

A 32-KB ePCM for Real-Time Data Processing in Automotive and Smart Power Applications

Marco Pasotti^{ID}, Riccardo Zurla^{ID}, Marcella Carissimi, Chantal Auricchio, Donatella Brambilla, Emanuela Calvetti, Laura Capecchi, Luigi Croce^{ID}, Daniele Gallinari, Cristina Mazzaglia, Vikas Rana^{ID}, Alessandro Cabrini^{ID}, and Guido Torelli^{ID}, *Senior Member, IEEE*

Abstract—In the frame of power electronics applications, the request for smart and reconfigurable devices is pushing integration technologies in the direction of embedded systems. In this scenario, microcontrollers play a key role, and the availability of embedded non-volatile memory (eNVM) to store the microcontroller code has become crucial to enable real-time customization and increase system flexibility. Among emerging NVMs, phase change technology is becoming a very attractive solution for the development of applications for smart power and automotive markets. In this paper, a 32-KB embedded phase change memory (ePCM) designed and manufactured in 0.11- μm smart power BCD technology with a specifically optimized Ge-rich Ge-Sb-Te (GST) alloy (supply voltage = 1.8 V) is presented. Thanks to the use of a differential sensing scheme, the proposed ePCM features 18-ns random access time with improved robustness against resistance drift. The word modify time under 32-cell programming parallelism was kept as low as 20 μs , thanks to enhanced programming circuits. The size of the 32-KB eNVM is about 0.7 mm^2 .

Index Terms—Automotive applications, differential sensing, embedded memory, germanium-rich alloy, nonvolatile memory, phase change memory, real-time systems, smart power devices.

I. INTRODUCTION

AMONG the emerging memories, phase change memory (PCM) [1], [2] is considered one of the most promising solutions for both stand-alone [3], [4] and embedded [5], [6] applications. In this respect, the possibility to integrate, with the minimum impact on process complexity (and thus on manufacturing costs), high-voltage and high-power components with non-volatile memory (NVM) is considered as a key enabling technology for the design of next-generation smart power applications. For this reason, embedded PCM (ePCM) (Fig. 1) is a good candidate since it requires only three additional masks in the back-end of the line [7] and offers a lean integration. Implementing the

Manuscript received December 3, 2017; revised February 17, 2018, March 13, 2018, and April 8, 2018; accepted April 8, 2018. Date of publication May 22, 2018; date of current version June 25, 2018. This paper was approved by Guest Editor Shidhartha Das. (*Corresponding author: Riccardo Zurla.*)

M. Pasotti, M. Carissimi, C. Auricchio, D. Brambilla, E. Calvetti, L. Capecchi, L. Croce, D. Gallinari, and C. Mazzaglia are with STMicroelectronics, 20864 Agrate Brianza, Italy.

R. Zurla, A. Cabrini, and G. Torelli are with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia, Italy (e-mail: riccardo.zurla@unipv.it).

V. Rana is with STMicroelectronics, Greater Noida 201308, India.

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

Digital Object Identifier 10.1109/JSSC.2018.2828805

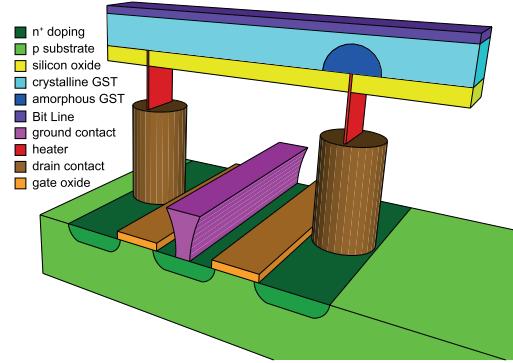


Fig. 1. 3-D view of two PCM cells that share the source diffusion of the selector and the ground contact. The cell on the left side is in the SET state, whereas the cell on the right side is in the RESET state.

memory cells in the back-end of the line is a very attractive feature for embedded applications since it allows minimizing the impact of the memory on the other integrated components that actually are the main contributors to the added value of the product.

It is worth to point out that electronic systems designed for industrial or automotive applications must typically operate under much more stringent conditions when compared to the case of products developed for the consumer market. Challenging issues arise when considering, for example, the operating temperature range that is required for automotive products. In this respect, the capability to guarantee the correct operation, over an extended temperature range, of an NVM based on a phase change material is a key feature. More specifically, in order to cope with the harsh constraints imposed by industrial and automotive standards [8], in this work, we took advantage of the characteristics of an optimized Ge-rich alloy [9] (which offers a higher crystallization temperature T_x with respect to the standard $\text{Ge}_2\text{Sb}_2\text{Te}_5$ alloy) and the use of a differential sensing scheme (where two memory cells store data in complementary states to cope with the resistance drift [10]–[13] of both the SET and the RESET state), thus enabling memory operations and data retention above 150 °C. The necessity to employ a phase change material with a higher crystallization temperature required the design of improved programming circuits aiming at minimizing the adverse effect of the higher T_x on programming throughput. Thanks to the proposed solutions, it has been possible to reduce power consumption, programming time, and costs of electrical wafer

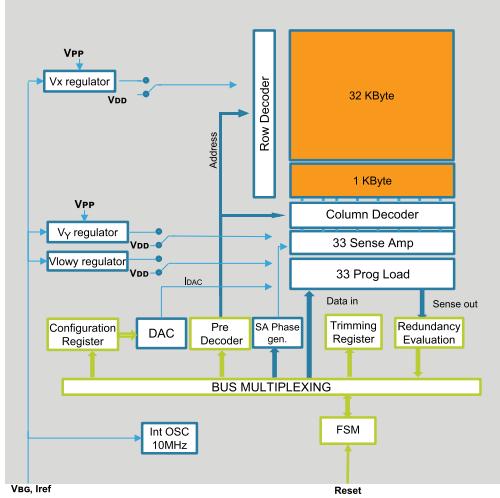


Fig. 2. Block diagram of the macrocell.

sort (EWS), as well as to cut-down microcontroller idle time when the ePCM is used to store real-time data.

The paper is organized as follows. Section II gives details of the architecture of the designed ePCM. Section III introduces the programming approach and presents the circuits that have been designed to efficiently implement the programming operations with the Ge-rich GST alloy used in this paper. Section IV describes the differential sensing scheme used to optimize read operation. Section V shows experimental results. Section VI provides considerations on IR voltage drop and electromigration in the conductive paths and, finally, Section VII draws concluding remarks.

II. MEMORY ARCHITECTURE

The ePCM elementary cell, which is based on an NMOS selector [7], was integrated in ePCM 0.11- μm BCD technology, and occupies $0.7 \mu\text{m}^2$ of silicon area. A 32-KB macrocell was designed and included in a test chip in 16 independent instances in order to increase the total number of cells in a single chip, to reduce testing cost and improve the consistency of the extracted statistical data. In addition to the 16 ePCM macrocells, the developed test chip also includes a built-in self-test (BIST) block, several test chip memory registers, a reference generator block, and the circuitry that manages the input–output interface. The reference generator block provides the macrocells with a bandgap voltage, $V_{BG} = 1.2 \text{ V}$, as well as with a reference current, $I_{ref} = 10 \mu\text{A}$. The schematic block of the macrocell is depicted in Fig. 2, and includes: the memory array (shown in Fig. 3), a finite-state machine (FSM), two decoders (namely, the row and the column decoder), three voltage regulators, an internal oscillator, a digital-to-analog converter (DAC, which sets the value of the programming current), and several internal registers. Two supply voltages are provided, namely, a low-voltage (LV) supply, $V_{DD} = 1.8 \text{ V}$, and a high-voltage supply, $V_{PP} = 5 \text{ V}$. V_{PP} feeds only the programming circuitry, whereas V_{DD} is used throughout the macrocell to supply all the other blocks (i.e., the FSM, the read circuitry, the DAC, the registers, and so on) in order to minimize power consumption.

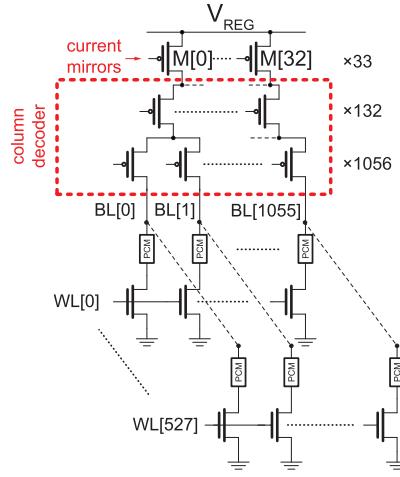


Fig. 3. Simplified schematic of the array architecture and the column decoder. The current mirrors driven by a DAC, set the value of the programming current (see Section III).

