Design-Technology Co-Optimizations (DTCO) for General-Purpose Computing In-Memory Based on 55nm NOR Flash Technology

Yang Feng[1,+], Bing Chen[2,+], Jing Liu[3], Zhaohui Sun[1], Hongyang Hu[3], Junyu Zhang[4], Xuepeng Zhan[1],Jiezhi Chen[1,*]
[1] School of Information Science and Engineering, Shandong University, Qingdao, China; [2] School of micro-nano electronics, Zhejiang University, Zhejiang, Hangzhou, China; [3] Key Laboratory of Microelectronic Devices and Integrated Technology, Institute of Microelectronics of Chinese Academy of Sciences, Beijing, China; [4] Neumem Co., Ltd, Hefei, China.
[*]Email: chen.jiezhi@sdu.edu.cn

***Abstract*—**In this work, based on 55nm NOR flash technology, a general-purpose computing in-memory (CIM) architecture is proposed for the first time. By using a device-aware DTCO approach, a flash-based high-precision partial differential equation (PDE) solver is constructed with the 32-bit floating point (FP) calculation ability. Memory cells (4bit/cell) work at the quasi-saturation region to balance the performances and reliabilities, and the hot hole injection (HHI) is utilized to tune each cell individually (negative V_{th} shift) in the matrix array, showing ultra-fast operation speed (~10ns) and ultra-low power dissipation. Comprehensive reliability characterizations are also done, including retention, read disturb, random signal noise (RTN) and endurance. It is witnessed that the proposed flash-based 32-bit CIM architecture can conduct high precision calculations with a good tolerance to the cells' fluctuations.

1. Introduction

With data scale-up, today's computation suffers from frequent data movement, including various tasks, such as games, climates and economics. To break the "memory wall", Computing in-memory (CIM) has been proposed and their superiority has been illustrated in some new computation area like artificial neuron network [1-7]. However, these solutions always focus on some specific applications. General purpose CIM should be developed for multi-tasks processing and thus high-precision calculations are required. That is, the memory cell unit should be highly reliable and controllable to construct the large matrix in CIM architecture. So far, researches on general-purpose CIM has been rarely reported except some work on partial differential equations (PDEs) solvers [8-10]. Recently, Michigan Univ. group demonstrated a memristor-based PDE solver by adopting the precision extension technique in small arrays (3x3 matrix), which helps to suppress the series resistance and sneak currents [8]. However, as noted by IBM group [9-10], memristor cells still suffer from significant inter- & intra- device variabilities as well as inhomogeneity across an array. In addition, considering the sense amplifier design and power consuming, it is still challenging to construct large arrays for vector-matrix multiply-and-accumulate (MAC) operations with memristor cells [7]. To address the strict requirements of general-purpose CIM, flash memory represents a great choice because of its ultra-high bit densities, robust reliabilities, and good controllability of cell variations in large arrays [7,11-12].

In this work, based on 55nm NOR flash, a general-purpose CIM architecture (Fig.1) that can conduct high-precision 32-bit floating point (FP) calculations is proposed. A design-technology co-optimizations framework is developed to explore the energy efficiency and accuracy of such an implementation. Hot-hole injection (HHI) method is adopted to provide fast and low-power solutions, and sub-saturation region is utilized for read operations to suppress the impacts of bias fluctuations as well as read disturbs (RD). Reliabilities are studied comprehensively, and the impacts of cells' fluctuations on the computing accuracy are discussed.

2. Proposed flash-based 32-bit FP PDE solver

The general-purpose CIM system requires a high-precision scheme. Here, we propose a 32-bit FP CIM architecture based on NOR flash memory array, as shown in Fig.2. For 32-bit FP calculations, the data includes 1-bit sign, 8-bit exponent, and 23-bit mantissa. To balance the accuracy and efficiency, the mantissa data are processed in flash-based CIM array while the sign and the exponent are processed in the register. Then, by threshold voltage (V_{th}) mapping via program and erase pulses, 6 flash cells (4bits/cell) are used to store one mantissa data for subsequent vector–matrix operations. As shown in Fig.2, one FP processor unit can handle the multiplication of two FP numbers at a time. The accumulated charge on each source line (SL) is the accumulation result of multiplying 4-bit by 4-bit binary data, and the result can be derived after the carry process. Above architecture is applied in NOR flash memory fabricated by 55nm node technology. Fig.3 shows the chip photograph together with cells' TEM figures.

