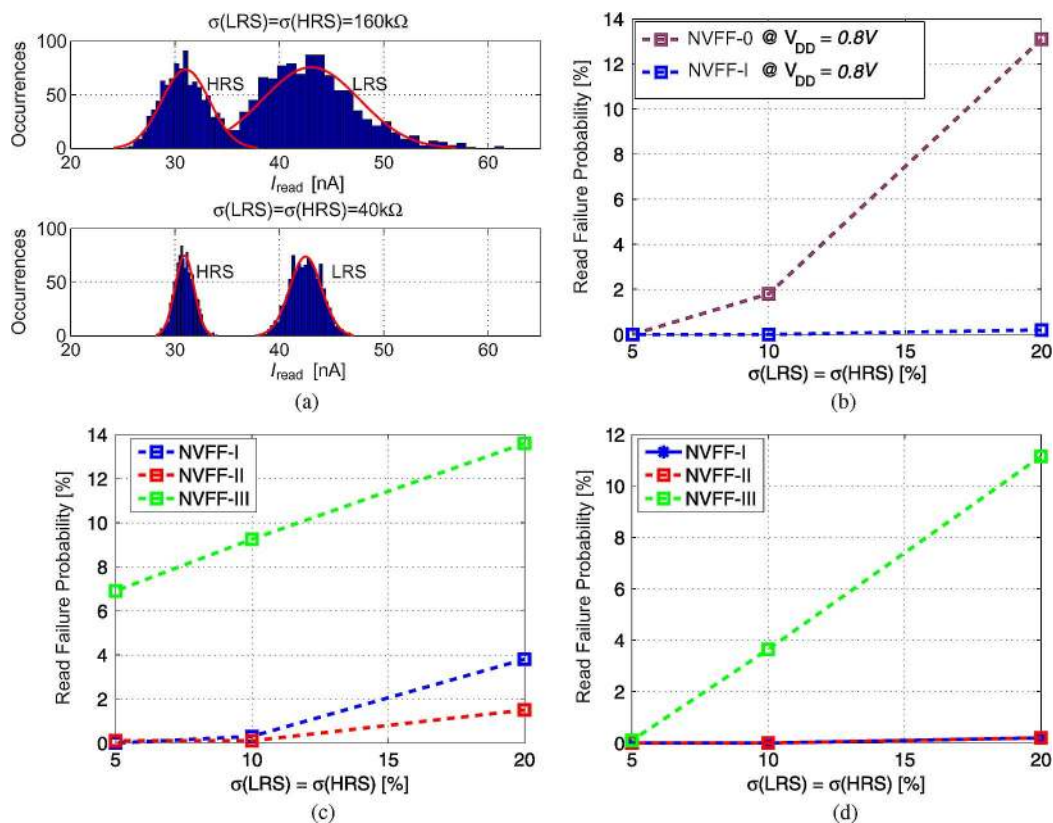


Fig. 5. (a): Distribution of the discharge current (I_{read}) through the two branches of the slave latch of the sub- V_T -optimized NVFF-I circuit, for 0.4 V, given for two different standard deviations of the ReRAM's resistance; (b)–(c): Read failure probabilities for a ReRAM resistance's standard deviation of 5%, 10%, and 20% of the nominal LRS value: (b) low-voltage optimization, NVFF-0 \rightarrow NVFF-I, 0.8 V; (c) comparison NVFF-I–NVFF-III, 0.4 V; and (d) comparison NVFF-I–NVFF-III, 0.8 V.

with reports about comparable state-of-the-art oxide stacks available in the literature [19], showing LRS (σ/μ) ratios of 20% under the same electrical conditions. Also, for mature CMOS technologies, the (σ/μ) ratios of most parameters of MOSFETs, such as the threshold voltage or the saturation current, typically range from 1% to 10% [20], [21]. Therefore, for our ReRAM devices, we are indeed considering both the case of a realistic, future, mature ReRAM process ($\sigma/\mu = 5\%$), comparable to CMOS), and the case of a currently still less mature ReRAM process ($\sigma/\mu = 20\%$).

B. Robustness at Ultra-Low Voltages

Among normal flip-flop sampling, ReRAM write, and ReRAM read operations, the ReRAM read operation is the most critical one. Indeed, studies have shown that normal operation of CMOS flip-flops can be robust in the sub- V_T domain [22]. In addition, in the present case, the write operation uses an elevated supply voltage of 2.4 V and, therefore, leads to a reliable operation.

