A Voltage-Controlled, Oscillation-Based ADC Design for Computation-in-Memory Architectures Using Emerging ReRAMs

MAHTA MAYAHINIA and ABHAIRAJ SINGH, Delft University of Technology
CHRISTOPHER BENGEL and STEFAN WIEFELS, RWTH Aachen University
MUATH A. LEBDEH, Delft University of Technology
STEPHAN MENZEL, Forschungszentrum Juelich GmbH, Peter-Gruenberg Institut (PGI-7)
DIRK J. WOUTERS, RWTH Aachen University
ANTENEH GEBREGIORGIS and RAJENDRA BISHNOI, Delft University of Technology
RAJIV JOSHI, IBM Thomas J. Watson Research Center
SAID HAMDIOUI, Delft University of Technology

Conventional von Neumann architectures cannot successfully meet the demands of emerging computation and data-intensive applications. These shortcomings can be improved by embracing new architectural paradigms using emerging technologies. In particular, *Computation-In-Memory (CiM)* using emerging technologies such as *Resistive Random Access Memory (ReRAM)* is a promising approach to meet the computational demands of data-intensive applications such as neural networks and database queries. In CiM, computation is done in an analog manner; digitization of the results is costly in several aspects, such as area, energy, and performance, which hinders the potential of CiM. In this article, we propose an efficient Voltage-Controlled-Oscillator (VCO)–based analog-to-digital converter (ADC) design to improve the performance and energy efficiency of the CiM architecture. Due to its efficiency, the proposed ADC can be assigned in a per-column manner instead of sharing one ADC among multiple columns. This will boost the parallel execution and overall efficiency of the CiM crossbar array. The proposed ADC is evaluated using a Multiplication and Accumulation (MAC) operation implemented in ReRAM-based CiM crossbar arrays. Simulations results show that our proposed ADC can distinguish up to 32 levels within 10 ns while consuming less than 5.2 pJ of energy. In addition, our proposed ADC can tolerate ≈30% variability with a negligible impact on the performance of the ADC.

CCS Concepts: • **Computer systems organization → Embedded systems**; Redundancy; *Robotics*; • **Networks** → Network reliability;

This research work is supported and funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 780215 in the project MNEMOSENE and by the Federal Ministry of Education and Research (BMBF, Germany) under Grant Agreement No. 16ES1133K/16ES1134 in the project NEUROTEC.

Authors' Addresses: M. Mayahinia, A. Singh, M. A. Lebdeh, A. Gebregiorgis, R. Bishnoi, and S. Hamdioui, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, Netherlands; emails: {M.Mayahinia-1, A.Singh-5, M.F.M.AbuLebdeh, A.B.Gebregiorgis, R.K.Bishnoi, S.Hamdioui}@tudelft.nl; C. Bengel, S. Wiefels, and D. J. Wouters, RWTH Aachen University, Sommerfeldstraße 24, 52074, Aachen, Germany; emails: {bengel, wiefels, wouters}@iwe.rwth-aachen.de; S. Menzel, Forschungszentrum Juelich GmbH, Peter-Gruenberg Institut (PGI-7), 52428, Juelich, Germany; email: st.menzel@fz-juelich.de; R. Joshi, IBM Thomas J. Watson Research Center, 200 Aqueduct Rd, Ossining, 10562, New York city, USA; email: rvjoshi@us.ibm.com.

1550-4832/2022/03-ART32 $15.00
https://doi.org/10.1145/3451212

Additional Key Words and Phrases: Computation-in-Memory, ReRAM, analog computation, analog to digital converter, voltage-controlled oscillator

ACM Reference format:
Mahta Mayahinia, Abhairaj Singh, Christopher Bengel, Stefan Wiefels, Muath A. Lebdeh, Stephan Menzel, Dirk J. Wouters, Anteneh Gebregiorgis, Rajendra Bishnoi, Rajiv Joshi, and Said Hamdioui. 2022. A Voltage-Controlled, Oscillation-Based ADC Design for Computation-in-Memory Architectures Using Emerging ReRAMs. *J. Emerg. Technol. Comput. Syst.* 18, 2, Article 32 (March 2022), 25 pages.
https://doi.org/10.1145/3451212

1 INTRODUCTION

Existing complementary metal-oxide-semiconductor (CMOS)–based von Neumann architectures, in which memory and computing unit are separated, are facing various device-, circuit-, and architecture-level challenges. These conventional architectures are severely impacted by the slow speed of memory accesses, their limited parallelism, and the stagnation of the clock frequency due to thermal issues, which are well known as memory wall, instruction-level parallelism wall and power wall terms, respectively [1–3]. On the other hand, devices are also facing challenges related to reliability, high leakage, and excessive manufacturing cost [4, 5]. These challenges are more pronounced for data-intensive applications such as neuromorphic computing in which energy efficiency and data movement minimization are of paramount importance [6]. For such application domains, the need to identify viable alternatives has gained growing attention. In this regard, with non-volatility, zero-leakage, and high-density properties, the resistive memory-based Computation-in-Memory (CiM) crossbar structure is a potential candidate to replace traditional architectures and accelerate data-intensive applications [7, 8].

CiM architectures can significantly improve both energy efficiency and performance of computing systems by performing the operation within the memory, which avoids expensive data movement [9–11]. To enable CiM, the memory module must support operations in addition to its functionality as data storage [12, 13]. CiM can be realized using different memory technologies, such as Dynamic Random Access Memory (DRAM) [14] and Static Random Access Memory (SRAM) [15], as well as emerging resistive-memory technologies such as Resistive Random Access Memory (ReRAM), Phase Change Memory (PCM) and Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM), in which data are represented by resistance states [16]. Since these memory technologies have their own specific properties, they can be used for specific CiM operations [17]. The physical attributes of these resistive memories make them inherently suitable for Multiply-and-Accumulate (MAC) operations in CiM architectures. For example, matrix multiplication (vector-matrix multiplication) using MACs is a frequent operation in several Artificial Intelligence (AI) and signal processing applications. This operation can be accelerated by CiM architecture organized in a crossbar structure, which is performed on the physical level via Ohm's Law for the multiplication and Kirchhoff's Law for the accumulation [18].

Since MAC acceleration in a crossbar structure is fundamentally an *analog* operation, interfaces for connecting digital and analog parts are required and represent a critical part of the design. Digital-to-Analog Converters (DACs) and Analog-to-Digital Converters (ADCs) are crucial components to handle the CiM inputs and outputs, respectively. Due to the complex process involved in analog-to-digital conversion and vice versa, DACs and ADCs usually occupy significantly large areas and consume significant power when compared with the total area of the crossbar array. Moreover, the ADC phase is considered to be the bottleneck in terms of performance and energy efficiency of CiM architectures [19, 20]. Hence, efficient ADC design is extremely important in order to improve the overall performance and energy consumption of CiM architectures.

Several efforts have been made to improve the efficiency of the ADC phase, either through different circuit methodology or different system architectures. For example, the authors of [19–22] proposed a shared ADC scheme in which the ADC is shared between *multiple* columns in order to improve area and energy efficiency. In their design, the number of columns that share the same ADC depends on both architecture and the ADC features. The authors of [23, 24], on the other hand, suggested a dedicated ADC design, in which one ADC interface is assigned per column to avoid time-multiplexing schemes to access a shared ADC. In order to amortize the design overheads, the authors of [25] designed for merging of ADC phases and activation computation phases. Due to the trade-off between different features of the ADC — such as resolution, power, latency, and area — the previous ADC designs are usually optimized for a specific feature. Moreover, the ADC designs in [19, 21–24] do not consider variability and reliability aspects, which have a severe impact on the resolution as well as energy efficiency of the ADC. Therefore, an efficient ADC design addressing the impact of variability and reliability issues is of decisive importance in order to improve the performance, area, and energy efficiency of CiM architectures.

In this article, we propose an efficient Voltage-Controlled-Oscillator (VCO)–based ADC design for ReRAM-based CiM crossbar arrays to alleviate the ADC phase bottleneck of analog computation. In our proposed ADC design, the bit-line current as an analog signal coming from the crossbar is first transformed into a voltage. Then, this voltage is transformed into a frequency with the help of the VCO, which is realized using a ring oscillator. Subsequently, the generated pulses are counted with a counter. The output of the counter must be unique for different input combinations. We also theoretically investigate different features of the proposed VCO-based ADC, such as its variability tolerance. To ensure the compatibility of the ADC and the ReRAM devices, we perform real device-level measurements and programming methods. Our contributions in this article are as follows:

- Design of a new VCO-based ADC using a ring-oscillator circuit for a crossbar array of resistive memories.
- Theoretical evaluation of reliability and variation tolerance of the proposed ADC design.
- Demonstration of the influence of variation based on real fabricated and characterized cells.
- Comprehensive comparison done with the state-of-the-art ADC designs.

Simulation results show that our proposed ADC can distinguish up to 32 levels within 10 ns while consuming less than 5.2 pJ of energy. In addition, our proposed ADC can tolerate $\approx 30\%$ variability with a negligible impact on the performance of the ADC. The voltage across the ReRAM devices does not exceed 0.3 V during operation. Thus, the data stored in the resistive memories will not suffer from read-disturb.

The rest of the article is organized as follows. Section 2 presents basic information about the CiM architecture and related work, followed by the details of the proposed VCO-based ADC design and its circuit-level theoretical analysis in Section 3. Section 4 presents the device and circuit level results and introduces a *Figure of Merit (FoM)* for further comparison with state-of-the-art designs. Section 5 contains our conclusions.

2 BACKGROUND

2.1 Computation-in-Memory

For data-intensive applications, the demand for high performance and energy efficiency is increasingly growing and existing architectures are incapable of fulfilling these requirements [26, 27]. *Computation-in-Memory (CiM)* architectures can significantly improve the performance and energy efficiency of data- and computation-intensive applications by performing operation and storage at the same physical location. Hence, CiM architectures remove the overheads associated with the costly data movements performed by the conventional von Neumann architectures by performing computing and storage at the same physical location [28].

Resistive memories — such as PCM [29], STT-MRAM [30], and ReRAM [31] — are promising alternatives to conventional memories such as DRAM or SRAM due to their attractive features, such as high density, almost zero leakage, and non-volatility. Also, due to their resistive characteristics, they are suitable for CiM architectures. In addition, various data-intensive applications and operations, such as MAC, can benefit from the computation in the crossbar organization of the resistive memories. Figure 1 shows the general architecture of the CiM unit (accelerator), built using a crossbar array of resistive memories. As shown by the digital input arrow in Figure 1, two sets of data are provided as an input to the CiM module. The first input set is stored in the resistive cells, which will be represented as the cell's conductance. This write operation is performed by using write buffer and driver circuits. The other input set is provided by applying input voltages to the crossbar. In this case, the decoder module selects the address for the correct location where the input voltage is provided. As the nature of the computation in the crossbar is analog, the digital data must first be converted into an analog signal. Converting the digital input to the corresponding analog voltage is done through DAC modules and drivers. Similarly, the analog output signal of the crossbar needs to be converted into a digital signal again using ADCs, as the CiM accelerator module is a part of the digital host unit. These stages are controlled via a control unit. Figure 1 also shows how a resistive memory–based crossbar array can efficiently perform MAC operations in an analog manner by using Ohm's Law and Kirchhoff's Current Law. According to Ohm's Law, current is the product of voltage (V; coming from the DAC modules) and conductance (G; these conductances are usually adjustable according to the training procedure [32]). According to Kirchhoff's Current law, the bit-line current $I_{\text{bit-line}}$ is equal to the current summation of all of the resistive cells (see Equation (1)) as follows:

$$I_{\text{bit-line}} = [V_1, V_2, V_3] \times [G_1, G_2, G_3]^T = V_1 \times G_1 + V_2 \times G_2 + V_3 \times G_3. \tag{1}$$

