SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2023, Vol. 66 222402:1–222402:6
https://doi.org/10.1007/s11432-022-3743-x

Near-threshold-voltage operation in flash-based high-precision computing-in-memory to implement Poisson image editing

Yang FENG[1], Bing CHEN[2*], Mingfeng TANG[1], Yuerang QI[1], Maoying BAI[1], Chengcheng WANG[1], Hai WANG[1], Xuepeng ZHAN[1], Junyu ZHANG[3], Jing LIU[4*], Jixuan WU[1*] & Jiezhi CHEN[1*]

[1]*School of Information Science and Engineering, Shandong University, Qingdao* 266200*, China;*
[2]*School of Micro-nano Electronics, Zhejiang University, Hangzhou* 310030*, China;*
[3]*Neumem Co., Ltd, Hefei* 230093*, China;*
[4]*Key Laboratory of Microelectronic Devices and Integrated Technology, Institute of Microelectronics of Chinese Academy of Sciences, Beijing* 100084*, China.*

Received 29 November 2022/Revised 29 January 2023/Accepted 18 April 2023/Published online 10 November 2023

Abstract We propose a NOR flash-based computing-in-memory (CIM) to implement high-precision (32-bit) Poisson image editing, including the gradient operations and Laplace operation. To meet the requirements of image processing, CIM operation schemes and reliabilities are carefully studied and optimized, showing that the power consumption at the near-threshold-voltage (NTV) region can be as low as 40 fJ/bit, which is two orders lower than working at the saturation region. High-temperature retention, as well as various disturbs, are also analyzed. The proposed CIM scheme can be applied as an energy-efficient approach to construct the high-precision image processing accelerator.

Keywords NOR flash memory, computing-in-memory, variation, Poisson image editing

Citation Feng Y, Chen B, Tang M F, et al. Near-threshold-voltage operation in flash-based high-precision computing-in-memory to implement Poisson image editing. Sci China Inf Sci, 2023, 66(12): 222402, https://doi.org/10.1007/s11432-022-3743-x

1 Introduction

Computing-in-memory (CIM) is an effective technical way to solve the "memory wall" and the performance bottleneck. Many emerging non-volatile memories (NVM), such as resistive random access memory (RRAM) [1–3], phase-change memory (PCM) [4, 5], ferroelectric RAM (FeRAM) [6, 7], and flash memory [8–13], have demonstrated their good capabilities in lots of artificial neural networks (ANN). Recently, the high-precision CIM architecture has attracted much more attention because it can provide a fundamental approach to meeting the strict requirements in diverse scientific calculations [2, 14]. Poisson image editing [15] is a seamless image editing algorithm that has been widely used to fuse the background image and the target image while retaining the gradient information of the source image. To implement high-precision Poisson image editing with CIM architecture, it is necessary to ensure the cells as well as the array have good stabilities and robust reliabilities. Compared with other NVMs, flash memory has significant benefits in terms of ultra-high On/Off ratio, good reliabilities, and strong controllability to cells' variations, thus it has great advantages in high precision computing. Importantly, flash memory has great compatibility with peripheral circuits and it is capable of designing large CIM arrays for large-scale processing.

So far, lots of work in CIM architectures for the neural network has been reported based on NOR flash memory. In 2017, the mixed signal neuromorphic classifier based on embedded NOR flash memory technology was demonstrated firstly by Guo et al. [8]. For more energy-efficient convolution operation,

[*] Corresponding author (email: bingchen@zju.edu.cn, liujing@ime.ac.cn, jixuanwu@sdu.edu.cn, chen.jiezhi@sdu.edu.cn)

