RZ 3948 (# ZUR1808-020) 08/15/2018
Electrical Engineering 4 pages

Research Report

8-bit Precision In-Memory Multiplication with Projected Phase-Change Memory

I. Giannopoulos, A. Sebastian, M. Le Gallo, V.P. Jonnalagadda, M. Sousa, M.N. Boon, and E. Eleftheriou

IBM Research – Zurich
8803 Rüschlikon
Switzerland

The is the accepted version of the article published by IEEE: I. Giannopoulos, A. Sebastian, M. Le Gallo, V.P. Jonnalagadda, M. Sousa, M.N. Boon, and E. Eleftheriou "**8**-bit Precision In-Memory Multiplication with Projected Phase-Change Memory," Proc. 2018 IEEE International Electron Devices Meeting (IEDM) doi: 10.1109/IEDM.2018.8614558

8-bit Precision In-Memory Multiplication with Projected Phase-Change Memory

I. Giannopoulos*, A. Sebastian*, M. Le Gallo, V.P. Jonnalagadda, M. Sousa, M.N. Boon, and E. Eleftheriou
IBM Research – Zurich, 8803 Rüschlikon, Switzerland, email: nno@zurich.ibm.com, ase@zurich.ibm.com

***Abstract*—** In-memory computing is an emerging non-von Neumann approach in which certain computational tasks such as matrix-vector multiplication are performed using resistive memory devices organized in a crossbar array. However, the conductance variations associated with the memory devices limit the precision of this computation. Here, we demonstrate that the so-called projected phase-change memory (Proj-PCM) devices can achieve 8-bit precision while performing scalar multiplication. The devices were fabricated and characterized using electrical measurements and STEM investigation. They are found to be remarkably immune to conductance variations arising from structural relaxation, 1/*f* noise and temperature variations. Moreover, it is possible to compensate for the temperature-dependent conductance variations in a crossbar array using a simple model. Finally, we experimentally demonstrate a neural network-based image classification task involving 30 such Proj-PCM devices.

I. Introduction

In-memory computing is an emerging computing paradigm that has the potential to increase the performance and area/energy efficiency of several artificial intelligence related computational tasks [1-4]. However, the limited precision of in-memory computing remains a key challenge. Much of the research effort is focused on system-level or architectural solutions to address this problem [5,6]. Here, we propose a device-level solution to address this challenge based on the concept of projected phase-change memory (Proj-PCM) [7,8]. In a Proj-PCM device, there is a non-insulating projection segment in parallel to the phase-change segment. By exploiting the highly non-linear *IV* characteristics of phase-change materials, we can ensure that during the write process, the projection segment has minimal impact on the operation of the device. However, during read, conductance values of programmed states are mostly determined by the projection segment that appears parallel to the amorphous phase-change segment. Hereby, we demonstrate the efficacy of these devices with respect to in-memory multiplications.

II. Device Fabrication and Characterization

We fabricated Proj-PCM devices based on a lateral device geometry (Fig. 1). GeTe serves as the phase-change layer, while the projection layer consists of a metal nitride. By applying appropriate programming pulses that cause a melt-quench process on the as-fabricated crystalline GeTe, it is possible to modulate the amorphous region's size. During the read process, one measures the resistance of the projection layer, which is in parallel to the amorphous region. The resulting programing curve is shown in Fig. 2. To validate our assumption, that we have a single contiguous amorphous region separated by crystalline GeTe, we performed extensive STEM studies on these devices (Fig. 3). These studies also indicate that the amorphous region is not perfectly centered, which is indicative of additional thermo-electrical effects that have been shown to play a significant role in these nanoscale devices [9]. Scalar multiplication using resistive memory devices rely on Ohm's law. Hence it is of interest to study the field dependence of electrical transport in Proj-PCM devices. It can be seen that compared with conventional PCM devices, Proj-PCM devices show a much weaker field dependence (Fig. 4). The precision associated with scalar multiplication and subsequently matrix-vector multiplication is strongly determined by the conductance variations associated with these devices. For example, conventional PCM devices exhibit a temporal evolution of conductance (drift) attributed to the structural relaxation of the amorphous phase [10]. Proj-PCM devices show a 50-fold reduction in drift (Fig. 5). Besides drift, there are conductance fluctuations arising from the 1/*f* noise which is also found to be substantially lower in our devices (Fig. 6). The variation in conductance arising from temperature fluctuations is another key challenge given the highly thermally activated nature of electrical transport in phase-change materials. What is even more detrimental is that the activation energy tends to vary for different programmed states and as can be seen later, this poses key challenges for developing effective temperature compensation schemes. The Proj-PCM devices on the other hand show a substantially weaker dependence on temperature variations (Fig. 7).

