A new reading mode based on balanced pre-charging and group decoding

Wenjuan Lu¹, Guoying Qin¹, Zhiting Lin¹, a), Xiulong Wu¹, and Chunyu Peng¹

Abstract In this paper, the balanced pre-charging and group decoding are presented to realize in-memory computing which can achieve continuous read operations after only one pre-charge operation. Compared with the analog computing of multi-row parallel reading, this structure adopts digital domain computing, which improves the accuracy of computing; compared with the non multi-row in-memory computing, it improves the computation speed. The simulation results of noise margin, read times, and read speed show that the proposed method can improve memory performance. The maximum stability can be improved by 13.8%, the read speed can be improved by 75%, and the power consumption can be reduced by 11.58%. Besides, it can be widely used in a variety of memory structures. The proposed method is applied to the existing in-memory computing structures to verify the effectiveness.

Key words: Static random-access memory, In-memory computing, Balanced pre-charging, Group decoding

Classification: Electron devices, circuits and modules (silicon, compound semiconductor, organic and novel materials)

1. Introduction

At present, the von-Neumann system [1] is widely used in the computer system. However, with the advent of the era of big data, in the face of such data-intensive applications as machine learning and artificial intelligence that need frequent read and write operations, the bottleneck of von-Neumann [2] gradually appears. New computing methods are proposed to alleviate the bottleneck of von-Neumann. In-memory computing [3, 4] is one of the most promising methods to achieve more effective data operation and higher data throughput. The difference between in-memory computing and traditional computation structures is that the computing module is embedded in the memory module in in-memory computing. The memory module not only plays the role of memory but also is a calculator, forming a tightly interweaving structure. The main feature is that it can complete the computation within the memory array without reading and transmitting the data. It can significantly reduce energy consumption.

In-memory computing has realized many arithmetic operations [5, 6, 7] and logic operations [8, 9, 10]. Usually, consecutive multiple data reads exist in many applications. There are two main methods to realize this kind of a large number of consecutive read operations. As shown in Fig. 1(a), the first method is the multi-row reading [11, 12, 13, 14], usually in the form of bit-line (BL / BLB) voltage, which opens multi-row at the same time through binary-weighted word-line (WL) pulse width [11, 12] or weighted height [13, 14]. The advantage of this method is that it can read multiple rows of data in the storage array at the same time, which increases the throughput and reduces the reading time. However, in practical operation, it is difficult to guarantee its linearity due to device mismatching and discharge nonlinearity. For example, the unit voltage drop (ΔV) of the larger binary number (such as 1111b) and that of the smaller binary number (such as 0001b) is not equal. It may not meet the accuracy requirements of some applications. Moreover, the A/D conversion module is needed, further increasing the area and power consumption. The second method uses traditional SRAM reading mode for non multi-row in memory computing, as shown in Fig. 1(b) [15, 16]. The advantage of it is the result can be obtained according to whether the BL voltage drops. As shown in Fig. 1(b), the ΔVBL is converted into 0 or 1, so the accuracy is very high; the disadvantage is that it can only read one word at a time. It requires many cycles for some applications with a large amount of data.

Therefore, this paper proposes a balanced pre-charging [17, 18, 19] and group decoding strategy. Fig. 1(c) shows the continuous read mode proposed in this paper. The BL voltage difference is also converted into 0 or 1, which is different from Fig. 1(b) in that ΔVBL=Vpre-VBL, and Vpre is the initial BL voltage in each reading. The structure enhances the speed compared with non multi-row in-memory computing and improves the accuracy compared with multi-row in-memory computing. Because the balance tube of the pre-charge circuit is reused in the proposed method, the area is not increased. The main merits of this work are as follows:

1. A balanced pre-charging method is proposed, and it is proved that the proposed structure can read...
continuously after once pre-charge. In traditional, the purpose of the balanced transistor is to pre-charge both the left and right bit-line voltage to $V_{DD}$ and ensure that the voltage of two bit-lines is the same. Here, we make full use of the existing balanced transistor and do not need to add any additional transistors to realize continuous read function. Different from the traditional method, the voltage bit-line does not need to be pre-charged to $V_{DD}$ in the proposed method, so the power consumption is saved.