2.1.1 ReRAM Technology.

Resistive switching devices have recently gained widespread attention due to their non-volatility, high integration density and their ability to overcome memory bandwidth issues by executing operations within the memory [1]. These properties make them attractive for various applications ranging from non-volatile memory [33], logic based around computation in memory [34], and neuromorphic computing [35]. Among the various restively switching technologies, bipolar resistive switching cells based on the Valence Change Mechanism (VCM) show great promise for various kinds of applications due to their properties such as non-linearity, multilevel capability, and stochastic switching behavior. VCM devices consist of a metal/mixed ion-electron conductor/metal structure. VCM switching has been recognized in devices based on various oxide materials such as HfO_x, TaO_x, ZrO_x, STO, or TiO_x [16, 36]. Oxide materials such as HfO_x are already compatible with conventional CMOS processes. If they are fabricated in a crossbar structure, they offer a high density of $4F^2$, where F is the minimum feature size of the process technology. It should be noted that the original memristor publication by HP, which was inspired by the memristor circuit theory of Leon Chua, also describes a VCM cell based on TiO_x [37]. In a VCM cell, the two metal electrodes possess different work functions towards the oxide layer, as shown in Figure 2(a). The electrode with the higher work and lower work function forms a Schottky-type and Ohmic type contact with the oxide, respectively. The Schottky-type electrode is then denoted as Active Electrode (AE) while the Ohmic type electrode is called Ohmic Electrode (OE) [38]. Directly after fabrication, the oxide is insulating, which makes a forming step necessary. During forming, a high voltage is applied, which locally reduces the oxide layer. This generates

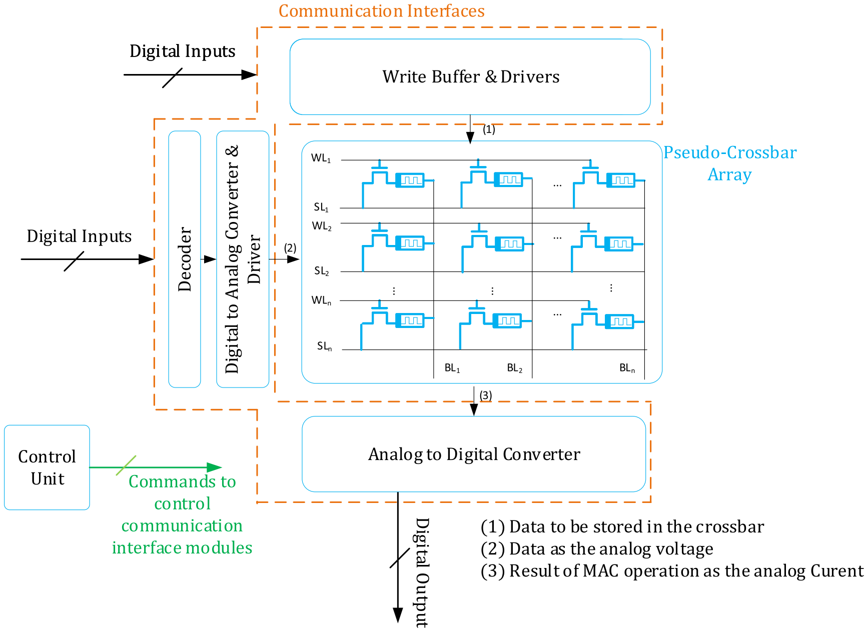

Communication Interfaces
Digital Inputs
Write Buffer & Drivers
(1)
Pseudo-Crossbar Array
Digital Inputs
Decoder
Digital to Analog Converter & Driver
(2)
WL$_1$
SL$_1$
WL$_2$
SL$_2$
WL$_n$
SL$_n$
BL$_1$
BL$_2$
BL$_n$
(3)
Control Unit
Commands to control communication interface modules
Analog to Digital Converter
Digital Output
(1) Data to be stored in the crossbar
(2) Data as the analog voltage
(3) Result of MAC operation as the analog Curent

Fig. 1. General structure of computation-in-memory architecture.

positively charged oxygen vacancies and reduces the resistance of the device. In filamentary switching systems, this reduction is confined to a small portion of the total cell area. During the SET operation, the concentration of oxygen vacancies increases in the vicinity of the AE, which reduces the resistance of the device. In this condition, the ReRAM cell is in the Low Resistance State (LRS). The SET occurs when a negative voltage is applied to the AE. The RESET occurs for a positive voltage applied to the AE, which repels oxygen vacancies from the AE and causes the ReRAM cell to be in the High Resistance State (HRS) [39]. Figure 2(b) shows these two states via an *I-V* curve as well and Figure 2(c) shows the ReRAM 1T1R bit-cell structure. The 1T1R structure is mainly required to program and verify read of the individual conductance values in each cell. During the read of individual cells, *sneak current paths* may cause errors [40] in which unwanted leaking current in the crossbar structure leads to deviation of results, especially in the analog computation. Besides solving the problem of sneak paths, 1T1R structures enable analog programming for the memristors with low standard deviation [41], since this enables accurate measurement of the programmed values. There are different ways to organise 1T1R bit-cells into 1T1R arrays, such as the 1T1R memory array or the Pseudo-Crossbar array structure [42]. In this article, we focus on the Pseudo-Crossbar array since it was specifically developed to enable MAC computation operation.

One of the biggest challenges for the usage of VCM-based ReRAM cells is their variability, which stems from their stochastic and atomistic switching behavior [43] and which is observed, for example, in the variability of the HRS and LRS [44, 45]. In the development of circuits and architectures, this variability has to be addressed as per the application requirements. The resistance distribution of the LRS usually follows a normal distribution while the HRS distribution follows a log-normal distribution [46]. Additionally, during read operation, it has to be ensured that the voltage dropping across the ReRAM device is not too high in order to prevent read disturb, which is the changing of the device state due to consecutive read operations. It has been shown that low voltages will

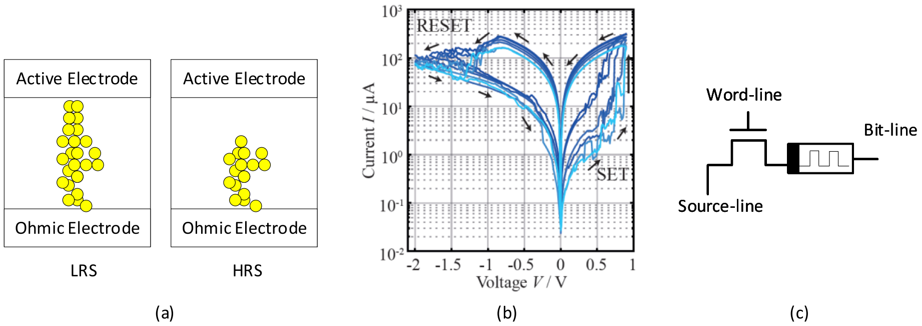


Fig. 2. (a) shows the ReRAM device structure for LRS and HRS, (b) shows exemplary *I-V* characteristics of the fabricated ReRAM devices measured with a sweep rate of $1\frac{V}{sec}$, and (c) shows a 1T1R ReRAM bit-cell.

also lead to a switching of the devices, albeit on a much longer time scale [47]. While read disturb cannot completely be eliminated as long as there is a voltage drop across the ReRAM cells during read, having this voltage small for a few nanoseconds per read means that read disturb will happen only after billions of read cycles.

2.1.2 Analog-to-Digital Converter (ADC). ADCs are generally used whenever an analog signal is needed as an input for digital modules. The CiM accelerator unit implemented with the ReRAM-based crossbar structure performs the operation in an analog manner. An ADC is required at the output of the crossbar to convert the analog result into the digital form, which can then be used by the digital host to which the CiM accelerator belongs. However, ADCs typically occupy a large area and consume significant energy compared with other components of CiM architectures [48]. Moreover, the major performance bottleneck of CiM architecture comes from the ADC phase. Thus, it is increasingly important to design highly efficient and compact ADCs for CiM operations.

ADCs mainly have two processing steps: (1) Sampling and Holding (S/H) and (2) Quantization and Encoding. The sampling frequency in the S/H step needs to meet the Nyquist rate and needs to amount to at least twice the highest data frequency. In the quantization phase, the reference signal is partitioned into a number of *quantas*. Then, the analog input is matched with one of these quantas in the encoding phase. A unique digital code will be assigned to the input reflecting which quanta it belongs to. Different designs of an ADC indeed have different ways for the quantization and encoding steps.

2.2 Related Works

Various previous works took different approaches to alleviate the ADC bottleneck. Each has its own pros and cons as well as common shortcomings. The authors of [21] proposed a shared ADC design in which an ADC is shared between multiple columns to tackle the bottleneck of the ADC phase. The authors used a Successive Approximation Register (SAR) ADC as their shared ADC and various columns have to access the ADC in a time-multiplexed scheme. The SAR-based ADC compares the output of the DAC module with the original analog signal and stores the result of the comparison in an SAR. Due to this closed-loop comparison, the digital output gets increasingly closer to the analog value. The advantage of the SAR-based ADC is that it is fast, has a high resolution, and it can produce all of the output bits in one activation cycle. Due to its large area and high power consumption, however, it has to be shared between multiple columns. The authors of [22] have selected a faster type of ADC: FLASH ADC. This ADC still needs to be shared between multiple columns. The Flash-type ADC is another design for an ADC in which the comparison of the analog input with all different reference signals is happening simultaneously. This makes the

Table 1. Summary of the Related Work

Type of ADC phase	Area	Power	Resolution	Speed	ADC assignment
SAR ADC [20, 21]	–	–	++	+	Shared
FLASH ADC [22]	-	-	+	++	Shared
IF [23, 25]	++	++	–	–	Dedicated
SA [24]	++	++	–	–	Dedicated
TDC [19]	–	–	+	+	Shared

Flash-type ADC very fast and also expensive in terms of power and area, as multiple comparators are required that all must work in parallel. However, the authors of [49] have proposed a memristor-based FLASH ADC with higher density than conventional FLASH ADCs, but a FLASH ADC still has a large area and is only useful for low resolutions [50]. Due to the low resolution of the FLASH ADC, it cannot produce all of the output bits in one activation cycle.

The authors of [23] and [24] have adopted a different approach: leveraging small-area and low-power ADC modules, to be assigned one per column. Both Integrate and Fire (IF) and Sense Amplifier (SA) ADC modules are slow and low resolution. To produce an n-bit output, these ADCs need a 2^n activation cycle; hence, the time complexity of these approaches is exponential $O(2^n)$. The authors of [23] have used an IF ADC module in which a capacitor is charging via the bit-line current. The voltage of this capacitor is accumulative and is compared with a predefined threshold voltage. When the voltage of the capacitor reaches the threshold voltage, spikes will be generated. Then, these spikes are counted by a counter. The number of pulses determines the equivalent digital value of the analog result of the CiM operation. The authors of [25] also used an IF as their ADC, but they have tried to hide the overhead of the ADC phase by merging it with the activation computation phase, which is required in convolutional or fully connected neural networks. On the other hand, the authors of [24] used an SA. This interface is designed and specified for memory-read operations. By modifying this module, it can also be used as an ADC interface. *Reuse* ability is the remarkable advantage of the SA. The required interface for ADCs in the area of computation in the ReRAM crossbar structure can be obtained through modifying an already designed SA with considerably less effort.

Architectural-level solutions for the ADC phase bottleneck have also been considered in [19, 20, 25]. For instance, the authors of [20] have used a shared ADC as well. To decrease the number of ADCs and their activity, they have added an analog dataflow to their architecture. In this method, analog partial summations for large kernels (which need to be mapped on multiple crossbars) can be substituted for the digital partial summations without requiring costly ADC phases. With this technique, they have been able to decrease the number of ADCs and their cycles. The authors of [19] have also used shared ADC interfaces and the idea of analog buffers. In contrast, the ADC interface in [19] is a Time-Digital Converter (TDC). Due to the low supply voltage and technology scaling, designing an ADC in the voltage domain has become more difficult. Time-based signals, on the other hand, can improve the resolution of the ADC with these restrictions [51]. To benefit from the properties of the time-based signals, the authors of [19] used a TDC and Digital-Time Converter (DTC) instead of an ADC and DAC, respectively. In this approach, signals are distinguishable from each other by a *delay* from an initial point. The TDC mechanism can be implemented in a fully digital flow [52–57] and not only has lower power consumption but is also less vulnerable to variations in noise, fabrication processes, voltage, and temperature [58]. Table 1 summarizes the advantages and disadvantages of the discussed related works.