In the implemented memory array, 528 Kcells are organized in 1024 columns or BitLines (BLs) and 528 rows or WordLines (WLs). The array is divided into two main sectors, namely, the user memory space and the reserved memory space (RMS). Thirty two additional columns are implemented as redundant BLs to enable readdressing one failed column every 32 BLs. Indeed, if a standard column (i.e., a data storage column) does not operate correctly after wafer fabrication, it is possible, during EWS, to redirect its address to the address of the corresponding redundancy column, thus increasing the manufacturing yield. The readdressing process is fully transparent to the user.

Both sectors have two sets of cells: direct cells (DCs), which store the plain data, and complementary cells (CCs), which store a complemented version of the information. We can therefore have a DC programmed in the SET state and its associated CC programmed in the RESET state (DC SET and CC RESET states, logic “1”) or a DC programmed in the RESET state and its associated CC programmed in the SET state (DC RESET and CC SET states, logic “0”). This approach allows compensating for the resistance drift of both the SET and RESET states in the used chalcogenide alloy by exploiting the above-mentioned differential sensing scheme to compare the current flowing through a DC to the current of the corresponding CC, as will be shown in better detail in Section IV. The 32-KByte user memory space occupies the first 512 WLs of the array and is entirely available to the final user, whereas the 1-KByte RMS is placed in the bottom 16 WLs and stores both the application data and the macrocell internal data (e.g., trimming configurations, voltage values, and malfunctioning columns to be readdressed). All the data stored in the reserved sector are written with a majority voting approach (and hence each information is copied in three different memory locations) to ensure an enhanced robustness of the system.

A memory location is defined as word, which consists of $32 + 1$ DCs and $32 + 1$ CCs (in both cases, 32 data and one redundant cells) and corresponds to one memory address. Therefore, the macrocell has 8448 addresses. The cells that

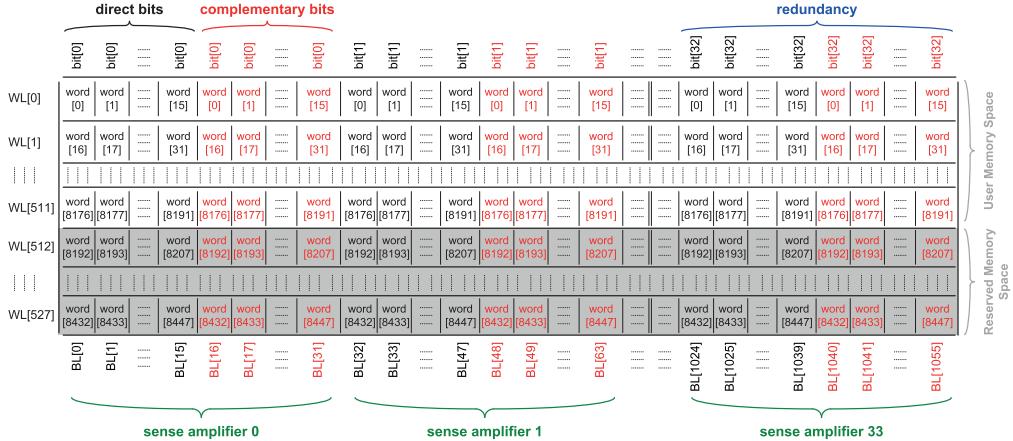


Fig. 4. Schematic of the ePCM array showing the physical bit scrambling. The RMS is highlighted in gray, whereas the rest of the array corresponds to the user memory space.

correspond to each word are physically scrambled throughout the array as shown in Fig. 4, where the DCs and the CCs are depicted in black and red colors, respectively. Each row contains 16 words plus the associated bits for redundancy. In each row, the first bits (bit[0]) of all words are stored in the 16 + 16 leftmost columns (16 DCs followed by 16 CCs), so that a sense amplifier can easily compare a DC with the corresponding CC, as will be explained in Section IV. Moving toward the right in the array, all the other bits (i.e., bit[1], bit[2], ..., bit[32]) occupy the successive columns following the same approach. The 16 + 16 rightmost columns are dedicated to store the redundant bits (one bit—i.e., one DC and one CC—for each of the 16 words), labeled as bit[32] in Fig. 4. This is the outcome of layout optimization aimed at minimizing parasitic effects (due to interconnections) and allowing easy BLs routing to the sense amplifiers and the column decoder.

The decoding logic required to access the memory array is implemented using LV transistors. The FSM controls read and write operations by setting adequate voltage and current amplitudes through voltage regulators and current generators, respectively.

III. PROGRAMMING CIRCUITS

Program operations are carried out by driving a current pulse through a PCM cell in order to first increase and then decrease the temperature of the chalcogenide material exploiting the thermal power provided, thanks to the Joule effect, by the programming current flowing through the heater. In particular, to achieve a crystalline state (i.e., SET operation, which sets the stored bit to 1), it is required to generate a long, low, and trapezoidal-shaped programming current pulse, whereas short, higher, and sharp rectangular-shaped programming current pulses determine an amorphous state (i.e., RESET operation, which sets the stored bit to 0) when applied to a PCM memory cell.

The basic idea to enhance the program throughput of our memory is to write multiple cells simultaneously. Therefore, the programming circuits have to be carefully designed in order to comply with the sudden and heavy request of

TABLE I
PULSE SPECIFICATIONS AND TOTAL CURRENT REQUIREMENTS

Pulse	I pulse (μ A)	T pulse (μ s)	Programming current (mA)		
			PAR 2**	PAR 16**	PAR 32**
SET	300	0.5–5	0.6	—	9.6
SET	350	0.5–5	0.7	—	11.2
SET	400	0.5–5	0.8	—	12.8
RESET	450	0.1–0.5	0.9	7.2	—
RESET	500	0.1–0.5	1	8.0	—
PS SET*	550	0.5–5	1.1	8.8	—
PS RES*	600	0.1–5	1.2	9.6	—
Forming*	600	0.5–5	1.2	9.6	—

* see Section V for the definition of these terms.

** PAR $n = n$ programming parallelism.

current (Table I) necessary to carry out a high-parallelism [e.g., parallelism (PAR) = 16 or 32] write operation. From this point of view, high-parallelism RESET is the most demanding operation since a sufficiently sharp pulse rising edge is required to reach the desired amplitude quickly (i.e., within a negligible fraction of the pulselength) so as to guarantee that an adequate volume of the phase change material is molten without the need for excessively increasing the pulselength. Indeed, on the one hand, the correct formation of the amorphous cap (i.e., the adequate random disposition of atoms with a minimum amount of crystalline grains in this cap) is fundamental to guarantee the desired data retention since, otherwise, the crystallization process that takes place after the programming operation can be significantly faster and can therefore corrupt the data. On the other hand, it is fundamental to keep the time length of RESET pulses short since PCM endurance decreases with increasing time spent in the molten phase of the chalcogenide alloy. A very sharp rising edge of the programming pulse allows reducing the overall pulselength in the design phase, thus allowing the best tradeoff between data retention and endurance.

A digitally programmable accurate voltage source is needed to supply the circuits that generate the programming current, which are implemented using mainly LV transistors in order to minimize area occupation. Indeed, since the voltage at the addressed BLs increases with increasing values of the current

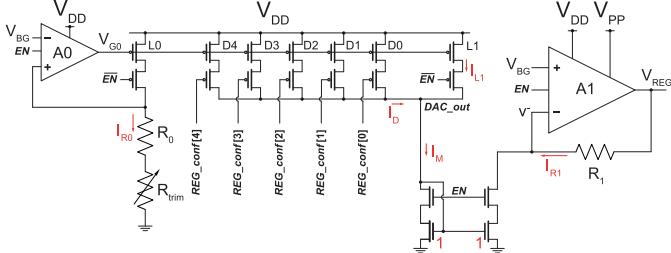


Fig. 5. Circuit schematic of the programmable voltage regulator that supplies the program circuitry. EN = enable signal (active high).

delivered to the cells, the supply voltage of these circuits must be a function of the programming current, so as to avoid any damage to LV devices and ensure, at the same time, the required accuracy of the generated pulses. Furthermore, the voltage regulator must ensure a limited output voltage drop when a large current is requested by the load, so as to allow the programming circuits to operate correctly even in the first part of the output current pulse.

