Reliability aspects of binary vector-matrix-multiplications using ReRAM devices

Bengel, Christopher; Mohr, Johannes; Wiefels, Stefan; Singh, Abhairaj; Gebregiorgis, Anteneh; Bishnoi, Rajendra; Hamdioui, Said; Waser, Rainer; Wouters, Dirk; Menzel, Stephan

DOI
10.1088/2634-4386/ac6d04

Publication date
2022

Document Version
Final published version

Published in
Neuromorphic Computing and Engineering

Reliability aspects of binary vector-matrix-multiplications using ReRAM devices

NEUROMORPHIC
Computing and Engineering

PAPER

Reliability aspects of binary vector-matrix-multiplications using ReRAM devices

Christopher Bengel[1,*], **Johannes Mohr**[1], **Stefan Wiefels**[2], **Abhairaj Singh**[3], **Anteneh Gebregiorgis**[3], **Rajendra Bishnoi**[3], **Said Hamdioui**[3], **Rainer Waser**[1,2,4], **Dirk Wouters**[1] and **Stephan Menzel**[2]

[1] Institute of Materials in Electrical Engineering and Information Technology II and Jülich Aachen Research Alliance (JARA)-Fit, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, Aachen, Germany
[2] Peter Grünberg Institute PGI 7, Forschungszentrum Jülich GmbH and JARA-FIT, Jülich, Germany
[3] Computer Engineering Department, Delft University of Technology, 2628 CD Delft, The Netherlands
[4] Peter Grünberg Institute PGI 10, Forschungszentrum Jülich GmbH and JARA-FIT, Jülich, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: christopher.bengel@rwth-aachen.de

Keywords: computing in memory, memristor, ReRAM, circuit design, compact modelling, vector-matrix-multiplication, dot-product

OPEN ACCESS

RECEIVED
23 December 2021

REVISED
31 March 2022

ACCEPTED FOR PUBLICATION
5 May 2022

PUBLISHED
22 June 2022

Abstract

Computation-in-memory using memristive devices is a promising approach to overcome the performance limitations of conventional computing architectures introduced by the von Neumann bottleneck which are also known as memory wall and power wall. It has been shown that accelerators based on memristive devices can deliver higher energy efficiencies and data throughputs when compared with conventional architectures. In the vast multitude of memristive devices, bipolar resistive switches based on the valence change mechanism (VCM) are particularly interesting due to their low power operation, non-volatility, high integration density and their CMOS compatibility. While a wide range of possible applications is considered, many of them such as artificial neural networks heavily rely on vector-matrix-multiplications (VMMs) as a mathematical operation. These VMMs are made up of large numbers of multiplication and accumulation (MAC) operations. The MAC operation can be realised using memristive devices in an analog fashion using Ohm's law and Kirchhoff's law. However, VCM devices exhibit a range of non-idealities, affecting the VMM performance, which in turn impacts the overall accuracy of the application. Those non-idealities can be classified into time-independent (programming variability) and time-dependent (read disturb and read noise). Additionally, peripheral circuits such as analog to digital converters can introduce errors during the digitalization. In this work, we experimentally and theoretically investigate the impact of device- and circuit-level effects on the VMM in a VCM crossbars. Our analysis shows that the variability of the low resistive state plays a key role and that reading in the RESET direction should be favored to reading in the SET direction.

1. Introduction

The need for new computing architectures to handle growing amounts of data in an energy-efficient manner has spurred efforts in utilizing novel memristive devices such as bipolar resistive switching devices based on the valence change mechanism (VCM) in these architectures [1–4]. VCM devices can be switched in a non-volatile manner between at least one high resistive state (HRS) and one low resistive state (LRS). The transition towards a lower resistive state is called SET while the opposite operation of increasing the resistance is called RESET. These devices are most often structured in a grid like manner in crossbar arrays. This structure allows for a direct mapping of certain mathematical operations such as multiplication and accumulation operations on the hardware [5–8]. More complex operations such as vector-matrix-multiplications (VMM) and matrix–matrix-multiplications can be built up from these basic operations. Neuromorphic algorithms such as deep neural networks heavily rely on VMM. However, vector and matrix operations are also relevant outside of the context of neural networks as they are widely used in benchmarks like the linear algebra kernel GEMM [9].

Previous works on the system-level have already demonstrated benefits of using these types of devices, showing energy improvements on the scale of a factor 1000 and latency improvements of a factor ∼60 [10–12]. However, these system-level analyses often assume idealised deterministic devices or only consider a completely stochastic undirected variability. As we will show in this work for VCM devices, the different device-level and circuit-level error sources are not completely stochastic but rather depend on the history of each device, the applied voltage or the interplay between device and circuit. This means that certain established error correction techniques will need to be adapted or new techniques will need to be developed to consistently reduce the error rates.

In this work, we study the VMM accuracy under the impact of programming variability, read disturb and read noise of the memristive devices. As the nomenclature of these device-level effects is not always clear, we provide here a short definition of our understanding of what these terms mean. A more detailed explanation of the physical mechanisms is provided in section 4. Programming variability describes the fact that programming both in the SET and RESET direction always shows a cycle-to-cycle (c2c) and device-to-device (d2d) spread in the resulting LRS and HRS states [13–15]. Read disturb describes the fact that over multiple readouts of a VCM device its resistance might drift from its initial value or even completely switch to the opposite state [16–18]. Read noise describes the fact that when reading out a VCM device the read-out current is not constant but will show random jumps whose height depends on the resistance value [19–22]. This noise is often referred to as random telegraph noise (RTN) [23–25]. To prevent confusion with the definition of RTN in other types of devices [26], however, we will refer to it as read noise. It has also been called program instability as it adversely affects write-verify algorithms [27].

Since all these variabilities have underlying physical mechanisms that arise from the devices themselves, it seems reasonable to also use device models that can capture these effects while simulating circuits based on VCM devices. VMMs using VCM cells are fundamentally analog operations, due to the resistive nature of the VCM cells, peripheral circuitry such as digital to analog converters (DACs) and analog to digital converters (ADCs) are required to co-integrate the crossbar arrays together with digital CMOS. This means that these peripheral elements can introduce additional errors during the VMM. Hence, in this work, while mainly focussing on the experimental and simulative description of device-level effects affecting the VMM, we will also consider the impact of the ADC using circuit-level simulations of the 1 transistor 1 resistor (1T1R) crossbar array together with DAC and ADC peripheral circuits.

Most analyses so far have focused on the impact of programming variability, read disturb and read noise on the performance of neural networks [21, 28]. In [29], the impact of different non-idealities such as line parasitics, read disturb, and read noise on the IMPLY logic was studied. Here, we focus on VMM as a mathematical operation since in neural networks it is not necessary to have a perfect accuracy or even a relatively good one to achieve convergence of the neural network. This means that the impact of non-idealities will not be as clear as when VMMs are treated as a purely mathematical operation.

In this work, using a multidisciplinary effort ranging from device-level experiments to circuit-level evaluation, we provide a comprehensive overview and understanding of relevant failure mechanisms and their individual effects on VMM. We have done an analysis of all three failure mechanisms by distinguishing them based on their impact on the VMM. They are classified as time-independent (programming variability) and as time-dependent (read disturb and read noise). The workloads that we are considering consist of a very large number of read operations per write operation (e.g. inference operations in artificial neural networks) or even writing the devices once and then never again. Therefore, the temporal variabilities read disturb and read noise become much more important than in use cases where read and write accesses are more balanced.

The remainder of the work is organized as follows, section 2 explains the experimental procedure explaining the device fabrication and measurement setups for single device and array characterization. Section 3 introduces the simulation setup introducing the device model and VMM architecture and section 4 deals with the device level error sources. Section 5 discusses error sources introduced by the peripheral circuitry and section 6 demonstrates the experimental results for the array-level characterization. Section 7 shows the simulation results of doing the VMM on the circuit-level and section 8 details the reliability challenges on this level. Section 9 concludes the paper.

2. Experimental procedure

In order to characterize these reliability aspects experimentally, we fabricated VCM ReRAM cells with a 30 nm Pt/5 nm ZrO_2/20 nm Ta/30 nm Pt stack. The ZrO_2 switching layer is deposited on a high work function (WF) or 'blocking' Pt bottom electrode (BE) via reactive RF sputtering. A low WF 'ohmic' Ta top electrode and an additional 30 nm Pt layer to prevent oxidation are deposited also by sputtering and are structured via a lift-off photolithography step. The Pt/ZrO_2 interface is then the electronically active electrode (AE) which means that it is the main originator of the resistance change, while the ZrO_2/Ta/Pt layers form the ohmic

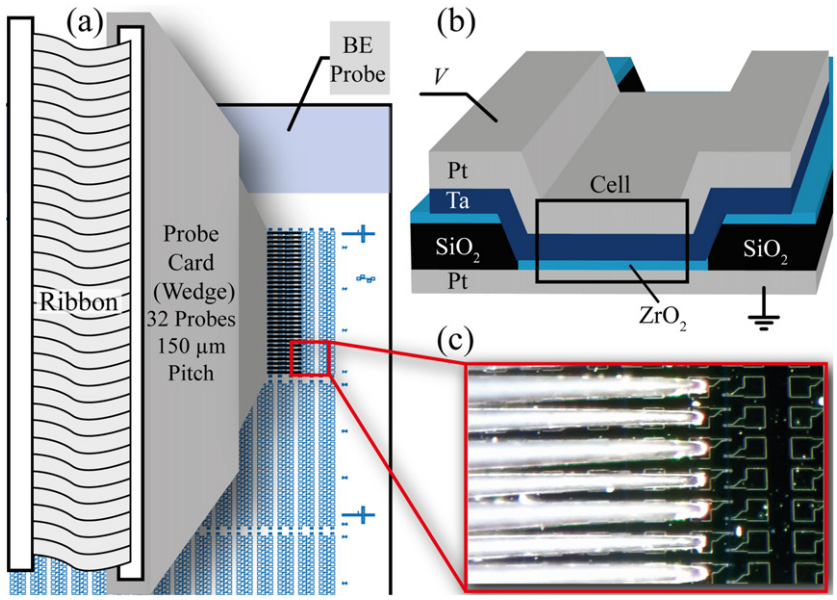

Figure 1. (a) Schematic of the experimental setup. (b) Device stack. (c) Microscopy of probes and sample.

electrode (OE). As illustrated in figures 1(a) and (c), the cells are arranged in a 32 wordlines by 1 bitline array structure comprised by 7 μm $\times$ 7 μm crossbar structures. The underlying Pt BE covers the whole substrate surface, but makes contact to the device only through an etched SiO_2 layer deposited on top via electron beam evaporation, as indicated in figure 1(b). This results in a comparatively low and homogenous series resistance along the array. Using a dedicated (wedge) probe card, all 32 cells within one line array are contacted. An additional probe connects the BE to the measurement system, which is a custom-built array tester based on the µcontroller module platform by aixACCT Systems. In all presented experiments, the Pt BE is set to GND and the bias is applied to the Ta top electrode. For comparison with our simulation models, however, all voltages in this paper are given with reference to 0 V at the ohmic Ta electrode resulting in negative forming and SET voltages and positive RESET voltages.