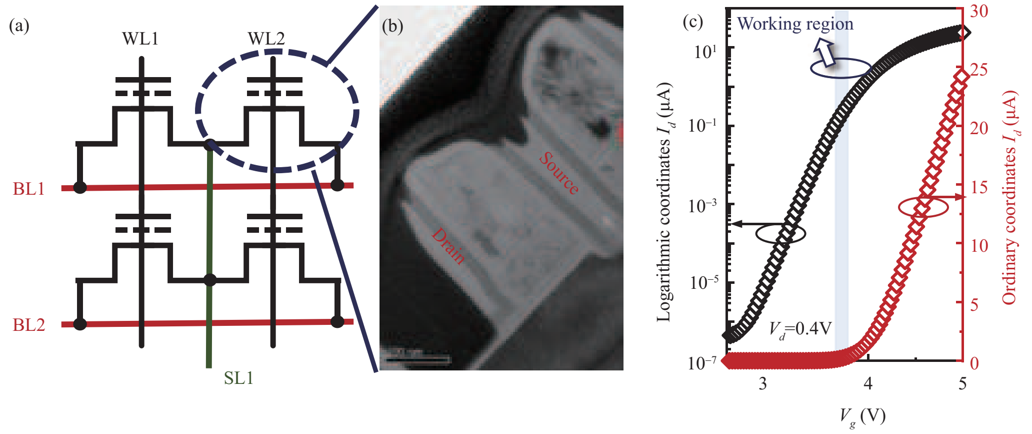


Figure 1 (Color online) (a) The schematic of NOR flash memory array; (b) the structure of fabricated NOR flash memory cell by 65 nm technology node; (c) cell current in the near-$V_{\rm th}$ region at $V_d = 0.4$ V.

Han et al. [10] proposed a novel NOR flash array computing scheme to execute the convolution. Xiang et al. [11,12] designed hardware implementation based on NOR flash computing array for analog deep neural network (DNN) and spike-driven convolutional neural network (SCNN). Benefiting from the mature process, NOR flash memory also has great performance in high-precision calculations, such as partial differential equations [16,17] and image processing [18].

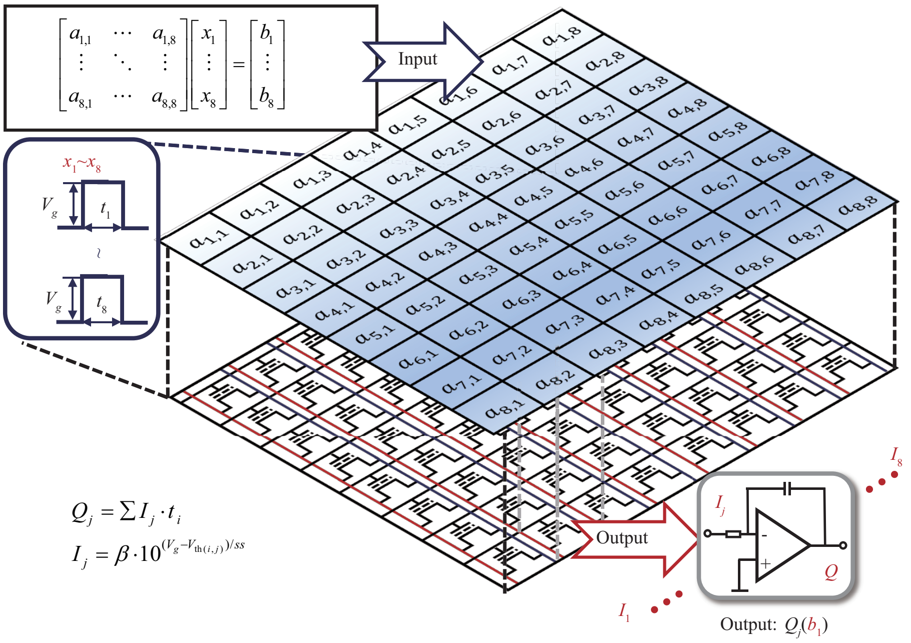
In this study, aiming at providing energy-efficient solutions to high-precision image processing accelerators, a NOR flash-based CIM scheme is proposed to implement Poisson image editing. By using the designed array and the optimized operation strategies, ultra-low energy consumption (40 fJ/bit) and high computing precision (32-bit) have been demonstrated.