Low performance is the common shortcoming between two general approaches of *shared* and *dedicated* ADC. The main factors for low performance are a time multiplexing scheme in the first

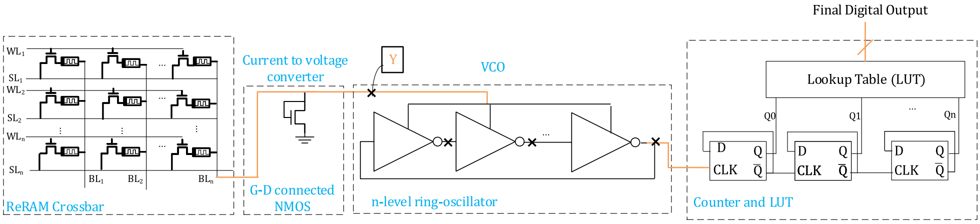


Fig. 3. Detailed schematic of the proposed VCO-based ADC.

approach and low resolution in the second approach. In addition to the known features of the ADC — namely, resolution, power, latency, and area — specific characteristics of the resistive memories, such as variability in the resistive states and restrictions on the maximum voltage across the device, must be taken into account in the ADC design. This motivates the need for a specific ADC design that fulfills the application requirements. In this article, we propose a VCO-based ADC. The VCO-based ADC is a type of time-based ADC (unlike SAR and FLASH ADCs) that produces signals with a unique frequency proportional to the analog signal [59]. In addition to the already mentioned properties of the time-based signals, the design of the VCO-based ADC is simpler since it does not need high-performance analog blocks such as amplifiers and DACs. There are several VCO-based ADCs available, such as those described in [60–62]. However, none of them is applicable for CiM due to their large area and high power consumption. Although the proposed VCO-based ADC does not offer high resolution, it is so small that it allows us to assign one single ADC to every single column. The VCO-based ADC also satisfies ReRAM device requirements, such as variability in their resistive states and the restriction on the maximum voltage across the device.

3 PROPOSED VCO-BASED ADC DESIGN

3.1 Design Implementation

In our design, we assume a restricted class of MAC operation in which input voltages and weights (the conductance of the ReRAM devices) can acquire only two distinct values. To convert the analog output signal of the crossbar to a digital signal, we decided to use a VCO-based ADC, which is time based. Therefore, it can provide the advantages of the time-based signals with a relatively easy design procedure. In the ADC phase period and in order to transfer an analog current into the digital signal with the help of VCO-based ADC, three stages are required. In the first stage, the analog bit-line current needs to be transformed into an analog voltage. In the next stage, the obtained analog voltage is transformed into pulses with the help of the VCO. In the last stage, the generated pulses are counted with a counter and mapped to the corresponding digital signal with the help of a Lookup Table (LUT). We modulate the power supply directly by regulating the read voltage (V_{read}) applied to the crossbar (as row voltages). Therefore, deactivating the row V_{read} disables the crossbar as well as the ADC. The output of this stage is the equivalent digital signal and can be processed by the digital host. The schematic of the whole system, including the ReRAM crossbar and the VCO-based ADC, is shown in Figure 3.

3.1.1 Linking Crossbar and ADC Using Transfer Functions. A transfer function is a mathematical function that describes the output of a system for each possible input. In our case, we will consider two systems: the 1T1R crossbar and the VCO-based ADC. For the crossbar, the input will be the specific resistance configuration of the cells that we read out, which is given as R_{eq} of the crossbar. We will consider the case in which all of the rows in a column are selected. R_{eq} is formed by the parallel connection of multiple series connections of ReRAM devices and access transistors (see

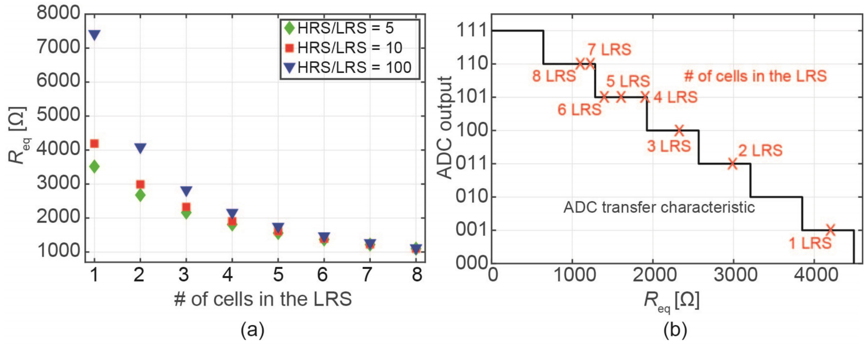


Fig. 4. (a) Equivalent resistance of the crossbar as a function of the number of cells in the LRS for different HRS/LRS ratios. (b) The black line denotes the ADC output as a function of R_{eq}. The red crosses display the R_{eq} values of Figure 4 for an HRS/LRS ratio of 10.

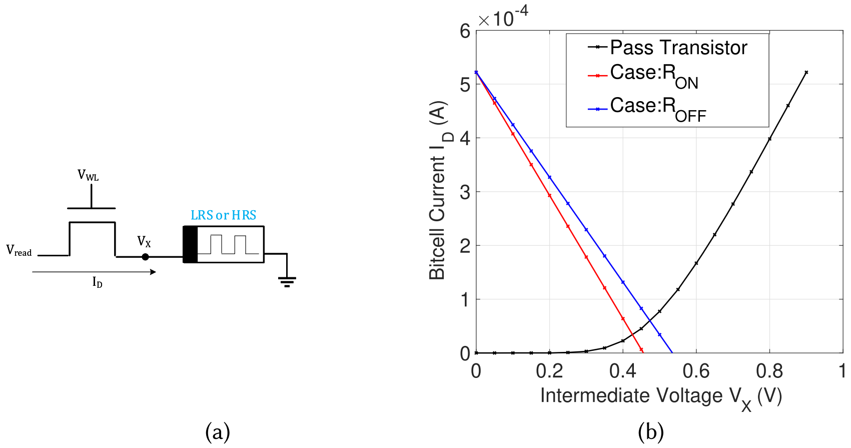


Fig. 5. Load-line characteristics of 1T1R bit-cell. (a) The schematic configuration for the ON and OFF state of the ReRAM device. Node V_X is shared between the pass transistor and ReRAM device. (b) The plot between bit-cell current (I_D) and intermediate voltage (V_X).

Figure 1). The equivalent resistance of this parallel connection can then be calculated using

$$R_{\mathrm{eq}} = \left(\frac{\#\text{ of cells in LRS}}{R_{\mathrm{LRS}} + R_{\mathrm{transistor,\,LRS}}} + \frac{n - \#\text{ of cells in LRS}}{R_{\mathrm{HRS}} + R_{\mathrm{transistor,\,HRS}}} \right)^{-1}, \tag{2}$$

where R_{LRS} and R_{HRS} denote the LRS and HRS resistance, respectively. $R_{\mathrm{transistor,HRS/LRS}}$ denotes the drain-source resistance of the transistor connected to a ReRAM cell in the LRS or HRS during readout, and n denotes the number of cells that are read in parallel with one ADC. The resulting transfer function of the crossbar can be seen in Figure 4(a). It follows from Equation (2) when the number of cells in the LRS is changed from one to eight.

In this case, we assume $R_{\mathrm{transistor,\,HRS}}$ to be 26 kΩ, $R_{\mathrm{transistor,\,LRS}}$ to be 5.8 kΩ, the LRS as 3 kΩ, and the HRS was varied from 15 kΩ to 300 kΩ to achieve various HRS/LRS ratios. The transistors were operating in the saturation region when connected to an LRS or HRS device. Figure 5(b) shows the load line characteristic of a 1T1R bit-cell with a LRS or HRS ReRAM cell. From this, it is obvious that the operating point of the transistor and its resistance will be different depending on the resistive state of the ReRAM cell.

For better clarity, a maximum of only 8 cells are considered to be read out at the same time. From this plot, it can be seen that a larger (smaller) HRS/LRS ratio leads to a less (more) linear relationship. This is because, for high HRS/LRS ratios, the equivalent resistance is almost exclusively determined by the number of cells in the LRS. The difference in the R_{eq} values decreases strongly for higher numbers of devices in the LRS, which increases the requirements on the ADC performance. It becomes more difficult to distinguish between different levels if more cells in the LRS state are connected. In addition, a smaller HRS/LRS ratio will lead to other issues because of an increased influence of Read Noise and Read Disturb. When the ratio between HRS state and LRS is decreased, the devices will be more susceptible to random variations and stress due to prolonged reading. If a sufficient number of cells are in the LRS, the equivalent resistance is very similar, independent of the HRS/LRS ratio. This means that the transfer function of the crossbar cannot be improved by optimizing the devices and has to be addressed by the ADC.

For an ADC, the transfer function displays its digital output value as a function of an analog input signal, usually the input voltage [63]. In our case, it is more useful to display the ADC transfer characteristic as a function of the resistances of the crossbar. Usually, in the design of ADCs, linearity of the input–output relationship is preferred. This means that the input levels corresponding to one ADC output have equal widths. Such a transfer function can be seen in Figure 4(b), illustrated as a black line. To determine whether a linear ADC characteristic is a reasonable choice for CiM using resistive devices, we combined the transfer functions of the crossbar with the transfer function of the ADC. This is possible, as the output of the crossbar transfer function is the same as the input of the ADC transfer function. The red crosses in Figure 4(b) are obtained using the R_{eq} values of the crossbar transfer function as presented in Figure 4(a) and an HRS/LRS ratio of 10. It should be noted here that the effective HRS/LRS ratio, when considering the serially connected transistors, is reduced from 10 to 6.36. The intersections between the ADC transfer characteristic (black line) and red crosses show to which ADC output the crossbar input is mapped. From Figure 4(b), it can be seen that the data in the crossbar does not map very well to the linear ADC characteristic since most of the possible outputs of the crossbar are mapped to the same ADC output (here, '101'). This is due to the fact that the R_{eq} for these crossbar outputs have very small differences between each other. This shows that the usually linear ADC transfer function is not an optimal solution for CiMs based on resistive devices. In summary, we can say that an ADC with a nonlinear transfer function should be better suited for CiM applications. The VCO-based ADC that is considered here has a strong nonlinear transfer characteristic, which makes it a suitable ADC candidate for CiM based on resistive memories. Yan et al. discussed a related issue. In [64], they compared different spacings of the resistive states (equal ΔR vs. equal ΔG) for a neural network using analog ReRAM devices. Their conclusion was that while both mappings delivered a comparable accuracy performance, the equal ΔR mapping was beneficial as it resulted in less severe constraints for the ADC. This result suggests that a matching between crossbar and ADC can be achieved on multiple levels, either in the ADC or in the spacing of the resistance values. Since we use the ReRAM devices in a binary way, different spacings are equivalent to choosing different $R_{\text{off}}/R_{\text{on}}$ ratios, which, as we showed in Figure 5(a), improves the linearity of the crossbars transfer function. However, it leads to other problems such as an increased susceptibility to device variability.

3.1.2 Current-to-Voltage Converter. In the VCO-based ADC, the bit-line current that contains the result of the MAC operation needs to be first transformed into a voltage. This analog voltage can be used as the input of the VCO element. For this purpose, four different methods can be selected and we have to explore all four possible connection options: capacitance, gate-drain connected NMOS (diode-connected structure), constant resistance, and none.

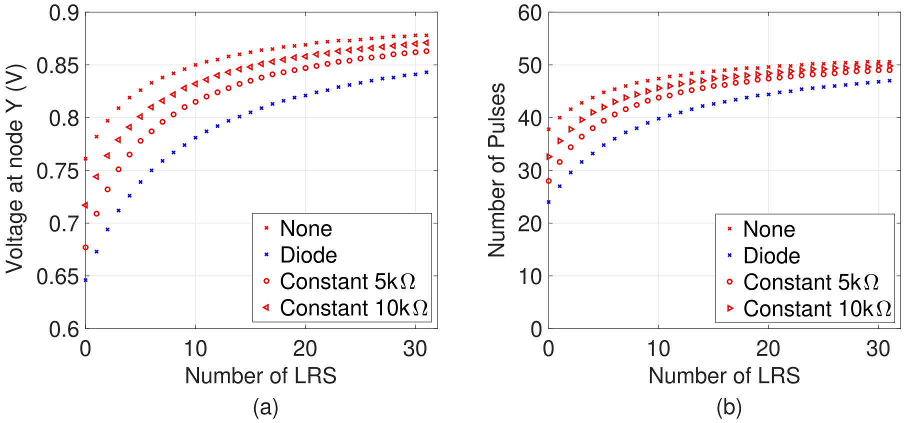


Fig. 6. (a) Voltage developed at node Y in Figure 3 and (b) resolution of the VCO-based ADC with different method ("None": no extra device; diode: diode-connected structure and 5 kΩ; 10 kΩ: two different values of constant resistances).