The experimental procedures comprise single cell as well as array mode operations. In both cases, an initial electroforming step is required. Here, a ramped voltage stress with a rise time of 2.5 ms towards a maximum voltage V_{FORM} is applied. According to a write-verify scheme, the cell resistances are measured after each electroforming attempt, by using a read pulse (20 μs at 0.2 V). The operation is repeated with $|V_{\mathrm{FORM}}|$ increased by 0.1 V until all cells are formed successfully or the maximum number of iterations is reached. Within this algorithm, V_{FORM} ranges between -1.8 V and -4 V. For the programming of single cells, a write-verify algorithm with 20 μs long square pulses ranging from 0.5 V to 2.5 V is used for the RESET procedure, while a negative polarity is used for SET.

For the array-level investigations, eight-bit words are considered, as using these all relevant multi-cell effects can be studied with a reasonable level of experimental effort. First, the desired pattern of logical '1's and '0's is stored, by applying the programming algorithm iteratively to all cells. The nominal LRS value (representing '1') was 3 kΩ, the nominal HRS value (representing '0') was 30 kΩ. Then, an eight-bit input vector is fed in by applying voltage pulses of V_{READ} or a high ohmic connection, corresponding to '1' and '0' bit values, to all wordlines simultaneously. To enable the characterization of noise, relatively long pulses ($\sim$2 s) are used. The measured bitline current represents the computation result. This procedure is repeated without re-programming for all 256 possible input vectors in section 6.1, as the focus is on showing all possible dot product results for a certain programmed pattern. In section 6.2, a worst-case input vector was used to study the impact of read disturb in more detail.

3. Simulation procedure

3.1. Device model

As a VCM device model, we used the physics-based compact model JART VCM v1b [30], which is written in Verilog-A. The compact model abstracts the device stack by splitting the ZrO_2 layer into a well conducting plug region and a disc region that modulates the cell resistance [31, 32]. The AE is described as a Schottky diode, while the OE is modelled through a series resistor. Another feature is the read noise extension [33] that

Table 1. Compact model parameters.

$A_{\mathrm{det}} = \pi {r_{\mathrm{det}}}^2 = 2.83 \times 10^{-15}$ m^2	$r_{\mathrm{det}} = 30$ nm
$N_{\mathrm{disc,min,det}} = 0.01 \times 10^{26}$ m^{-3}	$N_{\mathrm{disc,max,det}} = 150 \times 10^{26}$ m^{-3}
$l_{\mathrm{det}} = 0.8$ nm	$l_{\mathrm{cell}} = 5$ nm
$l_{\mathrm{plug}} = 4.2$ nm	$a = 0.25$ nm
$R_{\mathrm{TiO}x} = 50\ \Omega$	$\Delta W_{\mathrm{A}} = 0.85$ eV
$v_0 = 1 \times 10^9$ Hz	$\mu_n = 4 \times 10^{-6}$ m^2 V^{-1} s^{-1}
$R_{\mathrm{th0,SET}} = 1 \times 10^5$ K W^{-1}	$R_{\mathrm{th0,RESET}} = 8 \times 10^4$ K W^{-1}
$e\Phi_{\mathrm{Bn0}} = 0.52$ eV	$e\Phi_n = 0.1$ eV
$R_{\mathrm{th,line}} = 90\,471.47$ K W^{-1}	$R_0 = 50\ \Omega$
$\alpha_{\mathrm{line}} = 3.92 \times 10^{-3}$ 1/K	$m^* = 9.11 \times 10^{-31}$ kg
$e = 1.6 \times 10^{-19}$ C	$T_0 = 293$ K
$A^* = 6.01 \times 10^5$ A m^{-2} K^{-2}	$z_{\mathrm{vo}} = 2$
$k_{\mathrm{B}} = 1.38 \times 10^{-23}$ J K^{-1}	$\varepsilon_0 = 8.854 \times 10^{-12}$ As Vm^{-1}
$\varepsilon_{\Phi\mathrm{B}} = 5.5 \times \varepsilon_0$	$\varepsilon = 17 \times \varepsilon_0$
$h = 6.626 \times 10^{-34}$ Js	$N_{\mathrm{plug}} = 150 \times 10^{26}$ m^{-3}

Variability parameters

Symbol	Minimum/median/maximum
$N_{\mathrm{disc,min,var}}$ (10^{26} m^{-3})	0.005/0.01/0.02
$N_{\mathrm{disc,max,var}}$ (10^{26} m^{-3})	120/150/180
r_{var} (nm)	15/30/50
l_{var} (nm)	0.2/0.8/1.2
	Value
Relative standard deviation	1
c2c percentage	30%
Maximum step size	5%

models read noise by random jumps of single oxygen vacancies to and from the disc. The parameters used in this simulation are shown in table 1. While read noise is an effect that can only be observed in simulation if the corresponding model extension is used, read disturb is already a feature of the (basic) deterministic model. In the case of the deterministic model, read disturb would be deterministic as well, meaning that all devices would behave the same way during every evaluation cycle. The model parameters were adapted to describe the behavior observed in the different experiments better. For the simulations in section 4.2.3, we used $R_{\mathrm{th},0} = 4 \times 10^6$ K W^{-1} and $\Delta W_{\mathrm{A}} = 1.25$ eV to achieve a better agreement between experiment and simulation. For section 4.2.4, $\Delta W_{\mathrm{A}} = 1$ eV was chosen for the same reason.

3.2. VMM architecture

The ADC used in this work was proposed in [34]. It was designed with the aim of being small enough to fit in the column pitch of the 1T1R array. This would allow for providing one ADC per column, enabling an extremely parallel calculation of VMMs. A further focus of the design was that it should be tolerant to resistance variability. In [34] it was shown that it can tolerate programming variability of up to 30%. It is based on a voltage controlled oscillator (VCO), whose oscillation frequency depends on the data stored in the part of the crossbar that is read out. This is achieved by connecting the output of the 1T1R crossbar (bitline) to the supply voltage node of the VCO ($V_{\mathrm{DD,VCO}}$). Additionally, a diode connected NMOS transistor to ground is introduced at the common node of the crossbar and the ADC. The number of pulses that is generated by the VCO during a fixed evaluation time ($[0, t_{\mathrm{EVAL}}]$) is then counted using a JK flip-flop based synchronous counter. To improve the counter the output signal of the VCO (V_{out}) is amplified to V_{amp} which can be counted easier as it always oscillates between 0 V and 1 V instead of oscillating between 0 V and $V_{\mathrm{DD,VCO}}$. The schematic of the proposed circuit architecture is shown in figure 2. The row drivers are connected to the transistor gates (wordlines) and the column drivers are connected to the transistor contacts (sourcelines). The bulk contacts of the transistors are connected to ground. When considering the impact of device-level non-idealities, the circuits involved in the analog computing, i.e. row and column drivers, 1T1R crossbar and ADC are the most relevant. The used parameters to assess the effect of device non-idealities on the VMM are $V_{\mathrm{DD,VCO}}$, the number of pulses generated during the evaluation time and the digital output of the counter ($Q_0 \ldots Q_n$). We have introduced a few modifications over the original ADC design to improve the tolerance to variability, which however slightly increase the energy consumption and the area of the circuit. First, a signal restoration buffer stage was introduced at the output of the VCO to improve the signal shape and amplitude of the oscillating signal V_{out}. The buffered signal V_{amp} is then fed into a synchronous counter, which is clocked by V_{amp} instead of an asynchronous counter. The reason for using a synchronous counter is that it gives us a larger time window in which the output of the counter is valid. Further details on the design choices will be given in section 5. As the original design was simulated in a 28 nm technology by TSMC and the new design was simulated using

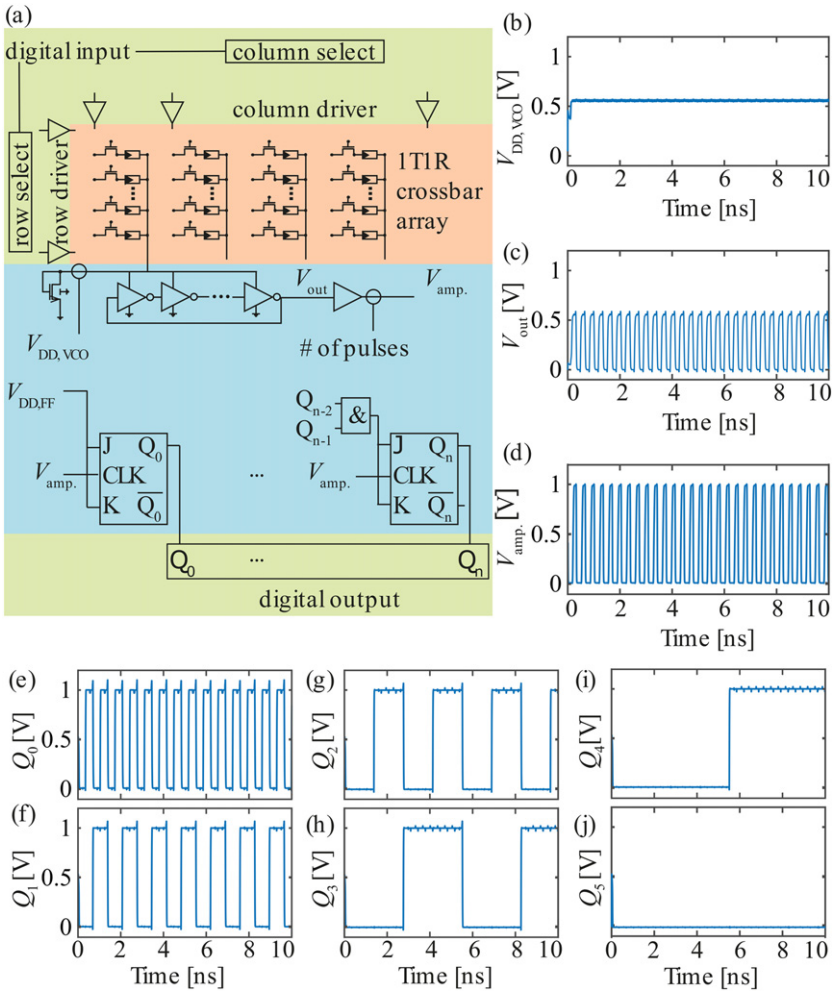

Figure 2. (a) Shows a schematic level overview of the proposed analog VMM architecture. The architecture consists of three parts, the digital select circuitry and output (green area), the analog 1T1R crossbar array (orange area) and the ADC circuit (blue area). (b) Shows the supply voltage of the ring oscillator, (c) shows the output signal of the VCO, (d) shows the amplified signal, (e)–(j) show the digital output of the counter from Q_0 to Q_5.