The implemented voltage regulator, shown in Fig. 5, generates an adequate programmable voltage, V_{REG} , exploiting the V_{BG} voltage provided by the above-mentioned reference generator block. The regulated voltage can be expressed as $V_{\text{REG}} = V_{\text{BG}} + R_1 I_{R1}$, where I_{R1} is the current flowing through the polysilicon resistor R_1 and is equal to $V_{\text{BG}}/(R_0 + R_{\text{trim}})$ multiplied by a programmable factor, N_{REG} , controlled by a DAC driven by signal $\text{REG_conf}[0 \text{ to } 4]$. Indeed, this signal sets the amount of current generated by the DAC, which is then copied with a unity-gain mirror and injected into R_1 , in order to set V_{REG} to the desired value

$$V_{\text{REG}} = V_{\text{BG}} \left(1 + N_{\text{REG}} \frac{R_1}{R_0 + R_{\text{trim}}} \right). \quad (1)$$

The undesired dependence of V_{REG} upon temperature variations is strongly mitigated by also implementing R_0 and R_{trim} as polysilicon resistors. A common centroid structure is used for R_0 , R_{trim} , and R_1 to improve matching. The resistance value of R_{trim} can be set by the FSM, through a 3-bit digital signal, in order to enhance accuracy. The optimal configuration of this signal (trim i with $i = 0$ to 7) is measured during EWS and stored into the RMS.

A key block of the designed voltage regulator is operational amplifier A1, since its characteristics are the major contributors to the voltage regulator overall performance. Hence, A1 must feature a sufficiently high unity-gain frequency and an adequately fast response to an abrupt load current change. In this respect, we designed a three-stage class-AB-output operational amplifier [14] (shown in Fig. 6) that exploits current buffer compensation [15] in order to achieve higher bandwidth and faster recovery with respect to standard nested Miller compensation [16]–[18]. Indeed, at sufficiently high frequency, the connection of compensator capacitor C_2 through transistor N_C generates a voltage gain equal to $C_2/C_{p,2}$ between node V_{REG} and V_2 , where $C_{p,2}$ is the parasitic capacitance associated with node V_2 . Under these conditions, transistor P_3 is connected in a super diode configuration, which decreases the impedance at node V_{REG} and, thus, moves the second non-dominant pole, ω_{p3} , toward higher

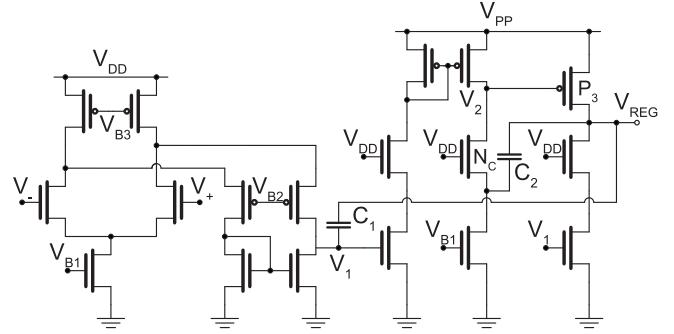


Fig. 6. Circuit schematic of the class-AB-output operational amplifier designed for the program load voltage regulator.

frequencies [19]:

$$\omega_{p3} \approx \frac{g_{m3}}{C_L} \frac{C_2}{C_{p,2}} \quad (2)$$

where g_{m3} is the transconductance of transistor P_3 and C_L is the total capacitive load at node V_{REG} . The obtained ω_{p3} was exploited to achieve a higher unity-gain frequency (about 10 MHz) with respect to the case of a direct connection of C_2 to node V_2 (i.e., the case of nested Miller compensation). In addition, the current buffer compensation greatly reduces the capacitive coupling between nodes V_2 and V_{REG} with respect to the case of a conventional nested Miller compensation, which significantly decreases the output voltage drop due to a sudden change of the load current and improves the positive V_{PP} power supply rejection ratio. Thanks to the class-AB output stage, the operational amplifier A1 can deliver up to 20 mA of current using a quiescent current of about 100 μA .

The implemented voltage regulator supplies the output section of the circuit, shown in Fig. 7, that generates and shapes the programming current pulses. This circuit exploits I_{ref} (provided by the reference generator block) as a reference current, which is then scaled with a programmable factor by a dedicated DAC to obtain I_{DAC} [20]. The scaling factor is controlled through a 6-bit digital signal, DAC_conf , whose value can be dynamically set by the FSM during a programming current pulse in order to shape it as required. Indeed, by controlling the amplitude of I_{DAC} and, hence, of I_{ctrl} [$I_{\text{ctrl}} = I_{\text{DAC}}$ thanks to the unity replica factor of current mirror (CM) N5, N6], the FSM can control the programming current I_{prog} that flows through the enabled transistor(s) M[k] (with $k = 0$ to 32) and is then fed to the memory cells selected by the column decoder (depicted in Fig. 3). In particular, by dynamically varying the value of signal DAC_conf during a SET operation, it is possible to achieve a close discrete approximation of the trapezoidal waveform needed to properly crystallize the chalcogenide layer. The falling slope of the pulse can be set by choosing an adequate variation of signal DAC_conf over time.

To comply with the above-mentioned requirements, the CMs must be able to rapidly provide the required RESET current when activated from idle state. Indeed, the conventional CM shown in Fig. 7 determines, in the current pulse generation, a delay that increases with increasing programming parallelism

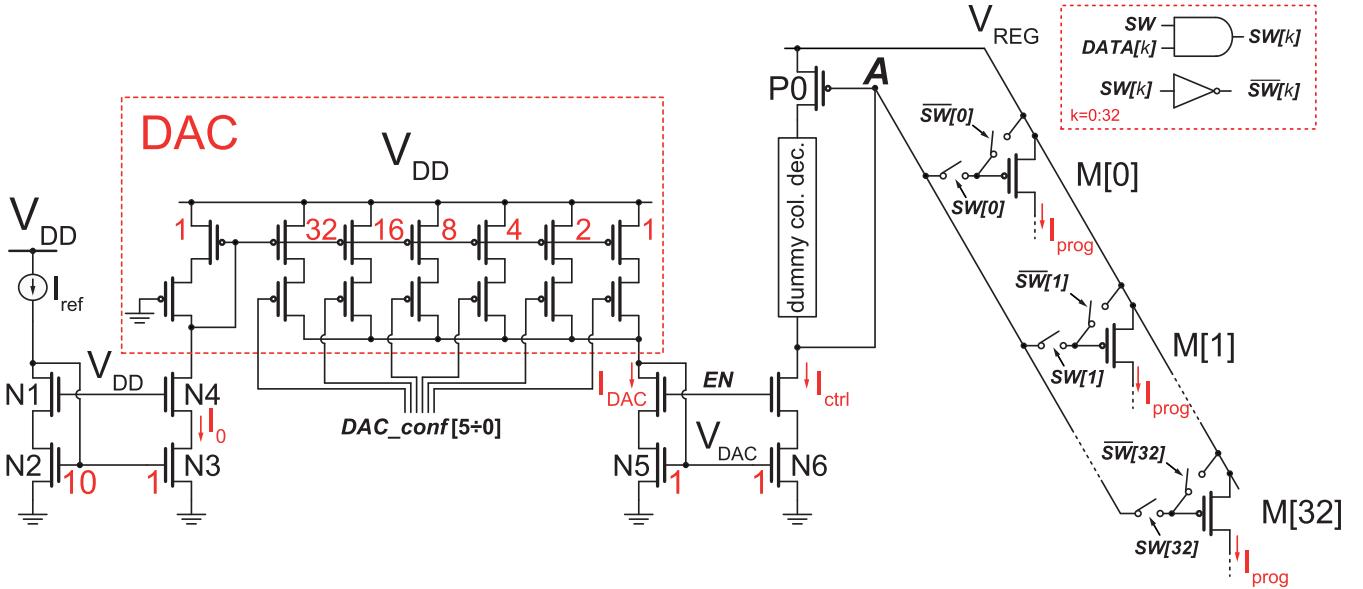


Fig. 7. Conventional current programming circuitry powered by the voltage regulator. Signal SW is driven high at the beginning of a programming pulse. Signal EN is set to 0 out of the programming phase to ensure zero current consumption. A dummy column decoder is included in the diode connection of transistor P0 to mimic for the voltage drop introduced by the column decoder in each selected programming path.

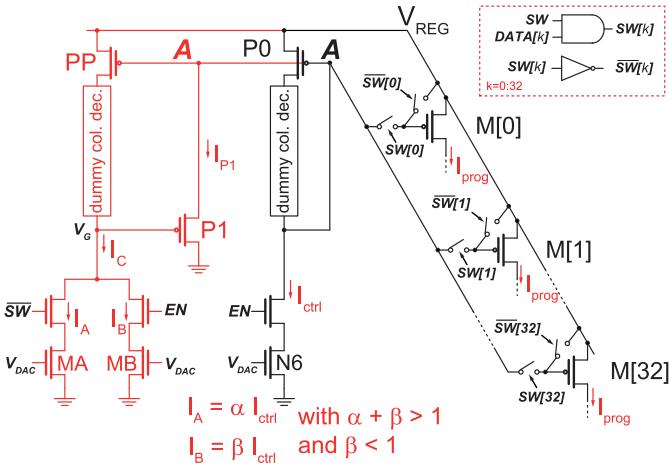


Fig. 8. Enhanced CM (red) applied to a conventional CM (black). Transistors M $[k]$ with $k = 0-32$ are the same transistors shown in Figs. 3 and 7 with the same names.

due to the capacitive load (gates of activated transistors M $[k]$) that is suddenly connected to the control branch by driving signals short wave (SW) $[k]$ high, thus increasing the voltage at node A and reducing the overdrive voltage of transistor P0 due to charge sharing.