2. A group decoding strategy is proposed. The decoding time is reduced, and the overall computing speed is improved. In traditional group decoding where memories have plural banks, complete decoding is required for every word readout. However, the proposed method only needs to decode the block at the beginning of multiple reads, and no decoding operation is needed in the subsequent decoding, which saves decoding time.

3. The performance of the memory is improved. The simulation analysis of the traditional 6T storage array shows that the stability can be improved by 13.8%. The read speed can be improved by 75%, and power consumption can be reduced by 11.58%.

Fig. 1 Memory read mode and BL swing in (a) multi-reading CIM, (b) multi-row reading CIM, and (c) consecutive read structure, where $d_i$ is the data stored in the cell.

2. Principle of balanced pre-charge

In traditional SRAM reading, all columns should be pre-charged to $V_{DD}$, and then the corresponding WL are opened according to the row decoder. When there is enough voltage difference between the BL and BLB, the data can be read out by the sense amplifier (SA). If the read operation is performed again, all columns need to be recharged. Read multiple words includes following steps: pre-charging, decoding, reading, pre-charging, decoding, reading... For many applications, they only require consecutive read, so much time is wasted. Thus, this paper intends to simplify the steps of consecutive reading as follows: pre-charging, decoding, reading, balancing, reading, balancing... With the pre-charge balancing transistor and group decoding circuit, the subsequent read operations can be performed without recharging and decoding between multiple reads, which reduces the time required for the traditional SRAM read.

The balanced pre-charging structure is shown in Fig. 2(a). Based on the general 6T memory array, this structure reuses the transistor $M_i(i=0,1,...,n)$ connecting BL and BLB of each column to achieve consecutive read operations. Take reading four rows consecutively as an example. Assume that the data $Q[0:3]$ stored in the bit cells are 1. When reading the first row, M0, M1, ..., Mn are on, and all BL / BLB voltages are pre-charged to $V_{DD}$, i.e., $V_{BL} / BLB = V_{DD}$, and then M0, M1, ..., Mn are off. The WL0 is activated. The BL voltage remains unchanged, and the BLB voltage starts to drop until the voltage can be identified by the SA. After reading the first row, there is a voltage difference between the source and drain of the PMOS $M_i$. $M_i$ is activated by the control signal $VA$ and, the source and drain voltage are balanced, i.e., $V_{BL} / BLB = (V_{BL} (1) + V_{BLB} (1)) / 2$. The average BL / BLB voltage is the initial voltage for the next reading operation. The remaining steps are similar to that of reading the first row. Without decoding, the duration of the pre-charge phase can be reduced, and so does the overall reading time. As the number of reads increases, the bit-line voltage decreases. Therefore, the pulse width of WL is gradually increased so that the discharge voltage of each reading is approximately the same. Only a small amount of delay standard cells are needed to generate different pulse widths. Fig. 2(b) shows the BL / BLB voltage waveform of four consecutive reads.

Fig. 2 (a) The balanced pre-charging structure memory array. (b) Sequence diagram of consecutive reading with one pre-charging.
3. Group decoding

Compared with the traditional decoding method [20, 21, 22], in this paper, a group decoding method is proposed for the consecutive read strategy. It adopts a "semi-random" decoding method. Decoding is performed by combining random decoding with fixed decoding. The array is divided into different blocks, and the blocks can be randomly selected by the control signal. The data inside the block can be decoded automatically. Take a 64 * 64 array shown in Fig. 3 (a) as an example to illustrate the group decoding method. Among them, A8A7A6A5A4A3A2A1A0 is the address signal. The group decoding circuit ignores A5~A0 during consecutive reading and takes A8A7A6 as the row group address. The 64 rows of the storage array are divided into eight large blocks, which are Block0, Block1, ..., and Block7. As shown in Fig. 3(b), the A8A7A6 controls which block to output. When performing an in-memory calculation, only A8A7A6 needs to be specified, and the group decoding module will continuously read all data in the group. The in-memory computing based on group decoding is divided into four steps.

Step 1. Read data in a single row. The WL of the first row is activated. Column selection signal A2A1A0 changes from 000 to 111 controlled by group decoding circuit. For example, As shown in Fig. 3(b), A2A1A0 = 000 is to select the first column in every eight columns, and the output data is detected by SAs and transferred to the operation module.