Effectiveness of Low-Voltage Optimizations (NVFF-0 \rightarrow NVFF-I): The effectiveness of the proposed low-voltage optimization techniques is investigated by comparing the read reliability of NVFF-0 and NVFF-I. This analysis is carried out in the near- V_T domain (at 0.8 V) and in the sub- V_T domain (at 0.4 V), as well as for different values of $\sigma(HRS) = \sigma(LRS)$.

An appropriate metric to assess the read robustness is the initial discharge current (I_{read}) flowing through the two branches of the slave latch, right after the de-assertion of

the EQUALIZE signal. Fig. 5(a) shows the distributions of I_{read} at 0.4 V for NVFF-I, with different standard deviations of HRS and LRS. For a well-controlled, repeatable ReRAM process with $\sigma(HRS) = \sigma(LRS) = 40 k\Omega$, the discharge current flowing through the branch containing the ReRAM in the HRS is clearly lower than the current flowing through the other branch (non-overlapping I_{read} distributions). This results in zero read failures out of 1000 Monte Carlo runs, as shown in Fig. 5(c). However, for a less precisely controlled ReRAM process with higher standard deviation of the resistance ($\sigma(HRS) = \sigma(LRS) = 160 k\Omega$), the distributions of I_{read} start to overlap, which results in a small read failure probability.

Fig. 5(b) shows the tremendous improvement in read robustness enabled by the low-voltage optimization techniques, i.e., by matching of the PDNs through the insertion of dummy transistors as well as transistor upsizing. More precisely, the figure shows the read failure probability (computed based on 1000 MC trials) versus the presumed values of $\sigma(HRS)$ and $\sigma(LRS)$, for a supply voltage of 0.8 V. Comparing NVFF-I with NVFF-0, the dummy transistors and the transistor upsizing decrease the read failure probability from 13% down to almost 0% for the presumed worst case ReRAM process ($\sigma(HRS) = \sigma(LRS) = 20\% \times \mu(LRS)$).

Comparison of NVFF-I–NVFF-III, and Single-Ended vs. Differential: Fig. 5(c) and (d) illustrate the robustness of NVFF-I, NVFF-II, and NVFF-III for operation at near- V_T and sub- V_T voltages, respectively. The differential NVFF-I and

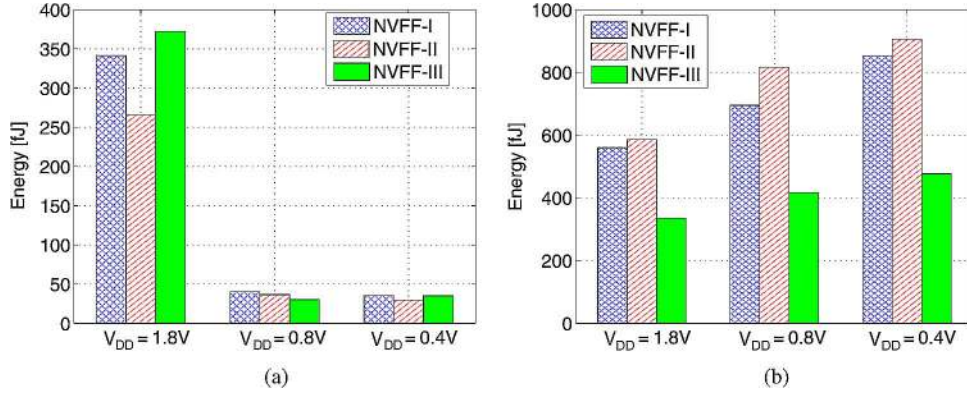


Fig. 6. (a) Read energy; and (b) write energy comparisons for different main supply voltages V_{DD} . Note that the voltage is always risen to 2.4 V for ReRAM write.

