ELectronics EXpress

LETTER

IEICE Electronics Express, Vol.12, No.20, 1–6

Methods to speed up read operation in a 64 Mbit phase change memory chip

Qian Wang*, Xi Li*, Houpeng Chen[a)], Yifeng Chen, Yueqing Wang, Xi Fan, Jiajun Hu, Xiaoyun Li, and Zhitang Song

State Key Laboratory of Functional Materials for Information, Shanghai Institute of Micro-system and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

a) *chp6468@mail.sim.ac.cn*

Abstract: A 64 Mbit phase change memory chip is fabricated in 40 nm CMOS technology. An improved fully-differential sense amplifier with a bias voltage instead of the reference resistor branch is proposed to diminish the chip area. The transient response capability of the proposed sense amplifier is improved by removing the large parasitic capacitance of bit line in the feedback network. Smaller parasitic capacitance is also obtained by the separated programming and reading transmission gates to speed up the read operation. The hierarchical bit line architecture is used to reduce the length of bit line, and thus favorable read performance can be achieved.

Keywords: phase change memory, sense amplifier, hierarchical bit line, read operation

Classification: Integrated circuits

References

[1] S. R. Ovshinsky: Phys. Rev. Lett. **21** (1968) 1450. DOI:10.1103/PhysRevLett.21.1450

[2] J. Maimon, E. Spall, R. Quinn and S. Schnur: IEEE Aerospace Conference Proc. (2001) 2289. DOI:10.1109/AERO.2001.931188

[3] M. Gill, T. Lowrey and J. Park: ISSCC (2002) 12.4. DOI:10.1109/ISSCC.2002.993006

[4] S. Hanzawa, N. Kitai, K. Osada, A. Kotabe, Y. Matsui, N. Takaura, M. Moniwa and T. Kawahara: ISSCC (2007) 26.2. DOI:10.1109/ISSCC.2007.373500

[5] M. Rizzi, N. Ciocchini, A. Montefiori, M. Ferro, P. Fantini, A. L. Lacaita and D. Ielmini: IEEE IEDM **13** (2013) 578.

[6] F. Bedeschi, E. Bonizzoni, O. Khouri, C. Resta and G. Torelli: ISCAS (2004) 625. DOI:10.1109/ISCAS.2004.1329349

[7] A. Redaelli, A. Pirovano, F. Pellizzer and A. L. Lacaita: IEEE Electron Device Lett. **25** (2004) 684. DOI:10.1109/LED.2004.836032

[8] N. Papandreou, H. Pozidis, T. Mittelholzer, G. F. Close, M. Breitwisch, C. Lam and E. Eleftheriou: IEEE IMW (2011). DOI:10.1109/IMW.2011.5873231

[9] X. Fan, H. Chen, Q. Wang, X. Li, Y. Zhang, J. Hu, R. Jin, Y. Chen and Z. Song: IEICE Electron. Express **11** (2014) 20141071. DOI:10.1587/elex.11.20141071

*These authors contributed equally to this work.

DOI: 10.1587/elex.12.20150792
Received September 18, 2015
Accepted September 24, 2015
Publicized October 8, 2015
Copyedited October 25, 2015

[10] B. D. Yang and L. S. Kim: IEEE J. Solid-State Circuits **40** (2005) 1366. DOI:10.1109/JSSC.2005.848032

1 Introduction

Phase change memory (PCM), based on the property of chalcogenide materials with two different resistances in amorphous and crystalline phase [1], has shown numerous advantages, such as non-volatility, high speed, low consumption, good scalability, high endurance and excellent compatibility with traditional silicon semiconductor fabrication [2]. Research activity of PCM has boomed since PCM shows commercial values [3, 4, 5]. Improving the performance of read-out speed is also a key issue for the PCM chip to be a commercial success. Basically, read operation is performed by measuring the resistance of the addressed storage device. By comparing the cell current with a reference one, the state of the programmed cell could be detected. Bedeschi etc. introduced a fully symmetrical sense amplifier (SA) in previous research. The reference cell is located within the memory array and is identically biased to avoid mismatch and improve sense accuracy [6]. However this choice could be a waste of area, especially for the memory chip with extremely large capacity. To prevent read disturbance due to the threshold switching mechanism [7], bit line biasing technique has been proposed by pervious researchers. The selected bit line is biased to a constant voltage by a voltage regulator [8]. Whereas this structure requires long time to charge the parasitic capacitance in the addressed bit line, and thus slows down the read operation. In order to improve read performance, the reduction of the parasitic capacitance in the bit line path could be an effective approach.