3. CHEI/HHI approach for individual cell tuning

Fig. 4(a) shows the process flow of tuning a cell to the target I_D (or V_{th}). As the standard operation in NOR flash memory, data program is done by channel hot electrons injection (CHEI) and data erase is done by electrons FN tunneling from FG. For FN erase, a positive bias is applied to the p-well and a negative bias is applied to the word lines (WLs), all cells at the same word-line are erased simultaneously. This is good for fast data easing while it cannot tune each cell individually. In this work, a new erasing scheme named as hot hole injection (HHI) [13] is adopted for the state tuning of matrix cells, as shown in Fig.4(b). The most attracting point of HHI method is that, it allows individual cell tuning with no necessary to re-design the cell array. In detail, as applying the positive bias in bit lines (BLs) and negative bias in WLs, band-to-band tunneling (BTBT) happens at the drain junction. Holes came from the drain flow toward the p-well gaining energy in the presence of the large reverse bias at the p-well/drain junction. In this way, some holes may reach kinetic energy high enough to overcome the silicon/oxide valence-band barrier and be injected into the FG in the presence of a highly negative electrostatic potential. Thus, injected holes will lower V_{th} and increase I_D.

Fig. 5 shows I_D-V_G currents of cells as performing different program/erase conditions, and Fig.6 summarizes the V_{th} shifts.

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720625

It is observed that, by using CHEI/HHI schemes, cell states can be tuned effectively and accurately, which benefits multi-states operations. More impressively, ~1.5V V_{th} adjustments can be realized by HHI in 10ns (Fig.7), and 50ns is almost enough for full cell erase. This is important for ultra-fast constructions of calculation matrixes as well as the requirements for ultra-low power dissipation. In comparison to the standard FN erase, the power dissipation by HHI is suppressed by five orders (Fig.8).

4. Reliability characterizations

In many cases, e.g. linear PDE, the calculation matrix is fixed and frequently invoked during iterative calculations, wherein RD is a key factor. Thus, for multi-bit operations in high-precision general-purpose CIM, optimal read conditions should be well studied. As shown in I_D-V_D curves (Fig.9), at higher V_D in the saturation region, it reaches the CHEI condition, starting to inject electrons to the FG and RD happens. While for the linear region at lower V_D, it is quite sensitive to V_D bias fluctuations, which can cause challenges for circuits design and error control. Thus, in this work, quasi-saturation region (V_G=4V, V_D=1V) is chosen as the read conditions, wherein lower sensitivity to V_D fluctuations as well as better immunity to RD are expected.

In Fig. 10(a), 16 discrete states in a flash cell unit (4-bit/cell) can be clearly distinguished, with the read I_D ranging from 0 to 15uA. To check the stability of cells when storing electrons, high/low temperature (HT/LT) data retention (DR) are tested. Here, cell retention properties at a wide temperature of -30℃ to 85℃ are measured. As shown in Fig.10(b), after tuning cells to different states, their currents are measured during the long-time retention at 85℃. For the low/high V_{th} cell with read currents of 5/15uA, the current drifts from electron lost caused lower V_{th} are kind of ignorable even after $1x10^4$ seconds retention. To study the cell-to-cell variations, Figs.10(c)~(d) summarizes 10 cells' results. As expected, retention at 85℃ is worse than that at 30℃. Still, it can be observed that current variations remain in a controlled range without accuracy loss. Besides the extremely good retention property in multiple cells, a further confirmation on chips' stabilities is necessary. As shown in Figs.10(e), after 48 hours baking at 250℃, the shift of V_{th} distributions (program state) is still as low as ~0.1V, which indicates that the impacts of HT DR are ignorable and a large matrix array can work stable even at high temperatures.