III. In-memory Multiplication

A. Scalar Multiplication

To perform the scalar multiplication operation, $\beta = \alpha \cdot \xi$, the variable ξ is mapped proportionally to a read voltage and α into a conductance state of the Proj-PCM device. Due to Ohm's law, one can obtain an approximate result $\hat{\beta}$ of β from the resulting read current. We performed 20,000 scalar multiply operations on 12 conductance states of a Proj-PCM device (Fig 8a). Due to the analog nature of the programming curve, it is possible to program the device to a desired conductance state with high precision using iterative programing. The achieved precision of the scalar multiply operation is comparable to 8-bit fixed point arithmetic at room temperature (Fig. 8b,c). This remarkable result is attributed to the significantly low conductance variations associated with the programmed states. The average dissipated power in a Proj-PCM device for scalar multiplication is 60 nW and the average energy consumption, assuming a 100 ns read time provided by an integrated readout circuit, is 6 fJ (Fig. 9). The latter is 33x lower than an 8-bit digital multiplication in 45 nm (0.2 pJ) [11]. A comparative study was done using non-

projected PCM devices and one could observe substantially reduced precision (Fig. 10). This is due to the significantly higher drift, $1/f$ noise and non-Ohmic transport behavior.

B. Temperature compensation method

Remarkably, the 8-bit precision can be retained at elevated temperatures aided by a simple compensation scheme. The projection material's resistivity exhibits a well-defined length-independent temperature dependence. A compensation scheme was devised by multiplying the read current with a single-variable equation that describes the temperature dependence of the projection material, that is $f(T) = 1 + \alpha_p(T - T_0)$. The same value of α_p (-3.0×10^{-3} K^{-1}) was used independent of the device conductance state. We repeated the scalar multiplication experiment while varying the ambient temperature as a sinusoidal profile between 25 and 55ºC. Fig. 11a shows the resulting error and its elimination by the compensation scheme in scalar multiplication, recovering the 8-bit fixed point arithmetic comparable precision (Fig. 11b). A slight shift to higher $\hat{\beta}$ at high temperatures arises from the fact that the compensation scheme assumes zero contribution of the amorphous PCM in the total temperature behavior. This could be tackled with a more complex compensation scheme.

C. Matrix-vector multiplication

By invoking the Kirchhoff's current summation rule in addition to the Ohm's law, one can multiply a matrix by a vector. If resistive memory devices are organized in a crossbar configuration, the matrix-vector multiplication $A \cdot x = b$ can be performed by mapping the elements of A to conductance values and the elements of x to read voltages applied to the rows of the crossbar (Fig. 12). Subsequently, the elements of b are computed from the column currents. The temperature compensation scheme proposed earlier can be applied to the column current. First, we simulated a 256×256 crossbar and tested the temperature compensation method in both PCM and Proj-PCM devices (Fig. 14a). The temperature dependence of the phase-change material was captured by a model based on the experimental results of Fig. 7. The activation energy for electrical transport was assumed to be normally distributed around the mean value $\overline{E}_\alpha$ = 0.2 eV with a standard deviation of 15 meV. For the Proj-PCM case, the projection material was modelled as a parallel resistor with a single temperature coefficient (Fig. 13). Because of the parallel current path with weaker temperature dependence, the Proj-PCM crossbar outperforms the PCM one. More importantly, it is impossible to compensate for ambient temperature variations at a crossbar array level with PCM due to the significant variations in the activation energy values (Fig. 14b). Proj-PCM devices, on the other hand, with their state/device-independent temperature dependence of electrical transport are much more amenable to such a compensation scheme. In addition to the simulation studies, we experimentally emulated 2000 matrix-vector multiplications employing 12 Proj-PCM devices arranged in a 4×3 virtual crossbar configuration. An equivalent experiment was repeated under temperature variations spanning from 25 to 55ºC. The column currents were translated to $\hat{b}_i$ and plotted against the exact result b_i for constant and varying temperature (Fig. 15 & 16, respectively). The precision loss in the latter case was recovered by the compensation scheme.