- *Capacitance:* In this case, the voltage across this element is exponentially reaching the V_{read} value with a time constant that is proportional to the ReRAM states. The frequency of the VCO's output is time dependent. After a specific time, it will be almost equal for all of the different ReRAM states. Thus, having a capacitor is not a wise converting method for this purpose.
- *G-D connected NMOS (diode-connected structure):* In this case, the gate and drain voltage is $\sqrt{\frac{2 \cdot I_{\text{bit-line}}}{\mu_n \cdot C_{OX} \cdot \frac{W}{L}}} + V_{\text{th}}$ (μ_n is the electron mobility, C_{OX} is the capacitance per area of the gate oxide, W and L are width and length of the NMOS, respectively, and V_{th} is the threshold voltage).
- *Constant resistance:* In this case, the voltage across the resistance is $R_s \cdot I_{\text{bit-line}}$.
- *None:* Relying on the input impedance of the VCO for the converting current into the voltage.

To select among these four methods, we use the following *resolution* criterion: a greater number of LRS for which the voltage at node Y in Figure 3 and the number of generated pulses become flattened indicates a higher resolution. As Figure 6 shows, using a diode-connected structure results in a better resolution compared with the other methods.

It is worth mentioning here that the parasitic bit-line capacitor is not necessarily required for the correct functionality of the circuit, but it exists due to the crossbar wires and junctions. This parasitic capacitor impacts different features of the circuit, as discussed in Section 3.2.

3.1.3 Voltage-Controlled Oscillator. The *VCO* is an abstract 2-terminal module that gets a DC *voltage* as the input and produces a periodic signal with a frequency F as the output. The frequency of the output signal is a function of the DC input voltage. To realize the VCO, we use a *ring-oscillator.* A ring-oscillator consists of n inverter gates in a loop (n: odd and greater than 1). The output of the last inverter is connected to the input of the first inverter. In Figure 3, by connecting input voltage to the bit-lines of the crossbar, all of the nodes of the circuit, including the node Y (which are marked with X), start to oscillate with the same frequency but different phases. By changing the bias voltage of the ring-oscillator node Y, the frequency of the oscillation is changing since the ring-oscillator can be considered to be a VCO. Simulations show that the change in the frequency (f) of the oscillation has a linear relation with the change in the bias voltage (V_{bias}): $\Delta f = K \cdot \Delta V_{\text{bias}}$. The constant K can be adjusted based on the number and size of the inverter's transistors. For different resistive states of the crossbar, the bias voltage will be different, but their voltage difference could be very small. These close voltage levels need to be transferred into different frequencies; thus, choosing a larger K value helps in improving the resolution. Moreover, the value of K has an impact on other features of the circuit, for instance, its ability to tolerate the variability in the resistive memories, which will be discussed in Section 3.2.

3.1.4 Counter and Lookup Table. The counter is also a crucial module in the proposed VCO-based ADC design. The bit-line current is eventually transformed into the number of pulses. To digitize these generated pulses, it can be counted in the certain period. As discussed earlier, high speed of the ring-oscillator (high K) is necessary to have an acceptable resolution for the ADC. The oscillation frequency, however, is limited by the speed of the counter, that is, how fast a counter can count. In sequential circuits as well as counters, two timing parameters should be considered to determine the speed of these circuits: data-path length and setup time. Data-path length is the delay of the circuit's critical path and setup time is the amount of time that the data needs to be stable before the active edge of the clock. To avoid timing errors in the counter module, the period of the clock (the signal produced by the ring-oscillator) must be greater than or equal to the sum of the data-path delay and setup time. In the fixed resolution, the speed of the ring-oscillator must be adjusted by considering the counting frequency of the counter. Therefore, in the fixed resolution, latency is determined only by the counter module. We have selected an asynchronous ripple counter as the counter module in our design because of its area efficiency. An *N-bit* ripple-counter can count 2^N states, which is the maximum number of countable states among counter structures. Moreover, the ripple-counter does not need any logic between its flip-flops for the correct functionality. Finally, the output of the counter must be mapped to a proper digital representation. This correspondence is done through the LUT.

3.1.5 Self-Timing Path (STP). The premise of the Self-Timing Path (STP) is to calculate the time a column with known resistance states takes to produce a known output and then use that time interval as the total time (T_{total}) of the required operation. To implement this technique, a dummy column is used (we take all resistive values as LRS in this dummy column) in the same crossbar array. As soon as the output number of pulses for this dummy column reaches the applied number of non-zero row voltages, it triggers the other columns to stop counting. This dummy column thus provides a variation-aware T_{total}. The acquired T_{total} is adaptive to global variations of CMOS and ReRAM devices, temperature, and fluctuations in the peripheral/core voltage supplies [65].

3.2 Theoretical Analysis

In this subsection, we theoretically investigate the impact of each element on the whole circuit.

- Access Transistor: As mentioned in Section 2, 1T1R bit-cells are beneficial in both removing the sneak paths and improving the writing process. We have chosen an NMOS access transistor located between the source-line and the ReRAM device. Different type (PMOS) or different location (between the ReRAM device and bit-line) has an impact on the features of the ADC since the resistance of the access transistor varies in these conditions. Further discussion on this matter is out of the scope of this article.
- VCO-Impedance: As discussed earlier, we have realized the VCO part with a ring-oscillator. During the oscillation, the gate voltage of the inverters is also changing. This change in the gate voltage results in varying the resistance of the inverters and the whole ring-oscillator. The oscillating nature of the ring-oscillator cannot be emulated by the mere resistance. A capacitive element, for instance, is also required for the correct oscillation modeling. Thus, we safely substitute a *varying impedance* for the ring-oscillator. Because of the oscillating nature in the output of the crossbar array (node Y in Figure 3), the source voltage of the

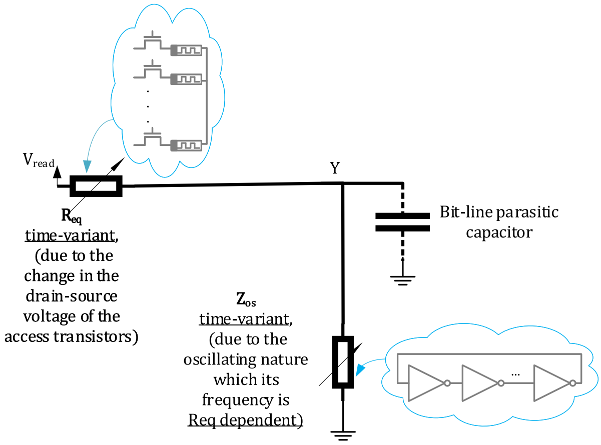

V_{read}
Y
Bit-line parasitic capacitor
R_{eq}
time-variant, (due to the change in the drain-source voltage of the access transistors)
Z_{os}
time-variant, (due to the oscillating nature which its frequency is Req dependent)
...

Fig. 7. Illustration of equivalent circuit for the proposed VCO-based ADC design, substitution of time-variant R_{eq} for the crossbar, and time-variant and R_{eq}-dependent impedance for the rest of the ring-oscillator.

access transistors is also changing, which ends up having time-variant equivalent resistance coming from the crossbar. To clarify the condition, the resistive state of the ReRAM devices is *data* dependent and determined independently (e.g., during the training phase of neural networks). Various data generate different time-variant equivalent resistance and different time-variant ring-oscillator impedance. Figure 7 shows a simplified schematic of the VCO-based ADC by substituting a time-variant R_{eq} for the crossbar and an R_{eq}-dependent and time-variant impedance (Z_{os}) for the ring-oscillator.

By considering Figure 7, we could theoretically evaluate the features of the VCO-based ADC, such as resolution, variability tolerance, power, energy, and voltage across the ReRAM devices.

Resolution: As Figure 4(a) also shows, by increasing the number of ReRAM devices in parallel, the difference between different resistive states is narrowing and they will be more difficult to distinguish. In the case of having N *parallel* ReRAM, the two resistive states of $\frac{R_{\text{LRS}}}{N}$ and $\frac{R_{\text{LRS}}}{N-1}||R_{\text{HRS}}$ has the minimum resistance difference. The resolution of the ADC is specified by these two closest resistive states. The ADC can distinguish between N levels as far as it is able to produce different outputs for these two different states. The ring-oscillator can generate a distinct number of pulses for these states if its bias voltage (node Y in Figure 3) has different corresponding values. The voltage at node Y is determined by voltage division between R_{eq} and Z_{os} and the voltage difference between these two resistive states is obtained via

$$V_{y_1} - V_{y_2} = V_{\text{read}} \cdot \left(\left(\frac{Z_{\text{os}_1}}{\frac{R_{\text{LRS}}}{N} + Z_{\text{os}_1}} \right) - \left(\frac{Z_{\text{os}_2}}{\frac{R_{\text{LRS}} \cdot R_{\text{HRS}}}{R_{\text{LRS}} + R_{\text{HRS}} \cdot (N-1)} + Z_{\text{os}_2}} \right) \right). \tag{3}$$

We must highlight that the bias voltage of the ring-oscillator (node Y) is also oscillating but to benefit from the abstract model of the VCO. We are considering the effective value (DC) of the voltage at node Y. In DC analysis, capacitors are behaving as the open circuit; thus, Z_{os} has only a resistive nature. As Equation (3) shows, by increasing the number of ReRAMs (N) the first terms in the denominators become negligible and the voltage difference between these two states approaches zero. As with the smaller $Z_{\text{os}_{1,2}}$ terms, N has a higher limit to become before the first terms, including N in denominator of Equation (3), to become negligible in comparison with $Z_{\text{os}_{1,2}}$. As discussed before, the bit-line current needs to convert into the voltage at the input of the VCO via different *impedance* elements. In Equation (3), the impedance of such a converter is in parallel to $Z_{\mathrm{os}_{1,2}}$ and makes these terms smaller. This results in a higher number of distinguishable levels. Thus, having a current-to-voltage converter is generally beneficial for the resolution. However, even without any converter and by relying on the impedance of the ring-oscillator, the bit-line current can transform into the voltage. Now, let us investigate the impact of both constant resistor and diode-connected element on the resolution. The diode-connected NMOS element can improve the resolution more than the constant resistor. It can be considered as a *voltage-controlled resistance* and its value is equal to $\frac{2}{\mu_n \cdot C_{\mathrm{OX}} \cdot \frac{W}{L}} \cdot \frac{V_{\mathrm{GS}}}{(V_{\mathrm{GS}}-V_{\mathrm{th}})^2}$ (V_{GS} is the gate-source voltage of the NMOS). By increasing N, V_{GS} gets larger and the resistivity of the diode-connected NMOS decreases. Hence, the terms $Z_{\mathrm{os}_{1,2}}$ get smaller, resulting in a higher resolution. It is worth mentioning here that the nominal values of the LRS and HRS also have an impact on the resolution. For larger values of the LRS and HRS, N could be larger before the two closest resistive states become indistinguishable.

Variability tolerance: Due to the stochastic nature of the switching in the ReRAM devices, the HRS and LRS resistance states are not fixed values; rather, they have a range. As a result, it is desirable for the ADC to tolerate the variability in the ReRAM resistive states. The proposed VCO-based ADC can tolerate this variability. To theoretically investigate this variability tolerance, first, we must emphasize that in contrast to the DC analysis, the capacitors also influence the resolution. The impedance of the ring-oscillator has a resistive and capacitive nature. In addition, the parasitic bit-line has an impedance equal to $\frac{1}{2\pi f C}$, in which f is the voltage frequency and C is the capacitance of the bit-line capacitor.