I_D-V_G curves before and after RD stressing on the cells are plotted in Fig. 11(a), and read current migration during/after 1×10^4s stressing is summarized in Figs.11(b)&(c). By these results, it can be concluded that RD has been well suppressed in the operations. In addition, read noise in nano-scale cells is another important reliability issue need to be cared about, like random telegraph noise (RTN). Fig.12 shows the measured RTN in one cell, showing a typical V_G dependence. Although RTN can be observed in some memory cells, it is confirmed that current fluctuations ratios caused by RTN are lower than 5% (Fig.13). Fig. 14 shows endurance test results. After 10k P/E cycling at the room temperature, as large as 6V memory window can be retained. Even for 85℃ operations, ~5 V memory window can be retained.

In addition, considering possible requirements for ultra-low power consumption with a sacrifice to the calculation precision, operations at the sub-threshold region is another promising candidate [14] because its supply voltages are very low in comparison to the saturation region and the linear region. Here, by setting 2.4V V_G and 0.8V V_D as the read biases, the cells are tuned to 2-bit/cell with four storage states, 1nA, 10nA, 100nA and 1uA. As shown in Fig.15, although the impacts of RD can be well controlled, the read current drift is very significant. Thus, for multi-bit data processing at the sub-threshold region, temperature compensation circuits are necessary. Fig.15(d) summarizes the comparisons between the sub-threshold region operation and the saturation region operation.

5. 32-bit FP CIM calculations and analysis

To verify the proposed flash-based CIM architecture (Fig.2), self-consistent calculations of Schrödinger/Poisson equations are used as examples, as shown in Fig.16. It can be solved accurately by the use of numerical methods. The PDE to be solved is shown in Fig. 17(a). Based on the triangular potential approximation, the solution of Schrödinger equation can be derived. By applying the result of Schrödinger equation to the Poisson equation as electron concentrations calculated by (4), the output from Poisson equation is acquired for the more accurate potential approximation used in Schrödinger equation rather than the original triangular potential approximation. Besides the iteration between the Schrödinger and Poisson equations, Poisson equation inside also needs iterations, which is solved by the finite differential method and Jacobi iteration, as shown in Fig. 17(b). It should be noted that, the Jacobi method is adopted because it can be directly mapped to the flash array using entirely iterative vector–matrix operations. The calculated results are shown in Fig. 18. Moreover, from the simulation results shown in Fig.19, it is observed that the tolerance to the cell current fluctuations in FP calculations are much better than that by using the conventional precision extension technology [8]. The accuracy is calculated by $1-\Sigma_{j=0}^{n-1}\left|x_j^E-x_j^N\right|/n$, where x^E is the exact solution, x^N is the numerically computed solution using the flash based PDE solver. Fig.20 shows the comparison with the previous work, and Table-I summarizes the difference between 32-bit floating point and 32-bit integer calculation.

6. Conclusions

To provide a low-power and high-precision solution to the general-purpose computing, for the first time, a novel 32-bit FP CIM scheme based on 55nm node NOR flash is proposed. HHI erase is adopted to tune individual memory cells for accurate calculations, showing ultra-fast tuning speed (~10ns) and ultra-low power dissipation. Subsequent comprehensive reliability (retention, read disturb, RTN noise and endurance) are characterized in cells with 16 storage states (4bit/cell). By taking the self-consistent calculations of Schrödinger and Poisson equations as an example, it is observed that the proposed FP CIM architecture has a high precision with a good tolerance to the cell current fluctuations in the arrays.

Acknowledgement: This work was supported by National Natural Science Foundation of China (Nos. 62034006, 91964105, 61874068), China Key Research and Development Program (No. 2016YFA0201802) and Natural Science Foundation of Shandong Province (No. ZR2020JQ28).
(+These authors contributed equally to this work)

[1] S. Slesazeck, et al., IEDM 2019; [2] S. Majumdar, et al., Adv. Elec. Mater., 2019, 5(3): 1800795.; [3] W. Kim, et al., VLSI 2019; [4] S. Angizi, et al., IEEE TCAD, 2019, 39(5): 1123-1136.; [5] W. Wan, et al., ISSCC 2020; [6] Y. C. Xiang, et al., IEDM 2019; [7] H. T. Lue, et al., IEDM 2019; [8] M. A. Zidan, et al., Nat. elec., 1(7), 2018; [9] M. L. Gallo, et al., Nat. elec., 1(4), 2018; [10] A. Sebastian, Nat. nanotech, 2020, 15(7): 529-544. [11] K. Ishimaru, et al., IEDM 2019; [12] T. H. Hsu, et al., IEDM 2020; [13] G. Malavena, et al., IEEE TED, 2019, 66(11): 4727-4732. [14] F. M. Bayat, et al., ISCAS, 2015.