NVFF-II topologies exhibit a much lower read failure probability than the single-ended NVFF-III topology. In NVFF-III, accuracy is also needed for the current source (transistor M3), to create an appropriate drop across the ReRAM device. If the value of the generated current deviates from its nominal value, the sense inverter might switch erroneously, giving rise to read failures. Conversely, in NVFF-I and NVFF-II only relative accuracy is needed. This is why they are more robust than NVFF-III, both at near- V_T and sub- V_T supply voltages. For a near- V_T supply voltage, NVFF-I and NVFF-II exhibit no read failures for $\sigma(\text{HRS}) = \sigma(\text{LRS}) \leq 10\% \times \mu(\text{LRS})$, whereas NVFF-III exhibits a read failure rate of almost 4% under the same conditions. Briefly, a differential structure using two complementary programmed ReRAM devices is more robust than a single-ended structure employing a single ReRAM device.

Finally, NVFF-II reads more reliably in the sub- V_T domain (at 0.4 V) than NVFF-I, since the former has fewer transistors and therefore less sources of mismatch in the PDN. In addition, NVFF-II has a slightly smaller area cost than NVFF-I since it uses two transistors less. In conclusion, the differential NVFF-II is selected as the best-practice circuit for robust operation at ultra-low voltages in the sub- V_T domain, whereas the smaller, single-ended NVFF-III circuit is a viable option only down to the near- V_T domain.

C. Energy Dissipation

Fig. 6(a) and (b) show the energy dissipation per single read and write operation, respectively, of the NVFFs at different supply voltages. The main power supply V_{DD} (used for read and normal operations) is swept from 1.8 V to 0.4 V. Prior to a write operation, the power supply is always risen to 2.4 V. For each V_{DD} , the read operation is performed at maximum speed, with the minimum required pulse widths for the EQUALIZE and READ signals, as given in Fig. 3(b) for NVFF-I and NVFF-II and in Fig. 4(b) for NVFF-III. We found that initially, voltage scaling from 1.8 V to 0.8 V considerably reduces the read energy; however, the active energy benefits of further voltage scaling are offset by longer pulse widths at 0.4 V (in the order of μs instead of tens of ns) and the associated integration of leakage currents.

Circuit NVFF-II can read faster than circuit NVFF-I thanks to the simpler PDN with a considerably reduced number of stacked transistors (only two instead of four). In fact, thanks to the lower resistance pull-down path, the minimum READ

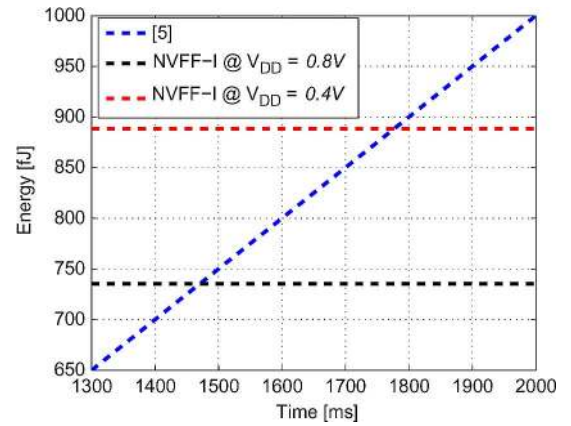


Fig. 7. Energy break-even sleep time of NVFF-I w.r.t. low-leakage CMOS latch [23].

pulse width for NVFF-II is reduced by almost $2\times$ compared to NVFF-I (see Fig. 4(b)), which leads to read energy savings in the sub- V_T regime where leakage plays a significant role. Note that the steady current flow during the entire duration of the READ pulse in the single-ended NVFF-III topology leads to a similar read energy compared to the differential topologies.

Regarding the write energy, the supply needs to be risen by 0.6 V and 2 V for a write operation, for a main V_{DD} of 1.8 V and 0.4 V, respectively. As illustrated in Fig. 6(b), the lower the main V_{DD} is, the larger the transition to 2.4 V, and the larger the write energy. As a main insight from Fig. 6(b), we conclude that the single-ended NVFF-III topology exhibits around half the write energy compared to the differential topologies, irrespective of the main supply voltage. In other words, the write energy is approximately linear to the number of ReRAM devices in the NVFF topology.

Irrespective of the NVFF topology, the write energy is always higher than the read energy. In fact, the write energy mostly depends on the ReRAM stack and cannot be improved by voltage scaling. As opposed to this, the read energy can be significantly lowered by voltage scaling, and becomes small compared to the write energy for low-voltage operation. The total write+read energy is approximately the same for both NVFF-I and NVFF-II, whereas this total energy is reduced to around half of the NVFF-III topology.