Step 2. The WLs in the Block are opened and closed in turn. As shown in Fig. 3(b), all the rows of data are selected in proper order. If the address of the next row still belongs to this group and the required data are not fully read, return to Step 1. Read the data and transfer the data into the operation module. Otherwise, go to Step 3.

Step 3. The data is calculated in the in-memory calculation module to get the result.

Step 4. Transfer the result back to the memory array.

Fig. 3(c) shows the circuit realizing the group decoding function and its sequence diagram. The input signal WL activated by the control signal A8A7A6 generates different WL pulses WL0, WL1,..., WL3 with a combination of delay and logic modules.

4. Simulation verification based on 6T SRAM

At first, the proposed structure is verified based on the general 6T SRAM with the 65nm technology. The Vdd of the SRAM is 1.2V. The simulation results including RSNM, read times, reading speed, and energy consumption are as follows.

4.1 RSNM

Due to the deviation of the device manufacturing process and external noise interference, SRAM may have reading errors [23, 24, 25]. When facing a large number of reading operations, the stability of the read is more important. The read static noise margin (RSNM) can be used as a critical reference index for reading stability [26, 27]. In traditional in-memory computing, the BL/BLB voltage needs to be pre-charged to the power supply voltage every time in the reading operation. While in the proposed method, the BL/BLB voltage only needs to be pre-charged to the power supply voltage at the first reading, and the initial BL/BLB voltage gradually decreases during consecutive reading. Therefore, it is necessary to measure the RSNM at different BL/BLB voltages to measure the bit cells' stability during the whole reading period. Fig. 4(a) shows the value of RSNM under different BL/BLB voltages. In the beginning, the value of RSNM increases with the decrease of BL/BLB voltage. When the BL / BLB voltage drops to about 0.75V, the value of RSNM is the largest. When the BL/BLB voltage continues to drop, the value of RSNM begins to decrease. As shown in Fig. 4 (a), selecting any voltage within the selected voltage range in the figure will not reduce the bit cell's stability.
Except for the first reading, the RSNM of other readings will be improved, up to 13.8% at most. Taking 8 consecutive reads as an example, the cell stability is improved and the overall reading error probability is reduced. The stability of the element is verified by Monte Carlo simulation, as shown in Fig. 4(b). Fig. 4(c) shows the performance does not change much as temperature varies, when the BL/BLB voltage is between 0.7V and 1.2V.

![Image](Image)

Fig. 4 (a) RSNM values at different BL/BLB voltages. (b) The results of Monte Carlo results of RSNM value under different BL/BLB voltage. (c) The effect of temperature change on RSNM.

4.2 The number of consecutive reads
Furthermore, the number of consecutive readings is related to the minimum sensing voltage of the SA. Enough voltage difference between BL and BLB is needed to make the SA work. Different SAs have different minimum differential voltages [28, 29, 30]. The maximum number of reading operations within the applicable voltage range (1.2V ~ 0.7V) mentioned above are shown in Table I. It can be seen that the more robust the SA is, the more times it can read continuously. When the bit-line voltage is above 0.7V, the worst RSNM is improved compared with that at 1.2V as shown in Fig. 4(b). Considering the worst case that may occur in the read operation, a relatively stable voltage (1.2V ~ 0.7V) is selected as the bit line voltage range based on the Monte Carlo simulation results. The stability of the cell is improved and the read error rate is reduced.

| Table I | The number of discharges at $V_{DD} = 1.2V$. |
|---------|-------------------------------------------|
| $\Delta V$ | 100(mV) | 125(mV) | 150(mV) |
| Number of times | 9 | 7 | 5 |

4.3 Analysis of voltage drop and speed
In order to test the effect of different initial BL/BLB voltage on the discharge voltage difference, we measured the variation of discharge voltage difference under three pulse width, i.e. $\Delta T_1$, $\Delta T_2$, and $\Delta T_3$. Fig. 5(a) shows the voltage difference only changes a little with the decrease of BL/BLB voltage in the range of 1.2V ~ 0.7V. At the same time, we fixed the discharge voltage difference and measured the opening time of the WL to reach the required discharge voltage difference. As shown in Fig. 5(b), the required opening time does not change much, which is within the tolerance of SA.