In this paper, a 64 Mbit PCM implemented in 40 nm CMOS technology is presented. An improved fully-differential SA is integrated to achieve fast response of sensing difference between the reference and the cell branch, resulting in fast read operation. Hierarchical bit line topology is proposed. The 64 Mbit memory is divided into 32 tiles of 2 Mbit memory blocks connected in parallel. Only one block is selected during read operation. The selected path is shortened, which reduces the parasitic capacitance and obtains favorable read speed.

2 Sense amplifier design

The structure of the proposed fully-differential SA is shown in Fig. 1(a). The generators of V_b and V_c are shown in Fig. 1(b) and Fig. 1(c), separately. *RBL* (Read Bit Line) is connected to the memory cell through reading transmission gate (RTG). V_b is the bias voltage which is generated from *PM9*, and is mirrored to *PM3* and *PM4*. Here, the width of *PM9* is m times of that of *PM3* and *PM4*. V_c is the clamp voltage to prevent the risk of read disturbance. *EN* is an Enable signal to shut down SA in idle, and also controls to discharge parasitic capacitance in bit line path through *NM9*.

I_{cell} represents cell current, while I_{ref} represents reference current. The difference between the two currents is integrated into the parasitic capacitor at node V_1

DOI: 10.1587/elex.12.20150792
Received September 18, 2015
Accepted September 24, 2015
Publicized October 8, 2015
Copyedited October 25, 2015

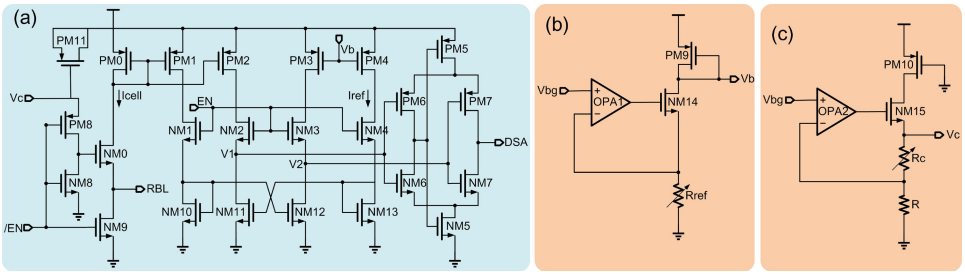


Fig. 1. (a) Structure of fully-differential SA. (b) Generator of bias voltage V_b. (c) Generator of clamp voltage V_c.

and V_2. When $I_{cell} > I_{ref}$, V_1 node will be charged into a high voltage while V_2 node will be discharged into a low voltage. A power voltage *DSA* will be obtained, which indicates the R_{cell} is in crystalline phase with lower resistance than R_{ref}, and vice versa. The cell and reference current equations are described in below:

$$I_{cell} = \frac{V_c - V_{th}}{R_{cell}}. \tag{1}$$

$$I_{ref} = \frac{V_{bg}/m}{R_{ref}}. \tag{2}$$

Here, V_{th} (~0.65 V) is threshold voltage of *NM0*. The size of *NM0* tends to be large (W/L = 8 um/2 um) to control the deviation of V_{th}, which is caused by the large difference of values of R_{cell} in different phases. V_{bg} (~1.25 V) is generated from bandgap block. R_{ref} is adjustable to satisfy process deviation. The numerators of the two equations, i.e. the voltages across cell and reference, could be balanced by choosing proper values of V_c and m. The clamp voltage V_c is generated in a resistance voltage dividing circuit, as shown in Fig. 1(c). The value of V_c could be calculated by following Eq. (3).