Comparison With Volatile CMOS Storage Element: Since most ReRAM technologies are still immature, due to partially

missing energy reports in the literature, and due to a multitude of different ReRAM technologies, we prefer to compare our approach with conventional, volatile, low-leakage CMOS storage elements. To this end, the total read+write energy of NVFF-I (or, equivalently, NVFF-II) is compared with the energy of a leakage-optimized latch in a mature CMOS technology [23]. The proposed NVFF topologies require a constant energy cost for sleep preparation and wake-up, i.e., the total read+write energy, which is independent of the sleep time. In contrast, the energy cost resulting from leakage currents increases with the sleep time for the conventional CMOS latch.

Fig. 7 shows that the proposed NVFF-I (or NVFF-II) circuit is more energy efficient for sleep times longer than 1.47 s and 1.77 s for operation at 0.8 V and 0.4 V, respectively, compared to the low-leakage CMOS latch.

V. DISCUSSIONS AND CONCLUSIONS

In this paper, we have proposed several non-volatile flip-flop (NVFF) circuits based on ReRAM technology, and compared their energy efficiency and robustness for operation at ultra-low voltages. These circuits leverage the use of sub- V_T operation enabling energy-efficient VLSI systems with zero-leakage sleep states. The manufactured oxide stacks switch their resistive state with a 0.18 μm CMOS-compatible voltage of 2 V and under a low current compliance of 10 μA . For all proposed NVFF topologies, the write energy is mostly ReRAM technology dependent. Conversely, thanks to sub- V_T and near- V_T operation, the read energy is brought down to 5.4% of the total read+write energy (in case of differential NVFF circuits). Under voltage scaling, the read energy improvement saturates between near- V_T and sub- V_T due to the increase in the READ pulse width. For the differential NVFF circuits, Monte Carlo simulations demonstrate a robust read operation at 0.4 V, accounting for parametric variations in both ReRAM devices and MOS transistors. Robustness can be further increased by having a larger ratio between the high and low resistance values of the ReRAM device.

Differential NVFF circuits with two complementary programmed ReRAM devices are more robust than their single-ended NVFF counterpart using only one ReRAM device, especially for operation at low voltages. However, the single-ended topology is around $2\times$ more energy efficient compared to the differential topologies. A differential NVFF circuit is required to ensure robust sub- V_T operation, while the more energy-efficient single-ended topology is a viable option only for operation at nominal or slightly scaled voltages. Among the considered differential circuits, it is clearly beneficial to connect the ReRAM read circuit to the internal nodes of the slave latch rather than to its current sink. In fact, the former architectural variant leads to smaller area (it avoids two transistors otherwise used for matching of PDNs), increased robustness at sub- V_T voltages (better matching due to less devices in PDN), higher ReRAM read speed (lower equivalent PDN resistance), and comparable total read+write energy.

In summary, the differential NVFF-II topology (with the ReRAM read circuit connected to the internal nodes of the

slave latch) is selected as the best-practice circuit for robust operation at ultra-low voltages in the sub- V_T domain, whereas the more energy-efficient and smaller, single-ended NVFF-III circuit is a viable option only down to the near- V_T domain.