IV. Pattern Classification

A single-layer neural network was experimentally emulated using 30 physical Proj-PCM devices arranged in a 10×3 crossbar (Fig. 17b). The network was trained to classify 3×3 pixel images to 3 classes [12]. It consists of 9 input nodes in which the pixel values are fed in, mapped as read voltages, while a 10^{th} input neuron serves as bias. Being a single layer network, the input and output neurons are directly connected, with the conductance of each of the 30 fully connected devices corresponding to the synaptic weight. The dot-product of the weights and inputs is calculated at each output neuron, which in our case is the column read current. Fig. 17a is a schematic representation of this network next to the training set, which consists of 3 images, the numbers 4, 1 and 0. The network was trained offline using the back-propagation algorithm. The resulting weights were mapped to conductance values and the ranges adjusted to match the dynamic range of the Proj-PCM devices (Fig. 17c). The classification accuracy was obtained on 2 test sets of 27 images under 2 scenarios of noise at the input neurons: 1) analogue noise introduced as a Gaussian distribution of pixel colors between 0 and 1 in the grayscale, and 2) digital noise that had one pixel of each original image flipped (Fig 18a). For the Gaussian noise a standard deviation value of 0.2 was chosen for the experiment (Fig. 17d). In both scenarios the classification accuracy was 100% at room and elevated temperature, even without the need to apply the temperature compensation method (Fig 18b & c).

V. Conclusions

We have conclusively shown 8-bit precise and low-power (60 nW) in-memory multiplication using Proj-PCM devices. We demonstrated scalar and matrix-vector multiplication with 8-bit precision, along with a method that corrects for the temperature variations and recovers the constant temperature precision. We also successfully implemented a single-layer neural network with 30 hardware Proj-PCM devices capable of errorless pattern classification at elevated temperatures. The 8-bit precision requires highly accurate conductance tuning as well as low-offset/low-noise analog circuitry, lest one of those factors become the actual limit on effective precision. Future work will aim at decreasing the absolute device conductance and enlarging the conductance window for easier integration in large neuromorphic crossbars, which should be achievable through material engineering and device scaling.

We acknowledge partial financial support from ERC grant 682675.

References

[1] M. Hu et al., Adv. Mater. 30.9, 1705914, 2018. [2] D. Ielmini et al., Nat. Electron. 1.6, 333, 2018. [3] G.W. Burr et al., Adv. in Phys.: X 2.1, 89-124, 2017. [4] A. Sebastian et al., Nat. Commun. 8.1, 1115, 2017. [5] M. Le Gallo et al., Nat. Electron. 1.4, 246-253, 2018. [6] I. Boybat et al., Nat. Commun. 9.1, 2514, 2018. [7] S. Kim et al., Proc. IEDM, 30.7.1-30.7.4, 2013. [8] W. Koelmans et al., Nat. Commun. 6, 8181, 2016. [9] J. L. M. Oosthoek et al., Rev. Sci. Instrum. 86, 033702, 2015. [10] M. Le Gallo et al., Adv. Electron. Mater., 1700627, 2018. [11] M. Horowitz, Proc. ISSCC, 10-14, 2014. [12] M. Prezioso et al., Nature 521, 61-64, 2015.