The proposed structure supports a wide range of BL/BLB voltages (VBL / BLB = 1.2V~0.7V) without significantly increasing WL’s opening time. Since the proposed strategy reduces the time of the pre-charge/decoding phase, the read speed is increased. This paper taking eight consecutive reads as an example, the traditional SRAM reading method needs to decode row by row. In contrast, the group decoding strategy mentioned in this paper only needs to consider the pre-charging time to read eight rows of data. A 128-row array is simulated. The simulation results show that the decoding time of the traditional SRAM is 1920ps (240ps at a time). In contrast, the pre-charging time of the equalization pre-charging structure proposed in this paper is about 480ps (60ps at a time), and the reading speed is increased by 75%.

![Image](Image)

Fig. 5 (a) Voltage Drop at different BL/BLB Voltage. (b) Opening time of the WL at different BL/BLB voltages.

4.4 Energy consumption
Besides, the proposed structure can reduce energy consumption. It can be proved that the energy consumption in the traditional reading method [31, 32] and the consecutive read strategy are:

$$ W_{\text{old}} = \frac{1}{2} \cdot n \cdot CU^2(2\alpha - \alpha^2) \quad (1) $$

$$ W_{\text{new}} = \frac{1}{2} \cdot n \cdot CU^2(2\alpha - n\alpha^2) \quad (2) $$
Where $W_{\text{read}}$ in Eq.(1) is the energy consumption when using the traditional SRAM reading method to perform $n$ times of reading operations. $n$ is the number of reads. $C$ is the parasitic capacitance of BL/BLB, and $U$ is the supply voltage $V_{\text{DD}}$. $\alpha=\Delta V/U$. $\Delta V$ is the bit-line voltage difference of each reading. $W_{\text{new}}$ in Eq. (2) is the energy consumption when using the continuous read method. If $n>1$, the total energy consumption is reduced.

Fig. 6 shows the power consumption of the continuous read and the traditional SRAM read under different $\Delta V$. The shadow part is reduced power consumption. When $\Delta V = 150\text{mV}$, 5 times of reading, 5.35% power consumption is saved; when $\Delta V = 125\text{mV}$, 7 times of reading, 6.27% power consumption is saved; when $\Delta V = 100\text{mV}$, 9 times of reading, 11.58% power consumption is saved.

![Fig. 6 Comparison of power consumption.](image)

5. Implementation of the proposed structure on 9T SRAM

To verify the effectiveness of the proposed strategy, it is also verified with the non-traditional 9T SRAM [15] based on 65nm technology. The $V_{\text{DD}}$ of the 9T SRAM is set to 1.2V.

In [15], 9T SRAM is used to perform XOR calculation. The data is stored in different rows. As shown in Fig. 7, the data in cell A is 1, and the data in cell B is 0. After reading the data in cell A, signal L is activated and R is disconnected. WL0 and B0 are both 0. When the A XOR B operation is needed, turn on the control signal B1 and WL1. At this time, the CMOS transistors drawn by the dotted line are all in the off state, and the current flow direction is shown in Fig. 7. $V_{\text{DD}}$ drops, representing A XOR B = 1. In this process, only continuous read operations are carried out. Thus, the proposed strategy in this paper can also realize the same function in-memory computing mode, further validating the effectiveness of the structure. The simulation results of 9T SRAM with the proposed method, including RSNM and reading speed are shown.

5.1 RSNM

Fig. 8(a) shows the RSNM of 9T cells at different BL/BLB voltages. Note that the trend of 9T SRAM RSNM is similar to that of traditional 6T SRAM. In the beginning, the value of RSNM increases with the drop of BL/BLB voltage. When the BL/BLB voltage drops to about 0.4V, the value of RSNM is the largest, and then the value of RSNM decreases. The value of RSNM in the range of 1.2V – 0.7V is higher than the original design, up to 6.2% at most.