To theoretically evaluate the variability (in the resistive memories) tolerance of the VCO-based ADC, the change in the frequency with respect to the change in the equivalent resistance must be checked, that is, $|\frac{\partial f}{\partial R_{\mathrm{eq}}}|$. The smaller the term $|\frac{\partial f}{\partial R_{\mathrm{eq}}}|$ is, the better is the variability tolerance. Please note that the *absolute value* of $\frac{\partial f}{\partial R_{\mathrm{eq}}}$ matters for the variability tolerance. According to the chain rule:

$$\left|\frac{\partial f}{\partial R_{\mathrm{eq}}}\right| = \left|\frac{\partial f}{\partial V_{\mathrm{bias}}} \cdot \frac{\partial V_{\mathrm{bias}}}{\partial R_{\mathrm{eq}}}\right|. \tag{4}$$

In Equation (4), the term $\frac{\partial f}{\partial V_{\mathrm{bias}}}$ is the speed of the ring-oscillator and equals K. The term $\frac{dV_{bias}}{dReq}$ is calculated according to

$$\left|\frac{\partial V_{\mathrm{bias}}}{\partial R_{\mathrm{eq}}}\right| = V_{\mathrm{read}} \cdot \left|\frac{\partial}{\partial R_{\mathrm{eq}}} \frac{Z_{\mathrm{os}} || \frac{1}{2\pi f C}}{Z_{\mathrm{os}} || \frac{1}{2\pi f C} + R_{\mathrm{eq}}}\right|. \tag{5}$$

Combining Equations (4) and (5), we get that

$$\left|\frac{\partial f}{\partial R_{\mathrm{eq}}}\right| = \left|K \cdot V_{\mathrm{read}} \cdot \frac{\left(\frac{\partial Z_{\mathrm{os}}}{\partial R_{\mathrm{eq}}} \cdot R_{\mathrm{eq}} - Z_{\mathrm{os}}\right) - 2\pi \cdot f \cdot C \cdot Z_{\mathrm{os}}^2}{(Z_{\mathrm{os}} + R_{\mathrm{eq}} \cdot Z_{\mathrm{os}} \cdot 2\pi f C + R_{\mathrm{eq}})^2 + (2\pi \cdot Z_{\mathrm{os}}^2 \cdot R_{\mathrm{eq}} \cdot C)}\right|. \tag{6}$$

Z_{os} and R_{eq} are in the range of kΩ[1], f is in the range of GHz, and C is in the range of fF. Thus, the terms $(R_{\mathrm{eq}} \cdot Z_{\mathrm{os}} \cdot 2\pi f C)$ and $(2\pi \cdot Z_{\mathrm{os}}^2 \cdot R_{\mathrm{eq}} \cdot C)$ are negligible. Since the term $(\frac{dZ_{\mathrm{os}}}{dR_{\mathrm{eq}}} \cdot R_{\mathrm{eq}} - Z_{\mathrm{os}})$ is negative and due to the impact of *absolute value*, the final formula for the variability tolerance is

$$\left|\frac{\partial f}{\partial R_{\mathrm{eq}}}\right| = K \cdot V_{\mathrm{read}} \cdot \frac{\left(Z_{\mathrm{os}} - \frac{\partial Z_{\mathrm{os}}}{\partial R_{\mathrm{eq}}} \cdot R_{\mathrm{eq}}\right) + 2\pi \cdot f \cdot C \cdot Z_{\mathrm{os}}^2}{(Z_{\mathrm{os}} + R_{\mathrm{eq}})^2}. \tag{7}$$

[1] For the correct functionality of the ADC, Z_{os} and R_{eq} must be in a same range. If Z_{os} is much larger, the voltage of node Y in Figure 7 would be V_{read} and if Z_{os} is much smaller, the voltage of this point would be 0. In either of these states, the change in the R_{eq} cannot be reflected on the node Y and the ADC would fail to distinguish between levels.

Equation (7) shows that by increasing the parasitic bit-line capacitor, variability tolerance decreases. The nominal values of the LRS and HRS also have an impact on variability tolerance. For instance, for higher resistances, the speed of the ring-oscillator (the constant K) must be higher since similar data patterns must result in similar digital outputs regardless of the resistive values of LRS and HRS. By increasing the nominal values of LRS and HRS, R_{eq} increases. This forces the bias voltage of the ring-oscillator to decrease (node Y in Figure 3). The number of generated signals, however, must remain the same to produce the same frequency with a lower bias voltage. A higher ring-oscillator speed is destructive for variability tolerance, but the larger LRS and HRS values contribute quadratically, which overcompensates for the former effect. To conclude, increasing the LRS and HRS value improves variability tolerance.

It is interesting to mention that adding an impedance element (such as the diode-connected structure in Figure 3) decreases variability tolerance by decreasing the Z_{os} term in the denominator of Equation (7). Confusion might result on this matter, as Z_{os} is present in both nominator and denominator with equal exponent. These two terms are not canceling each other out since the first term of the nominator ($Z_{\text{os}} - R_{\text{eq}} \cdot \frac{dZ_{\text{os}}}{dR_{\text{eq}}}$) is much larger than the second term ($2\pi f \cdot C \cdot Z_{\text{os}}^2$). Thus, the first term is dominant and the second term can be neglected here. However, if the only changing variable in the circuit is C, then the only changing term is ($2\pi f \cdot C \cdot Z_{\text{os}}^2$). In that case, the effect of C on variability tolerance must be considered.

Power consumption: In resistive elements, power is equal to $\frac{V^2}{R}$. Here, the term R indicates the total resistive load of the circuit, which is the series connection of the crossbar and the whole VCO-based ADC. Since the data manifest themselves as resistance states, the power consumption is data dependent. The maximum and minimum power consumption happens for *all LRS* and *all HRS*, respectively. The power is also affected by the nominal values of the LRS and HRS, as it is inversely proportional to R_{eq}. In the fixed resolution, the **latency** is determined only by the counter and is not affected by the data or LRS and HRS nominal values. The **Energy** is the power-latency product and is affected by the data and the nominal values of LRS and HRS in a similar way as the power.

Voltage across the memory cells: As discussed in Section 2, read-disturb phenomena are more likely to happen at high voltages across the ReRAM devices. This voltage can also be affected by the data and the nominal values of the LRS and HRS. The largest voltage across the device occur for *all HRS* since R_{eq} is maximal in this case. Larger (smaller) nominal values for LRS and HRS also results in a higher (lower) voltage across the ReRAM devices. A higher voltage across the ReRAM devices increases the probability of read disturb phenomena.

4 RESULTS

In this section, we present the device- and circuit-level results of the proposed VCO-based ADC design. The first part of this section discusses the ReRAM manufacturing results. The second part discusses the circuit-level results.

4.1 ReRAM Fabrication and Characterization

4.1.1 Device Fabrication. For the experimental investigation, we fabricated VCM ReRAM cells with a (30 nm Pt/5 nm ZrO_2/20 nm Ta/30 nm Pt) stack as shown in Figure 8(a).

The cells are arranged in a 7 μm × 7 μm crossbar structure, designed as a 32 × 1 cell array. A microscopic picture of this structure is given in Figure 8(b). Using a dedicated probe card, all 32 top electrodes of one array can be connected to the measurement device. The bottom electrode is common for all cells on the die and realized as a whole surface platinum layer underneath a structured SiO_2 layer, which separates the single array.

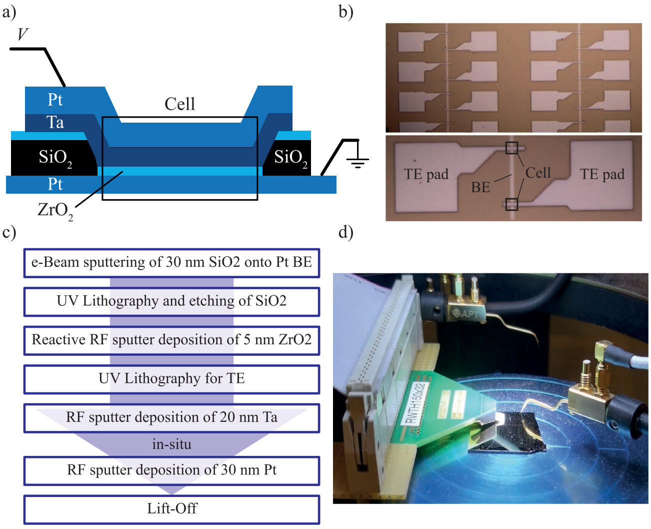

a)
V
Pt
Ta
Cell
SiO$_2$
SiO$_2$
Pt
ZrO$_2$
b)
TE pad
BE
Cell
TE pad
c)
e-Beam sputtering of 30 nm SiO2 onto Pt BE
UV Lithography and etching of SiO2
Reactive RF sputter deposition of 5 nm ZrO2
UV Lithography for TE
RF sputter deposition of 20 nm Ta
in-situ
RF sputter deposition of 30 nm Pt
Lift-Off
d)
RWTH150k02

Fig. 8. (a) Schematic of the cell stack used in this work. Note that the dimensions are not to scale. (b) Top view microscopy of the 32 × 1 array structure. Highlighted are bottom electrode (BE), top electrode (TE), and cells in their cross-section. (c) Fabrication process flow. (d) Photography of the measurement setup.

The fabrication workflow of the presented cells is illustrated in Figure 8(c). Onto the Pt bottom electrode covering the whole substrate surface, 30 nm SiO_2 is deposited via e-Beam sputtering. Using UV lithography, the array structure (represented by the vertical lines shown in Figure 8(b)) is transferred to the SiO_2 layer and etched free via hydrofluoric acid. After removing the photoresist, 5 nm ZrO_2 is deposited via reactive RF sputtering. Another UV lithography step is used to structure the top electrodes; subsequently, 20 nm Ta is deposited on the oxide via RF sputtering. To prevent oxidation of the Ta electrode, it is in situ covered by a 30 nm Pt layer. It may be noted that the SiO_2 layer is used to separate the single array and does not contribute to the resistive switching characteristics. This enables a large bottom electrode in the array structure, resulting in comparatively low series resistances. Since HfO_2 is more common as switching oxide in VCM devices, we would like to note that ZrO_2 and HfO_2 are almost identical with respect to their physico-chemical properties.

4.1.2 Device Measurement. The cells are characterized using a dedicated probe card providing 32 probes that are connected to a custom array tester based on the *μController Module* platform by *aixACCT Systems*. A photograph of the measurement setup is shown in Figure 8(d). All voltages are applied via the probe card shown on the left to the Ta top electrode. Using another probe shown on the right, the common Pt bottom electrode is connected to GND. Figure 9 sketches the general measurement flow for forming, SET, and RESET. Each cell requires an initial electroforming step to generate oxygen vacancies and develop a conducting filament [16]. Subsequently, the cell is cycled by alternating RESET and SET operations. The initial electroforming is performed by a triangular voltage pulse with a rise time of 20 ms. The forming stop voltage ranges from 3 V to 5 V. Here, a read-verify algorithm, implemented in the measurement software, is used. Starting with 3 V, the cell is read after each forming operation. In the case of failed forming, the pulse is repeated with increased stop voltage to a maximum of 5 V. During electroforming, the resistance of the cell is lowered by several orders of magnitude. Since this operation requires comparatively high voltages, the decreased resistance would result in a high current through the cell along with high temperature, which could cause irreparable damage to the cell [66]. Therefore, the cell current

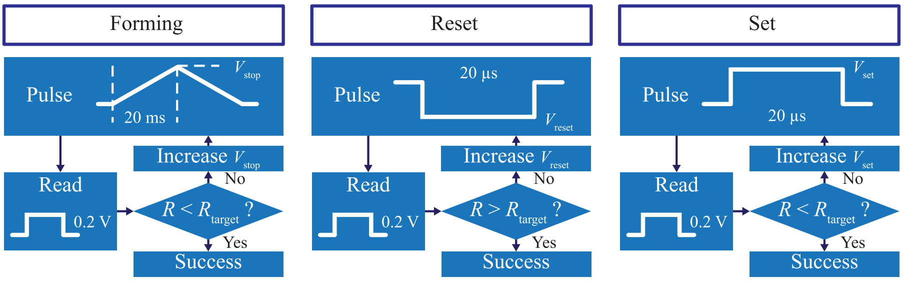


Fig. 9. Detailed flow of forming, RESET, and SET operation.

is limited during forming by adding a series resistance with $R_s = 10$ kΩ via the internal switch matrix of the measurement equipment.