2 Method and result

The structure of NOR flash memory is shown in Figure 1(a). The amount of charges stored in the floating gate determines the channel current at the same operating voltage, which can be delicately regulated by applying program pulses. In this work, the core processing unit is constructed with NOR flash memory arrays of 65 nm NOR flash memory technique, which is the prototype of the simulation and test in the work, as shown in Figure 1(b). The I_d-V_g curves of the log coordinate and linear coordinate are shown in Figure 1(c). Near-threshold-voltage (NTV) read scheme is adopted to significantly reduce energy consumption, where the supply voltage (V_{dd}) is massively decreased. Moreover, the read current is much smaller than the linear region and the saturation region. Compared with the sub-$V_{\rm th}$ region, the near-$V_{\rm th}$ region where the operating voltage is slightly higher than $V_{\rm th}$ has over 10 times less variation [19], thus the near-$V_{\rm th}$ region is easier to be applied in high precision CIM. The chosen operation region (V_g = 3.8 V, V_d = 0.4 V) is in the near-$V_{\rm th}$ region with a small I_d for low-power applications.

In previous work [8], the multiply-and-accumulate (MAC) operation is performed directly between an input vector, represented by V_g, and the coefficient matrix elements represented by $V_{\rm th}$. Considering the linearity requirements of high-precision applications, the input vector is represented by the pulse time. Figure 2 shows the way to implement MAC operations. Each source line (SL) connects a separate SA (sensing amplifier) and integrator to acquire the charge quantity as a result of MAC operations. The process of computing can be described as $Q = I \cdot T$.

The elements of the input vector and the coefficient matrix are represented by the pulse time and $V_{\rm th}$, respectively. The results represented by the charge quantity can be processed into the input vector of the next iteration. To make the newly inserted image match the style of the original image, image editing is utilized. By processing with the gradient field in the source image block, Poisson image editing makes the fusion boundary between the target scene and the source image smooth. The fused image block can be seamlessly fused into the target scene, and its hue and light can be consistent with the target scene. Poisson image editing includes the gradient operation and the Laplace operation, and there is an iterative process in the Laplace operation, which is suitable to be applied in CIM architecture. This can be replicated independently in each of the channels of a color image.