To overcome this limitation, which leads to unacceptable RESET pulse shapes when a high-parallelism programming is carried out, an enhanced CM for fast recovery was designed and included into the macrocell. For the sake of clarity, Fig. 8 shows the designed circuit (depicted in red) applied to the conventional CM (depicted in black) already shown in Fig. 7; the DAC circuit is omitted in Fig. 8. During the programming phase (enable signal EN high), before the beginning of a programming pulse (signal SW low), transistor P1 is kept in the ON-state thanks to the chosen value of the bias current, I_C ,

of the additional branch including transistor PP, which is generated by devices MA and MB driven by the same voltage, V_{DAC} , that drives device N6. More specifically, I_C corresponds to the sum of currents $I_A = \alpha I_{ctrl}$ and $I_B = \beta I_{ctrl}$, α and β being suitable factors imposed by the mirror ratios of transistors MA and MB, respectively. By choosing $\alpha + \beta > 1$, the negative feedback loop including devices P1, PP, and P0 forces the gate voltage, V_G , of transistor P1 to a value that makes this device to carry a current, I_{P1} , equal to $I_C - I_{ctrl} = I_{ctrl}(\alpha + \beta - 1)$, which is provided by transistor P0 (devices PP and P0 have the same size). This approach allows biasing transistor P1 with a well-defined current with no need for a particularly accurate matching of the NMOS transistors that set the values of the currents in the branches (i.e., transistors MA, MB, and N0). When a program pulse starts (signal SW driven high) and, as a consequence, the voltage at node A increases due to charge sharing, transistor P1 draws a current whose value is an increasing function of the (undesired) voltage increment at node A. This node is therefore quickly discharged: when it settles and, hence, the desired static value of programming current is reached, the current through P1 is automatically reduced to zero (indeed, the overall mirror ratio of MA and MB is reduced below unity since \bar{SW} is driven low and $\beta < 1$). By adequately sizing the values of α and β , it is possible to set the desired speed of the circuit and the time when P1 is turned off. This technique can be easily adapted to allow a higher parallelism (e.g., PAR = 64, 128, and so on) with no need of increasing either the static power consumption or the area occupation of the enhanced CM. Indeed, both factors α and β can be kept unchanged, whereas the channel width of transistor P1 can be increased to enhance its current in order to counteract the effect of the larger capacitive load.

Exploiting the presented circuits, it is possible to generate write pulses fully customizable in terms of amplitude

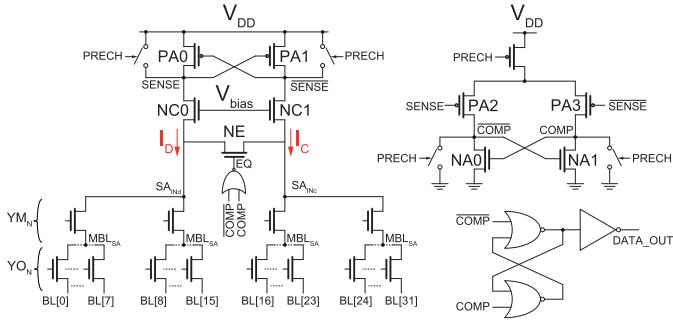


Fig. 9. Circuit schematic of a macrocell sense amplifier. The schematic depicts the connection of sense amplifier[0] to the corresponding BLs.

TABLE II

MEASURED RANDOM ACCESS TIME (EXPRESSED IN NANOSECONDS)
AT DIFFERENT TEMPERATURES AND POWER SUPPLY VALUES
(i.e., V_{DD} , $V_{DD,min} = 1.55$ V, AND $V_{DD,max} = 1.95$ V)

	Fresh program			After 24h bake at +190°C			After 96h bake at +190°C		
	$V_{DD,min}$	V_{DD}	$V_{DD,max}$	$V_{DD,min}$	V_{DD}	$V_{DD,max}$	$V_{DD,min}$	V_{DD}	$V_{DD,max}$
-40 °C	15.13	15.06	14.00	16.88	16.67	15.01	17.68	16.67	15.29
+25 °C	15.45	15.56	14.00	16.50	16.38	14.75	16.88	16.67	15.01
+150 °C	16.14	15.82	14.49	16.88	16.38	15.01	16.88	16.38	15.01

(DAC in Fig. 7), slope, duration, the number of pulses, and the verify level for the program-and-verify approach [21]. Moreover, the FSM controls the SET and RESET algorithms developed to achieve the desired cell resistance. Furthermore, at boot, the FSM loads, from the RMS, the trimming and the redundancy data that implement a redundant approach to tolerate single-bit failure.

IV. SENSING CIRCUITS

Read operations are performed through a differential sensing scheme (Fig. 9) that consists of 32 + 1 sense amplifiers (SAs) and a two-level addressing (transistors Y_{ON} and Y_{MN}), which selectively connects the cells that have to be read to the corresponding sense amplifiers. A sense amplifier sets an adequate biasing voltage (around 0.6 V) on the selected BLs and compares the current flowing through the selected DC, I_D , to the current flowing through the associated CC, I_C . The cell that shows a lower resistance (i.e., the cell in the SET state) discharges either node SENSE or node SENSE-bar faster than the associated cell, thereby turning on either transistor PA0 or transistor PA1. The current unbalance between the DC and the CC ($\Delta I = I_D - I_C$) is amplified using two cascaded latches, and the obtained digital data are provided to the output in less than 18 ns in all specified conditions. Table II shows the random access time measured at different temperatures (i.e., -40 °C, 25 °C, and 150 °C) using the minimum, the maximum, and the nominal specified LV supply values (i.e., $V_{DD} = 1.55$, 1.8, and 1.95 V) on the same array just after programming and after baking at +190 °C for 24 and 96 h in order to emulate the aging process. The worst-case scenario is represented by aged cells (which show the smallest reading window, i.e., the closest SET and RESET distributions, as will be shown in the following) read at the minimum allowed values of temperature and V_{DD} .

In order to ensure the accuracy of the read operation, a precharge pulse is delivered by the FSM to discharge and equalize

the internal SA nodes, so that the result of the current read procedure is not affected by the previous read operation. The differential sensing scheme helps to cope with both the crystalline and the amorphous resistance drift that takes place in the Ge-rich alloy used in this work, thanks to its intrinsic self-reference characteristic. This advantage can be clearly seen in Fig. 10, where the current distributions of cells in the logical “1” state (i.e., DCs in the SET state and CCs in the RESET state) measured at different temperatures just after programming and after bake at 190 °C for 96 h are depicted in the plots on the left side, and the corresponding distributions of the current difference between two associated cells are illustrated in the plots on the right side (very similar results are obtained for the case of the logical “0” state). It is particularly interesting to notice that, even though the single-cell reading window is severely reduced after baking and the resistance values of the cells vary significantly with temperature (left plots), the current difference between a DC and its complementary counterpart (right plots) remains sufficiently high (i.e., higher than 10% of the maximum cell current value, I_{max} , even in the worst case) to ensure data to be correctly retrieved in all the specified conditions, thus proving that this approach guarantees enhanced data retention. It is straightforward to understand that a larger current difference between two cells leads to a safer read operation. In this respect, it should be pointed out that performing sensing with no need for relying on a predetermined reference level result in safer reading even in the absence of any resistance drift, and is also beneficial in the case of other memory technologies.

The sensing circuitry is also used to carry out verify operations during a programming routine: after a programming pulse has been applied, the current flowing through the freshly written cell (either a DC or a CC) is compared to a fixed SET or RESET verify current (depending on the data to be written in the cell), which is generated by the DAC circuit shown in Fig. 7. The outcome of this single-ended reading determines the successive steps of the program algorithm: if the verify is successful, the write operation is considered concluded for that particular cell under programming, otherwise a higher current pulse is applied and then another verify operation takes place. This procedure is iterated until either the verify is successful or the number of applied pulses reaches a predetermined upper bound (in this case, the memory cell is considered to fail). The SET operation can require up to three different pulses, whereas in the case of the RESET procedure, the maximum number of pulses is two. The specifications of the different programming pulses are summarized in Table I.