![Fig. 7 Discharge diagram of BL when A = 1 and B = 0.](image)

5.2 Reading speed

In addition, the overall reading speed of 9T SRAM is similar to that of traditional 6T SRAM. As shown in Fig. 8(b), the abscissa is the initial voltage of BL / BLB, and the ordinate axis is the opening time of the WL required for BL and BLB to form sufficient voltage difference. In the beginning, the WL opening increases slowly with the decrease of BL/BLB voltage. When the BL/BLB voltage drops to about 0.7V, the opening time of WL increases rapidly, but within the selected BL/BLB voltage range (1.2V – 0.7V), the opening time of WL just changes a little bit. Since the proposed strategy reduces the time of the pre-charge/decoding phase, the read speed is increased.

6. Conclusion

This paper introduces the balanced pre-charging and group decoding methods and implements them to both 6T SRAM and non-traditional 9T SRAM, which proves the effectiveness of the methods. Simulation results show
that the proposed strategy can obtain higher reading stability and reading speed and lower power consumption.

Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No. 2018YFB2206202, the State Key Program of the National Natural Science Foundation of China under Grant 61934005, and the National Science and Technology Major Project under Grant 2017ZX01028-101-003.

References

[1] V. Neuman: The computer and the brain (Yale University Press, New Haven, 2003) 2nd ed. 327.

[2] J. W. Backus: “Can programming be liberated from the von Neumann style?: a functional style and its algebra of programs,” Communications of The ACM 21 (1978) 613 (DOI: 10.1145/35976.35979).

[3] D. G. Elliott, et al.: “Computational RAM: implementing processors in memory,” IEEE J. Design & Test of Computers 16 (1999) 32 (DOI: 10.1109/54.748803).

[4] D. A. Patterson, et al.: “Intelligent RAM (IRAM): chips that remember and compute,” ISSCC (1997) 224 (DOI: 10.1109/ISSCC.1997.585348).

[5] Y. Chen, et al.: “An efficient ReRAM-based inference accelerator for convolutional neural networks via activation reuse,” IEEE Electron. Express 16 (2019) 20190396 (DOI: 10.1587/exle.16.20190396).

[6] J. Zhang and Z. Wang: “In-memory computation of a machine-learning classifier in a standard 6T SRAM array,” IEEE J. Solid-State Circuits 52 (2017) 915 (DOI: 10.1109/JSSC.2016.2642198).

[7] M. Kang and N. R. Shanbhag: “In-Memory Computing Architectures for Sparse Distributed Memory,” IEEE J. Transactions on Biomedical Circuits and Systems 10 (2016) 855 (DOI: 10.1109/TBCAS.2016.254548).

[8] A. Agrawal, et al.: “X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories,” IEEE J. Transactions on Circuits and Systems 65 (2018) 4219 (DOI: 10.1109/TCSII.2018.2848990).

[9] S. Jeloka, et al.: “A 28 nm configurable memory (TCAM/BCAM/SRAM) using push-rule 6T bit cell enabling logic-in-memory,” IEEE J. Solid-State Circuits 51 (2016) 1009 (DOI: 10.1109/JSSC.2016.2515510).

[10] J. Zeng, et al.: “DM-IMCA: A Dual-mode In-Memory Computing Architecture for General Purpose Processing,” IEICE Electronics Express 17 (2020) 20200005 (DOI: 10.1587/exle.17.20200005).

[11] M. Kang, et al.: “Multi-functional in-memory inference processor using a standard 6T SRAM array,” IEEE J. Solid-State Circuits 53 (2018) 642 (DOI: 10.1109/JSSC.2017.2782087).

[12] M. Kang, et al.: “An In-Memory VLSI Architecture for Convolutional Neural Networks,” IEEE J. Emerging and Selected Topics in Circuits and System 8 (2018) 494 (DOI: 10.1109/JETCAS.2018.2829522).

[13] J. Zhang and Z. Wang: “In-memory computation of a machine-learning classifier in a standard 6T SRAM array,” IEEE J. Solid-State Circuits 52 (2017) 915 (DOI: 10.1109/JSSC.2016.2642198).

[14] A. Jaiswal, et al.: “8T SRAM Cell as a Multibit Dot-Product Engine for Beyond Von Neumann Computing,” IEEE J. Transactions on Very Large Scale Integration (VLSI) Systems 27 (2019) 2556 (DOI: 10.1109/TVLSI.2019.2929245).

[15] N. Jao, et al.: “Harnessing Emerging Technology for Compute-in-Memory Support,” ISVLSI (2018) 447 (DOI: 10.1109/isvlsi.2018.00087).