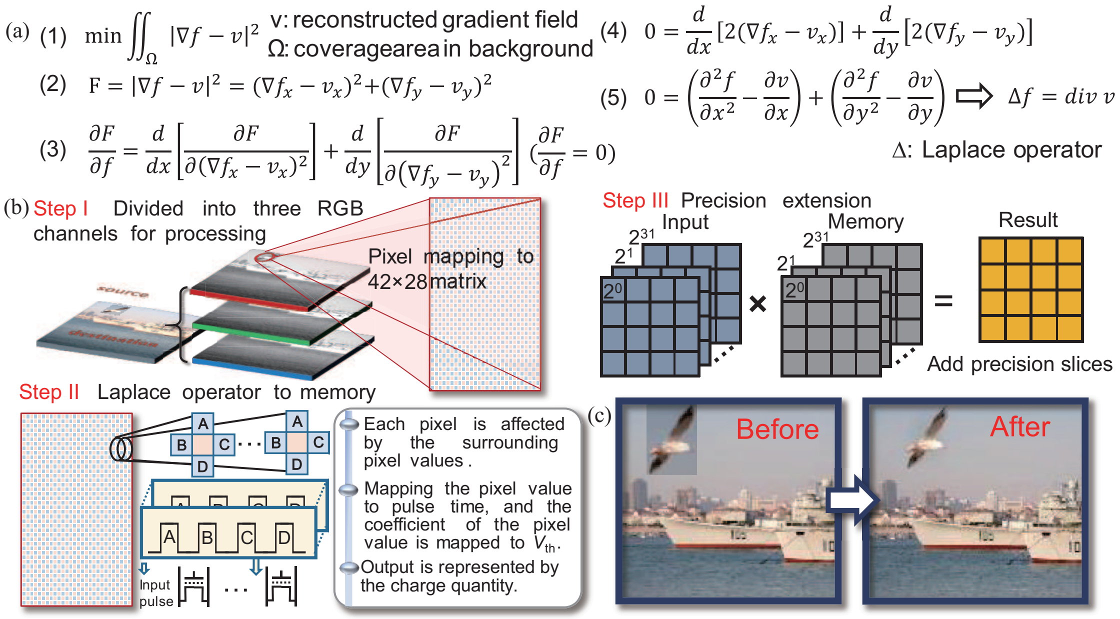


$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,8} \\ \vdots & \ddots & \vdots \\ a_{8,1} & \cdots & a_{8,8} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_8 \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_8 \end{bmatrix}$$

$$Q_j = \Sigma I_j \cdot t_i$$

$$I_j = \beta \cdot 10^{(V_g - V_{\text{th}(i,j)})/ss}$$

Figure 2 (Color online) Schematic of the method of scientific calculation. The matrix is mapped to the $V_{\rm th}$ of the cell, and the vector is mapped to the gate pulse time of the cell. By accumulating the current through the array cells adjusted by $V_{\rm th}$, the result of the integrator is the dot product of array cells with a common SL.



(a)

$$(1)\quad \min \iint_{\Omega} |\nabla f - v|^2$$

v: reconstructed gradient field
Ω: coveragearea in background

$$(2)\quad \mathrm{F} = |\nabla f - v|^2 = (\nabla f_x - v_x)^2 + (\nabla f_y - v_y)^2$$

$$(3)\quad \frac{\partial F}{\partial f} = \frac{d}{dx}\left[\frac{\partial F}{\partial(\nabla f_x - v_x)^2}\right] + \frac{d}{dy}\left[\frac{\partial F}{\partial\left(\nabla f_y - v_y\right)^2}\right] \quad \left(\frac{\partial F}{\partial f} = 0\right)$$

$$(4)\quad 0 = \frac{d}{dx}[2(\nabla f_x - v_x)] + \frac{d}{dy}[2(\nabla f_y - v_y)]$$

$$(5)\quad 0 = \left(\frac{\partial^2 f}{\partial x^2} - \frac{\partial v}{\partial x}\right) + \left(\frac{\partial^2 f}{\partial y^2} - \frac{\partial v}{\partial y}\right) \Rightarrow \Delta f = div\ v$$

Δ: Laplace operator

(b) Step I Divided into three RGB channels for processing

Pixel mapping to 42×28matrix

Step II Laplace operator to memory

- Each pixel is affected by the surrounding pixel values .
- Mapping the pixel value to pulse time, and the coefficient of the pixel value is mapped to $V_{\rm th}$.
- Output is represented by the charge quantity.

Step III Precision extension

Figure 3 (Color online) (a) Formula used in Poisson image editing. The Laplace operation is solved in CIM architecture. (b) The image is processed into RGB channels. The image is a 42×28 matrix, each value is affected by surrounding values presenting the pulse time of $V_{\rm BL}$. The coefficient of surrounding pixel values is stored in the cell as $V_{\rm th}$. The precision-extension technology is utilized in this architecture. Each flash memory cell stores 1 bit, thus 32-bit precision is achieved by 32 cells. (c) Comparison of results before and after Poisson image processing.