Table 2. ADC and crossbar parameters.

ADC parameter	Value	Crossbar parameter	Value
# of stages in ring oscillator	7	LRS	1.6 kΩ
V_{SL}	1 V	HRS	300 kΩ
V_{WL}	1.5 V	w_{1T1R}	250 nm
t_{EVAL}	5.1 ns	l_{1T1R}	32 nm

the 32 nm PTM [35] technology, a resizing of the circuit components was necessary. Table 2 shows the most important simulation parameters related to the ADC and the 1T1R crossbar.

4. Device-level error sources

Here, we show an analysis of all three failure mechanisms on a device-level. We distinguish them based on their impact on the VMM, where they are classified as time-independent (programming variability) and as time-dependent (read disturb and read noise). The time-dependant sources of variation can be grouped further by their influence on the device state. Since the random jumps in current due to read noise are equally likely to increase or decrease it, there is no tendency for a change in the average current level. On the other hand, read disturb is driven by an applied voltage, and therefore leads to a directional trend in the resistance levels. Depending on the polarity of the applied voltage (SET or RESET direction), and whether the affected cell is in the HRS or LRS, four cases can be distinguished which will all be investigated in section 4.2. For filamentary VCM cells, the change of the resistance is mainly connected with the movement of oxygen vacancies into or

out of the disc region, inside a metal-oxide filament. This movement is inherently stochastic between different devices as well as between multiple cycles on the same device.

4.1. Programming variability

One aspect of the programming variability is the stochasticity of the SET and RESET kinetics. It has been shown that one significant contributor for programming variability lies in the occurrence of multiple filaments in VCM cells [36]. While only one of those is the main contributor to the device resistance, this main filament can be a different one during multiple switching cycles. Each filament can have different radii as well as different oxygen vacancy concentrations. The variability model is also based on this assumption of varying filament geometries [30]. From device to device and from cycle-to-cycle, it is observed that the time required for successful switching at a given voltage can vary over orders of magnitude [37]. Furthermore, the initial resistive state has been shown to affect the switching kinetics [32]. In a previous study, a comprehensive model of this effect on array-level [30] was presented. Apart from the variability of the switching kinetics, the resistance state right after programming can vary over a broad range [38]. However, this programming variability can be addressed by program-verify algorithms as described in the experimental section.

4.2. Read disturb

During the read operation, a comparatively low voltage is applied to the device, which should ideally not affect the programmed state. Read disturb describes unintended changes of the resistive states during readout. In bipolar switching devices, both voltage polarities have to be considered. Reading in SET polarity might cause a resistance decrease in both LRS and HRS cells, whereas reading in RESET polarity might increase the resistance of a device in both the LRS or the HRS. This is because the mobile oxygen vacancies, whose concentration in the disc region determines the cell resistance, are two times positively charged and they will move into different directions depending on the voltage polarity. These effects are studied through experiment and simulation as discussed in the following subsections.

4.2.1. Reading a device in the HRS in the SET direction

In order to characterize the read disturb in SET direction experimentally, one line array (32 cells) was electroformed and each cell was switched 50 times. Then, the cells were programmed into an HRS resistance range of 15 kΩ to 25 kΩ. Subsequently, each cell was read by a voltage pulse with 20 k samples and a sampling rate of 0.8 MHz. In order to resolve the read disturb over seven orders of magnitude, the sampling rate was then reduced to 0.8 kHz and the read pulse was repeated five times. After that, all cells were programmed back into the HRS (15 kΩ to 25 kΩ) and the experiment was repeated several times to generate significant statistics. Using typical read voltages of −0.2 V to −0.35 V, only minor effects could be resolved within reasonable measurement times (10 s). To gain an estimation about the long term stability against read disturb, we repeated the experiment at elevated voltages (−0.5 V to −0.8 V). The resulting experimental resistance distributions are depicted in figure 3(a), (c), (e) and (g). It can be seen that the initial distribution broadens significantly over time. Furthermore, a tail of cells that flipped to the LRS evolves where the number of flipped cells increases over time. As expected, the effect is accelerated with increasing read voltage. Using the compact model to repeat the experiment under the same conditions (resistance range, applied voltages and pulse duration), a much larger statistic can be generated as shown in figure 3(b), (d), (f), (h). As in the experiment, the simulation also shows a broadening over time. There is first a broadening of the distribution due to read noise and with increasing time the read disturb leads to cells switching into the LRS.

From the data in figure 3(a), (c), (e) and (g), the SET time was extracted as the time when the cell resistance drops below 7 kΩ. As in this case the SET is unintended and appears at low voltages, we classify it as a read disturb in SET direction. The cumulative distribution of the extracted SET times is shown for the different read voltages in figure 4(a). Here, the distributions for −0.5 V and −0.6 V are observed to follow log-normal statistics. The left border of the distributions is slightly bended due to the limited resolution at short times ($<1.25\ \mu s$). As expected, higher voltages shift the distributions towards shorter disturb times and closer to the resolution limit. By applying log-normal fits to the distributions, a trend for different voltages and percentiles could be extracted, which is depicted in figure 4(b). The disturb time exponentially increases with decreasing read voltage. Using this observation, an estimation of the disturb behaviour at a typical read voltage of −0.2 V can be made by extrapolation. The dotted, horizontal lines indicate reading times of 1 year (black) or 10 years (gray). This suggests that a typical cell ($\sigma = 0$) is stable for 10 years at up to −0.35 V and the worst out of 1000 cells ($\sigma = -3$) is similarly stable if the voltage is limited to -0.1 V. Please note, that this only accounts for cells in an HRS resistance range of 15 kΩ to 25 kΩ. We would expect a higher stability starting from a higher resistance as it has been previously shown that a higher initial resistance state leads to a larger delay in the SET process [32].

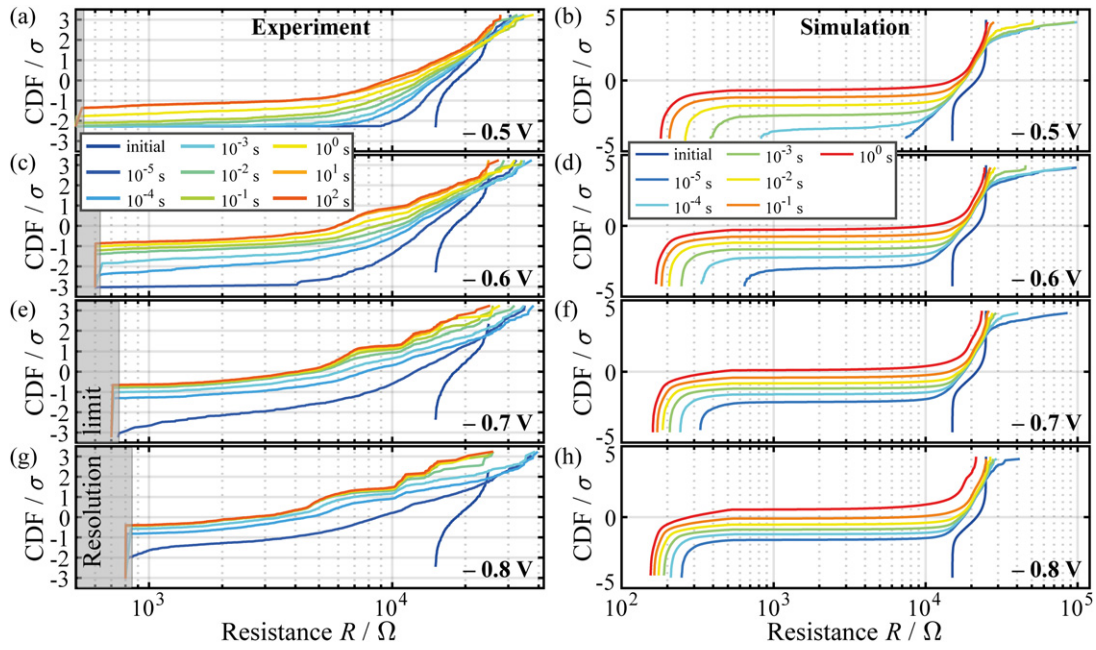

Figure 3. Experimental (a), (c), (e) and (g) and simulated (b), impact of four different read voltages applied in the SET direction on cells in an initial HRS of 15 kΩ to 25 kΩ. The distributions broaden over time and continuously shift towards LRS. Furthermore, a growing tail of flipped cells occurs. The effects scale with the applied (read) voltage.

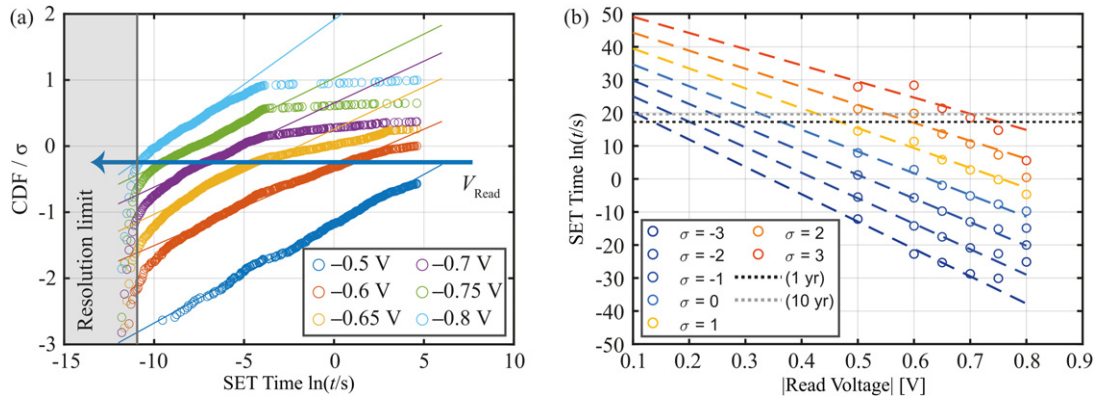

Figure 4. Read disturb phenomena in the SET direction. (a) CDF of the set time for different voltages. (b) Extracted SET time for different percentiles versus read voltage. An empirical, linear fit with respect to the logarithmic time axis is used in order to predict the behavior at lower voltages as those are more relevant for the VMM but are not easily accessible for the experiment.

