Temperature Compensation Schemes for In-Memory Computing using Phase-Change Memory

Iason Giannopoulos, Manuel Le Gallo, Vara Prasad Jonnalagadda, Evangelos Eleftheriou and Abu Sebastian
IBM Research - Zurich, 8803 Rüschlikon, Switzerland
Email: nno@zurich.ibm.com; ase@zurich.ibm.com

***Abstract*—The explosive growth in data-centric artificial intelligence related applications necessitates exploration of non-von Neumann computing paradigms such as in-memory computing. The ability to perform certain computational tasks within the memory unit will reduce dramatically the time and energy that is spent into shuttling the data from the memory to the processing unit. However, the nanoscale resistive memory devices that are useful for these technologies suffer from non-ideal characteristics. In this work we deal with the computational precision loss due to the strong and inhomogeneous temperature dependence of resistive devices and in particular phase-change memory. We describe a temperature compensation method that applies to resistive crossbar arrays and its realization as a peripheral circuit. We derive array-level temperature compensation functions that are remarkably effective for projected phase-change memory devices. We simulate the system and experimentally validate its efficacy in the task of matrix-vector multiplications. The computational precision is found to be equivalent to an 8-bit multiplier at elevated temperatures.**

I. Introduction

In-memory computing is a promising non-von Neumann computing paradigm, in which nanoscale resistive memory devices are simultaneously storing data and performing basic computational tasks such as logical and arithmetic operations [1]–[3]. For example scalar multiplication can be performed using such devices using Ohm's law. Moreover, if these devices are organized in a crossbar configuration, the same concept extends to compute matrix-vector products by utilizing Kirchhoff's current law in addition to Ohm's law. However, inter-device variability, non-ideal characteristics of resistive memory such as electronic noise, drift and temperature dependence limit the computational precision and have to be addressed. One of the key challenges is that of the temperature dependence of the conductance states, that causes conductance distortions at the stored matrix elements leading to erroneous calculations [4]. Note that resistive memory devices such as metal-oxide resistive random access memory (ReRAM) and phase change memory (PCM) typically exhibit thermally activated electrical transport. The inter-device variation of activation energy makes an array-level compensation scheme very challenging. In this article we propose a temperature compensation method that applies to specifically designed resistive-memory devices, namely the projected phase-change memory, and exploits their unique temperature dependence of conductance.

II. In-Memory Multiplication

Information is stored in resistive memory devices in terms of their conductance states. PCM has been extensively studied in particular for in-memory computing applications [5]–[8]. The phase-change material in a PCM device undergoes transitions from the highly conductive crystalline to the significantly less conductive amorphous phase. The volume ratio between amorphous and crystalline portions determines the conductance state. Therefore by gradually increasing the amorphous volume the device shows multilevel storage capabilities. Yet the vast majority of non-ideal electronic characteristics are attributed to the amorphous phase [9]. The *projected PCM* is an emerging nanoscale device technology, in which the programming and the read-out mechanisms are decoupled [10]. A lateral projected line-cell consists of a phase-change material layer in parallel with the projection material, which is typically a metal nitride (Fig. 1(a)). Due to the highly non-linear voltage-current characteristics of the amorphous phase, the high-amplitude programming pulse can modulate the amorphous volume, while the low-amplitude read pulse bypasses this volume and current flows through the more conductive projection segment. Therefore the information retrieval signal is marginally affected by the amorphous phase non-idealities. This design is found to be remarkably immune to conductance variations arising from structural relaxation, 1/f noise and temperature variations. Due to the multilevel programming capabilities of PCM one can set the device to the desired conductance state with high precision using iterative programming. In a projected cell these states are temporally stable, showing drift coefficients up to 50 times reduced compared to conventional PCM (Fig. 1(b)).