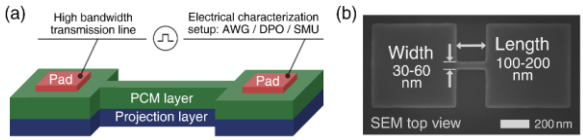

Fig. 1. (a) Schematic 3D view of a lateral Proj-PCM device. The bi-layer stack is encapsulated by SiO2 and 2 pads connect it to a characterization setup. (b) SEM image of a device during fabrication. Different active-area dimensions within the noted width/length range were fabricated and characterized.

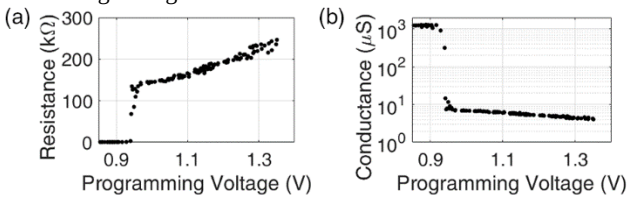

Fig. 2. (a) Programming curve of a Proj-PCM device. Higher voltage amplitudes increase the amorphous volume. (b) Any desired conductance state within the dynamic range can be achieved using iterative programing.

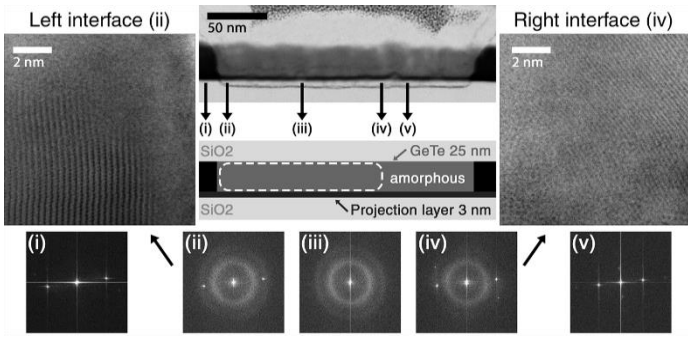

Fig. 3. Cross sectional STEM studies on the active area of a programmed cell. The color-coded and in-scale schematic explains the components of the device. High resolution images were obtained at the 5 marked areas (i-v) revealing the size and the location of the amorphous volume. In the left and right interface images the transition from ordered to disordered atomic configuration is clearly shown. FFT analysis points out that the characteristic spectra of the crystalline phase (discrete points: i,v) and the amorphous phase (ring-shaped pattern: iii) coexist at the interfaces (ii,iv).

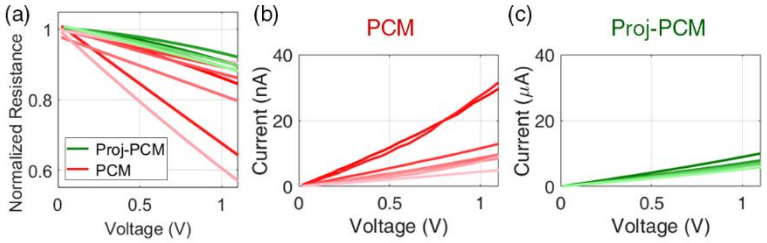

Fig. 4. (a) Normalized resistance versus voltage for different programmed states in PCM and Proj-PCM, showing the strong field dependence of PCM, and its weakening in Proj-PCM. Proj-PCM has consistently an Ohmic behavior over a wider range in the low-field regime, i.e. during read. (b) Current-voltage characteristics of PCM. (c) Current-voltage characteristics of Proj-PCM.

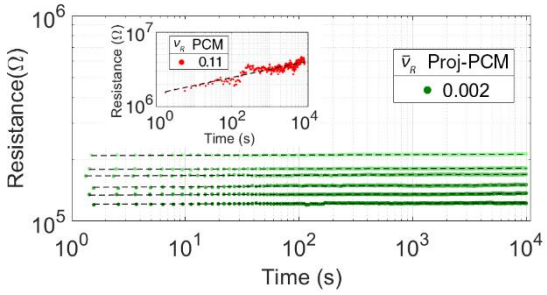

Fig. 5. Resistance drift of 6 resistance states in Proj-PCM. The drift coefficient ν_R is determined by a power-law fit and shows a 50-fold reduction in Proj-PCM compared to PCM (inset).