4.2.2. Reading a device in LRS in RESET direction

Regarding the opposite direction, the read disturb behaviour is measured analogously. The cells are initially programmed into a range of 1 kΩ to 3 kΩ. Subsequently, elevated read voltages (0.5 V to 0.8 V) are applied in RESET direction. The resulting resistance distributions are shown in figure 5(a), (c), (e) and (g) for the experiment and figure 5(b), (d), (f) and (h) for the simulation. It can be seen that no flipped bits occur within the tested read duration of 10 s. In the RESET direction, the disturb behaviour is rather characterized by a broadening and continuous drift of resistance distributions towards higher resistances.

However, at 0.5 V all cells remained under 5 kΩ within 10 s. Even at 0.8 V, the worst cell only degraded to 8.5 kΩ. This means that the shift occurs rather in the range of LRS resistances.

4.2.3. Reading a device in LRS in SET direction

The state of a device in the LRS could be read with a voltage in SET polarity. This might lead to a further reduction in its resistance. To test this, a number of cells programmed to a 3 kΩ LRS were stressed with a series of read pulses of 2 s length each. Different amplitudes of the pulses were evaluated. For each voltage, the average of ten measurements is shown in figure 6(a). The values are presented as a percentage of the initial value for clarity. Error bars indicate the amount of variation between the measurements.

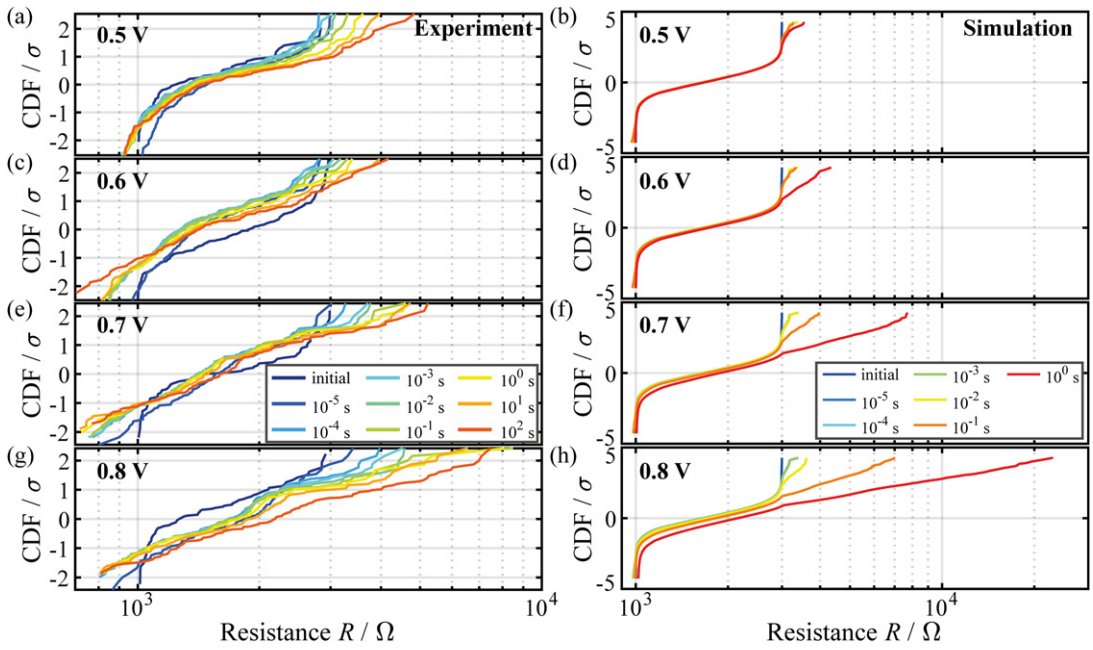

Figure 5. Experimental (a), (c), (e) and (g) and simulated (b), (d), (f) and (h) impact of four different, elevated read voltages applied in the RESET direction on cells in an initial LRS of 1 kΩ to 3 kΩ. No flipped cells are observed. The distributions broaden over time and continuously shift towards HRS. Compared to the SET direction, the impact of the read operation is significantly reduced at comparable voltages.

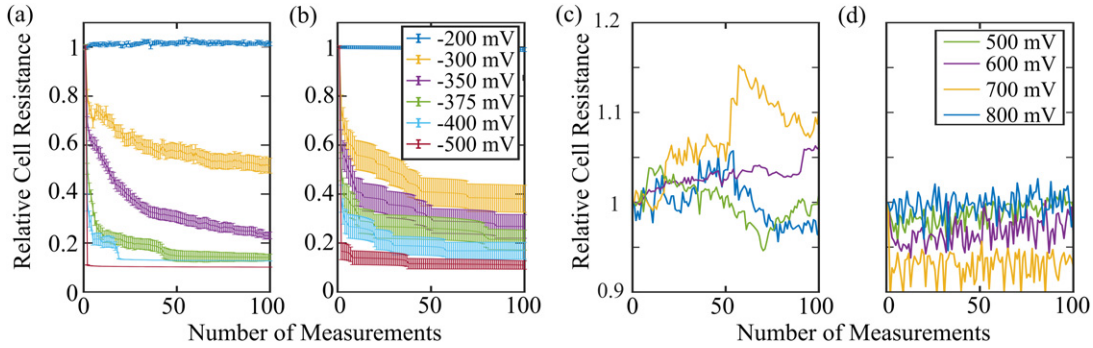

Figure 6. Experimental (a) and simulated (b) impact of reading devices in the LRS state in the SET direction at different voltages and experimental (c) and simulated (d) results for reading a device in the HRS in the RESET direction at different voltages.

Up to about −200 mV, the cell resistance remains stable for the entire measurement. For higher voltages however, it rapidly drops to 20%–30% of its original value. This transition becomes more abrupt for higher absolute read voltages, with the majority of the change occurring during the first few pulses, followed by a more gentle drift towards lower values. For −400 mV and −500 mV the resistance seems to approach a fixed resistance value. This value strongly depends on the series resistance connected to the device. Generally, this series resistance can be due to the measurement setup or other connected periphery, it can be due to the crossbar parasitic elements or it can be a built-in feature of the device as in the case of [38].

This behaviour can be understood from the device kinetics. It has been demonstrated that the switching time of VCM devices depends on the initial state [39]. The lower the resistance, the faster the switching for a given voltage. Consequently, the LRS can be disturbed by voltages well below V_{SET}, used for switching from the HRS. Since during the read operation no current compliance is active, the resistance decreases until the current is limited by the series resistance of the circuit. Therefore, the final resistance settles, and is somewhat lower for higher V_{READ}, as seen in figure 6(a). Due to this limitation by external resistances, the intrinsic variability of the devices plays less of a role, as evidenced by the reduced variability between measurements. Simulation results produced with the compact model (figure 6(b)) demonstrate that it captures the essential features well. At −200 mV only a minor drift is observed, while −325 mV leads to a halving of the devices resistance. The

Table 3. Summary of the different types of read disturb.

Device state/reading direction	Relevance and impact
LRS/SET	Relevant above $\|V\| > 200$ mV further decrease of the device resistance
LRS/RESET	Relevant above $\|V\| > 500$ mV small drift observable
HRS/SET	Relevant above $\|V\| > 100$ mV drift of the resistance state towards the LRS, bit flips possible
HRS/RESET	No effect found up to 800 mV

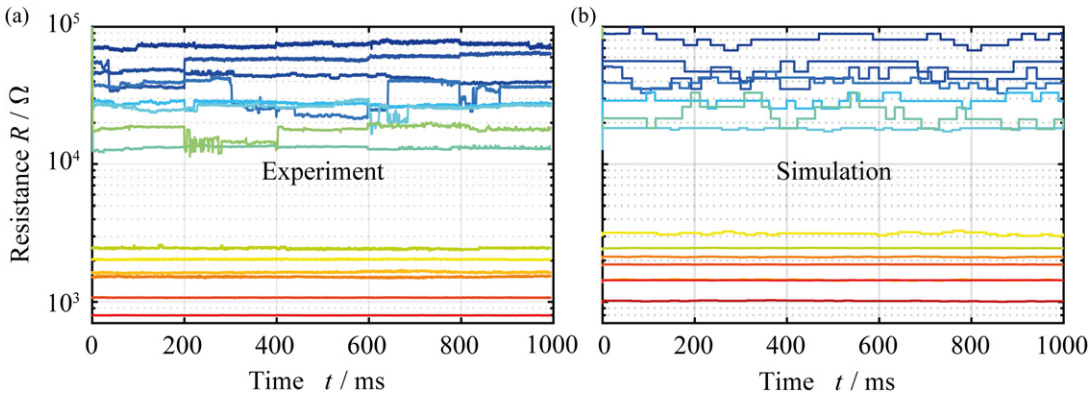

Figure 7. (a) Experimental read traces in different resistance regions at a constant voltage of −0.35 V. Read noise is observed in HRS whereas the LRS states are comparatively stable. (b) Representative simulated read noise curves in different resistance regions.

resistance settles at about 10% for −500 mV. Note that the reduced variability for this state is correctly reproduced. In general, the simulated variation appears to be larger. This is due to the fact that a much larger number of experiments can be simulated than physically performed.

4.2.4. Reading a device in HRS in RESET direction

Of course, it is interesting to investigate if a similar drift exists in the opposite direction. To evaluate this, cells were programmed to a 30 kΩ HRS and different voltages with RESET polarity were applied. Besides this, the same procedure was followed as in the previous case.

The resulting normalized resistances are shown in figure 6(c, d). Over the course of the measurement, the resistances deviate from the initial value by about 5%–10%, with no obvious trend towards higher or lower values. The applied voltages do not seem to have a significant influence on the magnitude of the deviation. This indicates that for the HRS the behaviour is dominated by the read noise, leading to a random distribution of the observed resistances, rather than a voltage driven drift. While in [40] it has been reported, that a SET disturb could occur when applying multiple RESET pulses, this was not observed in this study. One reason for this might be the different device dimensions as the devices used in this study have a cross sectional area of 49 μm^2 range while the devices in [40] were in the range of a few hundred nm^2. Larger device sizes could lead to more stable and relaxed vacancy distributions when compared to smaller device cross sections where the volume for vacancy redistribution is much smaller.

In conclusion, read disturb has to be considered as reliability aspect for VMM applications. In order to determine the tolerated maximum read voltages, we recommend to characterize the read-disturb characteristics of the devices to be used comprehensively. If possible, reading in RESET direction should be preferred since the read-disturb effects generally appear at higher absolute voltages. In section 6, we will relate these findings to the dot product calculation and parallel readout. Our findings are summarized in table 3. We would like to emphasize that these findings are only valid for the resistive ranges that we have considered so having different LRS or HRS states might lead to different results. However the general differences between reading in the SET and reading in the RESET direction should still hold as they are a consequence of the device physics themselves.

