(19) **United States**

(12) **Patent Application Publication** (10) Pub. No.: US 2017/0301405 A1
ZHANG   (43) Pub. Date: Oct. 19, 2017

(54) **MULTI-BIT-PER-CELL THREE-DIMENSIONAL ONE-TIME-PROGRAMMABLE MEMORY**

(71) Applicant: **ChengDu HaiCun IP Technology LLC**, ChengDu (CN)

(72) Inventor: **Guobiao ZHANG**, Corvallis, OR (US)

(73) Assignee: **ChengDu HaiCun IP Technology LLC**, ChengDu (CN)
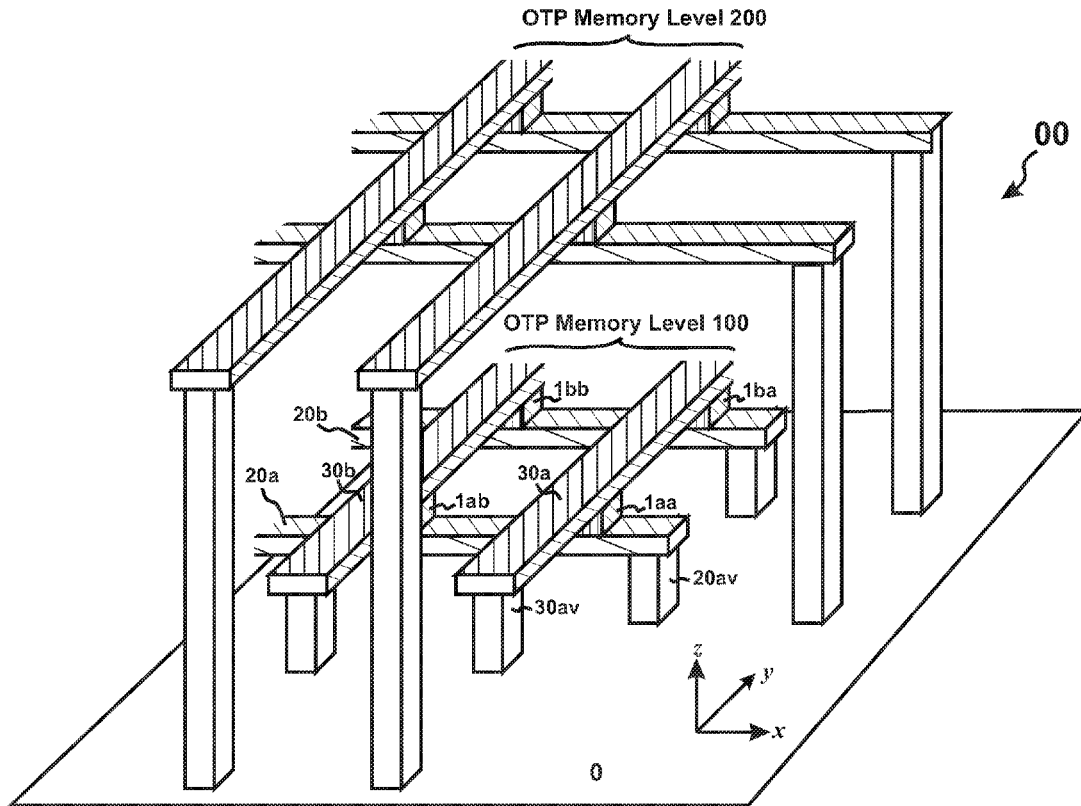
(21) Appl. No.: **15/488,435**

(22) Filed: **Apr. 14, 2017**

(30) **Foreign Application Priority Data**

Apr. 14, 2016   (CN) 201610238012.7

(57)   **ABSTRACT**

The present invention discloses a multi-bit-per-cell three-dimensional read-only memory (3D-OTP$_{MB}$). It comprises a plurality of OTP cells stacked above a semiconductor substrate. Each OTP cell comprises an antifuse layer, which is irreversibly switched from a high-resistance state to a low-resistance state during programming. By adjusting the programming current, the programmed antifuses have different resistances.