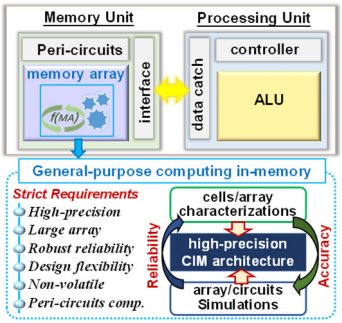


Fig.1. A schematic design of general-purpose CIM system by co-optimizing memory cells and operations modes.

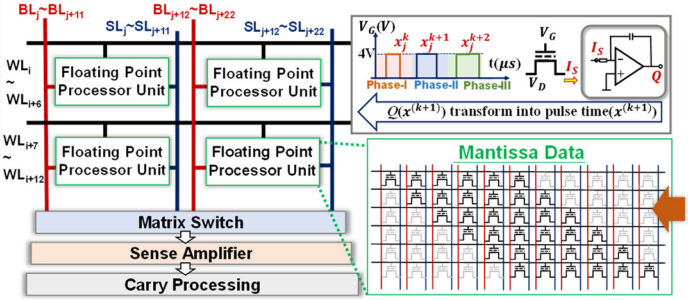


Fig.2. Proposed 32-bit FP computing architectures based on NOR flash memory array. In the FP processer units, each memory cell is precisely tuned to 16 states (4-bit/cell) and two FP data mutilation can be processed.

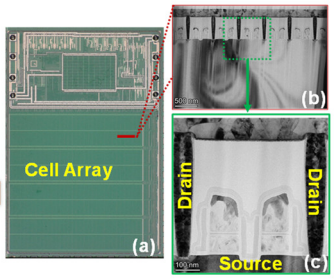


Fig.3. NOR flash memory chip by 55nm technology node: (a) chip photograph; (b) cells' string; (c) two cells with a common source.

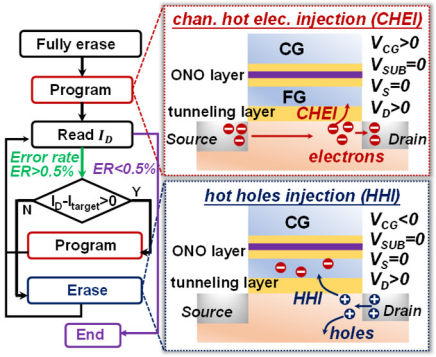


Fig.4. Flow chart of individual cell tuning for multi-bit operations. Here, CHEI programming and HHI erasing are adopted for accurate V_{th} modulations.

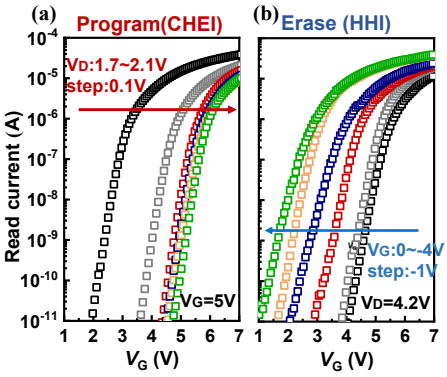


Fig.5. (a) Cell currents as sweeping V_G at various program conditions (V_G=5V); (b) Cell currents as sweeping V_G at various HHI condition (V_D=4.2V).

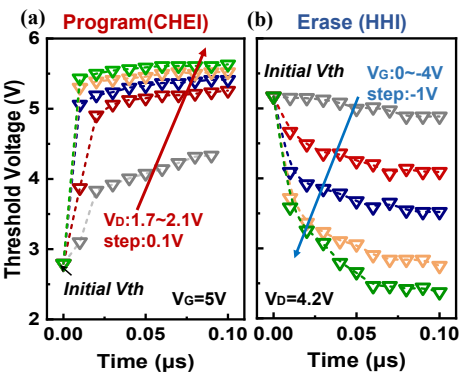
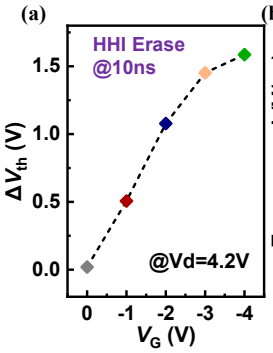


Fig.6. Summarized threshold voltages' modulation by (a) programming pulses and (b) erasing pulses.