Scalar multiplication can be performed on a resistive memory device when the one variable is mapped proportionally to the read voltage and the other to the device conductance state. According to Ohm's law, the read-current corresponds to the product value, from which an approximation of the exact result is deciphered (Fig. 1(a)). By arranging these devices in a crossbar configuration and by invoking Kirchhoff's current law, one can multiply a matrix by a vector (Fig. 1(c)). The product $Ax = b$ can be computed by mapping the elements of matrix A to conductance values within the dynamic range of the devices and the elements of x to read voltages applied to the rows of the crossbar. The resulting column currents are proportional to the elements of b (Fig. 1(d)). Alternatively,

978-1-7281-4922-620$31.00

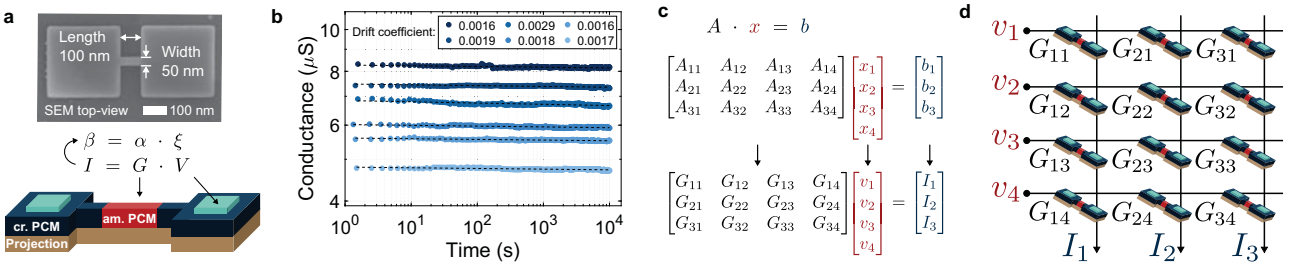

Fig. 1. (a) Schematic illustration of a lateral projected PCM device and a top view SEM image during fabrication. An amorphous volume is created in the 50 nm thin line and the device conductance can be mapped as one of the multiplication variables. (b) Multilevel programming capabilities of projected PCM. The programmed conductance is temporally stable with remarkably reduced drift coefficients compared to conventional PCM. (c) The mapping process in a matrix-vector multiplication. The matrix elements are encoded to conductance values and the vector to row-bias voltages. The result is decoded from the column current according to Ohm's and Kirchhoff's laws. (d) A crossbar arrangement of projected PCM devices that is used to perform the multiply operation.

the vector x can be encoded to pulse duration and then the result occurs from the integrated charge. Previous works have shown that the achieved precision of matrix-vector multiply operation at room temperature is comparable to 4-bit fixed point arithmetic for conventional PCM [11] and to 8-bit for the projected memory [12]. The latter remarkable result is attributed to the significantly low conductance variations associated with the programmed states.

In Section III, we present an overview of array-level temperature compensation methods. We will particularly focus on a method that involves the use of temperature compensation functions that are analytically derived. Subsequently, these compensation functions are derived for the projected PCM devices. In Section IV, we validate the efficacy of this compensation method using simulations and experiments.

III. Temperature Compensation Method

A. Approaches towards array-level temperature compensation

Column currents are measured by ADCs and the output vector is fed to a decoding unit that provides the final result. However, variations of the operating conditions such as temperature alter the stored matrix elements, leading to erroneous calculations. A compensation unit is required, in order to correct the temperature-induced distortions of the output vector. A system-level design of this unit consists of an adjustment circuit configured for receiving current values at an actual operating condition and multiplying them by an appropriate correction factor (Fig. 2). The fastest and most efficient solution is to multiply the whole vector with a single number that is only a function of temperature and has no dependence on either the device or the conductance state. In other words adjusting the output current values to the ones that would have been measured, if the system was at an operating condition that is pre-defined as the reference one. One approach comprises a dedicated column in the crossbar, which is programmed to reference conductance values [6]. This method is directly probing the temperature effect on the devices and the correction factor occurs as the ratio of the total column conductance measured at the actual operating condition and the reference one. Alternatively, a single-variable

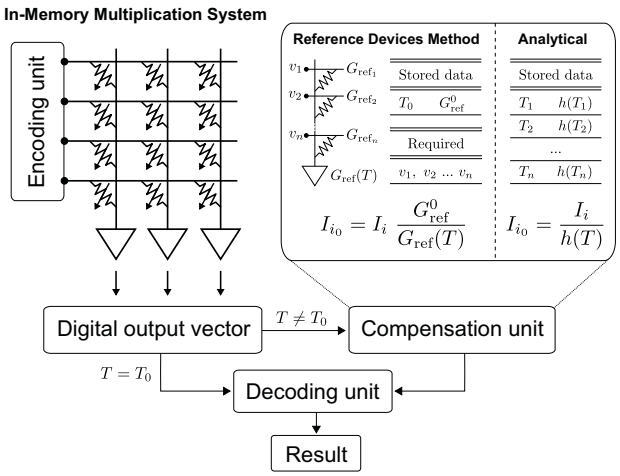

Fig. 2. Block-diagram of the main components of an in-memory multiplication system with temperature compensation. The encoded input is applied at the crossbar rows and the column currents are converted to a digital array via ADCs at the periphery. When the ambient temperature differs from the reference one, this vector is multiplied by a correction factor within the compensation unit using either of the proposed methods. The final result is the deciphered corrected vector.