$$V_c = V_{bg} \cdot \frac{R_c}{R_c + R}. \tag{3}$$

Here, R is a fixed resistor, while R_c is adjustable. V_c drop will occur when *EN* is activated. To reduce the voltage drop, low dropout regulator is used to generate V_c, and furthermore, a decoupling capacitor *PM11* is added between V_c and power supply. Simulation waveforms of read operation are shown in Fig. 2(a) and (b). Power supply is 2.5 V. R_{ref} is 200 Kohm. R_{cell} is 1 Mohm in Fig. 2(a), and 30 Kohm in Fig. 2(b). The simulated sense time of our SA is 42 ns. The difference between this SA and the one we described in another literature [9], is that, we propose a bias voltage V_b to replace the reference resistor branch. Since there are 384 SAs shared one group of V_b and V_c generators in the chip, this method could reduce the total area of SAs. Another difference is that, V_c is connected to the gate of *NM0* through transmission gate (TG), instead of the drain of *NM0*. The long bit line with large parasitic capacitance is excluded in the feedback network, and the stable time of the amplifier system is shortened. Moreover, we separate the programming transmission gate (PTG) and RTG. RESET current (~mA level) flows through PTGs from write driver into the memory cell, while read current (~μA level) flows through RTGs from SA into the selected cell. The size of RTG could be smaller than that of PTG, and the smaller parasitic capacitance of RTG is easy to be

DOI: 10.1587/elex.12.20150792
Received September 18, 2015
Accepted September 24, 2015
Publicized October 8, 2015
Copyedited October 25, 2015

charged, which speeds up the operation of reading. The comparison between the two SAs is shown in Table I.

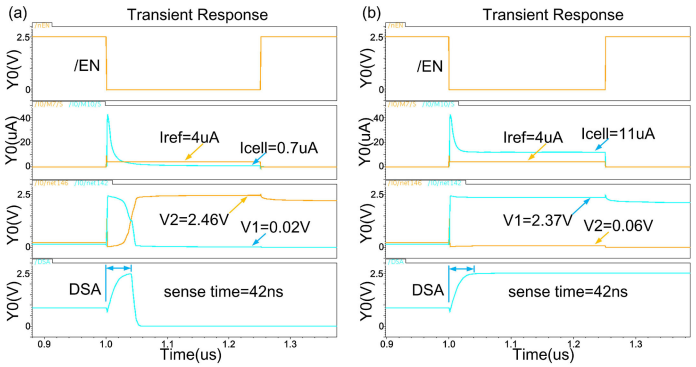


Fig. 2. (a) Simulation results of read operation for $R_{cell} > R_{ref}$. (b) Simulation results of read operation for $R_{cell} < R_{ref}$.

Table I. The comparison between the proposed SA and that in Reference [9]

	The proposed SA	SA in Reference [9]
Process technology	40 nm PCM	130 nm PCM
Power supply	2.5 V	3.3 V
Block area	120 um^2 (46080 um^2 of 384 SAs)	235 um^2 (1880 um^2 of 8 SAs)
Sense time	42 ns @27°C, simulated	98 ns@27°C, simulated

3 Hierarchical bit line architecture

The 64 Mbit memory with 32 Mbit redundancy cells for ECC computing is divided into 4 banks. Each bank consists of 8 tiles of sub-banks (1024 rows × 3072 columns). Fig. 3(a) shows the architecture of a 2 Mbit array, including 8 tiles of 256 Kbit arrays (1024rows × 256columns) organized in parallel. Every 1024rows × 1column shares a *BBL* (Block Bit Line). Every 16 *BBLs* in 8 tiles are connected to a *LBL* (Local Bit Line) through 16 BTGs (Block Transmission Gates). Every 16 *LBLs* are connected to a *RBL* in SA through 16 RTGs. Every 128 *LBLs* are connected to a *GBL* (Global Bit Line) in write driver through 128 PTGs. This structure is called hierarchical bit line. Yang etc. proposed a hierarchical bit line structure based on TGs [10]. To save the area, we exclude PTGs, RTGs, SAs, column decoders and write drivers in sub-bank, and share them in bank level.