For both RESET and SET operations, rectangular voltage pulses with a length of 20 µs are applied. Using an equivalent read-verify algorithm, the pulse height varies from −0.5 V to −5 V for the RESET process and from 0.5 V to 5 V for the SET operations. After each programming pulse, a read pulse is applied and the cell resistance is determined. Unless the state is within the required margins, further programming pulses with increased voltage are applied, as shown schematically in Figure 9. It was observed that no series resistance is required for the pulsed SET and RESET schemes. It may be noted that the typical pulse voltage for a successful SET is approximately 1 V. The typical RESET voltage is −1.8 V. The high maximum voltages of the algorithm are usually not necessary. The algorithm ensures reliable programmability of each cell with a preferably low voltage. Furthermore, it enables programmability of cells into variability margins specified by the application.

4.1.3 Measurement Results. As outlined in Table 2, the circuit design desires ReRAM resistances of $R_{\text{HRS}} = 30$ kΩ and $R_{\text{LRS}} = 3$ kΩ. To model the ReRAM devices, we use fixed resistances, as 3 and 30 kΩ, which are relatively low resistances. For such low resistances, the kind of ReRAM devices we used behaves quite linearly and shows little RTN. As later discussed in detail in Section 4.2.2, the VCO-based ADC can tolerate variation in resistive states of the ReRAM devices up to <30%. Figure 10 shows experimentally obtained cumulative distributions of 7000 HRS and LRS states each. The data are acquired by cycling 32 cells of one array with the read-verify algorithm into the highlighted variability margins of 2.2 kΩ to 3.8 kΩ for the LRS and 22 kΩ to 38 kΩ for the HRS. These margins are well within the tolerated maximum of 30%. Thus, it can be stated that the fabricated devices match the specifications given in Table 2 and, therefore, are appropriate for the proposed application.

4.2 Circuit Level Results

The electrical schematic simulation of this design has been done in TSMC 28 nm technology. The Analog parts of the design, the crossbar and the ring-oscillator, are simulated using the Cadence Spectre simulator, whereas the digital parts, the counter and the LUT, are first described with Verilog code and then synthesized with the Cadence GENUS tool. Table 2 shows the simulation parameters.

4.2.1 Design Parameters Tuning and Exploration. To improve the circuit from the resolution point of view, we first need to *tune* different parameters of the circuit, such as V_{read} (Figure 11(a)), the transistor sizes in the inverters (Figure 11(b)), the number of the inverter gates in the ring-oscillator (Figure 11(c)), and the width of the diode-connected structure (Figure 11(d)). To select these parameters, a trade-off analysis between different aspects of the circuit must be performed.

Table 2. Simulation Parameters

Parameters	Specifications
RRAM Device	ZrO_2/Ta
HRS	30 kΩ
LRS	3 kΩ
CMOS Technology	28 nm TSMC
Read Voltage	0.9 V with ±10% variations
CMOS Specs	TT, 27°C
CMOS Variation	3 σ
Counter/LUT Voltage	0.9 V with ±10% variations
Latency of the counter	40 ps

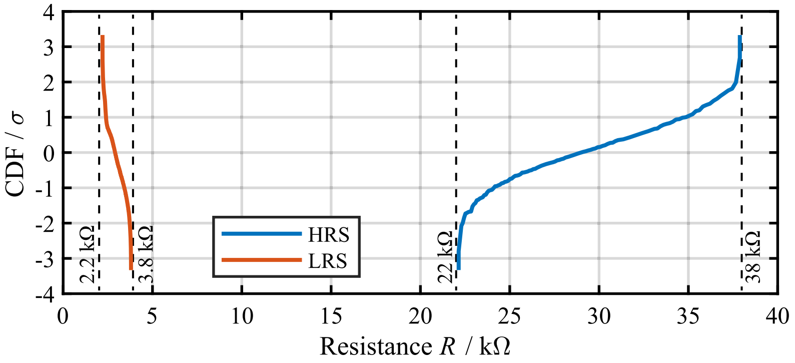
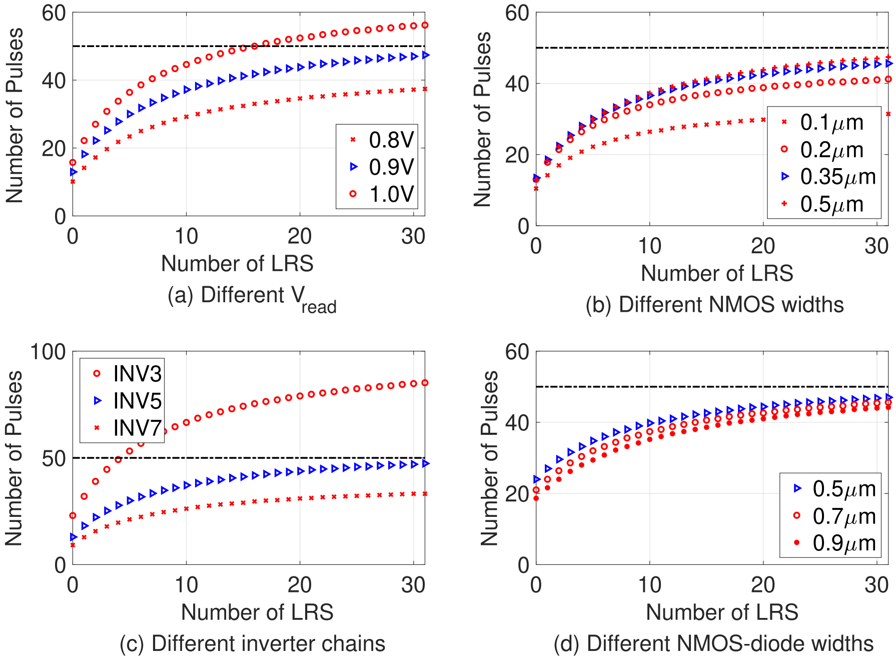


Fig. 10. Cumulative distributions of experimental HRS and LRS data covering 7000 states each.

On the one hand, for V_{read}, lower values are beneficial from a power consumption point of view. On the other hand, higher values for V_{read} increase the resolution (more dynamic range for the number of generated pulses). Due to the timing constraints of the counter, the number of pulses cannot be more than a specific number, and increasing V_{read} beyond a certain voltage (0.9 V) does not have any impact on the resolution. Regarding the size of the transistors in the inverter, smaller widths are beneficial for the area (for similar rise and fall times of the pulses, the size of the PMOS is kept twice the size of the NMOS), whereas larger widths result in a higher speed of the ring-oscillator, which is beneficial for the resolution. Increasing the width from 0.35 μm to 0.5 μm does not have any specific impact on the resolution. Thus, the NMOS width has been selected to 0.35 μm. The number of inverter gates in the ring-oscillator also has to be investigated. A shorter ring results in a higher speed of the ring-oscillator, which is beneficial for the resolution. Moreover, a shorter ring improves area efficiency. However, increasing the speed of the ring-oscillator beyond timing constraints of the counter will not improve the resolution. Rather, this will increase the probability of timing errors. In the proposed VCO-based ADC, the number of inverters is 5. As mentioned before, we have selected a diode-connected NMOS structure as current-voltage converter. The width of this device also needs to be determined. The simulations show that the width of the diode-connected NMOS does not have a tangible impact on the resolution. Thus, we can select it as low as possible to improve area efficiency. As lower widths are more prone to CMOS variations, we have selected the width of diode-connected NMOS as 0.5 μm, as increasing beyond this width does not help further. For all four plots of Figure 11, we increase the number of LRS from 0 to 31



(a) Different V_{read}

(b) Different NMOS widths

(c) Different inverter chains

(d) Different NMOS-diode widths

Fig. 11. Number of pulses versus different parameters. (a) V_{read}, (b) the inverter's NMOS width (here, the width of the PMOS is twice the width of the NMOS), (c) the number of the inverter stages in the ring-oscillator, and (d) the width of the diode-connected NMOS. In all graphs, the horizontal dashed line shows the maximum countable pulses in 2 ns, which is 50 (the minimum period of generated pulse must be at least 40 ps to be countable with the counter and $\frac{2\ ns}{40\ ps} = 50$).

and count the number of pulses. *A more dynamic range for the number of generated pulses while not exceeding the counter's maximum frequency* is our criterion for selecting the parameters.

4.2.2 VCO-Based ADC Evaluation. To investigate the latency of the ADC, two points must be considered: (1) the number of generated pulses must be unique (not necessarily linear) for a different number of LRS devices and (2) the number of generated pulses must be countable by the counter. Figure 12(a) shows the number of pulses for different numbers of LRS, from 0 to 31. In 2 ns evaluation time, the proposed VCO-based ADC can generate different numbers of pulses for 13 different levels. For a higher number of levels, the number of generated pulses is not unique anymore. To consider the impact of process, voltage and temperature variability, 100 Monte Carlo analysis has been performed with and without STP technique to reduce the impact of global variations in CMOS devices (with normal distribution). Figure 12(b) also shows the voltage of the node Y in Figure 3 (bias voltage of the ring-oscillator) with 100 Monte Carlo analysis. The minimum voltage at node Y is 0.6 V; as V_{read} is 0.9 V, *the maximum voltage across the ReRAM devices* is less than 0.3 V. Thus, the probability read-disturb in the ReRAM is rather low.

By increasing the evaluation time, the resolution of the ADC increases at the cost of latency and energy. Figure 13(a) shows the impact of evaluation time on the number of generated pulses with 100 times Monte Carlo analysis with STP technique. Figure 13(b) shows the impact of evaluation time on the energy and resolution of the ADC. The latter is almost linear, as the number of generated pulses is more or less proportional to the evaluation time.

The *area* of the VCO-based ADC contains four elements, the diode-connected structure to convert current into voltage, the ring-oscillator, the counter, and the LUT. As already discussed, the

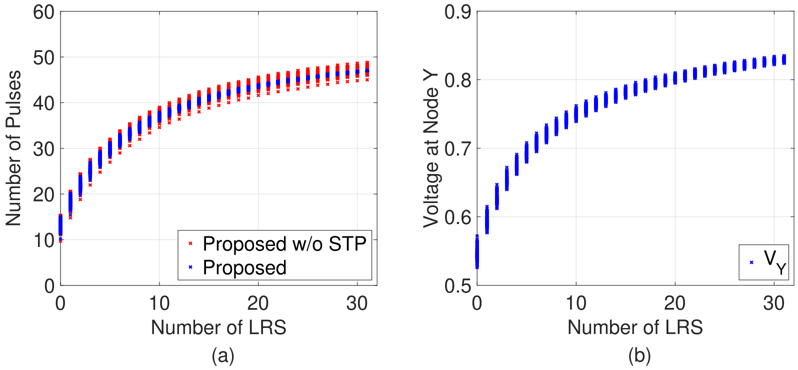


Fig. 12. (a) The number of generated pulses without and with STP feature for 2 ns evaluation period for different numbers of LRS devices.(b) The bias voltage or the voltage developed at node Y (V_Y) of the ring-oscillator.

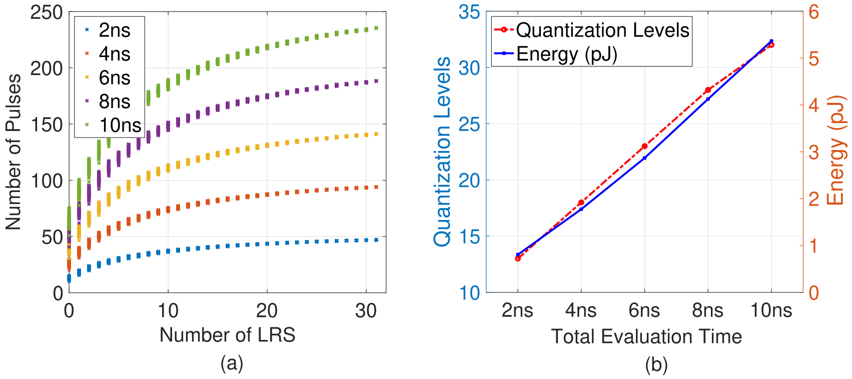


Fig. 13. Impact of evaluation time on (a) the number of generated pulses and (b) the energy and resolution of the ADC.

Table 3. The Area of the Different Parts of the VCO-Based ADC

Circuit Design	Area (μm^2)
NMOS (diode-connected)	0.022
Ring-oscillator	0.22
Counter	0.81
LUT	0.69
Total area	1.74

number of generated pulses of the ADC can be adjusted by changing the evaluation time. To report the area, we consider a 6-stage ripple counter. Table 3 shows the area of each element and the total area of the whole design. As shown in the Table 3, the proposed VCO-based ADC has a small enough area to assign one to each column.