Figure 3(a) shows the formula used in Poisson image editing. Eq. (1) in Figure 3(a) shows that the divergence of the target gradient field (Laplace operation) should be consistent with the reconstructed gradient field as much as possible to smooth the interface, where the reconstructed gradient field (v) is the sum of the background gradient field and picture gradient field. Eqs. (2)–(5) in Figure 3(a) show further derivation. The schema of CIM is shown in Figure 3(b). Firstly, the RGB channels of the color image are handled separately. Then the Laplace operation is solved by the iterative method when it is transformed into matrix format by using the finite difference method. In this way, it can be easily calculated in a flash

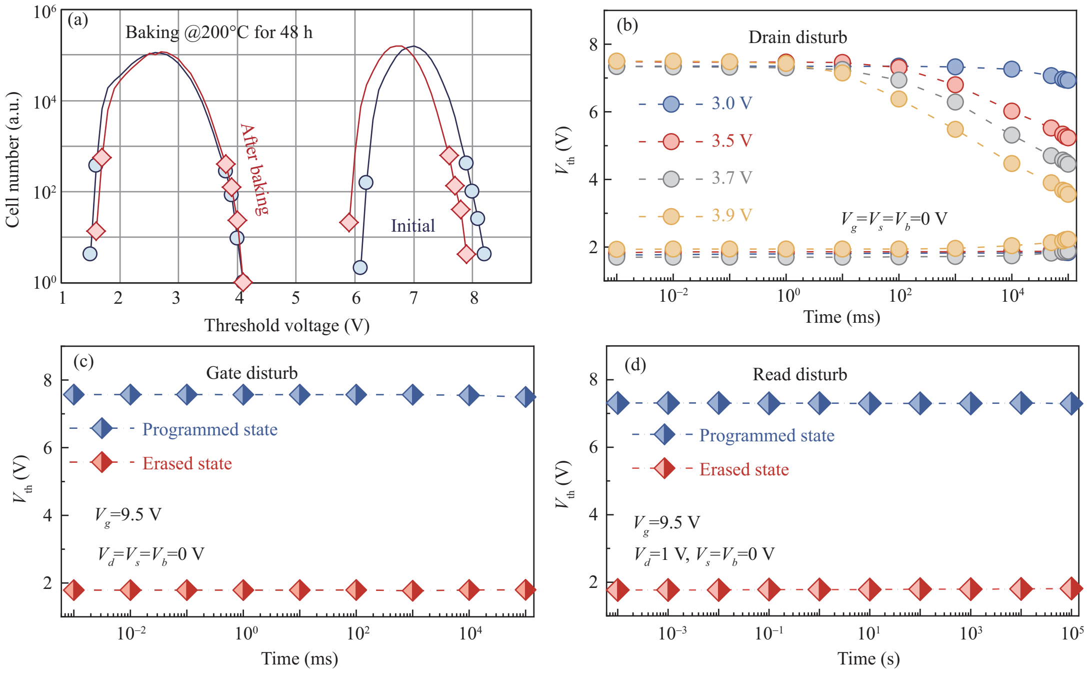


Figure 4 (Color online) (a) Retention is characterized at 200°C to study the stabilities of stored data: cell number distributions after baking at high temperature in the erase state and the program state. Disturbs are characterized to study the stabilities of stored data during the repeated read operations, the V_{th} varies with disturb time plotted in different disturb conditions: (b) drain disturb; (c) gate disturb; (d) read disturb (RD).

array. In the process of solving the Laplace operation, each pixel is affected by the surrounding four-pixel values, which are represented by four pulse time inputs in sequence. The V_{th} stored in cells represents the coefficient to be multiplied by the four input pulse values. The output is represented by the integral of the current. Based on Ohm's law and Kirchhoff's current law, the array can realize the MAC operation. The output from the last iteration can be transformed to the input represented by a pulse time of V_g, and then this process is repeated iteratively by feeding the output to the system as the next input vector until the expected result is acquired. The precision-extension technology is utilized to improve computing precision. The input vector and coefficient matrix are sliced and thereby one number can be represented by multiple cells. Single-level-cell (SLC, 1-bit/cell) is used for processing for more precise computing, each cell stores one bit (32-bit data can be presented by 32 cells). 42×28 image needs 112896 cells in total. The results in Figure 3(c) show how well the edges blend after Poisson image editing.