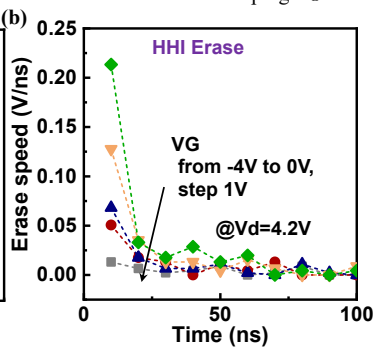


Fig.7. (a) Large memory window adjustment (~1.5V) can be realized by 10ns HHI erase; (b) estimated HHI erase speed. Higher V_G can achieve faster speed while lower V_G is helpful for fine-tuning.

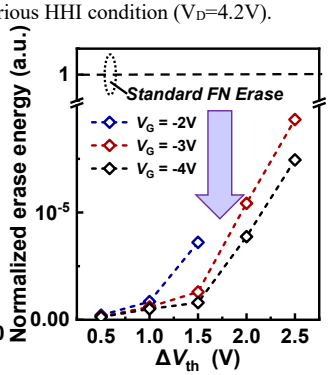


Fig.8. Comparisons of erase power consumption between HHI erasing and standard FN erasing.

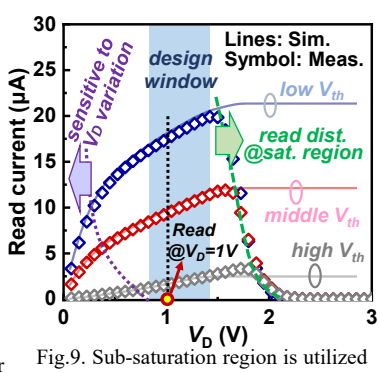


Fig.9. Sub-saturation region is utilized for cells' discrete states tuning to avoid read disturbs and impacts of supply voltage fluctuations.

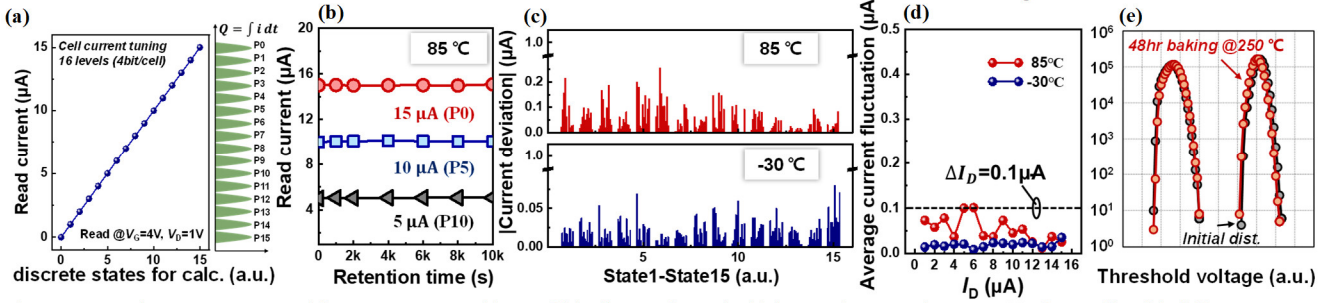


Fig.10. (a) Read currents are tunned into 16 states to achieve 4-bit/cell operations; (b) high retention retentions at 85℃, three cells with different states are demonstrated; (c) read current fluctuations after 1×10^4 seconds baking at 85℃ and -30℃, ten cells for each state are characterized here; (d) the average current fluctuations of each state; (e) chip-base high temperature characterizations (250℃) on data retention properties of NOR flash memory array.