compensation function $h(T)$ can be computed for a set of temperature values and be stored as a look-up table. Although the former approach is simpler and may capture additional effects e.g. deviations emerging from device aging, it is a more computationally intensive routine and would yield to a system with significantly lower efficiency, due to the fact that it requires an additional calculation between the input voltage vector and the measured current to get the correction factor. Next we analytically derive the temperature compensation function, $h(T)$, for projected PCM devices.

B. Compensation functions for projected PCM

The compensation functions, $h(T)$, have to be derived individually for the employed resistive technology. For example in conventional PCM the amorphous phase exhibits thermally activated electrical transport, that is described by:

$$G_\alpha(T) = G^*_\alpha \; e^{-\frac{E_\alpha}{k_B T}} \tag{1}$$

The activation energy E_α has a minor temporal dependence (due to structural relaxation of the amorphous phase) and shows inter-device variability. These effects limit dramatically the ability of a compensation scheme to precisely predict the stored matrix elements at any temperature. E_α is a material property, that typically follows Gaussian distribution around a mean value. On the other hand, the projected PCM has significantly weaker temperature dependence, which can be described by a linear approximation.

$$G_p(T) = \frac{G_{p0}}{1 + \alpha_p\,(T - T_0)} \tag{2}$$

A unique advantage of these devices is that the temperature coefficient of conductance α_p is a property of the projection material, a fact that makes their behavior state-independent and the system is much more amenable to an effective compensation scheme. The amorphous phase and the projection segment can be modeled as 2 resistors in a parallel configuration. In an ideal device the conductance value should be completely dominated by the projection segment, therefore it should be at least 100 times more conductive than the amorphous volume. In that case the compensation function is directly given as the temperature dependence of the projection segment.

$$h(T) = \frac{1}{1 + \alpha_p\,(T - T_0)} \tag{3}$$

The scheme that uses the compensation function given by (3) is referred to as the *first order* compensation scheme. However, neglecting the contribution of the amorphous volume limits the achievable precision without offering any efficiency or complexity advantage to the system. Moreover, due to the much weaker temperature dependence of the projection material compared to the amorphous phase, their conductance ratio reduces significantly as the temperature deviates from the reference T_0. Therefore to maximize precision there is a need for an analytic model.

Activation energy is fundamentally uncertain and the best approximation is given by the mean experimental value $\bar{E_\alpha}$. The projection/amorphous conductance ratio (4) as a function of temperature is given by combining (1) and (2).

$$\lambda(T) = \frac{G_{p0}}{1 + \alpha_p\,(T - T_0)} \; \frac{1}{G^*_\alpha} \; e^{\frac{\bar{E_\alpha}}{k_B T}} \tag{4}$$

We define $\lambda_0 = \lambda(T_0)$ as the conductance ratio at the reference temperature. Combining (4) with the fact that the total conductance is the sum of the conductance values associated with the amorphous PCM and projection segments, we derive a compensation function (5) that has temperature as the only input variable. This is the *second order* compensation scheme.

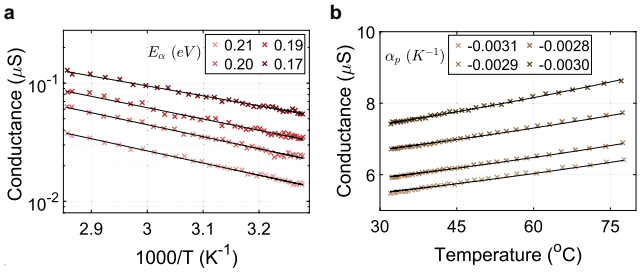

Fig. 3. (a) The experimentally measured temperature dependence of conductance in the amorphous phase is fitted by an Arrhenius equation. E_α is determined by the slope as shown for 4 different conductance states. (b) The experimental temperature dependence of projected PCM can be described using a linear equation. The coefficient α_p is extracted from a linear fit.