3 Reliability analysis and discussion

Although Poisson image editing in NOR flash memory shows excellent results, it is necessary to have comprehensive studies on the reliabilities of the constructed CIM to assure high precision computing in different conditions. High-temperature data retention (HTDR), RD and endurance are characterized. The maximum variation range in the near-V_{th} region is also simulated and analyzed.

3.1 HTDR and RD

In the CIM process, when the threshold voltage is set, the electrons need to be maintained in the cell to ensure current stability, which makes the retention characteristic particularly important, especially when the chip works in a special high-temperature environment. To check the stability of cells in different technology nodes when storing electrons, HTDR properties are tested in 65 nm NOR flash cells by characterizing the retention at 200°C, as shown in Figure 4(a). The V_{th} remains a small variation after 48 h baking, especially the data at the erased state, indicating good retention properties that can benefit high accuracy control. The calculation of Poisson image editing contains a large number of iterations, which means that multiple times of read operations are necessary at intervals. Thus, the properties of read stabilities that are strongly correlated to the stored electrons in the cells need to be considered.

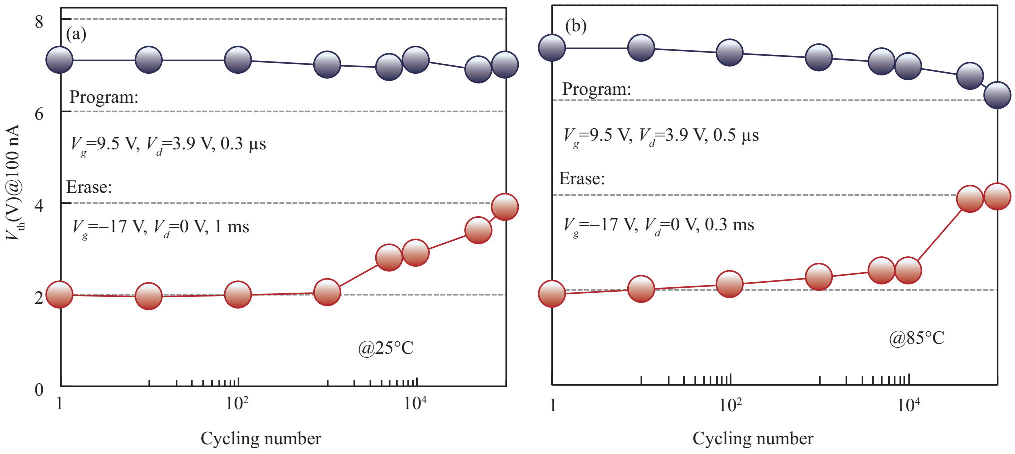


Figure 5 (Color online) Endurance (program/erase cycling) is tested at (a) 25°C and (b)85°C.

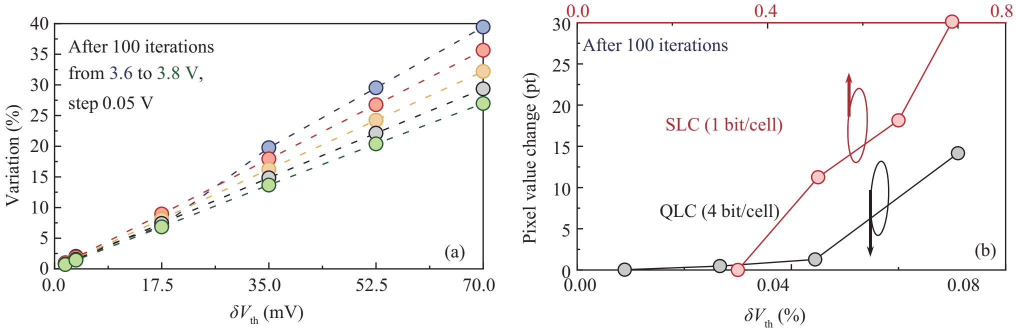


Figure 6 (Color online) (a) Variations after 100 iterations for computing in near-$V_{\rm th}$ of NOR flash at different V_g, from 3.6 to 3.8 V with 0.05 V step, wherein the impacts of $\delta V_{\rm th}$ are considered; (b) simulation results of pixel value change after 100 iterations by including the impacts of $\delta V_{\rm th}$ variations.