REFERENCES

- [1] A. Chan *et al.*, "Low power wireless sensor node for human centered transportation system," *Systems, Man and Cybernetics Tech. Dig.*, 2012.
- [2] J. Abouei, J. Brown, K. Plataniotis, and S. Pasupathy, "Energy efficiency and reliability in wireless biomedical implant systems," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 456–466, 2011.
- [3] J. Constantin *et al.*, "TamaRISC-CS: An ultra-low-power application-specific processor for compressed sensing," *VLSI-SoC Tech. Dig.*, 2012.
- [4] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3163–3173, Nov. 2009.
- [5] C.-M. Jung, K.-H. Jo, E.-S. Lee, H. M. Vo, and K.-S. Min, "Zero-sleep-leakage flip-flop circuit with conditional-storing memristor retention latch," *IEEE Trans. Nanotechnol.*, vol. 11, no. 2, pp. 360–366, 2012.
- [6] G. W. Burr *et al.*, "Overview of candidate device technologies for storage-class memory," *IBM J. Res. Devel.*, vol. 52, no. 4.5, pp. 449–464, 2008.
- [7] D. Strukov, G. Snider, D. Stewart, and S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, 2008.
- [8] S. Onkaraiyah *et al.*, "Bipolar ReRAM based non-volatile flip-flops for low-power architectures," *NEWCAS Tech. Dig.*, 2012.
- [9] Y. Jung *et al.*, "MTJ based non-volatile flip-flop in deep submicron technology," *Int. J. Circuit Theory Applicat.*, 2012.
- [10] W. Zhao, E. Belhaire, and C. Chappert, "Spin-MTJ based non-volatile flip-flop," *NANO Tech. Dig.*, 2007.
- [11] Y. Jung *et al.*, "MTJ based non-volatile flip-flop in deep submicron technology," *ISOC Tech. Dig.*, 2011.
- [12] I. Kazi *et al.*, "A ReRAM-based non-volatile flip-flop with sub- V_T read and CMOS voltage-compatible write," *NEWCAS Tech. Dig.*, 2013.
- [13] Y. S. Chen *et al.*, "Challenges and opportunities for HfOX based resistive random access memory," *IEDM Tech. Dig.*, 2011.
- [14] C. Yoshida, K. Tsunoda, H. Noshiro, and Y. Sugiyama, "High speed resistive switching in Pt/TiO₂/TiN film for nonvolatile memory application," *Appl. Phys. Lett.*, vol. 91, p. 223510, 2007.
- [15] Y.-B. Kim *et al.*, "Bi-layered RRAM with unlimited endurance and extremely uniform switching," *VLSI Tech. Dig.*, 2011.
- [16] S.-G. Park *et al.*, "A non-linear ReRAM cell with sub-1 μA ultralow operating current for high density vertical resistive memory (VRRAM)," *IEDM Tech. Dig.*, 2012.
- [17] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, "Nonvolatile magnetic flip-flop for standby-power-free SoCs," *Proc. IEEE CICC*, 2008.
- [18] P.-F. Chiu *et al.*, "Low store energy, low VDDmin, 8T2R nonvolatile latch and SRAM with vertical-stacked resistive memory (memristor) devices for low power mobile applications," *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1483–1496, Jun. 2011.
- [19] A. Chen and M.-R. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," *IEEE Reliability Physics Symp. Dig.*, 2011.
- [20] S. Saxena *et al.*, "Variation in transistor performance and leakage in nanometer-scale technologies," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 131–144, Jan. 2008.
- [21] C. M. Mezzomo *et al.*, "Characterization and modeling of transistor variability in advanced CMOS technologies," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2235–2248, Aug. 2011.
- [22] P. Meinerzhagen, S. Sherazi, A. Burg, and J. Rodrigues, "Benchmarking of standard-cell based memories in the sub-domain in 65-nm CMOS technology," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 1, no. 2, pp. 173–182, 2011.
- [23] P. Meinerzhagen *et al.*, "A 500 fW/bit 14 fJ/bit-access 4kb standard-cell based sub- V_T memory in 65 nm CMOS," *Proc. ESSCIRC*, 2012.



Ibrahim Kazi was born in Pakistan in 1987. He received the B.S. degree in electronics from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, and the M.S. degree in system on chip design from KTH, Stockholm, Sweden. His M.S. thesis was *Low Power Current Mode Sigma-Delta ADC using a Ring Oscillator Based Quantizer*. Currently, he is a doctoral student at MICAS, working towards the Ph.D. degree in flexible mixed digital-RF modulators.



Pascal Meinerzhagen (S'10–M'14) received the M.Sc. and B.Sc. degrees, both in electrical engineering, from EPFL, Lausanne, Switzerland, in 2008 and 2006, respectively. He also received the M.Sc. degree in micro- and nanotechnologies for integrated systems jointly from Grenoble INP, Politecnico di Torino, and EPFL in 2008. He received the Ph.D. degree in electrical engineering from EPFL in February 2014.

He is currently a senior research scientist at Intel Labs, Intel Corporation, Hillsboro, OR, USA. In 2014, he was a post-doctoral fellow and a lecturer at the Bar-Ilan University, Ramat-Gan, Israel, where he established the Advanced Digital VLSI Design course. His current research interests are broad, ranging from conventional to emerging memories, to ultra-low power VLSI (biomedical), to error-resilient systems (wireless communications), to power delivery and power management techniques. He has authored/co-authored one invited book chapter, 24 journal articles and international conference papers (two under review), and three patent applications.