4.3. Read instability

Read noise, often also called RTN, refers to random jumps of the read current in a frequency range between 10^0 and 10^4 Hz. It has been attributed to either the charging/activation and discharging/deactivation of traps [21, 41], or the random redistribution of oxygen vacancies in the oxide layer [33]. The amount of read noise was found to be stronger for higher resistance states [28, 33, 41], as well as for increased temperatures and read voltages [19]. Experimentally obtained read noise characteristics are demonstrated in figure 7(a). Shown is the measured resistance of several cells in HRS and LRS, recorded with a constant voltage pulse of −0.35 V for 1 s.

It can be seen that the HRS states (10 kΩ to 100 kΩ) exhibit significant jumps whereas the LRS (< 5kΩ) is comparatively stable.

In a previous study, it was demonstrated that these jumps lead to a stochasticity of the programmed state according to an intrinsic distribution, where its width typically exceeds the programming variability, if an appropriate program verify algorithm is used [33]. Further, it was shown that these jumps can be grouped based on different jump heights and identified jumps of single vacancies at different locations within the conducting filament as origin [33].

The effect of the resistance state on the amount of read noise is a feature of the used compact model as demonstrated in figure 7(b). The resistances were evaluated at −0.1 V applied to the AE while the OE was grounded. As expected, the amount of read noise is proportional to the resistance range with higher resistances leading to higher amounts of read noise. The compact model here only considers the large jumps between plug and disc.

5. Peripheral circuitry

For the VMM operation, an ADC is required in addition to the 1T1R crossbar array and the input drivers for VMM. This ADC can be an additional source of errors, which will be investigated in the following. As the used ADC is based on counting pulses generated by a ring oscillator during a fixed time interval, we will focus on errors associated with this functionality. A Monte Carlo analysis was already shown in the original publication [34]. To study possible error sources of the ADC, we simulated the complete VMM system as shown in figure 2 using a deterministic device model. In this way, we can investigate the VMM in a complete system while neglecting device errors. Additionally, for the analysis in this chapter, we only consider one evaluation cycle, meaning that we can neglect read disturb and all errors are induced by the ADC.

The evaluation process of the ADC is split into three parts. First, the bitline current coming from the crossbar is converted into a supply voltage for the VCO. While the crossbar delivers a constant current, if the devices are deterministic, the supply voltage nevertheless shows small oscillations (a few tens of mV) as the load represented by the ring oscillator changes. As a ring oscillator consists of an odd number of inverters $\#m$, in one cycle $\left\lfloor \frac{\#m}{2} \right\rfloor$ inverters have their output connected to $V_{\mathrm{DD,VCO}}$ and in the other cycle $\left\lceil \frac{\#m}{2} \right\rceil$ inverters. The brackets with the edges at the bottom represent a rounding to the nearest smaller integer, while the brackets having an edge at the top represent a rounding towards the nearest higher integer. Therefore, this oscillation is linked to the functionality of the ADC itself and is not an error. In the second step, the supply voltage is converted by the ring oscillator into a pulse train, which oscillates between zero and $V_{\mathrm{DD,VCO}}$. To simplify the counting, we amplify this voltage through the use of a buffer amplifier made up of two inverters resulting in the signal V_{amp}, oscillating between 0 V and V_{DD}. In the third step, this pulse train is fed into the clock inputs of a negative-edge triggered JK flip-flop based synchronous counter. The process of counting the pulses can thereby show several errors, some of which are more generic and some of which are specific and/or worsened by the application of doing VMM with resistive devices. The general errors that can occur are that the pulse train is too fast for the counter, which can lead to miscounting. In the VMM application, this can, however, be mitigated by looking at the configuration that gives us the fastest pulse train. This configuration is found when all read out devices are in the LRS and all transistor gates are selected. This means that the circuit is designed to be fast enough to count at this frequency, or, alternatively that this case is made slower by, for example, increasing the LRS of the VCM cells or increasing the resistance of the 1T1R select transistors. More specific errors are that pulses are too close to t_{EVAL} and that t_{EVAL} happens at the same time as the state change of the counter. Both errors are related to finding a suitable value for t_{EVAL}. Further, the different resistance configurations of the crossbar and the different number of activated wordlines lead to different oscillation frequencies, which means that the frequency of the counter becomes a variable value as well.

In the original publication [34], it was suggested to use a self timing path (STP) to determine this time point as the STP would contain the same type of variability as the evaluated array part. However, for the types of variability that we are considering here, this is not true or cannot be assumed in general. For instance, read noise stems from a stochastic process inside each device, which makes it unlikely that there will be a correlation of the amount or direction of read noise between multiple devices. Additionally, read disturb is a cumulative voltage dependent process, meaning that the amount of read disturb a device expresses depends on the number of read operations it has seen and on the voltage dropping over the cell. As the number of read operations is not known for individual devices in the array, it cannot be assumed that there will be a correlation of the read disturb of devices in the STP and of the arbitrary read out devices in the array. One way of mitigation would be wear levelling [42, 43]. Regarding read disturb, it should also be noted that the voltage dropping over a certain device depends on the state of the devices being read out in parallel, as well as the number of activated wordlines. This would even be a challenge for wear levelling. So, new ideas might be necessary. In summary, the physical nature of read noise and read disturb makes it difficult to mitigate variability based on

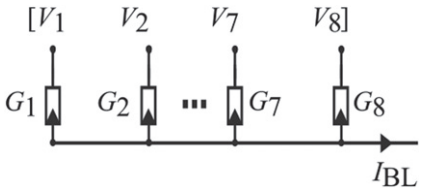

Figure 8. Shows the corresponding equivalent circuit diagram for the measurement setup on the passive arrays. A zero in the input vector is realized as an high ohmic output of the measurement system.

these effects through an STP necessitating a deeper understanding of the impact of various reliability effects on VMM accuracy.

6. Experimental eight-bit dot-product results

Experimentally, the dot-product operation is validated on a selected subset of the possible conductance vectors, as testing all possible combinations of input voltages (2^8) and resistance states (2^8) would result in $2^8 \times 2^8 = 65.536$ possible operations. As we want to focus here on the impact of device-level non-idealities, we specifically selected our measurements to showcase these. Fortunately, it can be assumed that the order of the '1's in the conductance vector is irrelevant, as due to the large area of the BE the series resistance of the bitline can be neglected, and therefore, all cells occupy equivalent locations in the circuit. Only their number and whether they are paired with a '1' in the input vector determines the current flowing through the bitline, corresponding to the computation result. This is of course the behavior expected from the mathematical definition of the dot-product. For clarity, the bitline current value can be rewritten as

$$\begin{aligned} I_{\mathrm{BL}} &= G_{\mathrm{LRS}} \sum_{\{i|G_i=G_{\mathrm{LRS}}\}} V_i + G_{\mathrm{HRS}} \sum_{\{i|G_i=G_{\mathrm{HRS}}\}} V_i \\ &= G_{\mathrm{LRS}} \cdot V_{\mathrm{READ}} \cdot d + G_{\mathrm{HRS}} \cdot \left(\sum_i V_i - d \cdot V_{\mathrm{READ}} \right), \end{aligned} \tag{1}$$

where d is the numerical result of the dot product and G_{LRS} and G_{HRS} are the target conductance values of the states.

Even though it is not strictly necessary, for the input vector V, all possible combinations were considered. This was realized by incrementing an eight-bit integer and using its binary representation to apply the voltages V_{READ}/high ohmic to the cells. Figure 8 schematically shows the measured passive line array configuration. It should be noted, that while a '1' in the input vector is encoded as V_{READ} a zero has to be realised as a high ohmic output of the measurement setup, as applying 0 V would lead to parallel current paths.

6.1. Effect of LRS and HRS variability on the dot product

Figure 9 shows the results of these dot product measurements. To demonstrate the deviations from the ideal case, each measured current is plotted against the value expected for this input pattern, given the nominal resistances for HRS and LRS. V_{READ} was chosen as -0.2 V. As shown in figures 3 and 6, this voltage amplitude is low enough not to cause read disturb over the duration of the measurement, independent of the device state. The solid black line has a slope of one and therefore ideally all points should lie on this line. The slight vertical deviations are mainly attributed to the programming variability. The distribution of measured bitline currents is also plotted as a histogram on the right side of (a)–(c). The different colours represent different dot product results.

Figure 9(a) shows the resulting I_{BL} for a case with two '1's in the conductance vector. The allowed variability of the LRS and HRS during the write verify operation was 10%. From the histogram it can be seen that the measurements fall in one of three groups of currents. The means of these groups are separated by roughly 50 μA. For each group, there is an additional spread of about 40 μA. These features can be understood from equation (1). The major contributions to the current come from the first term, as $G_{\mathrm{LRS}} \approx 10\ G_{\mathrm{HRS}}$. These represent the '1's in the conductance vector. Since in this example there are two LRS cells, the summation can result in zero, one or two $V_{\mathrm{READ}}{}^{*}G_{\mathrm{LRS}}$ contributions. This leads to the three groups of currents, each with a mean current of $I_{\mathrm{LRS}} = G_{\mathrm{LRS}}{}^{*}V_{\mathrm{READ}}$ is higher than the previous one. They correspond to the three possible dot product results (0, 1, 2) for this resistance state vector. The spread within each group is caused by currents through the selected HRS cells, described by the second term in equation (1). These contributions are much

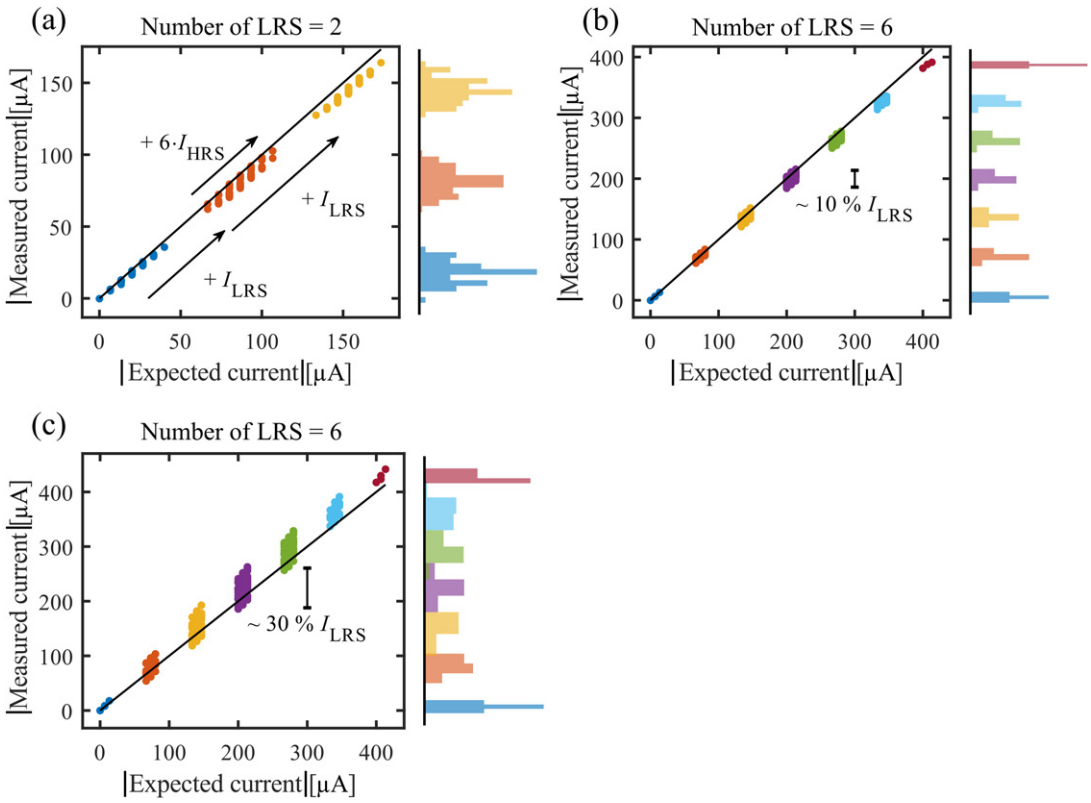

Figure 9. Dot product results for different programmed resistance patterns consisting of eight VCM devices under all possible input voltage vectors. (a) Shows the results for two devices in the LRS and 6 devices in the HRS state. (b) and (c) Shows the results for six devices in the LRS and two devices in the HRS state. The LRS variability for (a) is 10% for (b) its 10% and for (c) its 30%.