Figure 4(b) shows the $V_{\rm th}$ varies with disturb time when the applied different voltage in the drain. The $V_{\rm th}$ of programmed state cells changes greatly, while the erased state cells remain constant. Figure 4(c) depicts the stable gate disturb characteristic both in the programmed state cells and erased state cells. The RDs characteristic is shown in Figure 4(d), with the applied V_g and V_d of 5 and 1 V, respectively. The minimal $V_{\rm th}$ variation fully meets the requirement for iteration numbers of CIM calculation.

Figure 5 shows the results of the endurance test. After 100K P/E cycling at room temperature as well as 85°C, the 3 V memory window is still retained. Furthermore, the coefficient matrix in this work is constant, meaning that this computing does not need frequent P/E cycling. The usage count is not limited by the P/E cycling number.

3.2 $V_{\rm th}$ variation analysis

HT DR and RD characteristics show the cell current variations can remain in a controlled range. A detailed simulation is still needed to confirm the maximum threshold voltage drift that the calculation can tolerate. The impacts of operating voltage drifts are analyzed. To select the optimal operating voltage, the pixel change of image editing varies with $V_{\rm th}$ and is tested at various V_g from 3.6 to 3.8 V. Figure 6(a) shows the current variations increase with the decrease of V_g (closer to $V_{\rm th}$). To balance the variation and power consumption, we choose $V_g = 3.8$ V and $V_d = 0.4$ V as the optimal operating voltage. NTV region has more serious variations than other regions. It is necessary to study the impact of $V_{\rm th}$ variations. The comparison between quad-level-cell (QLC, 4-bit/cell) and SLC is shown in Figure 6(b). A tiny disturbance can change the pixel value a lot and destroy the whole calculation. While SLC can tolerate up to 0.4% variation, which is 7–8 times higher than QLC. Therefore, SLC is chosen to accomplish the calculations. According to the results, the variations of $V_{\rm th}$ still need to be controlled within ±17.5 mV (0.5% variation). It is acceptable that the pixel value variation is within 10 according to the image editing effect.

Table 1 Comparison of cell types, precision, and energy

Computing target	Cell type	Precision (bit)	Energy (fJ/bit)
This work	SLC NOR	32	40
Convolution [10]	SLC NOR	8	10.4
Neural network [20]	SLC NAND	4	8

When calculating in the saturation region, the cell can adopt the QLC scheme, the cell number reduces to a quarter of the SLC scheme and the variations in the saturation region will be greatly eased. Despite these advantages, the key factor with CIM is power consumption, which is lower when operating in the near-$V_{\rm th}$ region than when working in saturation. The energy is calculated by multiplying the cell current in the NTV region, the pulse time (100 ns), and drain voltage, then dividing it by the bit numbers. By this scheme, the array energy in the near-$V_{\rm th}$ region is 40 fJ/bit, while the energy in the saturation region is 4 pJ/bit. The comparison of the proposed scheme with other related work is summarized in Table 1.

4 Conclusion

We propose a low power-consumption solution to implement high-precision Poisson image editing in the flash-based CIM. The reliability characteristics including RD and HTDR, as well as the impacts of cell variations, are comprehensively investigated. Characterization results show that flash-based CIM is suitable to accomplish precise image processing by working at the near-threshold region, and the power consumption can be suppressed to as low as 40 fJ/bit. The proposed energy-efficient approach could shed light on CIM engineering to construct the high-precision image processing accelerator.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62034006, 92264201, 91964105), Natural Science Foundation of Shandong Province (Grant Nos. ZR2020JQ28, ZR2020KF016), and Program of Qilu Young Scholars of Shandong University.