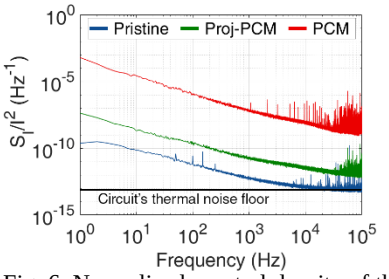

Fig. 6. Normalized spectral density of the read-current noise in the crystalline phase and in a programmed state for PCM and Proj-PCM, where it is 10^4 times lower.

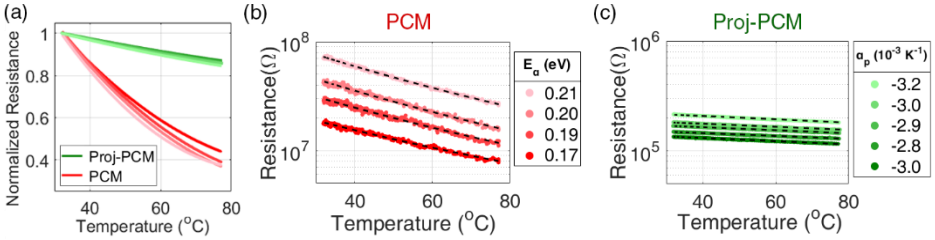

Fig. 7. (a) Normalized resistance versus temperature for different annealed states in PCM and Proj-PCM. (b) The exponential temperature dependence of resistance in PCM is fitted by an Arrhenius equation $R(T) = R^*\exp(E_\alpha/k_BT)$. Activation energy E_α is determined by the slope. (c) The temperature dependence of Proj-PCM can be described by a simple linear approximation $R(T) = \rho[1 + \alpha_p(T - T_0)]$. Temperature coefficient of resistance α_p is extracted via linear fit.

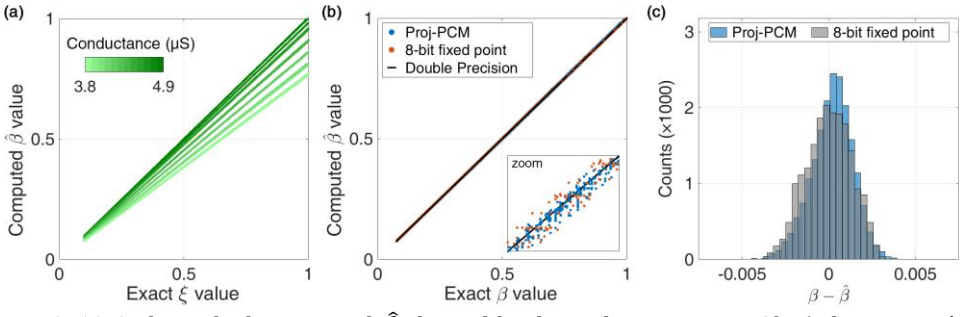

Fig. 8. (a) Scalar multiplication result $\hat{\beta}$ obtained by the read current using Ohm's law against ξ which is mapped in read voltage values, for Proj-PCM conductance states between G_{min} (3.8 μS) and G_{max} (4.9 μS). (b) Scalar multiplication result $\hat{\beta}$ computed using the Proj-PCM against both 8-bit fixed point arithmetic and the exact result β. (c) Error distribution of the 20000 scalar multiplication results for the Proj-PCM compared with the 8-bit fixed point arithmetic.

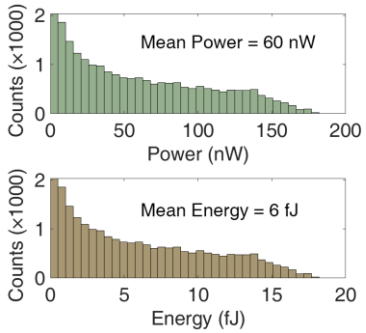

Fig. 9. Distribution of power per scalar multiply operation. Energy is calculated assuming 100 ns read time provided by an integrated readout circuit.