smaller ($I_{HRS} \approx 0.1\ I_{LRS}$), but because there are six HRS they can add up to a contribution in the same order of magnitude as those from one LRS cell.

An additional contribution to the vertical spread in the groups is due to the finite programming accuracy. The measured I_{BL} then depends on which cells are used to realize the dot-product result. For example, the result $d = 1$ can be realized with each of the cells in LRS. In principle, the variation in the HRS cells also has an influence, but due their small current contribution, this is less significant.

In figure 9(b), the measurement is shown for the inverted pattern of the conductance vector. Six cells are programmed to the LRS, leaving two in the HRS. As expected, there are now seven main groups of I_{BL}, corresponding to the possible dot-product results zero to six. Note that the groups corresponding to zero to two are still at the same current level as in the first case. The spread within each group is now much smaller as there are only two HRS cells that can at most lead to an additional current of 2 I_{HRS}. In both cases, all groups are clearly separated. For a larger programming variability, this may no longer be possible. To illustrate this, the same configuration of the conductance values is shown when programmed with a tolerance allowing for up to 30% deviation in the device resistances figure 9(c). As expected, some distributions are spread out vertically much more than in the previous case, mainly the central ones. These correspond to the dot-products that can be realized with many different combinations of inputs and subsets of the LRS cells. Due to the large programming tolerances, the resulting I_{BL} will differ significantly between these. Less affected are the higher dot-products, as the number of possible realizations is low. The same applies for the result 0, for which no LRS cells are involved.

6.2. Read disturb in parallel operation

To characterize the influence of prolonged read stress on the dot-product reliability, the array was programmed to a state with four LRS bits and four HRS bits. Then, a read operation was performed with an input vector containing only '1's, therefore applying voltage pulses to all cells simultaneously. I_{BL} then represents the effective resistance of all cells in parallel. This was repeated 100 times. A range of different amplitudes and polarities for V_{READ} was tested.

Figure 10(a) shows the resulting resistances for the SET direction. In the left plot, the transient response during the first read pulse is shown, whereas the average values for all following pulses are shown on the right. Clearly, a low read voltage of −300 mV leaves the resistance largely unchanged, apart from a small drift

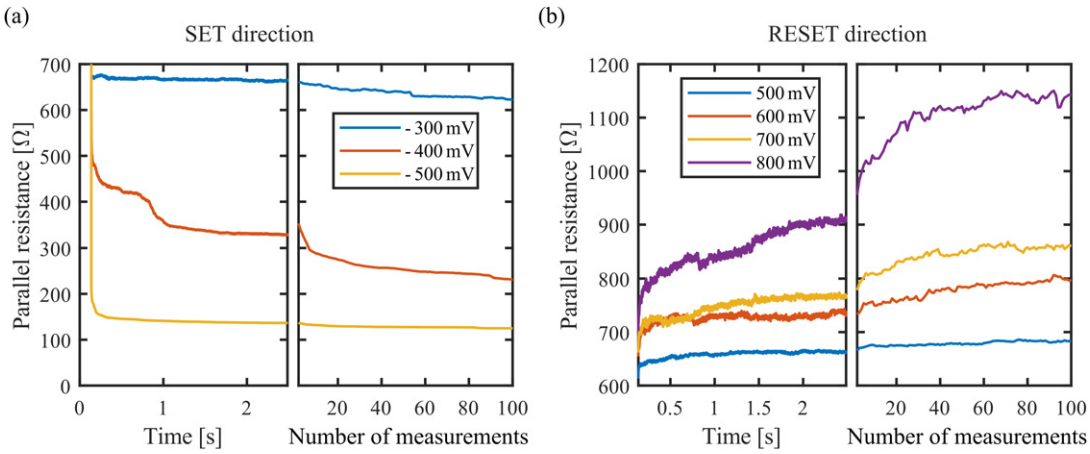

Figure 10. Experimental read disturb at different voltages in SET (a) and RESET (b) direction. One measurement takes 2.5 s.

towards lower resistances. For higher absolute voltages, however, a significant drop appears. The majority of this happens rapidly already at the beginning of the first pulse. This change is not due to a switching of some HRS cells to LRS, as might be expected, but due to a further drop in resistance of the LRS cells as shown in figure 6. There, a voltage above −300 mV leads to a large drop in resistance in a few seconds. For the read disturb from the HRS state on the other hand, a typical SET time on the order of years was found for this voltage. Therefore, in an array application, the impact of read disturb on LRS devices is the limiting factor. In the case with the read voltage applied in RESET polarity (figure 10(b)), a much smoother drift in resistance towards higher values is seen. This is caused by an increase in the LRS resistances at voltages above ∼500 mV. The read disturb measurements for single cells confirm this view, as a moderate increase in resistance of a few kΩ was observed. Considering the effect of a RESET direction voltage on HRS devices, as shown in figure 6, the results indicate no significant drift for this voltage range. Therefore, in this case too, the LRS cells are the limiting factor. Regarding the reliability of the dot-product computation, these results indicate that it is advisable to read in the RESET direction, to minimize the disturbance of the programmed states.

7. Vector-matrix-multiplication

The usual way of how VMM using resistive devices is defined in equation (2) for the case of an 8 × 8 matrix. In this case, the voltage input vector $\boldsymbol{V} = (V_1 \ldots V_8)$ and the current output vector $\boldsymbol{I} = (I_1 \ldots I_8)$ both have eight components, while the 8 × 8 matrix $\boldsymbol{G}$ is represented by the VCM device conductances.

$$\boldsymbol{I} = \boldsymbol{G} * V \Leftrightarrow$$

$$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_7 \\ I_8 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} & \ldots & G_{17} & G_{18} \\ G_{21} & G_{22} & \ldots & G_{27} & G_{28} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ G_{71} & G_{71} & \ldots & G_{77} & G_{78} \\ G_{81} & G_{82} & \ldots & G_{87} & G_{88} \end{bmatrix} * \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_7 \\ V_8 \end{bmatrix}. \tag{2}$$

The component I_1 of the result vector is then calculated as

$$\begin{aligned} I_1 = {} & G_{11} * V_1 + G_{12} * V_2 + G_{13} * V_3 + G_{14} * V_4 \\ & + G_{15} * V_5 + G_{16} * V_6 + G_{17} * V_7 + G_{18} * V_8, \end{aligned} \tag{3}$$

or more generally as

$$I_i = \sum_{j=1}^{8} G_{ij} * V_j, \tag{4}$$

where each component of the result vector is the dot-product between input voltage vector and corresponding row of the conductance matrix which is produced in one evaluation phase of a single ADC. In our case, as both the vector components and the matrix are binary, the resulting multiplication of vector element and matrix element will mathematically be zero if one or both of the multiplicands are zero and one if both the

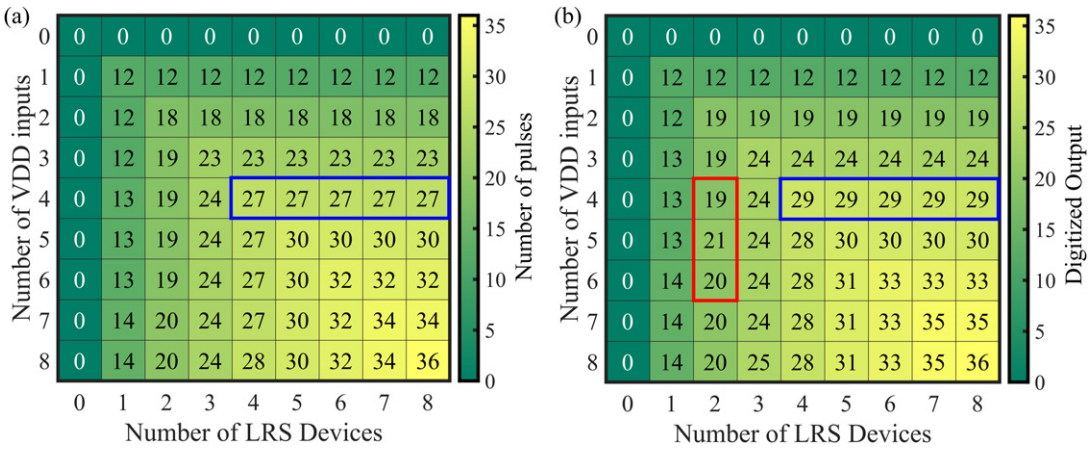

Figure 11. Number of generated pulses (a) and digitized output (b) for all possible combinations of the eight-bit voltage input vector and the eight-bit resistance column.

multiplicands are one. In the case of the voltage input vector, zero volt represents a mathematical zero and VDD represents a mathematical one, while in the case of the 1T1R matrix elements the mathematically zero is represented by the HRS of the VCM cell and the one is represented by the LRS of the VCM cell.

7.1. VMM in 1T1R arrays

To map this operation onto hardware, we apply the voltage vector ($\boldsymbol{V}$) to the different rows of the 1T1R crossbar. This implies, that one row of the conductance matrix ($G_{i,1\ldots8}$) is stored in the devices of one column in the 1T1R crossbar. The elements of the current output vector (I_i) are then produced column-wise and fed into the ADC. Following this definition of the VMM, each dot product operation can produce $n+1$ different results for n-bit vectors.