V. EXPERIMENTAL RESULTS

The micrograph of the implemented test chip is depicted in Fig. 11, where the main components are highlighted and the 16 32-KB macrocells are easily recognizable. The eight sets of measurements shown in Fig. 12 (each one depicted with a different marker) represent the output voltage of the regulator (V_{REG}) obtained, as a function of control signal REG_conf, for different configurations of the 3-bit signal that controls the value of resistor R_{trim} (see Fig. 5), i.e., signal trim i with $i = 0$ to 7. The blue dashed line depicts the

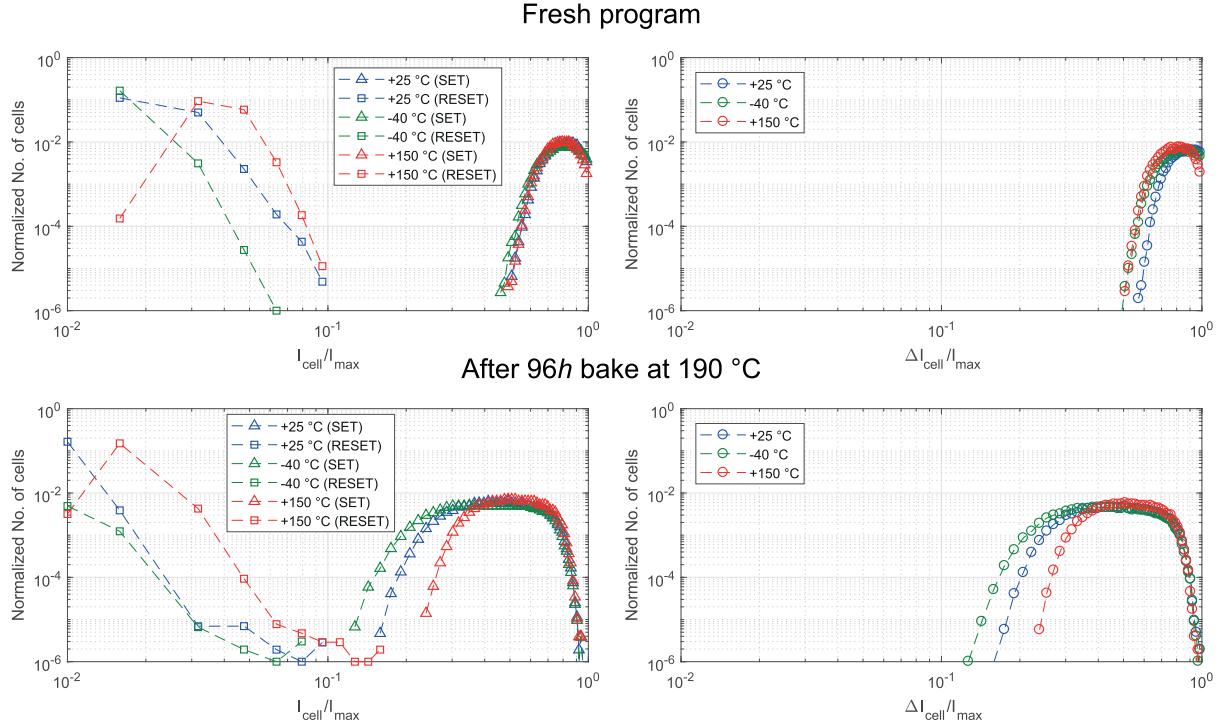


Fig. 10. Measured distributions of cell current (left plots, I_{cell}) and corresponding current difference distributions (right plots, ΔI_{cell}) obtained at different temperatures just after programming (top plots) and after bake at 190 °C for 96 h (bottom plots). 264 Kcells per distribution.

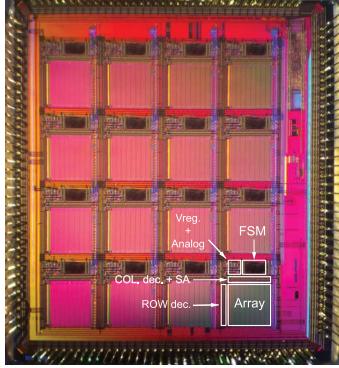


Fig. 11. Test chip micrograph (main blocks are identified). Sixteen identical 32-KB ePCMs are integrated in the device.

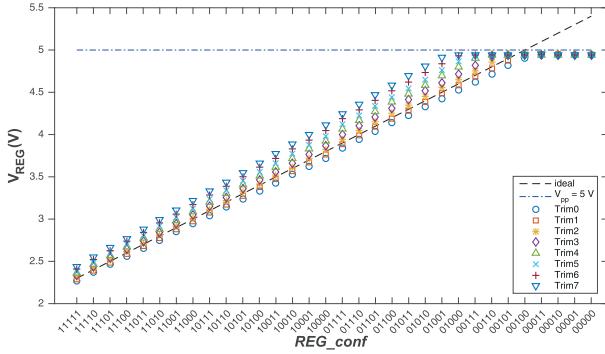


Fig. 12. Measured V_{REG} as a function of signal REG_conf with different trimming configurations (trim i with $i = 0\text{--}7$).

desired value of V_{REG} , whereas the black dash-dotted line represents the high-voltage power supply V_{PP} (which was set to 5 V in these measurements). Fig. 13 shows the measured

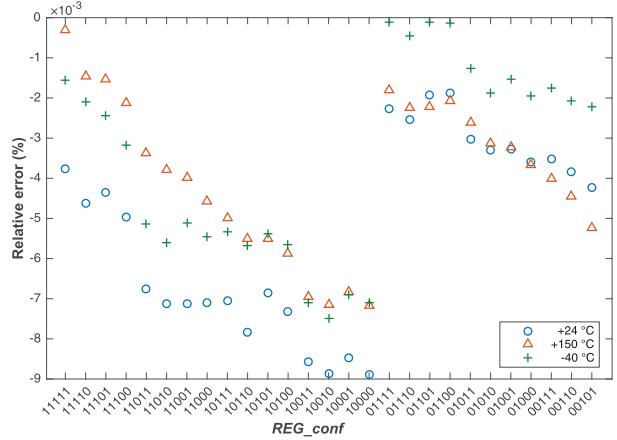


Fig. 13. Measured relative error of V_{REG} with respect to the desired value as a function of signal REG_conf when using the optimal trimming configuration (i.e., trim1) at different temperatures: $T = -40$ °C, $T = +24$ °C, and $T = +150$ °C.

relative error of voltage V_{REG} (generated with the optimal trimming configuration, i.e., from Fig. 12, trim1) with respect to the desired value collected at the minimum specified operating temperature (i.e., $T = -40$ °C), at room temperature (i.e., $T = +24$ °C), and at the maximum specified operating temperature (i.e., $T = +150$ °C). It is worth to notice that the values of signal REG_conf used for these particular measurements are in the range from the minimum V_{REG} configuration (i.e., 11111) to the configuration that corresponds to $V_{REG} = 4.9$ V (i.e., 00101), since V_{PP} was set equal to 5 V. The abrupt change in the error observed when passing from configuration 10000 to 01111 is typical in binary-weighted current-steering DACs. In all cases, the relative error is comprised between

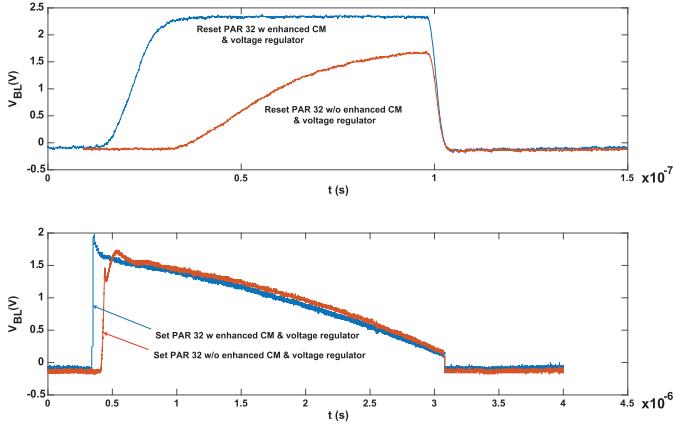


Fig. 14. Measured RESET (top) and SET (bottom, linear staircase down [22]) pulses at high-parallelism (32 PCM cells) with (blue curves) and without (red curves) the EPC shown in Figs. 6 and 8.

0% and -1% , and no significant correlation with temperature variations is observed.