The proposed VCO-based ADC can tolerate variation in the ReRAM devices up to a certain percentage. As already discussed, the write-verify algorithm at the device level can ensure the variability within the acceptable margin. However, as Figure 14 shows, the resolution of the ADC tends to decrease with increasing the variability. To analyze the impact of variability in ReRAM,

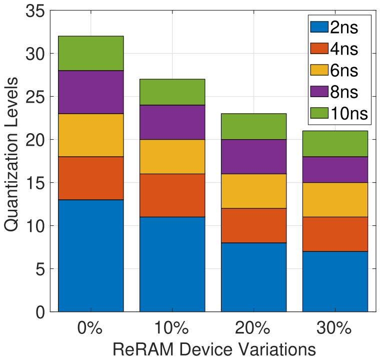


Fig. 14. The impact of variability in ReRAM devices (worst-case conditions have been taken into account) on the resolution of the ADC by also considering CMOS variation.

Table 4. Features of the Proposed VCO-Based ADC

Feature	Value
Resolution (bit)	3 to 5
Latency (ns)	2 to 10
Voltage across the ReRAM device (V)	less than 0.3
Energy (pJ)	0.8 to 5.2
Area (μm^2)	1.74
ReRAM variation awareness	≈30%

we have considered the worst case, that is, all of the ReRAM devices either in LRS or HRS. As outlined in Table 2, CMOS variation is also considered in this simulation. The 10% variability, for instance, means that the resistance of the ReRAM devices can either be 2700 Ω or 3300 Ω for the LRS and 27 kΩ or 33 kΩ for the HRS.

Table 4 summarizes the features of our VCO-based ADC. The small area of the proposed VCO-based ADC enables us to assigned one ADC per column, as it is able to distinguish between more than one level per activation cycle. Moreover, the proposed ADC meets the requirements of the ReRAM devices, such as voltage across the device and variability.

4.2.3 Comparison with State-of-the-Art. Our proposed ADC provides an efficient solution that implements MAC operation in a group of memristor-based crossbar columns. In this manner, we have included a subset of a fully fledged general purpose vector-matrix multiplication (VMM), which is the most fundamental computational unit in today's hardware systems implementing machine learning algorithms. In other words, we presented a column that is a block computed in a parallel manner in a crossbar of arbitrary number of columns. This implies that this sub-block can be repeated many times, and similar efficiency can be easily scaled for larger crossbars.

The essential step towards comparing the efficiency of the design is defining a *Figure of Merit* (FoM). To define a proper FoM, let us first investigate the energy of the ADC phase. The energy of this phase can be determined by

$$E = P \times N_{\mathrm{ADC}} \times L \times \left(\frac{N_{\mathrm{C}}}{N_{\mathrm{ADC}}}\right) \times \left(\frac{N_{\mathrm{S}}}{N_{\mathrm{D}}}\right). \tag{8}$$

Here, P is the power consumption of the ADC, N_{ADC} is the number of ADCs, L is the latency of the interface, N_C is the number of columns in the crossbar, N_S is all possible states covered by the DAC and the crossbar's rows, and N_D is the number of levels distinguishable by the ADC.

Table 5. Comparison of Different ADC Interfaces

Feature	SAR ADC [21, 67]	SA [24, 68]	This work
Technology node	32 nm	65 nm	28 nm
Area	9600 μm^2	78.3 μm^2	1.74 μm^2
Latency	1 ns	5 ns	2 to 10 ns
Energy	16 uJ	114 to 130 fJ	0.8 to 5.2 pJ
Resolution (# of levels)	128	1	8 to 32
Variability in memristors	Not considered	Not considered	Considered
Voltage across the memristors	Not considered	Not considered	less than 0.3 V
FoM	1.25×10^{-13} J	1.14 to 1.30×10^{-13} J	1 to 1.6×10^{-13} J

In general, the total energy is the power-latency product. The whole power of the ADC phase is equal to the power of one ADC multiplied by the number of ADCs. The latency of the ADC is determined by three factors: the *latency of one ADC*, the *degree of the ADC sharing*, and the *total number of activations per column*. The degree of time-sharing is simply obtained by $\frac{N_C}{N_{ADC}}$. The total number of activations per column is obtained by $\frac{N_S}{N_D}$, which shows the number of cycles an ADC must be activated in order to convert the analog data of a column to a digital representation.

The FoM must reflect the efficiency of the design; hence, it must include parameters that are directly controllable by the designer. In the energy formula of the ADC (Equation (8)) P, L, and N_D are 100% controllable by the ADC designer, whereas N_C and N_S are determined by *crossbar* and *DAC modules and number of crossbar's rows*, respectively. N_{ADC} is semi-controllable by the ADC designer, since it is derived from *min* $\left(\frac{\text{Area budget of the system}}{\text{Area of the ADC}}, \frac{\text{Power budget of the system}}{\text{Power of the ADC}}\right)$. The area and power budget of the system is not controllable by the designer, but the area and power of the system can be fully controlled by the designer.

By defining FoM as $\frac{\text{energy per module}}{N_D}$, it includes all of the parameters that are controllable by the ADC designer and reflects the efficiency of the circuit. A smaller FoM implies a more efficient circuit design. Table 5 includes different features for the SAR ADC [67], SA [68], and VCO-based ADC. The results reported for the SAR ADC [67] and SA [68] are based on the actual measurements while, for this work, results are gathered through electrical schematic simulation. As shown in the table, the proposed VCO-based design in 28 nm technology node has almost the same (even better for lower resolution) circuit design efficiency as previous designs and also considers two reliability concerns of the memristor devices: the variability and voltage across the device.

5 CONCLUSION AND FUTURE WORK

CiM architectures using emerging memory technologies have the potential to overcome the data transfer and performance challenges of conventional von Neumann–based designs. However, due to analog computation, the efficiency of CiM architecture is highly limited by the ADC phase. In order to address this issue, a VCO-based ADC design is presented in this article. In the proposed ADC design, the bit-line current coming from the crossbar is first converted into voltage. Then, the voltage is used to drive a VCO, which generates pulses with a frequency proportional to the voltage. The proposed ADC is evaluated using a ReRAM-based CiM crossbar array. Simulation results show that it can distinguish up to 32 levels within 10 ns while consuming less than 5.2 pJ of energy. In addition, our proposed ADC can tolerate ≈30% variability of the resistive device state with a negligible impact on the performance of the ADC. A direction for further work is to improve the resolution of our ADC design while maintaining a compact and low-power design. Moreover, efficient programming circuits need to be further explored to realize the write-verify operations. Other future works include enhancing the implementation of the currently high area and power-consuming counter and LUT designs.

REFERENCES

[1] S. Hamdioui, S. Kvatinsky, G. Cauwenberghs, L. Xie, N. Wald, S. Joshi, H. M. Elsayed, H. Corporaal, and K. Bertels. 2017. Memristor for computing: Myth or reality?. In *Design, Automation and Test in Europe Conference and Exhibition (DATE'17)*. 722–731.

[2] O. Mutlu. 2018. Processing data where it makes sense in modern computing systems: Enabling in-memory computation. In *7th Mediterranean Conference on Embedded Computing (MECO'17)*. IEEE, 8–9.

[3] C. David Wright, Peiman Hosseini, and Jorge A. Vazquez Diosdado. 2013. Beyond von-Neumann computing with nanoscale phase-change memory devices. *Advanced Functional Materials* 23, 18 (2013), 2248–2254.

[4] Sasikanth Manipatruni, Dmitri E. Nikonov, and Ian A. Young. 2018. Beyond CMOS computing with spin and polarization. *Nature Physics* 14, 4 (2018), 338–343.

[5] J. Yu, M. Abu Lebdeh, H. Du Nguyen, M. Taouil, and S. Hamdioui. 2020. The power of computation-in-memory based on memristive devices. In *25th Asia and South Pacific Design Automation Conference (ASP-DAC'20)*. 385–392.

[6] Qian Wang, Youjie Li, Botang Shao, Siddhartha Dey, and Peng Li. 2017. Energy efficient parallel neuromorphic architectures with approximate arithmetic on FPGA. *Neurocomputing* 221 (2017), 146–158.

[7] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, and Y. Xie. 2016. Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories. In *53nd ACM/EDAC/IEEE Design Automation Conference (DAC'16)*. 1–6.

[8] S. Hamdioui, H. A. Du Nguyen, M. Taouil, A. Sebastian, M. L. Gallo, S. Pande, S. Schaafsma, F. Catthoor, S. Das, F. G. Redondo, G. Karunaratne, A. Rahimi, and L. Benini. 2019. Applications of computation-in-memory architectures based on memristive devices. In *Design, Automation and Test in Europe Conference and Exhibition (DATE'19)*. 486–491.

[9] Shuang Gao, Fei Zeng, Minjuan Wang, Guangyue Wang, Cheng Song, and Feng Pan. 2015. Implementation of complete Boolean logic functions in single complementary resistive switch. *Scientific Reports* 5 (2015), 15467.

[10] Yaxiong Zhou, Yi Li, Lei Xu, Shujing Zhong, Huajun Sun, and Xiangshui Miao. 2015. 16 Boolean logics in three steps with two anti-serially connected memristors. *Applied Physics Letters* 106, 23 (2015), 233502.

[11] Zhong Sun, Elia Ambrosi, Alessandro Bricalli, and Daniele Ielmini. 2018. Logic computing with stateful neural networks of resistive switches. *Advanced Materials* 30, 38 (2018), 1802554.

[12] Giacomo Indiveri and Shih-Chii Liu. 2015. Memory and information processing in neuromorphic systems. *Proc. IEEE* 103, 8 (2015), 1379–1397.

[13] Sandeep Kaur Kingra, Vivek Parmar, Che-Chia Chang, Boris Hudec, Tuo-Hung Hou, and Manan Suri. 2020. SLIM: Simultaneous logic-in-memory computing exploiting bilayer analog OxRAM devices. *Scientific Reports* 10, 1 (2020), 1–14.

[14] Vivek Seshadri, Kevin Hsieh, Amirali Boroum, Donghyuk Lee, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry. 2015. Fast bulk bitwise AND and OR in DRAM. *IEEE Computer Architecture Letters* 14, 2 (2015), 127–131.

[15] Mingu Kang, Min-Sun Keel, Naresh R. Shanbhag, Sean Eilert, and Ken Curewitz. 2014. An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. IEEE, 8326–8330.

[16] Rainer Waser, Regina Dittmann, Georgi Staikov, and Kristof Szot. 2009. Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges. *Advanced Materials* 21, 25-26 (2009), 2632–2663.

[17] Hoang Anh Du Nguyen, Jintao Yu, Muath Abu Lebdeh, Mottaqiallah Taouil, Said Hamdioui, and Francky Catthoor. 2020. A classification of memory-centric computing. *ACM Journal on Emerging Technologies in Computing Systems* 16, 2 (2020), 1–26.

[18] Daniele Ielmini and H.-S. Philip Wong. 2018. In-memory computing with resistive switching devices. *Nature Electronics* 1, 6 (2018), 333–343.

[19] Weitao Li, Pengfei Xu, Yang Zhao, Haitong Li, Yuan Xie, and Yingyan Lin. 2020. TIMELY: Pushing data movements and interfaces in PIM accelerators towards local and in time domain. *arXiv:2005.01206*

[20] Teyuh Chou, Wei Tang, Jacob Botimer, and Zhengya Zhang. 2019. Cascade: Connecting RRAMs to extend analog dataflow in an end-to-end in-memory processing paradigm. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. 114–125.

[21] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar. 2016. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In *ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA'16)*. 14–26.

[22] Shihui Yin, Xiaoyu Sun, Shimeng Yu, and Jae-sun Seo. 2019. High-throughput in-memory computing for binary deep neural networks with monolithically integrated RRAM and 90 nm CMOS. *arXiv:1909.07514*

[23] L. Song, X. Qian, H. Li, and Y. Chen. 2017. PipeLayer: A pipelined ReRAM-based accelerator for deep learning. In *IEEE International Symposium on High Performance Computer Architecture (HPCA'17)*. 541–552.

[24] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie. 2016. PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In *ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA'16).* 27–39.