Dr. Meinerzhagen is a reviewer for 15 international journals and conferences, including IEEE TCAS-I, IEEE TCAS-II, the Elsevier *Microelectronics Journal*, and IEEE ISCAS. He has received an Intel Ph.D. fellowship and two best paper nominations.



Pierre-Emmanuel Gaillardon (S'10–M'11) received the Electrical Engineer degree from CPE, Lyon, France, in 2008, the M.Sc. degree from INSA, Lyon, France, in 2008, and the Ph.D. degree in electrical engineering from the University of Lyon, France, in 2011.

He works for EPFL, Lausanne, Switzerland, as a research associate at the Laboratory of Integrated Systems (LSI). Previously, he was a research assistant at CEA-LETI, Grenoble, France. Involved in the Nanosys project, his research activities and interests

are currently focused on emerging nanoscale devices for digital circuits and systems.

Dr. Gaillardon was a recipient of the C-Innov 2011 Best Thesis Award and the Nanoarch 2012 Best Paper Award. He has served as a TPC member for the Nanoarch'12–'14, CMOSETR'13–'14, and ISVLSI'14 conferences and is a reviewer for several journals (AIP APL, IEEE TNANO, IEEE TVLSI, ACM JETC), conferences (ICECS, ISCAS), and funding agencies (ANR, Chairs of Excellence program of Nanosciences Foundation).



Davide Sacchetto (S'12–M'13) received the B.S. degree in physics engineering from Politecnico di Torino, Italy, in 2007. In 2008 he received the joint M.S. degree in micro and nano technologies for integrated systems from École Polytechnique Fédérale de Lausanne (EPFL), the Institut National Polytechnique de Grenoble (INPG), and the Politecnico di Torino (POLITO). He received the Ph.D. degree in microsystems and microelectronics in September 2013 from EPFL.

His research interests focus on novel devices, investigating issues ranging from solid-state microfabrication to circuit implementation. He is interested in the post-processing fabrication of CMOS technology that can enable integration of CMOS with the additional functionality

that emerging technology can bring. For instance, resistive RAM devices integrated with CMOS can bring performance benefits to diverse CMOS applications, such as FPGAs, artificial neural networks, and standalone memories.



Yusuf Leblebici (S'88–M'90–SM'98–F'10) received the B.Sc. and M.Sc. degrees in electrical engineering from Istanbul Technical University, Istanbul, Turkey, in 1984 and 1986, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), IL, USA, in 1990.

Between 1991 and 2001, he worked as a faculty member at UIUC, at Istanbul Technical University, and at Worcester Polytechnic Institute (WPI). In 2000–2001, he also served as the Microelectronics

Program Coordinator at Sabanci University, Turkey. Since 2002, he has been a Chair Professor at the Swiss Federal Institute of Technology in Lausanne (EPFL), and Director of the Microelectronic Systems Laboratory. His research interests include design of high-speed CMOS digital and mixed-signal integrated circuits, computer-aided design of VLSI systems, intelligent sensor interfaces, modeling and simulation of semiconductor devices, and VLSI reliability analysis. He is the coauthor of six textbooks, namely, *Hot-Carrier Reliability of MOS VLSI Circuits* (Kluwer Academic, 1993), *CMOS Digital Integrated Circuits: Analysis and Design* (McGraw Hill, 1st ed., 1996, 2nd ed., 1998, 3rd ed., 2002), *CMOS Multichannel Single-Chip Receivers for Multi-Gigabit Optical Data Communications* (Springer, 2007), *Fundamentals of High Frequency CMOS Analog Integrated Circuits* (Cambridge University Press, 2009), *Extreme Low-Power Mixed Signal IC Design* (Springer, 2010) and *Reliability of Nanoscale Circuits and Systems* (Springer, 2011), as well as more than 250 articles published in various journals and conferences.

Dr. Leblebici has served as an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II, and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATED (VLSI) SYSTEMS. He has also served as the general co-chair of the 2006 European Solid-State Circuits Conference and the 2006 European Solid State Device Research Conference (ESSCIRC/ESSDERC). He was elected as a Distinguished Lecturer of the IEEE Circuits and Systems Society for 2010–2011.