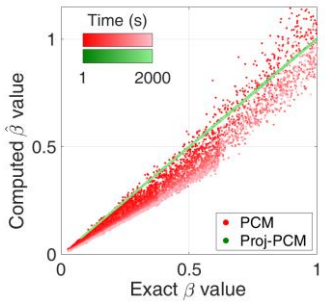

Fig. 10. Scalar multiplication result $\hat{\beta}$ in PCM and Proj-PCM against exact result β. Drift and 1/*f* noise in PCM cause errors.

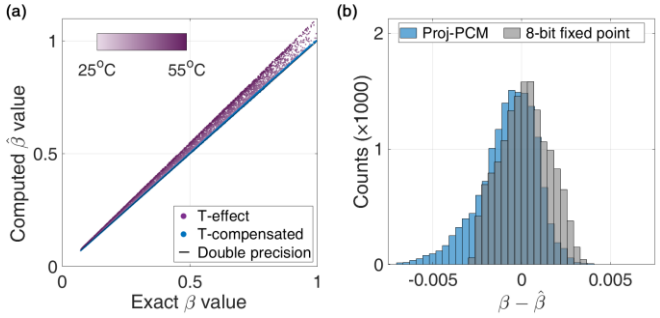

Fig. 11. (a) The effect of temperature in the experiment of Fig 8b. The temperature compensation model is used to correct read current and recover precision loss. (b) Error distribution for the temperature compensated Proj-PCM compared with 8-bit fixed point arithmetic.

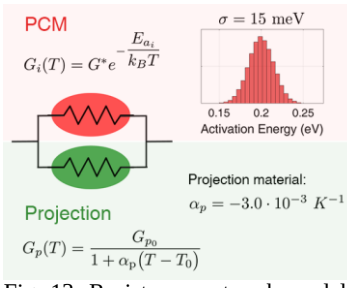

Fig. 13. Resistance network model of the amorphous PCM in parallel with the projection layer. Based on experimental data, E_α was normally distributed, while α_p was the same.

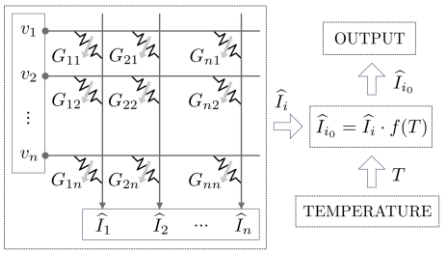

Fig. 12. Temperature compensation procedure in a crossbar that comprises Proj-PCM devices. Column current is corrected by the compensation equation: $f(T) = 1 + \alpha_p(T - T_0)$, in which temperature is the only required input.

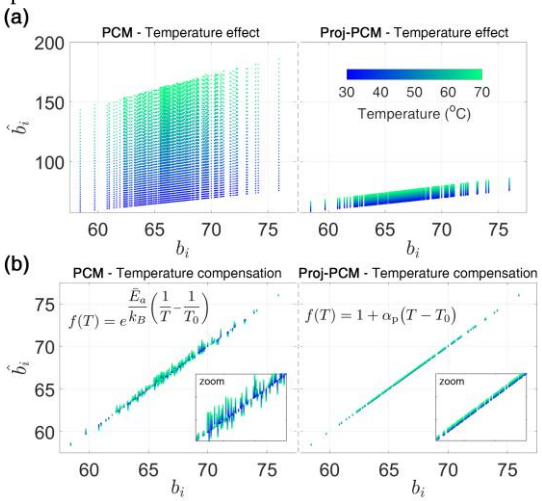

Fig. 14. (a) Simulations of the temperature effect on a matrix-vector multiplication in a 256x256 crossbar. (b) The inapplicability of a crossbar array level compensation scheme in PCM due to significant variations in the activation energy values, compared to Proj-PCM. For each case we used the corresponding compensation equation $f(T)$, as described in Fig. 12.