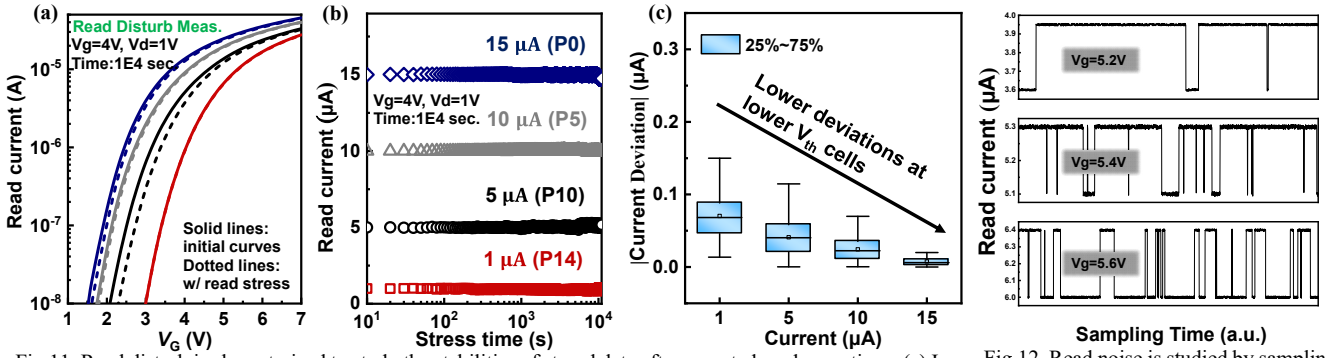


Fig.11. Read disturb is characterized to study the stabilities of stored data after repeated read operations: (a) I_D-V_G properties; (b) read current stabilties during long-time read bias stressing; (c) the statistical distributions of cells' current drifts after 1×10^4 second read bias stressing.

Fig.12. Read noise is studied by sampling measurements, wherein RTN caused read current fluctuations can be observed.

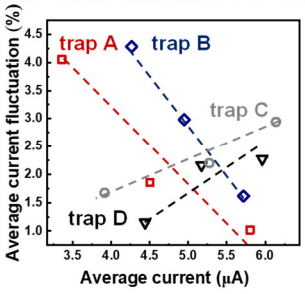


Fig.13. Summarized read current fluctuations of several traps.

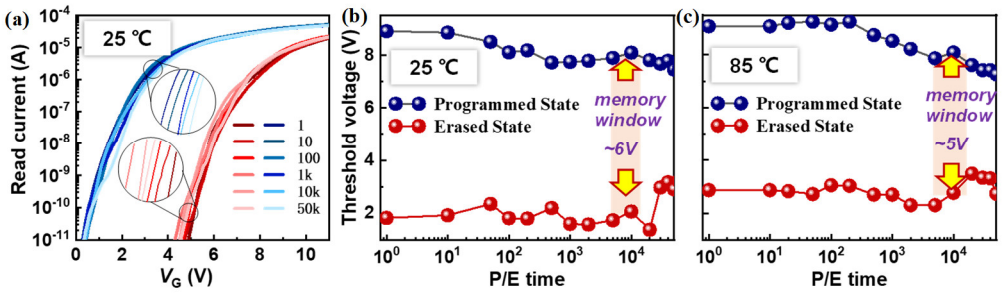


Fig.14. (a) I_D V_G curves during P/E cycling, the shifts of curves are ignorable during cycling; (b) a large memory window (~6V) retains after 10K cycling at 25℃ and (c) ~5V retains after 10K cycling at 85℃.

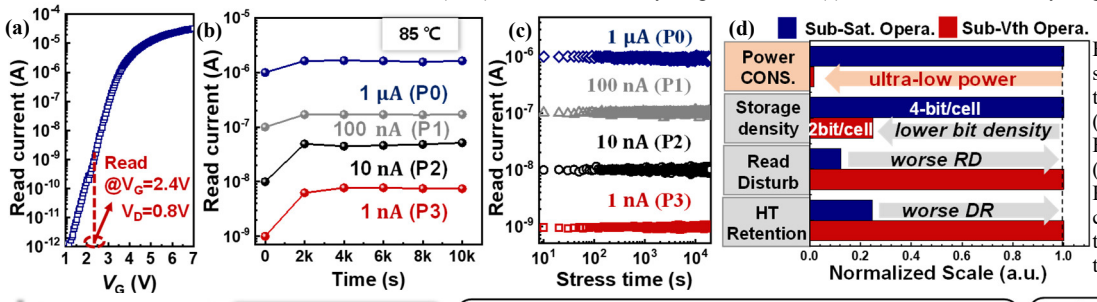


Fig.15. For operations at the sub-threshold region, cells are tuned to 4 states (2-bit/cell): (a) Read bias conditions, (b) HT retention behaviors and (c) long-time read disturbs. (d) Reliability and performance comparisons of operations at the sub-threshold region and the sub-saturation region.