[25] B. Yan, Q. Yang, W. Chen, K. Chang, J. Su, C. Hsu, S. Li, H. Lee, S. Sheu, M. Ho, Q. Wu, M. Chang, Y. Chen, and H. Li. 2019. RRAM-based spiking nonvolatile computing-in-memory processing engine with precision-configurable in situ nonlinear activation. In *Symposium on VLSI Technology.* T86–T87.

[26] Handel Jones. 2015. Semiconductor industry from 2015 to 2025 [White paper]. International Business Strategies.

[27] H. A. D. Nguyen, J. Yu, L. Xie, M. Taouil, S. Hamdioui, and D. Fey. 2017. Memristive devices for computing: Beyond CMOS and beyond von Neumann. In *IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC'17).* 1–10.

[28] G. Indiveri and S. Liu. 2015. Memory and information processing in neuromorphic systems. *Proc. IEEE* 103, 8 (2015), 1379–1397.

[29] G. De Sandre, L. Bettini, A. Pirola, L. Marmonier, M. Pasotti, M. Borghi, P. Mattavelli, P. Zuliani, L. Scotti, G. Mastracchio, F. Bedeschi, R. Gastaldi, and R. Bez. 2010. A 90 nm 4MB embedded phase-change memory with 1.2V 12ns read access time and 1MB/s write throughput. In *IEEE International Solid-State Circuits Conference (ISSCC'10).* 268–269.

[30] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, T. Kishi, T. Kai, M. Amano, N. Shimomura, H. Yoda, and Y. Watanabe. 2010. A 64MB MRAM with clamped-reference and adequate-reference schemes. In *IEEE International Solid-State Circuits Conference (ISSCC'10).* 258–259.

[31] Meng-Fan Chang, Shin-Jang Shen, Chia-Chi Liu, Che-Wei Wu, Yu-Fan Lin, Ya-Chin King, Chorng-Jung Lin, Hung-Jen Liao, Yu-Der Chih, and Hiroyuki Yamauchi. 2013. An offset-tolerant fast-random-read current-sampling-based sense amplifier for small-cell-current nonvolatile memory. *IEEE Journal of Solid-State Circuits* 48, 3 (2013), 864–877.

[32] Loai Danial and Shahar Kvatinsky. 2020. Analog to Digital Converter Using Memristors in a Neural Network. (Aug. 27 2020). US Patent App. 15/931,690.

[33] A. Chen. 2016. A review of emerging non-volatile memory (NVM) technologies and applications. *Solid-State Electronics* 125 (2016), 25–38.

[34] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams. 2010. "Memristive"switches enable "stateful" logic operations via material implication. *Nature* 464, 7290 (2010), 873–876.

[35] J. Jang, S. Park, Y. Jeong, and H. Hwang. 2014. ReRAM-based synaptic device for neuromorphic computing. In *IEEE International Symposium on Circuits and Systems (ISCAS'14) Melbourne, VIC, Australia.* IEEE, 1054–1057.

[36] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai. 2012. "Metal" oxide RRAM. *Proc. IEEE* 100, 6 (2012), 1951–1970.

[37] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams. 2008. The missing memristor found. *Nature* 453, 7191 (2008), 80–83.

[38] C. Funck, A. Marchewka, C. Baeumer, P. C. Schmidt, P. Mueller, R. Dittmann, M. Martin, R. Waser, and S. Menzel. 2018. A theoretical and experimental view on the temperature dependence of the electronic conduction through a Schottky barrier in a resistively switching $SrTiO_3$-based memory cell. *Adv. Electron. Mater.* 4, 7 (2018), 1800062.

[39] F. Cueppers, S. Menzel, C. Bengel, A. Hardtdegen, M. von Witzleben, U. Boettger, R. Waser, and S. Hoffmann-Eifert. 2019. Exploiting the switching dynamics of HfO2-based ReRAM devices for reliable analog memristive behavior. *APL Mater.* 7, 9 (2019), 91105/1–9.

[40] H. Zheng, T. Chang, K. Xue, Y. Su, C. Wu, C. Shih, Y. Tseng, W. Chen, W. Huang, C. Chen, X. Miao, and S. M. Sze. 2018. Reducing forming voltage by applying bipolar incremental step pulse programming in a 1T1R structure resistance random access memory. *IEEE Electron Device Letters* 39, 6 (2018), 815–818.

[41] Xia Sheng, Catherine E. Graves, Suhas Kumar, Xuema Li, Brent Buchanan, Le Zheng, Sity Lam, Can Li, and John Paul Strachan. 2019. Low-conductance and multilevel CMOS-integrated nanoscale oxide memristors. *Advanced Electronic Materials* 5, 9 (2019), 1800876.

[42] W. Qian, P. Chen, R. Karam, L. Gao, S. Bhunia, and S. Yu. 2017. Energy-efficient adaptive computing with multifunctional memory. *IEEE Transactions on Circuits and Systems II: Express Briefs* 64, 2 (2017), 191–195.

[43] C. Baeumer, R. Valenta, C. Schmitz, A. Locatelli, T. O. Mentes, S. P. Rogers, A. Sala, N. Raab, S. Nemsak, M. Shim, C. M. Schneider, S. Menzel, R. Waser, and R. Dittmann. 2017. Subfilamentary networks cause cycle-to-cycle variability in memristive devices. *ACS Nano* 11, 7 (2017), 6921–6929.

[44] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini. 2013. Understanding switching variability and random telegraph noise in resistive RAM. In *IEEE International Electron Devices Meeting (IEDM'13).* IEEE, 31.5.1–31.5.4.

[45] H. Aziza, M. Bocquet, J.-M. Portal, and C. Muller. 2011. Bipolar OxRRAM Memory Array Reliability Evaluation Based on Fault Injection. In *IEEE 6th International Design and Test Workshop (IDT'11).* 78–81.

[46] S. Wiefels, U. Böttger, S. Menzel, D. J. Wouters, and R. Waser. 2020. Statistical modeling and understanding of HRS retention in 2.5 MB HfO2 based ReRAM. In *IEEE International Memory Workshop (IMW'20).* 1–4.

[47] S. Menzel, U. Böttger, M. Wimmer, and M. Salinga. 2015. Physics of the switching kinetics in resistive memories. *Adv. Funct. Mater.* 25, 40 (2015), 6306–6325.

[48] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams. 2016. Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication. In *53rd ACM/EDAC/IEEE Design Automation Conference (DAC'16).* 1–6.

[49] K. Humood, B. Mohammad, H. Abunahla, and A. Azzam. 2020. On-chip tunable memristor-based flash-ADC converter for artificial intelligence applications. *IET Circuits, Devices Systems* 14, 1 (2020), 107–114. DOI: http://dx.doi.org/10.1049/iet-cds.2019.0293

[50] B. Razavi. 2017. The flash ADC [A Circuit for All Seasons]. *IEEE Solid-State Circuits Magazine* 9, 3 (Summer 2017), 9–13. DOI: http://dx.doi.org/10.1109/MSSC.2017.2712998

[51] R. B. Staszewski, K. Muhammad, D. Leipold, Chih-Ming Hung, Yo-Chuol Ho, J. L. Wallberg, C. Fernando, K. Maggio, R. Staszewski, T. Jung, Jinseok Koh, S. John, Irene Yuanying Deng, V. Sarda, O. Moreira-Tamayo, V. Mayega, R. Katz, O. Friedman, O. E. Eliezer, E. de-Obaldia, and P. T. Balsara. 2004. All-digital TX frequency synthesizer and discrete-time receiver for Bluetooth radio in 130-nm CMOS. *IEEE Journal of Solid-State Circuits* 39, 12 (2004), 2278–2291.

[52] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury. 2018. A 55 nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots. In *IEEE International Solid-State Circuits Conference (ISSCC'18).* 124–126.

[53] D. Miyashita, R. Yamaki, K. Hashiyoshi, H. Kobayashi, S. Kousai, Y. Oowaki, and Y. Unekawa. 2013. A 10.4pJ/b (32, 8) LDPC decoder with time-domain analog and digital mixed-signal processing. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers.* 420–421.

[54] S. Levantino, G. Marzin, and C. Samori. 2014. An adaptive pre-distortion technique to mitigate the DTC nonlinearity in digital PLLs. *IEEE Journal of Solid-State Circuits* 49, 8 (2014), 1762–1772.

[55] A. Mantyniemi, T. Rahkonen, and J. Kostamovaara. 2009. A CMOS time-to-digital converter (TDC) based on a cyclic time domain successive approximation interpolation method. *IEEE Journal of Solid-State Circuits* 44, 11 (2009), 3067–3078.

[56] D. Miyashita, R. Yamaki, K. Hashiyoshi, H. Kobayashi, S. Kousai, Y. Oowaki, and Y. Unekawa. 2014. An LDPC decoder with time-domain analog and digital mixed-signal processing. *IEEE Journal of Solid-State Circuits* 49, 1 (2014), 73–83.

[57] C. Zhang, J. Gu, L. Gao, T. Ouyang, and B. Wang. 2017. Time-domain computing circuits for addition and multiplication computation. In *International Conference on Electron Devices and Solid-State Circuits (EDSSC'17).* 1–2.

[58] Sicong Liu, Yingyan Lin, Zimu Zhou, Kaiming Nan, Hui Liu, and Junzhao Du. 2018. On-demand deep model compression for mobile devices: A usage-driven model selection framework. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services.* 389–400.

[59] J. Kim, T. Jang, Y. Yoon, and S. Cho. 2010. Analysis and design of voltage-controlled oscillator based analog-to-digital converter. *IEEE Transactions on Circuits and Systems I: Regular Papers* 57, 1 (2010), 18–30.

[60] S. Rao, B. Young, A. Elshazly, W. Yin, N. Sasidhar, and P. K. Hanumolu. 2011. A 71dB SFDR open loop VCO-based ADC using 2-level PWM modulation. In *Symposium on VLSI Circuits — Digest of Technical Papers.* 270–271.

[61] J. Daniels, W. Dehaene, M. Steyaert, and A. Wiesbauer. 2010. A 0.02mm2 65nm CMOS 30MHz BW all-digital differential VCO-based ADC with 64dB SNDR. In *2010 Symposium on VLSI Circuits.* 155–156.

[62] P. Prabha, S. J. Kim, K. Reddy, S. Rao, N. Griesert, A. Rao, G. Winter, and P. K. Hanumolu. 2015. A highly digital VCO-based ADC architecture for current sensing applications. *IEEE Journal of Solid-State Circuits* 50, 8 (2015), 1785–1795.

[63] S. Rapuano, P. Daponte, E. Balestrieri, L. De Vito, S. J. Tilden, S. Max, and J. Blair. 2005. ADC parameters and characteristics. *IEEE Instrumentation Measurement Magazine* 8, 5 (2005), 44–54.

[64] B. Yan, C. Liu, X. Liu, Y. Chen, and H. Li. 2017. Understanding the trade-offs of device, circuit and application in ReRAM-based neuromorphic computing systems. In *IEEE International Electron Devices Meeting (IEDM'17).* 11.4.1–11.4.4.

[65] B. S. Amrutur and M. A. Horowitz. 1998. A replica technique for wordline and sense control in low-power SRAM's. *IEEE Journal of Solid-State Circuits* 33, 8 (1998), 1208–1219.

[66] Jingjia Meng, Bingyuan Zhao, Qiyun Xu, Jonathan M. Goodwill, James A. Bain, and Marek Skowronski. 2020. Temperature overshoot as the cause of physical changes in resistive switching devices during electro-formation. *J. Appl. Phys.* 127, 235107 (2020), 127.

[67] L. Kull, T. Toifl, M. Schmatz, P. A. Francese, C. Menolfi, M. Braendli, M. Kossel, T. Morf, T. M. Andersen, and Y. Leblebici. 2013. A 3.1mW 8b 1.2GS/s single-channel asynchronous SAR ADC with alternate comparators for enhanced speed in 32nm digital SOI CMOS. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers.* 468–469.

[68] T. Na, B. Song, J. P. Kim, S. H. Kang, and S. Jung. 2017. Offset-canceling current-sampling sense amplifier for resistive nonvolatile memory in 65 nm CMOS. *IEEE Journal of Solid-State Circuits* 52, 2 (2017), 496–504.

Received August 2020; revised November 2020; accepted February 2021