To better appreciate the improvement introduced by the enhanced programming circuitry (EPC), the BL[0] voltage, $V_{BL[0]}$, collected during a RESET pulse as well as during a SET pulse (both carried out with parallelism 32) is plotted in Fig. 14: the red curves correspond to programming pulses obtained with a standard CM and a voltage regulator that includes a conventional nested Miller compensated operational amplifier, whereas the blue curves correspond to programming pulses carried out with the proposed voltage regulator (Figs. 5 and 6) and the improved CM (Fig. 8). It can be clearly seen in the top plot that the rising edge of the RESET pulse obtained with the standard scheme is too slow, which hinders the full pulse amplitude from being reached and, hence, the chalcogenide alloy from being properly molten, whereas the RESET pulse generated by the proposed circuit guarantees the required sharpness of the rising edge and, hence, allows the full pulse amplitude to be attained. The proposed EPC also ensures a faster rising edge of SET pulses (bottom plot): the case of a trapezoidal-shaped SET programming pulse is shown in the figure. It is important to point out that, in the fabricated test chip, it is not possible to directly measure the current pulses: the collected BL voltage pulses are a function not only of the generated programming current but also of the resistance of the chalcogenide layer, which changes during the falling edge of the SET pulse, thus giving rise to non-linearities in this portion of the pulse.

The effect of the proposed EPC (in particular, of the enhanced CM in Fig. 8) can be easily appreciated in Fig. 15, where the evolution over time of voltages at nodes A (top) and BL[0] (bottom) is depicted in a time frame that corresponds to the beginning of a RESET pulse with parallelism 32 in two cases, namely, with (blue curves) and without (red curves) the presented improved programming circuits (the signals in Fig. 15, as well as the internal signals shown in the following, were picked up by means of active micro-probes, which feature an input capacitance of about 20 fF and, therefore, have a negligible impact on the measured waveforms). In both cases, the voltage at node A (top curves) is first increased due

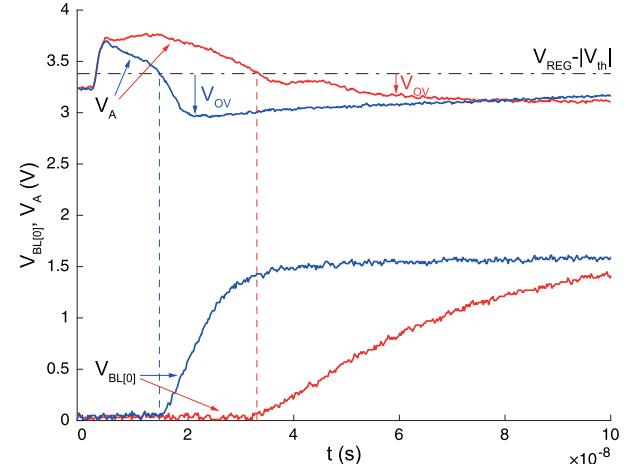


Fig. 15. Measured voltages at node A, V_A (top), and at BitLine BL[0], $V_{BL[0]}$ (bottom), at the beginning of two high-parallelism (32 cells) RESET operations carried out with (blue curves) and without (red curves) the EPC shown in Figs. 6 and 8. The overdrive voltage, V_{OV} , of transistor P0 in Fig. 8 is also highlighted.

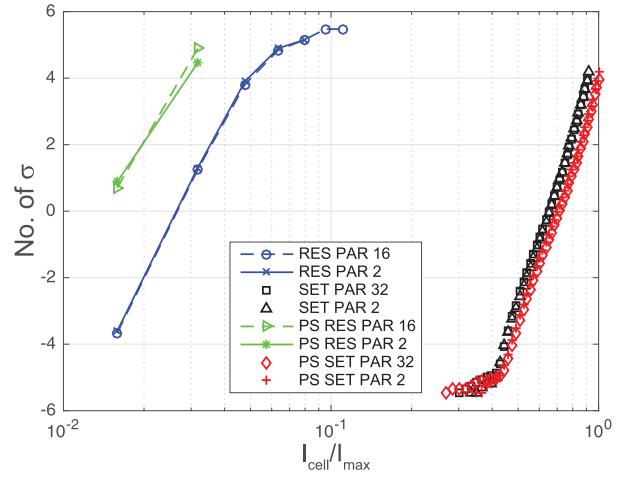


Fig. 16. Measured normal inverse distributions of cell current after different programming pulses carried out with different parallelisms as shown in the legend. 8 Mcells per distribution. The data corresponding to low-parallelism programming (i.e., PAR = 2) were obtained without the EPC, whereas the high-parallelism programming distributions (i.e., PAR = 16, 32) were achieved with the EPC.

to the connection of the gates of 32 M[k] transistors (as pointed out in Section III), then the enhanced CM sinks the charges added to this node due to charge sharing much faster with respect to the conventional mirror and, hence, turns the M[k] transistors on much earlier (vertical blue dashed line, to be compared with the vertical dashed red line).

The effectiveness of the proposed high-parallelism programming approach is illustrated in Fig. 16, where the measured cell current distributions obtained after different kinds of programming pulses with RESET parallelism of 16 and SET parallelism of 32, are compared to the cell current distributions obtained using the standard programming circuit with low parallelism (two for both SET and RESET operations). It is apparent that the increased programming parallelism does not degrade the obtained current distributions, and that it is therefore possible to achieve an equivalent programming accuracy while increasing program throughput.

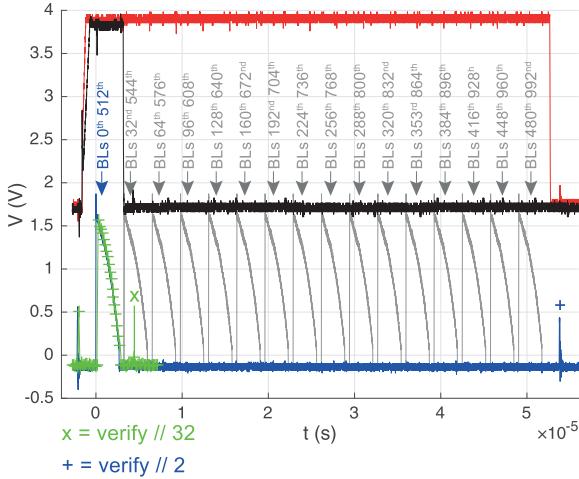


Fig. 17. Measured voltages V_{REG} (top—red and black traces) and $V_{\text{BL}[0]}$ (bottom—blue and green traces) during two “all 1s” (from “all 0s”) programming algorithms implemented with parallelism of 2 (red and blue traces) and 32 (black and green traces). In the left part of the lower graphs, the green curve is superimposed to the blue curve. The gray traces are included for illustrative purpose: these traces show the pulses on the other BLs, which could not be measured.

In order to illustrate different time durations of the high- and low-parallelism programming algorithms, the measured evolution over time of voltages V_{REG} and V_{BL} is illustrated in Fig. 17. The traces at the top depict the measured output voltage of the proposed regulator when a SET program procedure is carried out on all the DCs of a word (i.e., when 32 PCM cells are programmed to the SET state) exploiting a parallelism equal to two (red curve) and 32 (black curve). Analogously, the voltage at BL[0] during the same operations is represented at the bottom with green and blue traces for the case of 32- and 2-parallelism, respectively. The gray waveform illustrates the time position of the SET pulses on the other BLs, which are not provided with a micro-pad (due to area constraints) and, therefore, could not be measured. It is apparent that the proposed solution requires much shorter programming time than the conventional one: this implies that the standard solution has to keep the programming circuitry active for a longer time. Therefore, the proposed solution not only provides substantial time saving but also allows a more efficient power usage.

In the proposed memory, DCs and CCs are programmed separately. Since, obviously, SET and RESET pulses are also applied separately, a complete program cycle for a number of bits programmed simultaneously consists of four steps, namely, DC SET, DC RESET, CC SET, and CC RESET. Table III summarizes the measured performance in terms of word modify time with different degrees of parallelism in a variety of cases of standard SET and RESET algorithms and during the preliminary conditioning of PCM cells carried out at EWS, namely, during pre-soldering (PS) SET, PS RESET (which are applied before soldering the die to achieve stronger—and, hence, more stable—SET and RESET state, respectively), and forming (which is applied at the beginning of the device life prior to any programming operation, aiming at optimizing the physical structure of the chalcogenide alloy

TABLE III
MEASURED MODIFY TIME FOR DIFFERENT INITIAL AND FINAL CONDITIONS. (PAR X-Y REFERS TO RESET AND SET PARALLELISM)

	No. bits to program	Data (hex)		Time (μs)		Saved time
		Initial	Final	PAR 2–2	PAR 16–32	
User mode	0	0000 0000	0000 0000	4.28	4.28	0.0%
	32 (best)	0000 0000	FFFF FFFF	74.41	15.39	79.3%
	32 (worst)	5555 AAAA	AAAA 5555	138.69	20.64	85.1%
	16 (best)	0000 0000	F0F0 F0F0	44.50	15.39	65.4%
	16	F0F0 F0F0	AAAA AAAA	78.77	20.24	74.3%
	16	AAAA AAAA	5555 AAAA	78.86	20.24	74.3%
	16	F0F0 F0F0	F0F0 0F0F	78.85	20.65	73.8%
	16 (worst)	00FF OFF0	F00F 00FF	78.86	20.65	73.8%
EWS	33 PS RES	—	—	10.21	4.25	58.4%
	33 PS SET	—	—	91.01	8.54	90.6%
	33 Forming	—	—	436.25	28.76	93.4%