To study the analog VMM operation, the circuit structure from figure 2(a) was simulated using the deterministic version of the JART VCM compact model with the parameters from table 1.

Figure 11(a) shows the number of generated pulses and figure 11(b) shows the digitized output of the counter for all possible combinations of input voltage vector and conductance matrix. Figure 11(a) is calculated by counting the signal crossings of V_{amp} in software. The colour bar indicates the number of counted pulses. Essentially, the result of the dot product is determined by the minimum of the number of ones in the input vector and the number of LRS devices in the selected rows (dot product = min{number of V_{DD} inputs, number of LRS devices}). This is due to the special initialisation of the stored resistances and applied voltage vectors which was needed since it was not feasible to simulate all $2^8 \times 2^8 = 65.536$ different cases. As in section 6 we neglected the position of the '1' and '0' in the stored and input vector. For example in the case of four devices in the LRS state the top four cells were set to the LRS and having three V_{DD} inputs means that the first three voltage inputs were V_{DD}. As can be expected the number of generated pulses, as well as the digitized output increase with a larger number of V_{DD} inputs and with a larger number of devices in the LRS. As explained in section 5, the digitized output does not have to be the same as the number of pulses, e.g., because signal crossings appear too close to t_{EVAL}. This will lead to a larger number of pulses compared to the digitized output. Nevertheless, figure 11(b) shows that the ADC can differentiate the different dot products from each other since the difference in the digital outputs between different dot product results is at least one. In figure 11(a), the increase in the number of pulses due to devices in the HRS being selected can be observed, e.g. in column two in the rows two to nine. This is also observed in the experimental results in figure 9(a). In figure 11(b) we can also see some of the errors discussed in section 5, e.g. from column three row five to row seven (red rectangle) the digitized output increases first from 19 to 21 and then decreases back to 20. This is due to the fact that the counter switches its state very close to the evaluation time t_{EVAL} for this dot product result. Going from row five to row six, one additional input is set to V_{DD}, which means that an HRS is selected (as in this column there are only two devices in the LRS). This slightly increases the oscillation frequency of the VCO to a point, where at t_{EVAL} the first output bit of the counter Q_0 is still one and has not changed to zero as it is supposed to do at the transition from 19 ($Q_4Q_3Q_2Q_1\boldsymbol{Q_0} = 10011$) to 20 ($Q_4Q_3Q_2Q_1\boldsymbol{Q_0} = 10100$). This type of error leads to the digitized output being one larger than the number of pulses. If this type of error appears at the second bit position, the digitized output is larger by two. This can be seen in the blue rectangle in the fifth row in the columns five to nine between figures 11(a) and (b).

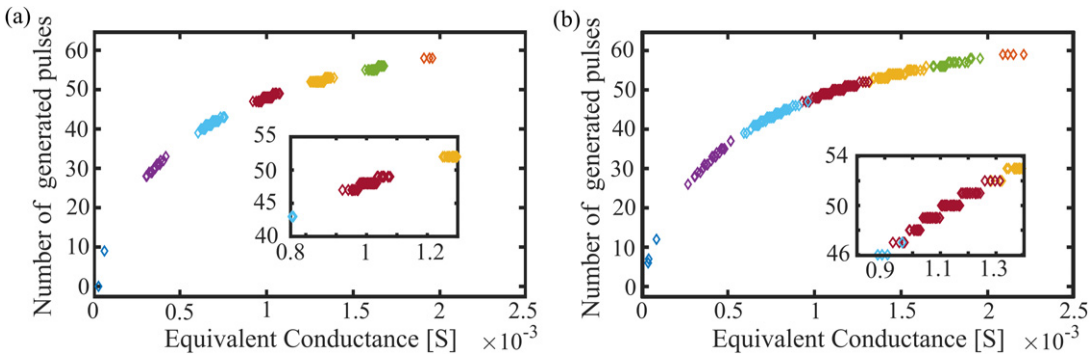

Figure 12. The number of generated ADC pulses as a function of the equivalent conductance of the crossbar for a LRS programming variabilities of 10% (a) and 30% (b) from figures 9(a) and (b).

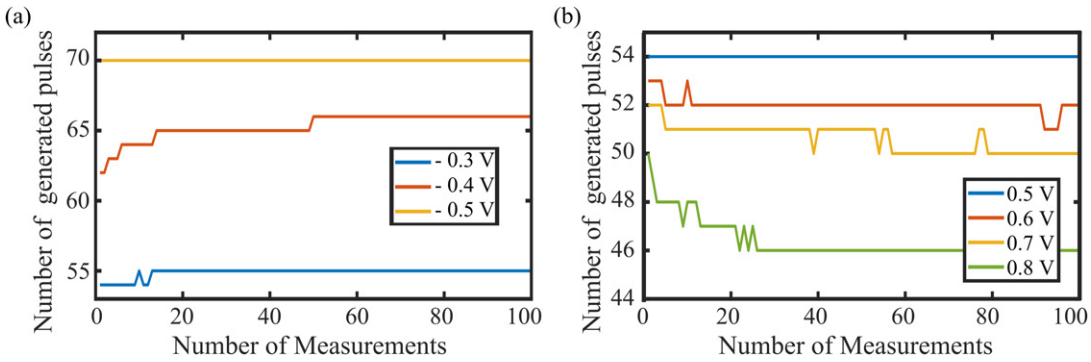

Figure 13. Shows the number of generated pulses over multiple measurements at different read voltages in the SET direction (a) and RESET direction (b). Each measurement is 2.5 s long.

8. Device-level effects on the VMM

8.1. Effect of LRS and HRS variability on the VMM

To study the impact of device non-idealities, we utilised the measurement results from section 6. As a first step, we studied the effect of different amounts of LRS variability. This was done by calculating the equivalent conductors conductances of the measured currents from figures 9(b) and (c) and then by replacing the 1T1R crossbar with these equivalent conductances. For these measurements, six devices were programmed to the LRS and two devices to the HRS. The design of the ADC was the same as in table 2. The resulting number of generated ADC pulses can be seen in figures 12(a) and (b). As shown in figure 11(b) the digitized output of the ADC can vary from the generated number of pulses due to different ADC related reasons as explained in sections 5 and 7. Therefore, to focus on the impact of the device effects on the VMM, we use here the number of generated pulses as key figure of merit as it is not affected by ADC counting errors. The different colours in figures 12(a) and (b) indicate different dot product results of the measured currents. One important feature is the nonlinearity that can be observed in figure 12 when compared with the linearity of the currents. This was also shown previously in [34] and is due to the different transfer functions of 1T1R crossbar and ADC [44, 45]. While the non-overlapping currents from figure 9(a) can also be distinguished at the stage of the ADC in the number of generated pulses, the overlap in the measured currents in figure 9(b) also translates into an overlap in the number of generated pulses in figure 12(b). It should be noted, however, that the maximum overlap for different dot products is only one, as can be seen in the zoom in, which means that the results still represent a good approximation.

8.2. Effect of read disturb and read noise on the VMM

The effects of read disturb and read noise are investigated using the same method as in section 8.1. Here, we utilize the measurements from figure 10, which represent read disturb measurements of four devices in the LRS and four devices in the HRS at different voltages. Each measurement was ∼2.5 s long, which represents 500 000 000 read cycles at an evaluation time of 5.1 ns. Again, we plotted the number of generated pulses to divide the ADC counter errors from the device errors. For both the SET direction, figure 13(a), and the

RESET direction, figure 13(b), the change over time increases at higher read voltages and the biggest change happens already in the first cycle. The different colours represent the mean behaviour at different voltages, while the bands represent the edges of the single device behaviours at the corresponding voltages. As can be seen from figures 10(a) and (b), at small voltages the drift is relatively weak (SET direction) or almost non-existent (RESET direction). For the RESET, however, the cells are stable for much higher voltages when compared with the SET direction. For higher voltages, the largest change already happens during the first measurements, which explains the different initial numbers of pulses.

To summarize, if the voltages are kept small enough the impact of read disturb can be more or less neglected on the timescale that was considered here (∼250 s). These safe voltage ranges are below 300 mV for the SET direction and below 500 mV for the RESET direction.

9. Conclusion

In summary, our paper shows the effect of device-level non-idealities of VCM based ReRAM devices on the accuracy of the VMM on a device and circuit-level considering binary devices and binary inputs. It was found, that the LRS variability plays a larger role than the HRS variability when reading out multiple devices in parallel. As the LRS variability can be well controlled through the current compliance of the series transistor in the 1T1R structure, this suggests the possibility of reading out more devices in parallel. Additionally, we could show that it is reasonable to read with the RESET polarity as this leads to a much slower and more gradual change of the resistance state when compared to reading in the SET direction. Moreover, this drift only appears at much higher voltages.

Regarding the optimisation of the energy efficiency of the VMM operation, small voltages are of course favourable as they result in lower currents. As we have shown here, for the case of read disturb, smaller voltages will also lead to a significantly longer resistance stability, due to the non-linear switching kinetics. Therefore, as long as the signal to noise ratio can be kept at an acceptable level, the read voltages should be reduced as far as possible. From an energy point of view, the LRS would need to be increased the most as devices in the LRS are the main current contributors. Since the resistance ratio of HRS to LRS needs to be kept at a certain level when multiple devices are read out in parallel this will require the HRS to be increased as well. As a rule of thumb, the smallest HRS state divided by the number of devices in parallel must be larger than the highest LRS. Of course, increasing the HRS to LRS ratio simplifies the analog to digital conversion as the current contribution of the devices in the HRS can be neglected ideally. Using higher resistances also has disadvantages as they generally require higher voltages to program, show a higher read noise and exhibit larger programming variability of their resistance distributions.

An important question for future work will be the impact of scaling up the array sizes. Larger array sizes will lead to larger parasitic elements, especially series resistances, which will impact the programming of devices based on the position of a device in an array, i.e. how far it is placed from the driving and read out circuitry. One likely impact of this on the circuit level is that the resistance window between LRS and HRS will close as the parasitic resistances represent a series resistance being added to LRS and HRS. Parasitic capacitances will also play a role as they will affect the timing of the ADC and might require increasing the evaluation period until the array output is charged to the correct voltage level for the VCO based ADC. A partial solution to this issue would be to increase the driver strength by increasing the driver circuits or even to use multiple drivers at different positions around the array. In that case, the most critical device will be in the middle of the 1T1R array. This will of course complicate the architecture but it should allow for larger array sizes while not affecting the evaluation cycles of the ADC.