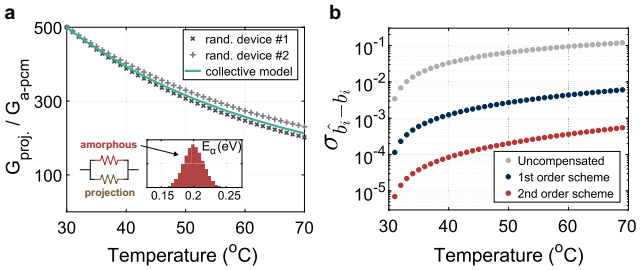

Fig. 4. (a) Two random devices of the simulated crossbar with activation energies E_{α_1} and E_{α_2} are selected to show their deviation from the collective model employed by the compensation scheme that uses $\bar{E_\alpha}$. The inset shows the simple two-parallel-resistor network that was used for simulations as well as the Gaussian distribution of E_α. (b) The simulated temperature dependence of computational precision in projected PCM is expressed as the standard deviation of the error between the crossbar-computed vector elements $\hat{b}_i$ and the exact ones computed with double precision b_i.

$$h(T) = \frac{1}{1 + \lambda_0} \left[\frac{\lambda_0}{1 + \alpha_p\,(T - T_0)} + e^{-\frac{\bar{E_\alpha}}{k_B}\left(\frac{1}{T} - \frac{1}{T_0}\right)} \right] \tag{5}$$

IV. Simulations and Experimental Validation

Although the compensation functions need only the ambient temperature measurement as input, material properties such as the temperature coefficient of conductance and the activation energy of the amorphous phase have to be experimentally determined. To experimentally validate the temperature compensation schemes, we fabricated and characterized conventional and projected PCM line-cells according to the design in Fig. 1(a). The activation energy of the amorphous phase can be extracted by the slope in an Arrhenius plot representation (Fig. 3(a)). The temperature coefficient of conductance in projected PCM is the slope of the linear dependence (Fig. 3(b)).

A. Large-scale crossbar simulation

To study the benefits of both compensation schemes, we simulated a 256x256 crossbar based on the experimental temperature characteristics of Fig. 3. The crossbar was populated with projected PCM devices and was used to compute a

matrix-vector multiplication $b = Ax$. The normally distributed between 0 and 1 elements of A and x were mapped to conductance values and read voltage amplitudes respectively. The activation energies were distributed in a Gaussian manner around the mean of $E_\alpha = 0.2\ eV$ with a standard deviation of $15\ meV$. The projection segment was modeled as a resistor in parallel to the amorphous phase volume, having a single thermal coefficient of conductance set to $\alpha_p = -0.003\ K^{-1}$. The crystalline phase was in series with this network and its contribution to the total conductance was considered negligible. For initialization we selected a representative conductance ratio of 500 between the projection segment and the amorphous phase at the reference temperature of 30^oC. The ratio reduces as a function of temperature depending on the exact E_α of each device. The collective model described by (4) is using the mean activation energy and was plugged in the compensation scheme (Fig. 4(a)). The simulated result $\hat{b}$ is compared with the exact (double precision) result b and the standard deviation of the error is plotted against temperature (Fig. 4(b)). We notice that the 1st order compensation scheme reduces the error by a factor of 30 and 20 at the low and high temperature range respectively. An additional factor of 15 is gained by the 2nd order scheme, which decreases to 10 at high temperatures.

B. Experimental validation

We confirmed the thermal model simulations with experiments on a small scale emulated crossbar. An example operation was set up based on Fig. 1(d) schematics. Projected PCM devices were programmed to values that correspond to a given 4x3 matrix and were measured individually for 5000 random input vectors. The column currents were added up in high precision. During these multiplications the ambient temperature was varied in a sinusoidal manner between 25 and 55^{o}C and was accurately measured with a calibrated thermocouple sensor. With the 1st order scheme employed, we notice that although the achieved precision is comparable with the 8-bit fixed-point arithmetic, the error is shifted towards the negative side of the chart. The amorphous phase contribution is not taken into account and this results to measured current that was higher than the model could predict (Fig. 5(a)). Data are much better centered around the mean, when the 2nd order scheme is used. A small group of points extends beyond the 8-bit error margin and towards the positive side of the chart (Fig. 5(b)). This is because temperature accelerated the state-relaxation of these devices making them less conductive than the thermal model estimates. According to the simulations one should expect to compute at least 10 times more precisely using the 2nd order scheme. Nevertheless, experiments showed only marginal benefits, because the expected gains were overshadowed by additional non-idealities of PCM devices such as electronic noise. This would not have been the case if the projection segment was more resistive and the contribution of the amorphous phase to the total conductance was stronger.