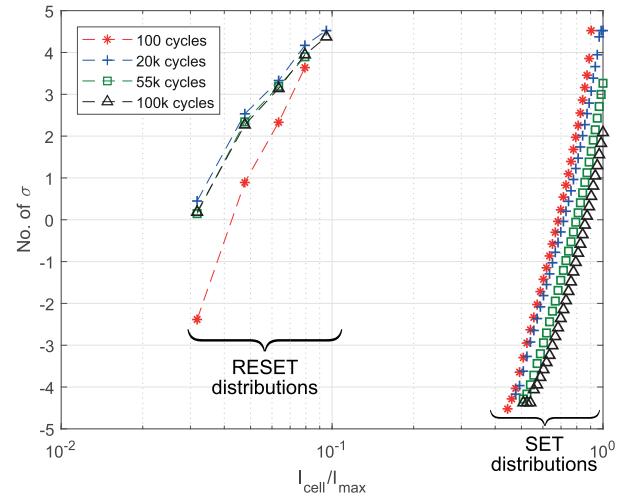


Fig. 18. Measured normal inverse distributions of cell current in both the RESET (left curves) and the SET state (right curves) after 100, 20000, 55000, and 100000 programming cycles. Each cycle consists of a RESET algorithm followed by a SET algorithm (see Section IV).

by means of fully crystallizing the active volume). It is worth to point out that, in the last cases, it is possible to operate with the maximum programming parallelism of 33 (32 standard array cells + 1 redundancy cell). As given in Table III, the word modify time is reduced by different factors depending on both the data that has to be written and the data currently stored in the cells. However, the most significant gain, from an engineering point of view, is the 85.1% decrease in the worst-case time window during programming operation. Moreover, a substantial decrease (up to 93.4%) of the word modify time is also obtained when the EWS pulses are applied, thus achieving a significant cost reduction.

Fig. 18 shows the normal inverse distribution of the cell current in both the SET and the RESET state measured after up to 100000 programming cycles (where one cycle corresponds to two write operations, i.e., from logical “0” to logical “1” and vice versa), performed in order to demonstrate the memory endurance. It can be clearly seen that not only the distributions do not present any evidence of decay but also that the SET distribution improves with increasing number of programming cycles. This enhancement is ascribed to a better crystallization of the chalcogenide volume above the active

TABLE IV
GENERAL CHARACTERISTICS OF THE EMBEDDED MEMORY

Technology	ePCM 0.11 μm BCD	Memory cell size	$0.7 \mu\text{m}^2$
Memory capacity	32 KByte	Area	0.7 mm^2
LV supply (V_{DD})	$1.55\text{--}1.95 \text{ V}$	HV supply (V_{PP})	$4.5\text{--}5.5 \text{ V}$
Temp. range	-40°C to 150°C	Data bus	32 bit
Random access time	18 ns	Endurance	100 k cycles

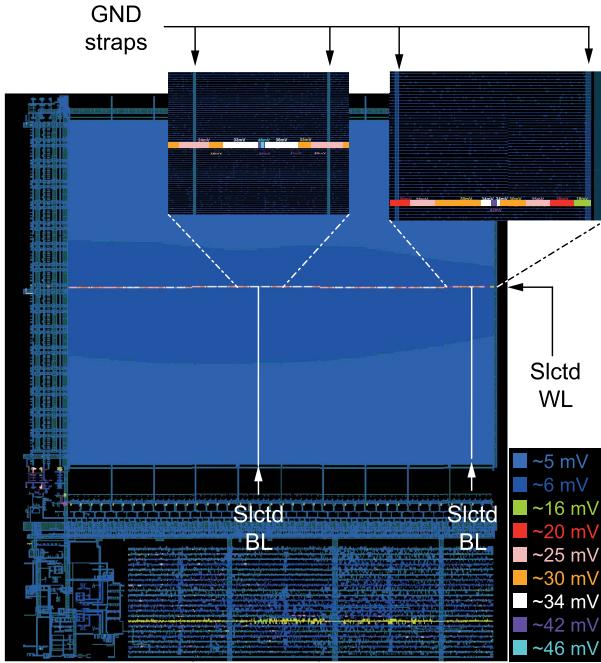


Fig. 19. Simulated voltage drop on the ground path during a high-parallelism write operation.

portion of the phase change material, which acts as a parasitic series resistance in the reading path. Table IV summarizes characteristics and specifications of the memory.

VI. RELIABILITY ANALYSES: IR DROP AND ELECTROMIGRATION

In order to achieve a more robust design, improve reliability, and guarantee the correct behavior of the proposed circuit, a dynamic IR voltage drop analysis was run on the entire memory focusing, in particular, on the internal power-supply paths (i.e., V_{DD} , V_{PP} , and ground) during a high-parallelism write operation when the selected cells are driven with the maximum programming (RESET) current. In addition, the simulation conditions (i.e., $T = +150^\circ\text{C}$ and fast transistor models) were chosen to investigate the worst-case scenario for reliability aspects. The simulated IR voltage drop (determined by the parasitic resistance of the selected path) on the ground distribution tree of the entire memory is highlighted in Fig. 19 with the aid of a color scale. In particular, the two insets depict the IR drop on the source line of the selected row: it is interesting to observe the variation of the drop as a function of the distance between the cell under programming and the closest ground straps as well as the asymmetry introduced by boundary effects. Indeed, the maximum voltage drop (around 46 mV) is reached midway between two ground straps, which, on the opposite, are the source line points that show the minimum voltage drop (around 25 mV).

Moreover, several simulations were run on the entire memory, during both read and write operations, to estimate the maximum current density that flows in the conductive paths, aiming at minimizing the electromigration phenomenon and, thus, at increasing the lifespan of the memory circuit. Thanks to this approach, it has been possible to properly size both the width of metal lines and the number of parallel vias in order not to exceed the maximum allowed current density even when the maximum programming parallelism is employed.

VII. CONCLUSION

The key features of the embedded PCM presented in this paper are intrinsic robustness against the resistance drift of both the SET and the RESET states, a random read access time of 18 ns in the whole temperature (-40°C to 150°C) and supply ranges (V_{DD} : 1.55–1.95 V, V_{PP} : 4.5–5.5 V) that meet automotive specifications, and enhanced program throughput (average program throughput equal to 3.6 Mcells/s, which corresponds to 1.8 Mbit/s, when using SET pulses of few microseconds, as required to achieve proper crystallization in the Ge-rich chalcogenide alloy used in this work). The two first features are achieved thanks to a differential sensing scheme based on two-cell per bit storage (two memory cells store data in complementary states and feed the two inputs of a differential sense amplifier during reading), whereas the third feature is obtained by means of improved programming circuits that allow high program parallelism (in particular, an improved operational amplifier for the voltage regulator that supplies the program circuits and a CM provided with very fast recovery when a high capacitive load is connected to its control branch). The proposed programming circuits can also be used in the conventional PCMs (1 bit/cell storage and shorter programming pulses), thus resulting in higher program throughput.

REFERENCES

- [1] S. Tyson, G. Wicker, T. Lowrey, S. Hudgens, and K. Hunt, "Nonvolatile, high density, high performance phase-change memory," in *Proc. IEEE Aerosp. Conf.*, vol. 5, Mar. 2000, pp. 385–390.
- [2] G. W. Burr *et al.*, "Recent progress in phase-change memory technology," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 6, no. 2, pp. 146–162, Jun. 2016.
- [3] Y. Choi *et al.*, "A 20 nm 1.8 V 8 Gb PRAM with 40 MB/s program bandwidth," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2012, pp. 46–47.
- [4] C. Zambelli, G. Navarro, V. Sousa, I. L. Prejbeanu, and L. Perniola, "Phase change and magnetic memories for solid-state drive applications," *Proc. IEEE*, vol. 105, no. 9, pp. 1790–1811, Sep. 2017.
- [5] H. Y. Cheng *et al.*, "Novel fast-switching and high-data retention phase-change memory based on new Ga-Sb-Ge material," in *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 2015, pp. 3.5.1–3.5.4.
- [6] W. C. Chien *et al.*, "Reliability study of a 128 Mb phase change memory chip implemented with doped Ga-Sb-Ge with extraordinary thermal stability," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2016, pp. 21.1.1–21.1.4.
- [7] G. De Sandre *et al.*, "A 90 nm 4 Mb embedded phase-change memory with 1.2 V 12 ns read access time and 1 MB/s write throughput," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2010, pp. 268–269.
- [8] Y. Taito *et al.*, "A 28 nm embedded SG-MONOS flash macro for automotive achieving 200 MHz read operation and 2.0 MB/S write throughput at T_i of 170°C ," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2015, pp. 1–3.