Acknowledgments

This work was in part funded by the German Research Foundation (DFG) under Grant No. SFB 917, in part by the Federal Ministry of Education and Research (BMBF, Germany) in the Project NEUROTEC (Project Nos. 16ME0398K and 16ME0399), in part by the European Union's Horizon 2020 Research and Innovation Program through the Project MNEMOSENE under Grant No. 780215 and it is based on the Jülich Aachen Research Alliance (JARA-FIT).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Christopher Bengel https://orcid.org/0000-0002-2892-9837
Johannes Mohr https://orcid.org/0000-0003-0815-3047
Stefan Wiefels https://orcid.org/0000-0003-2820-9677
Abhairaj Singh https://orcid.org/0000-0002-2729-7057
Anteneh Gebregiorgis https://orcid.org/0000-0001-5909-4927
Rajendra Bishnoi https://orcid.org/0000-0002-1590-0365
Said Hamdioui https://orcid.org/0000-0002-8961-0387
Rainer Waser https://orcid.org/0000-0002-9080-8980
Stephan Menzel https://orcid.org/0000-0002-4258-2673

References

[1] Wang Z, Wu H, Burr G W, Hwang C S, Wang K L, Xia Q and Yang J J 2020 Resistive switching materials for information processing *Nat. Rev. Mat.* **5** 173–95
[2] Burr G W *et al* 2017 Neuromorphic computing using non-volatile memory *Adv. Phys.:* X **2** 89–124
[3] Dittmann R and Strachan J P 2019 Redox-based memristive devices for new computing paradigm *APL Mater.* **7** 110903
[4] Zidan M A, Strachan J P and Lu W D 2018 The future of electronics based on memristive systems *Nat. Electron.* **1** 22–9
[5] Hu M *et al* 2018 Memristor-based analog computation and neural network classification with a dot product engine *Adv. Mater.* **30** 1705914
[6] Prezioso M, Merrikh-Bayat F, Hoskins B D, Adam G C, Likharev K K and Strukov D B 2015 Training and operation of an integrated neuromorphic network based on metal-oxide memristors *Nature* **521** 61–4
[7] Grenouillet L *et al* 2021 16k bit 1T1R OxRAM arrays embedded in 28 nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors *IEEE Int. Memory Workshop (IMW)* pp 1–4
[8] Mahadevaiah M K, Perez E, Wenger C, Grossi A, Zambelli C, Olivo P, Zahari F, Kohlstedt H and Ziegler M 2019 Reliability of CMOS integrated memristive HfO_2 arrays with respect to neuromorphic computing *IEEE Int. Reliability Physics Symp. (IRPS)*
[9] Dongarra J J, Du Croz J, Hammarling S and Duff I S 1990 A set of level 3 basic linear algebra subprograms *ACM Trans. Math. Softw.* **16** 1–17
[10] Ankit A *et al* 2019 PUMA: a programmable ultra-efficient memristor-based accelerator for machine learning inference *Proc. 24th Int. Conf. Architectural Support for Programming Languages and Operating Systems* pp 715–31
[11] Chi P, Li S, Xu C, Zhang T, Zhao J, Liu Y, Wang Y and Xie Y 2016 PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory *ACM/IEEE 43rd Annual Int. Symp. Computer Architecture (ISCA)* pp 27–39
[12] Song L, Qian X, Li H and Chen Y 2017 PipeLayer: a pipelined ReRAM-based accelerator for deep learning *IEEE Int. Symp. High Performance Computer Architecture (HPCA)* pp 541–52
[13] Chen A and Lin M R 2011 Variability of resistive switching memories and its impact on crossbar array performance *49th Annual IEEE Int. Reliability Physics Symp. (IRPS)*
[14] Chen Y S *et al* 2009 Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity *IEEE Int. Electron Devices Meeting* pp 95–8
[15] Molas G, Sassine G, Nail C, Alfaro Robayo D, Nodin J-F, Cagli C, Coignus J, Blaise P and Nowak E 2018 Resistive memories (RRAM) variability: challenges and solutions *ECS Trans.* **86** 35–47
[16] Shim W, Luo Y, Seo J-S and Yu S 2020 Investigation of read disturb and bipolar read scheme on multilevel RRAM-based deep learning inference engine *IEEE Trans. Electron Devices* **67** 2318–23
[17] Yang C *et al* 2020 Industrially applicable read disturb model and performance on mega-bit 28 nm embedded RRAM *IEEE Symp. VLSI Technology*
[18] Diokh T *et al* 2013 On the impact of the oxide thickness and reset conditions on activation energy of HfO_2 based ReRAM extracted through disturb measurements *IEEE Int. Integrated Reliability Workshop Final Report* pp 106–9
[19] Huang P *et al* 2017 RTN based oxygen vacancy probing method for Ox-RRAM reliability characterization and its application in tail bits *IEEE Int. Electron Devices Meeting (IEDM)* pp 21.4.1–4
[20] Huang P, Xiang Y C, Zhao Y D, Liu C, Gao B, Wu H Q, Qian H, Liu X Y and Kang J F 2018 Analytic model for statistical state instability and retention behaviors of filamentary analog RRAM array and its applications in design of neural network *IEEE Int. Electron Devices Meeting (IEDM)*
[21] Du Y, Jing L, Fang H, Chen H, Cai Y, Wang R, Zhang J and Ji Z 2020 Exploring the impact of random telegraph noise-induced accuracy loss on resistive RAM-based deep neural network *IEEE Trans. Electron Devices* **67** 3335–40
[22] Raghavan N, Degraeve R, Fantini A, Goux L, Strangio S, Govoreanu B, Wouters D, Groeseneken G and Jurczak M 2013 Microscopic origin of random telegraph noise fluctuations in aggressively scaled RRAM and its impact on read disturb variability *Reliability Physics Symp. (IRPS)*
[23] Chai Z *et al* 2018 The over-reset phenomenon in Ta_2O_5 RRAM device investigated by the RTN-based defect probing technique *IEEE Electron Device Lett.* **39** 955–8
[24] Puglisi F M, Larcher L, Padovani A and Pavan P 2015 A complete statistical investigation of RTN in HfO_2-based RRAM in high resistive state *IEEE Trans. Electron Devices* **62** 2606–13
[25] Balatti S, Ambrogio S, Cubeta A, Calderoni A, Ramaswamy N and Ielmini D 2014 Voltage-dependent random telegraph noise (RTN) in HfO_x resistive RAM *IEEE Int. Reliability Physics Symp. (IRPS)*

[26] da Silva M B, Tuinhout H, Zegers-van Duijnhoven A, Wirth G I and Scholten A 2014 A physics-based RTN variability model for MOSFETs *IEEE Int. Electron Devices Meeting* pp 35.2.1–4
[27] Fantini A *et al* 2015 Intrinsic program instability in HfO_2 RRAM and consequences on program algorithms *Electron Devices Meeting (IEDM)* (7–9 December 2015)
[28] Kang J, Yu Z, Wu L, Fangf Y, Wang Z and Cai Y 2017 Time-dependent variability in RRAM-based analog neuromorphic system for pattern recognition *IEEE Int. Electron Devices Meeting (IEDM)* pp 6.4.1–4
[29] Zanotti T, Puglisi F M and Pavan P 2020 Circuit reliability analysis of RRAM-based logic-in-memory crossbar architectures including line parasitic effects, variability, and random telegraph noise *IEEE Int. Reliability Physics Symp. (IRPS)* pp 1–5
[30] Bengel C, Siemon A, Cuppers F, Hoffmann-Eifert S, Hardtdegen A, von Witzleben M, Hellmich L, Waser R and Menzel S 2020 Variability-aware modeling of filamentary oxide based bipolar resistive switching cells using SPICE level compact models *IEEE Trans. Circuits Syst.* I **67** 4618–30
[31] JART 2019 Juelich Aachen Resistive Switching Tools (JART) *Techreport* Juelich Aachen Research Alliance
[32] Cüppers F, Menzel S, Bengel C, Hardtdegen A, von Witzleben M, Böttger U, Waser R and Hoffmann-Eifert S 2019 Exploiting the switching dynamics of HfO_2-based ReRAM devices for reliable analog memristive behavior *APL Mater.* **7** 091105
[33] Wiefels S, Bengel C, Kopperberg N, Zhang K, Waser R and Menzel S 2020 HRS instability in oxide based bipolar resistive switching cells *IEEE Trans. Electron Devices* **67** 4208–15
[34] Mayahinia M *et al* 2021 A novel voltage controlled oscillation based ADC design for computation-in-memory using emerging ReRAMs *ACM J. Emerg. Technol. Comput. Syst.* **18** 32
[35] Zhao W and Cao Y 2006 New generation of predictive technology model for sub-45 nm early design exploration *IEEE Trans. Electron Devices* **53** 2816–23
[36] Baeumer C *et al* 2017 Subfilamentary networks cause cycle-to-cycle variability in memristive devices *ACS Nano* **11** 6921–9
[37] Fleck K, Böttger U, Waser R, Aslam N, Hoffmann-Eifert S and Menzel S 2016 Energy dissipation during pulsed switching of strontium-titanate based resistive switching memory devices *Proc. 46th European Solid-State Device Research Conf. (ESSDERC)* (Lausanne, Switzerland 12–15 September 2016) pp 160–3
[38] Schoenhals A, Waser R and Wouters D 2017 Improvement of SET variability in TaOx based resistive RAM devices *Nanotechnology* **28** 465203
[39] Strachan J P, Torrezan A C, Miao F, Pickett M D, Yang J J, Yi W, Medeiros-Ribeiro G and Williams R S 2013 State dynamics and modeling of tantalum oxide memristors *IEEE Trans. Electron Devices* **60** 2194–202
[40] Degraeve R *et al* 2012 Dynamic hour glass model for SET and RESET in HfO_2 RRAM *Proc. 2012 Symp. VLSI Technology* p 75
[41] Puglisi F M, Zagni N, Larcher L and Pavan P 2018 Random telegraph noise in resistive random access memories: compact modeling and advanced circuit design *IEEE Trans. Electron Devices* **65** 2964–72
[42] Li W, Shuai Z, Xue C J, Yuan M and Li Q 2019 A wear leveling aware memory allocator for both stack and heap management in PCM-based main memory systems *2019 Design, Automation Test in Europe Conf. Exhibition (DATE)* pp 228–33
[43] Liao J, Zhang F, Li L and Xiao G 2015 Adaptive wear-leveling in flash-based memory *IEEE Comput. Architect. Lett.* **14** 1–4
[44] Khaddam-Aljameh R *et al* 2021 2021 HERMES core—a 14 nm CMOS and PCM-based in-memory compute core using an array of 300 ps/LSB linearized CCO-based ADCs and local digital processing *Symp. VLSI Technology* pp 1–2
[45] Singh A, Abu Lebdeh M, Gebregiorgis A, Bishnoi R, Joshi R V and Hamdioui S 2021 SRIF: scalable and reliable integrate and fire circuit ADC for memristor-based CIM architectures *IEEE Trans. Circuits Syst.* I **68** 1917–30