References

1 Li C, Hu M, Li Y N, et al. Analogue signal and image processing with large memristor crossbars. Nat Electron, 2018, 1: 52–59

2 Ma W, Zidan M A, Lu W D. Neuromorphic computing with memristive devices. Sci China Inf Sci, 2018, 61: 060422

3 Hao Y X, Zhang Y, Wu Z H, et al. Uniform, fast, and reliable CMOS compatible resistive switching memory. J Semicond, 2022, 43: 054102

4 Raoux S, Wełnic W, Ielmini D. Phase change materials and their application to nonvolatile memories. Chem Rev, 2009, 110: 240–267

5 Cheng C D, Tiw P J, Cai Y M, et al. In-memory computing with emerging nonvolatile memory devices. Sci China Inf Sci, 2021, 64: 221402

6 Mikolajick T, Dehm C, Hartner W, et al. FeRAM technology for high density applications. MicroElectron Reliab, 2001, 41: 947–950

7 Jiang Y N, Huang P, Zhou Z, et al. Circuit design of RRAM-based neuromorphic hardware systems for classification and modified Hebbian learning. Sci China Inf Sci, 2019, 62: 062408

8 Guo X, Bayat F M, Prezioso M, et al. Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells. In: Proceedings of Custom Integrated Circuits Conference (CICC), Austin, 2017. 1–4

9 Bavandpour M, Mahmoodi M R, Strukov D B. Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond. IEEE Trans Circ Syst II, 2019, 66: 1512–1516

10 Han R Z, Huang P, Xiang Y C, et al. A novel convolution computing paradigm based on NOR flash array with high computing speed and energy efficiency. IEEE Trans Circ Syst I, 2019, 66: 1692–1703

11 Xiang Y C, Huang P, Zhou Z, et al. Analog deep neural network based on nor flash computing array for high speed/energy efficiency computation. In: Proceedings IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, 2019. 1–4

12 Xiang Y C, Huang P, Han R Z, et al. Efficient and robust spike-driven deep convolutional neural networks based on NOR flash computing array. IEEE Trans Electron Dev, 2020, 67: 2329–2335

13 Lee S T, Yeom G, Hwang J, et al. Utilization of unsigned inputs for NAND flash-based parallel and high-density synaptic architecture in binary neural networks. IEEE J Electron Dev Soc, 2021, 9: 1049–1054

14 Joshi V, Le Gallo M, Haefeli S, et al. Accurate deep neural network inference using computational phase-change memory. Nat Commun, 2020, 11: 1

15 Pérez P, Gangnet M, Blake A. Poisson image editing. In: Proceedings of ACM SIGGRAPH, San Diego, 2003. 313–318

16 Feng Y, Chen B, Liu J, et al. Design-technology co-optimizations for general-purpose computing in-memory based on 55nm NOR flash technology. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), San Francisco, 2021

17 Feng Y, Wang F, Zhan X P, et al. Flash memory based computing-in-memory system to solve partial differential equations. Sci China Inf Sci, 2021, 64: 169401

18 Zhang D, Wang H, Feng Y, et al. Implementation of image compression by using high-precision in-memory computing scheme based on NOR flash memory. IEEE Electron Dev Lett, 2021, 42: 1603–1606

19 Jiang X B, Guo S F, Wang R S, et al. New insights into the near-threshold design in nanoscale FinFET technology for sub-0.2V applications. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), San Francisco, 2016

20 Lue H T, Hsu P K, Wei M L, et al. Optimal design methods to transform 3D NAND flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN). In: Proceedings of IEEE International Electron Devices Meeting (IEDM), San Francisco, 2019