Fig. 3(b) shows the simulation results of read time with or without hierarchical bit line architecture. The read time for a 32 Kbit array (1024 rows × 32 columns) is 39 ns, and that for a 256 Kbit array (8192 rows × 32 columns) without hierarchical bit line architecture is 235 ns. While for a 256 Kbit array consisting of eight 32 Kbit array (8 × 1024 rows × 32 columns), the read time is 42 ns, which is close to that for a 32 Kbit array. The reason is that only one of the 32 Kbit arrays is activated simultaneously during read operation, thus the load of *RBL* is mainly determined by the length of *BBL*. With the proposed hierarchical bit line architecture, the parasitic

DOI: 10.1587/elex.12.20150792
Received September 18, 2015
Accepted September 24, 2015
Publicized October 8, 2015
Copyedited October 25, 2015

capacitance of bit line is reduced enormously, and the read operation is speeded up significantly.

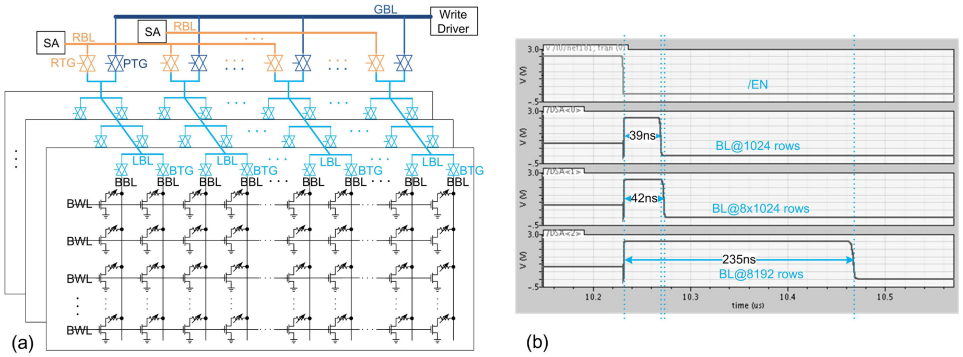
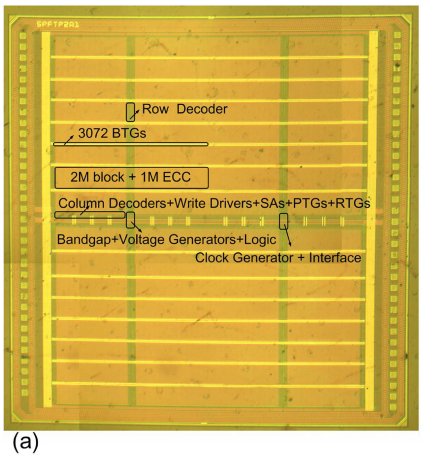


Fig. 3. (a) Hierarchical bit line topology of a 2 Mbit array. (b) Simulation results of read time with or without hierarchical bit line architecture.

4 Chip measurement results

Our 64 Mbit PCM chip with hierarchical bit line architecture was fabricated in 40 nm CMOS process. The microphotograph of the chip is shown in Fig. 4(a). The main characteristics of the chip are summarized in Fig. 4(b).



Chip size	15.42um^2
Cell size	0.087um^2
ECC	8+4 scheme
Interconnects	1 poly + 4 Cu + 1 Al
Interface	Test vehicle
Supply voltage	2.5V
RESET current	1.0mA/10ns
SET current	0.2mA/200ns

(b)

Fig. 4. (a) The microphotograph of the chip. (b) The main characteristics of the chip.