[16] F.Hsheh, et al.: “TSV-free FinFET-based Monolithic 3D + IC with computing-in-memory SRAM cell for intelligent IoT devices,” IEDM (2017) I (DOI: 10.1109/iedm.2017.8268380).

[17] B. Yang, et al.: “A Low-Power SRAM Using Bit-Line Charge-Reusing for Read and Write Operations,” IEEE J. Journal of Solid-State Circuits 45 (2010) 2173 (DOI: 10.1109/JSSC.2010.2063950).

[18] B. Yang, et al.: “A low power charge-recycling ROM architecture,” IEEE J. Transactions on Very Large Scale Integration Systems 11 (2013) 590 (DOI: 10.1109/TVLSI.2003.1816138).

[19] K. Kim, et al.: “A Low-Power SRAM Using Bit-Line Charge-Reusing,” IEEE J. Solid-State Circuits 43 (2008) 446 (DOI: 10.1109/JSSC.2007.914294).

[20] B. N. Bagamma, et al.: “Implementation of 5–32 address decoders for SRAM memory in 180nm technology,” ICEEECOT (2017) 110 (DOI: 10.1109/ICEEECOT.2017.8286449).

[21] D. Balobas, et al.: “Design of Low Power, High Performance 2:4 and 4:16 Mixed-Logic Line Decoders,” IEEE J. Transactions on Circuits and Systems II-express Briefs 64 (2017) 176 (DOI: 10.1109/TCISSL.2016.2555020).

[22] M. A. Turi, et al.: “High-Performance Low-Power Selective Precharge Schemes for Address Decoders,” IEEE J. Transactions on Circuits and Systems II: Express Briefs 55 (2008) 917 (DOI: 10.1109/TCSII.2008.923435).

[23] E. Seevinck: “Static-noise margin analysis of MOS SRAM cells,” IEEE J. Solid-State Circuits 22 (1987) 748 (DOI: 10.1109/JSSC.1987.1052809).

[24] K. Takeda, et al.: “A read-static-noise-margin-free SRAM cell for low-V/sub dd/ and high-speed applications,” ISSCC 41 (2006) 113 (DOI: 10.1109/JSSC.2005.859030).

[25] S. Raghayeh, et al.: “Statistical Analysis of Read Static Noise Margin for Near/SubThreshold SRAM Cell,” IEEE J. Transactions on Circuits and Systems 61 (2014) 3386 (DOI: 10.1109/TCSII.2014.2327334).

[26] M. Wieckowski, et al.: “A black box method for stability analysis of arbitrary SRAM cell structures,” IEEE DATE 2010 (2010) 795 (DOI: 10.1109/DATA.2010.5456943).

[27] R. Giterman, et al.: “Towards a black-box methodology for SRAM stability analysis,” IEEE EEEI (2014) (DOI: 10.1109/EEL.2014.7005777).

[28] Y. Lai, et al.: “A Resilient and Power-Efficient Automatic-Power-Down Sense Amplifier for SRAM Design,” IEEE J. Transactions on Circuits and Systems II: Express Briefs 55 (2008) 1031 (DOI: 10.1109/TCSII.2008.926797).

[29] V. P. Hu : “Reliability-Tolerant Design for Ultra-Thin-Body GeOI 6T SRAM Cell and Sense Amplifier,” IEEE J. Electron Devices Society 5 (2017) 107 (DOI: 10.1109/JEDS.2016.2644724).

[30] B. Liu, et al.: “A Low-Voltage SRAM Sense Amplifier With Offset Cancelling Using Digitized Multiple Body Biasing,” IEEE J. Transactions on Circuits and Systems II: Express Briefs 64 (2017) 442 (DOI: 10.1109/TCSI.2016.2563660).

[31] S.B. Lokesh, et al.: “Design of Read and Write Operations for 6t Sram Cell,” J. IOSR Journal of VLSI and Signal Processing 8 (2018) 43 (DOI: 10.9790/2400-080101434).

[32] S.B. Lokesh, et al.: “Design of Read and Write Operations for 6t Sram Cell,” J. IOSR Journal of VLSI and Signal Processing 8 (2018) 43 (DOI: 10.9790/2400-080101434).