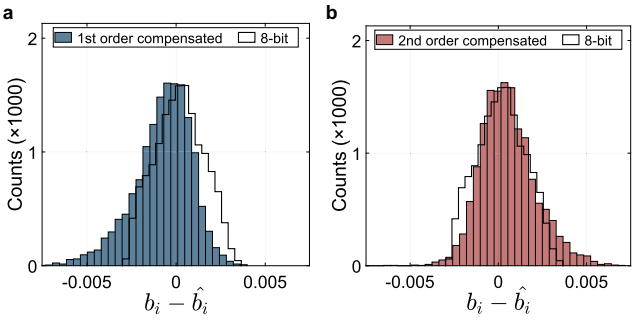

Fig. 5. Error distribution of 5000 matrix-vector multiplications at temperature conditions varying within the range 25-55^{o}C. The error of the temperature-compensated results using the 1st (a) and the 2nd (b) order compensation schemes are plotted along with the error of an 8-bit fixed point arithmetic multiplier computing the exact same tasks.

V. Conclusion

High precision in-memory computing is essential for future non-von Neumann systems targeting AI applications. In this work we demonstrated the difficulties of an effective temperature compensation method in resistive memory crossbars arising mainly from inter-device variability. We proposed a compensation method that applies to specific resistive devices, in which the component that dominates the device conductance has near-homogeneous temperature dependence. Focusing on the task of in-memory multiplication, we designed a system that can perform matrix-vector multiplications with high precision at elevated temperatures. We derived temperature compensation functions that capture the temperature dependence of projected PCM devices and simulated the compensation capabilities on a 256x256 crossbar. Finally we experimentally evaluated the compensation schemes at temperatures up to 55^{o}C and we matched the 8-bit equivalent precision for multiplication using projected PCM.

Acknowledgment

We acknowledge partial financial support from European Research Council (ERC) grant 682675.

References

[1] D. Ielmini and H.-S. Philip Wong, "In-memory computing with resistive switching devices", Nat. Electron., vol. 1, pp. 246–253, Apr. 2018.
[2] M.A. Zidan *et al.*, "A general memristor-based partial differential equation solver," Nat. Electron., vol. 1, pp. 411–420, Jul. 2018.
[3] A. Sebastian *et al.*, "Computational memory-based inference and training of deep neural networks", 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019, pp. T168-T169.
[4] Majed Valad Beigi and Gokhan Memik, "Thermal-aware Optimizations of ReRAM-based Neuromorphic Computing Systems", Proceedings of the 55th Annual Design Automation Conference, San Francisco, CA, June 2018, pp. 39:1-39:6.
[5] A. Sebastian *et al.*, "Temporal correlation detection using computational phase-change memory", Nat. Commun., vol. 8, no. 1115, 2017.
[6] M. Le Gallo *et al.*, "Mixed-precision in-memory computing," Nat. Electron., vol. 1, pp. 246–253, Apr. 2018.

[7] I. Boybat *et al.*, "Neuromorphic computing with multi-memristive synapses", Nat. Commun., vol. 9, no. 2514, 2018.
[8] A. Sebastian, M. Le Gallo and E. Eleftheriou, "Computational phase-change memory: beyond von Neumann computing", J. Phys. D: Appl. Phys., vol. 52, no. 44, 2019.
[9] M. Le Gallo, D. Krebs, F. Zipoli, M. Salinga and A. Sebastian, "Collective Structural Relaxation in Phase-Change Memory Devices", Adv. Electron. Mater., vol.4, 1700627, 2018.
[10] W. W. Koelmans *et al.*, "Projected phase-change memory devices", Nat. Commun., vol. 6, no. 8181, 2015.
[11] M. Le Gallo, A. Sebastian, G. Cherubini, H. Giefers and E. Eleftheriou, "Compressed Sensing With Approximate Message Passing Using In-Memory Computing," IEEE Transactions on Electron Devices, vol. 65, no. 10, pp. 4304-4312, Oct. 2018.
[12] I. Giannopoulos *et al.*, "8-bit Precision In-Memory Multiplication with Projected Phase-Change Memory," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 27.7.1-27.7.4.