The resistance distribution of the experimental chip is shown in Fig. 5(a). A resistance difference of about two orders of magnitude between the SET and RESET states is observed. Fig. 5(b) shows the experimental transient waveforms of read operation. *DSA* is in tri-state before SA enabled. With */EN* falling, *PM8* and *NM0* shown in Fig. 1(a) are turned on, and the selected bit line is charged by relatively high current I_{cell}, thus *DSA* rises from ground level. After finished the charge procedure, measuring of cell resistance begins. If the selected memory cell is in amorphous phase, then $I_{cell} < I_{ref}$, and *DSA* will be pulled down to ground level. In contrast, if $I_{cell} > I_{ref}$, which means the cell is in crystalline structure, *DSA* will

DOI: 10.1587/elex.12.20150792
Received September 18, 2015
Accepted September 24, 2015
Publicized October 8, 2015
Copyedited October 25, 2015

continue to rise to power level. Here, sense time, which corresponds to the delay from the falling edge of */EN* to the falling edge of *DSA* when reading "0" state, is about 50 ns, which is close to the simulation result. The delay time between the completing of sensing and the output of ground or power level is caused by discharging or charging of the parasitic capacitance in the testing path, including an output TG with */EN* as the enable signal and a metal connection to the test PAD. When */EN* is negated, the output TG is switched off, and the voltage of the test PAD is pulled down to ground. The deviation of sense time for different cells is small (+/−5%). One reason is that sense time is mainly determined by the parasitic capacitance of *BBL*, the length of which is same for all the cells. Another reason is that the deviation of $(I_{cell} - I_{ref})$ is small due to about 1 order of magnitude difference between R_{cell} and R_{ref}. The smaller the $(I_{cell} - I_{ref})$ is, the longer the sense time will be. Tested results show that all the sense times are within 50 ns.

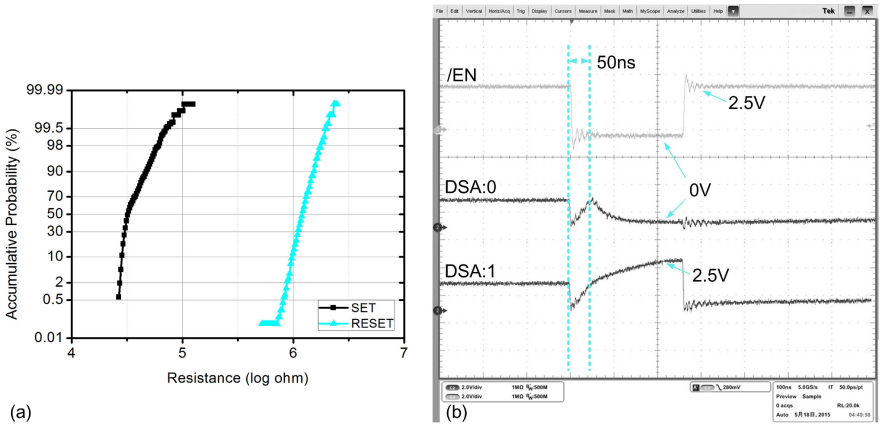


Fig. 5. (a) Resistance distribution for RESET and SET operation of the experimental PCM chip. (b) Measured voltage waveforms of the SA during read operation.

5 Conclusion

In this paper, an improved fully-differential sense amplifier with a bias voltage instead of the reference resistor branch has been presented. It is integrated in a 64 Mbit PCM chip, which is fabricated in 40 nm CMOS process. The hierarchical bit line topology has also been used in the chip. Experimental results show that the improved sense amplifier and the proposed architecture are able to speed up the read operation (~50 ns). The current pulses of RESET and SET programming are 1.0 mA/10 ns and 0.2 mA/200 ns, respectively. The difference between memory cell resistances after RESET and SET reaches about 2 orders of magnitude.

Acknowledgments

This work was financially supported by National Key Basic Research Program of China (2013CBA01900, 2011CBA00607, 2011CB932804), Science and Technology Council of Shanghai (13DZ2295700, 13ZR1447200, 14ZR1447500).

DOI: 10.1587/elex.12.20150792
Received September 18, 2015
Accepted September 24, 2015
Publicized October 8, 2015
Copyedited October 25, 2015