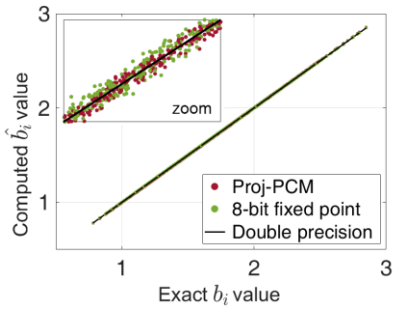

Fig. 15. 2000 experimental 4×3 matrix-vector multiplication results $\hat{b}_i$ computed with Proj-PCM against the exact result b_i and the 8-bit fixed point arithmetic.

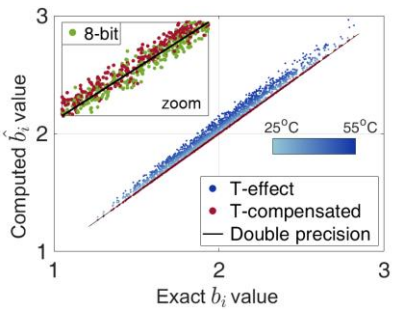

Fig. 16. 2000 experimental 4×3 matrix-vector multiplications at various elevated temperatures. Precision loss is recovered by the compensation scheme.

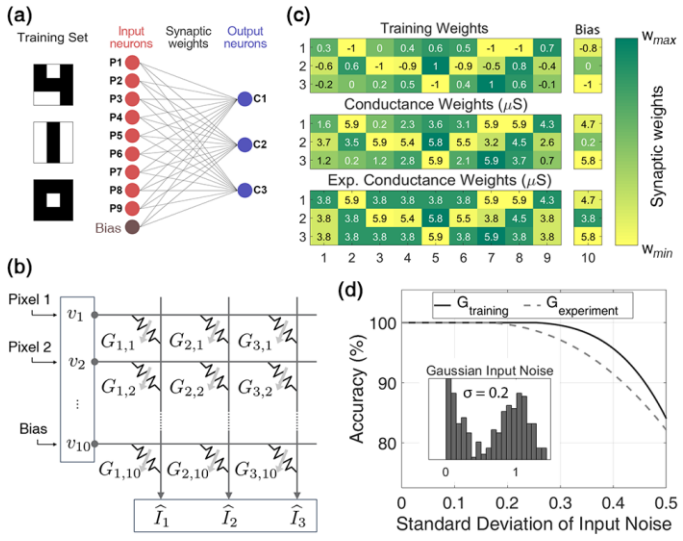

Fig. 17. (a & b) A set of 30 Proj-PCM cells make up a 10×3 crossbar that serves as a neural network that is trained on a set of 3 images. (c) Training weights are mapped to conductance values and ranged to match the dynamic range of the devices. (d) The latter affects the classification accuracy in the Gaussian noise scenario.

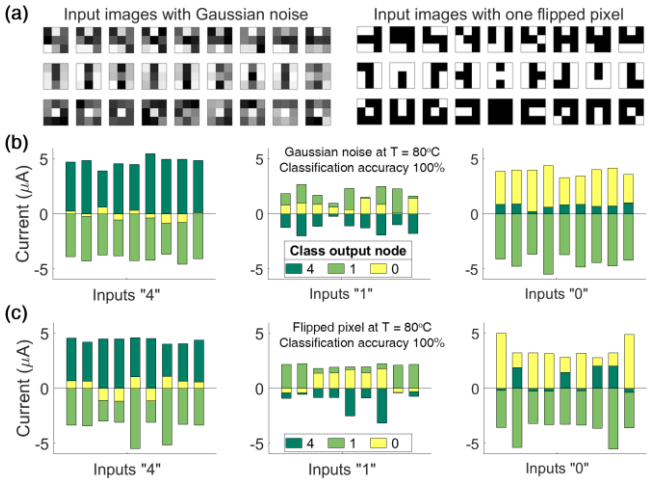

Fig. 18. Classification of 27 images in 2 scenarios of noise at the input neurons. (a) Analogue noise was introduced as a Gaussian distribution of pixel colors between 0 and 1 in the grayscale (σ = 0.2) whereas the set for digital noise had one pixel of each original image flipped. (b & c) In both scenarios the classification accuracy was 100% at room and elevated temperatures, without the need to apply the temperature compensation scheme.