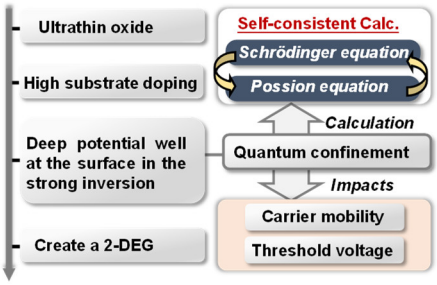


Fig. 16. Self-consistent solution of Schrödinger and Poisson equation of quantum confinement.

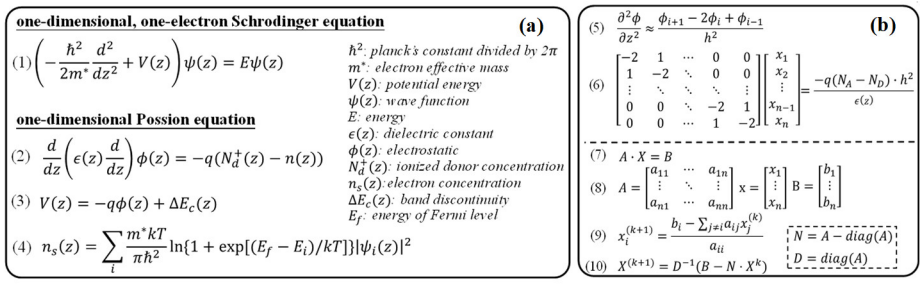


Fig.17. (a) Self-consistent solution of Schrodinger equation and Poisson equation, the electron concentration can be obtained; (b) Finite difference method and Jacobi iteration.

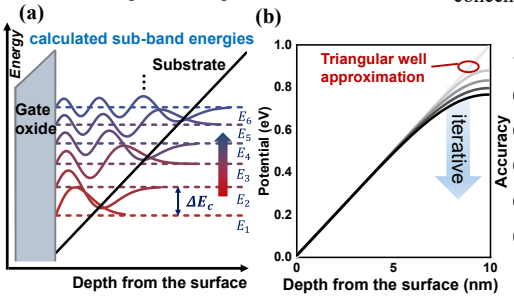


Fig. 18. (a) Self-consistent solution of Schrödinger and Poisson equation of quantum confinement. (b) Potential changes with iterations of Poisson equation.

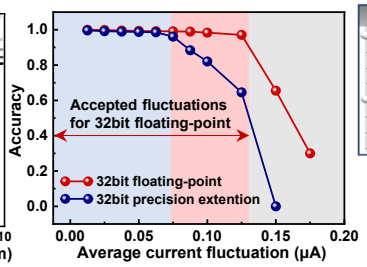


Fig. 19. Accuracy comparison between 32bit float-point and 32bit precision extension technology.

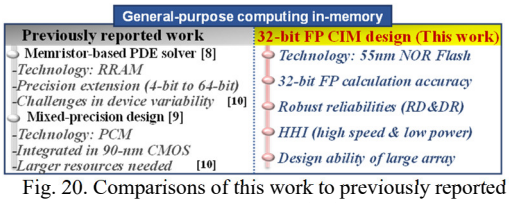


Fig. 20. Comparisons of this work to previously reported work on general-purpose CIM calculations [8-10].

Table I. Comparisons of 32-bit FP calculation to the reported 32-bit precision extension technology.

To improve precision	Range	Cell number	Accuracy
32 bit precision-extention [8]	$[\pm2^{32}, \pm2^{32}]$ int	16	×0.6
32 bit floating point	**~[±3.4^±38,±1.4e^±45] float**	**36 in use**	**1**