Andreas Burg (S'97–M'05) was born in Munich, Germany, in 1975. He received the Dipl.-Ing. degree from the Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland, in 2000. He then joined the Integrated Systems Laboratory of ETH Zurich, from which he graduated with the Dr. sc. techn. degree in 2006.

In 1998, he worked at Siemens Semiconductors, San Jose, CA, USA. During his doctoral studies, he worked at Bell Labs Wireless Research for one year. From 2006 to 2007, he held positions as postdoc-

toral researcher at the Integrated Systems Laboratory and at the Communication Theory Group of the ETH Zurich. In 2007, he co-founded Celestrus, an ETH-spinoff in the field of MIMO wireless communication, where he was responsible for the ASIC development as Director for VLSI. In January 2009, he joined ETH Zurich as SNF Assistant Professor and as head of the Signal Processing Circuits and Systems group at the Integrated Systems Laboratory. Since January 2011, he has been a tenure-track Assistant Professor at the École Polytechnique Fédérale de Lausanne (EPFL) where he is leading the Telecommunications Circuits Laboratory in the School of Engineering. In his professional career, he has been involved in the design of more than 35 ASICs. He has published more than 120 papers in peer-reviewed conferences and journals.

In 2000, Dr. Burg received the Willi Studer Award and the ETH Medal for his diploma and his diploma thesis, respectively. He was also awarded an ETH Medal for his Ph.D. dissertation in 2006. In 2008, he received a 4-years grant from the Swiss National Science Foundation (SNF) for an SNF Assistant Professorship. In 2013, he received a Best Paper Award from the *EURASIP Journal on Image and Video Processing*. He has served on the TPC of various conferences on VLSI, signal processing, and communications and was a TPC chair for VLSI-SoC 2012. He also served as General Chair for D43D 2012 at EPFL. He served as Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART I in 2013, as Editor for the *MDJI Journal on Low Power Electronics and Applications* since 2012, and as Editor for the Elsevier *Microelectronics Journal* since 2014.



Giovanni De Micheli (M'83–SM'89–F'94) received the Nuclear Engineer degree from Politecnico di Milano, Italy, in 1979, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California at Berkeley, CA, USA, in 1980 and 1983, respectively.

He is Professor and Director of the Institute of Electrical Engineering and of the Integrated Systems Centre at EPF Lausanne, Switzerland, and is program leader of the Nano-Tera.ch program. Previously, he was Professor of Electrical Engineering at Stanford University, Stanford, CA, USA. His research interests include several aspects of design technologies for integrated circuits and systems, such as synthesis for emerging technologies, networks on chips and 3D integration. He is also interested in heterogeneous platform design including electrical components and biosensors, as well as in data processing of biomedical information. He is author of *Synthesis and Optimization of Digital Circuits* (McGraw-Hill, 1994), and co-author and/or co-editor of eight other books and of over 600 technical articles. His citation h-index is 83 according to Google Scholar.

Prof. De Micheli is a Fellow of ACM and a member of the Academia Europaea. He is a member of the Scientific Advisory Board of IMEC (Leuven, B), CfAED (Dresden, D) and STMicroelectronics. He was the recipient of the 2012 IEEE/CAS Mac Van Valkenburg award for contributions to theory, practice and experimentation in design methods and tools, and of the 2003 IEEE Emanuel Piore Award for contributions to computer-aided synthesis of digital systems. He also received the Golden Jubilee Medal for outstanding contributions to the IEEE CAS Society in 2000, the D. Pederson Award for the best paper in the IEEE TRANSACTIONS ON CAD/ICAS in 1987, and several Best Paper Awards including DAC (1983 and 1993), DATE (2005), and Nanoarch (2010 and 2012). He has served IEEE in several capacities, including Division 1 Director (2008–2009), co-founder and President Elect of the IEEE Council on EDA (2005–2007), President of the IEEE CAS Society (2003), and Editor-in-Chief of the IEEE TRANSACTIONS ON CAD/ICAS (1997–2001). He has been Chair of several conferences, including DATE (2010), pHealth (2006), VLSI SOC (2006), DAC (2000), and ICCD (1989).