- [9] P. Zuliani *et al.*, "Overcoming temperature limitations in phase change memories with optimized Ge_xSb_yTe_z," *IEEE Trans. Electron Devices*, vol. 60, no. 12, pp. 4020–4026, Dec. 2013.
- [10] G. F. Close *et al.*, "A 256-mcell phase-change memory chip operating at 2+ bit/cell," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 6, pp. 1521–1533, Jun. 2013.
- [11] N. Ciocchini, E. Palumbo, M. Borghi, P. Zuliani, R. Annunziata, and D. Ielmini, "Modeling resistance instabilities of set and reset states in phase change memory with Ge-Rich GeSbTe," *IEEE Trans. Electron Devices*, vol. 61, no. 6, pp. 2136–2144, Jun. 2014.
- [12] A. Athmanathan, M. Stanisavljevic, N. Papandreou, H. Pozidis, and E. Eleftheriou, "Multilevel-cell phase-change memory: A viable technology," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 1, pp. 87–100, Mar. 2016.
- [13] M. Le Gallo, A. Sebastian, D. Krebs, M. Stanisavljevic, and E. Eleftheriou, "The complete time/temperature dependence of I–V drift in PCM devices," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Pasadena, CA, USA, Apr. 2016, pp. MY-1-1–MY-1-6.
- [14] X. Peng and W. Sansen, "Transconductance with capacitances feedback compensation for multistage amplifiers," *IEEE J. Solid-State Circuits*, vol. 40, no. 7, pp. 1514–1520, Jul. 2005.
- [15] B. K. Ahuja, "An improved frequency compensation technique for CMOS operational amplifiers," *IEEE J. Solid-State Circuits*, vol. SSC-18, no. 6, pp. 629–633, Dec. 1983.
- [16] J. H. Huijsing, "Multi-stage amplifier with capacitive nesting for frequency compensation," U.S. Patent 4 559 502, Dec. 17, 1985.
- [17] R. G. H. Eschauzier, L. P. T. Kerklaan, and J. H. Huijsing, "A 100-MHz 100-dB operational amplifier with multipath nested-Miller compensation," *IEEE J. Solid-State Circuits*, vol. 27, no. 12, pp. 1709–1716, Dec. 1992.
- [18] S. O. Cannizzaro, A. D. Grasso, R. Mita, G. Palumbo, and S. Pennisi, "Design procedures for three-stage CMOS OTAs with nested-Miller compensation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 5, pp. 933–940, May 2007.
- [19] R. Castello, "CMOS buffer amplifier," in *Analog Circuit Design*, J. Huijsing, R. van der Plassche, and W. Sansen, Eds. Boston, MA, USA: Kluwer Academic, 1993, pp. 113–138.
- [20] F. Bedeschi *et al.*, "A bipolar-selected phase change memory featuring multi-level cell storage," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 217–227, Jan. 2009.
- [21] C. Villa, D. Mills, G. Barkley, H. Giduturi, S. Schippers, and D. Vimercati, "A 45 nm 1 Gb 1.8 V phase-change memory," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2010, pp. 270–271.
- [22] F. Bedeschi, C. Boffino, E. Bonizzoni, C. Resta, and G. Torelli, "Staircase-down SET programming approach for phase-change memories," *Microelectron. J.*, vol. 38, nos. 10–11, pp. 1064–1069, Oct./Nov. 2007.



Marco Pasotti joined STMicroelectronics, Agrate Brianza, Italy, in 1994. His current research interests include high-performance embeddable flash and phase change memories, multi-level flash memories, cost-effective NVM solutions, new technologies for non-volatile memories, and all related application aspects.



Riccardo Zurla received the Ph.D. degree in microelectronics from the University of Pavia, Pavia, Italy, in 2018.

He is currently a Research Fellow with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia. His research activity was carried out in the frame of a cooperation, started in 2015, with the Smart Power Technology Group of STMicroelectronics. His current research interests include the design of analog circuits for non-volatile memories and the experimental characterization of phase change memory cell arrays.



Marcella Carissimi received the Laurea Engineering degree (*laude*) in electronic engineering from the Politecnico of Milano, Milan, Italy.

She joined STMicroelectronics, Agrate Brianza, Italy, in 2001, where she has been involved in the design of flash memory test vehicles aimed at technology development. Since 2015, she has been leading the project of silicon demonstrators for embedded phase change memory (PCM) for BCD technology. Her current research interests include the field of digital and mixed-mode design for non-volatile memories in the context of embedded applications.



Chantal Auricchio received the Laurea degree in electronic engineering from the Politecnico of Milano, Milan, Italy, in 1995.

In 1994, she joined STMicroelectronics, Agrate Brianza, Italy, for an internship in central research and development process development and later she moved toward both analog and digital design. She is currently involved in the design of high-performance embeddable phase change memories.



Donatella Brambilla received the Laurea degree in electronic engineering from the Politecnico of Milano, Milan, Italy, in 1996, with an internship in STMicroelectronics, Agrate Brianza, Italy.

She has been with STMicroelectronics as an Analog Designer for hard disk driver products in BCD technology, a memory designer of volatile memories in both CMOS and BCD technologies and of non-volatile memories. She is currently working on digital controller implementation of embedded phase change memory for BCD technology development.



Emanuela Calvetti received the Laurea degree in electronic engineering and the Ph.D. degree in organic thin-film transistor modeling from University of Brescia, Brescia, Italy.

In 2006, she joined STMicroelectronics, Agrate Brianza, Italy, where she is working on FTP memories, flash and currently on phase change memories.



Laura Capecchi received the Laurea degree from the Politecnico of Milano, Milan, Italy, in 2002.

In 2003, she joined STMicroelectronics, Agrate Brianza, Italy, where she is involved in the design and development of high-performance embedded non-volatile memories.



Luigi Croce joined STMicroelectronics, Agrate Brianza, Italy, in 1989. His current research interests include high-performance embeddable flash and phase change memories, new technologies for non-volatile memories, and all related testing and application aspects.



Daniele Gallinari received the Laurea degree in electronic engineering from the University of Parma, Parma, Italy, in 2003.

He joined STMicroelectronics, Agrate Brianza, Italy, in 2004, where he is mainly involved in eNVM testing and characterization for smart power applications.



Cristina Mazzaglia joined STMicroelectronics, Agrate Brianza, Italy, in 2014, where she is working as a Digital Design and Test Engineer. Her current research interests include the development of a test environment for silicon validation, EWS flow qualification, and DFT of phase change memory (PCM) test chips.



Vikas Rana joined STMicroelectronics, Greater Noida, India, in 2003. His current research interests include high-performance embeddable flash and phase change memories, cost-effective NVM solutions, new technologies for non-volatile memories, and all related application aspects.



Alessandro Cabrini received the Laurea degree (*summa cum laude*) and the Ph.D. degree in electronic engineering from the University of Pavia, Pavia, Italy, in 1999 and 2003, respectively. During his Ph.D. work, he focused on the development of circuits for non-volatile semiconductor memories with high-density storage capability.

From 2004 to 2011, he was a Research Fellow with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, where he is currently an Assistant Professor. His current research interests include non-volatile memories and, in particular, the design and characterization of innovative storage devices such as multi-level phase change and flash memories and ReRAM, and also the design of analog circuits for the de–dc power conversion (especially fully integrated architectures based on charge pump structures), non-linear electronic systems and the development of circuits and algorithms for the analysis of electroencephalographic records (EEG).



Guido Torelli (M'90–SM'96) was born in Rome, Italy, in 1949. He received the Laurea degree (*hons.*) in electronic engineering from the University of Pavia, Pavia, Italy, in 1973.

After graduating, he worked one year at the Institute of Electronics, University of Pavia, on a scholarship. In 1974, he joined SGS-ATES (now part of STMicroelectronics), Milan, Italy, where he was a Design Engineer for MOS IC's and was involved in both digital and mixed analog–digital circuit development, and where he became the Head of the MOS IC's Design Group for Consumer Applications and was appointed Dirigente. Since 1987, he has been with the University of Pavia, where he is currently a Full Professor. His current research interests include MOS IC design. In this frame, he is currently involved in the fields of non-volatile memories (including phase change memories) and CMOS analog circuits.

Prof. Torelli was a co-recipient of the Institute of Electrical Engineers Ambrose Fleming Premium (session 1994–1995).