00

OTP Memory Level 200

OTP Memory Level 100

1ba

1bb

1aa

20av

1ab

30a

30av

20b

30b

20a

z

y

x

0

Fig. 1

`0`

30a

1aa

22

24

20a

Fig. 2A

`1`

30b

1ab

25x

22

24

20a

Fig. 2B

`2`

30c

1ac

25y

22

24

20a

Fig. 2C

`3`

30d

1ad

25z

22

24

20a

Fig. 2D

Fig. 3

Fig. 4A

Fig. 4B

30a

1aa

22

24

20a

Fig. 5A

30a

1aa

22

26

24

20a

Fig. 5B

30a

1aa

22

20a
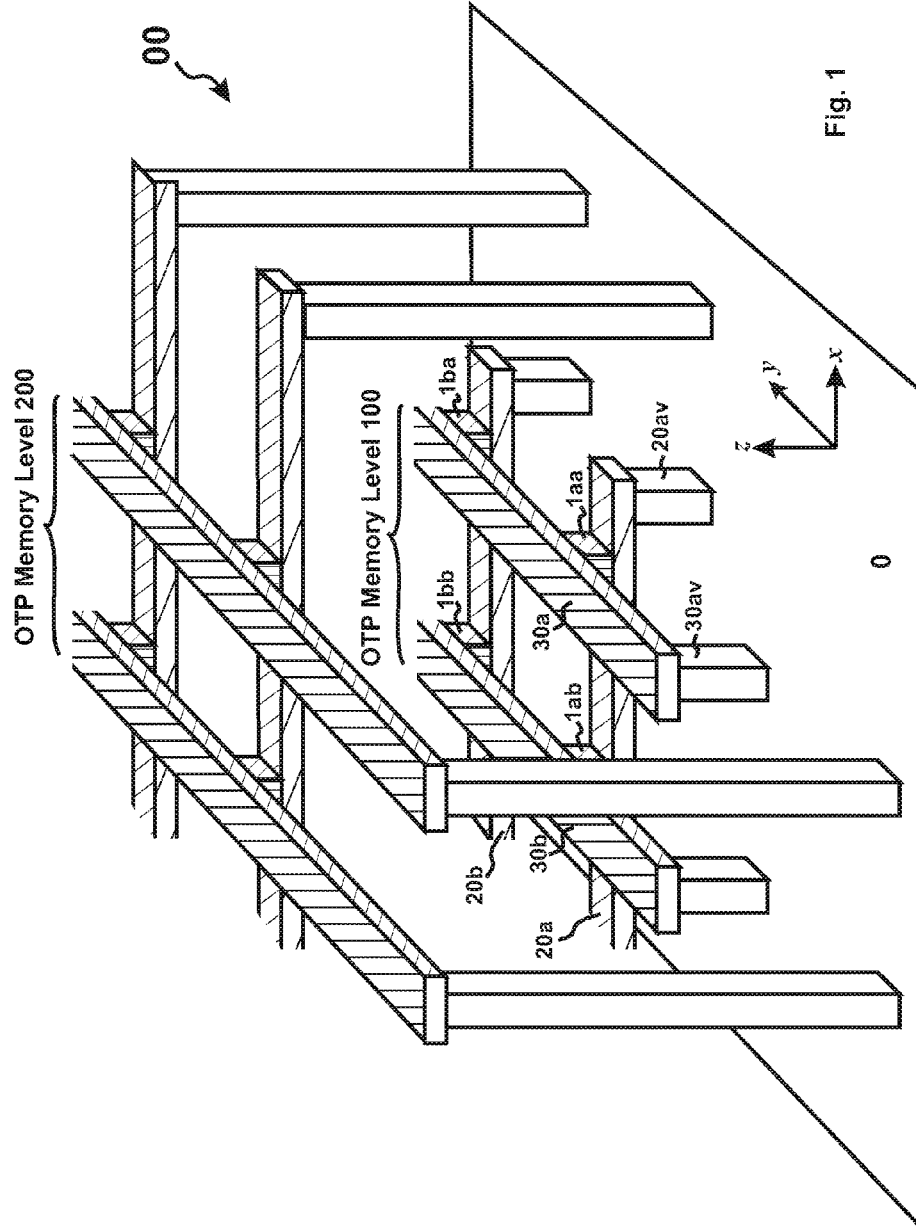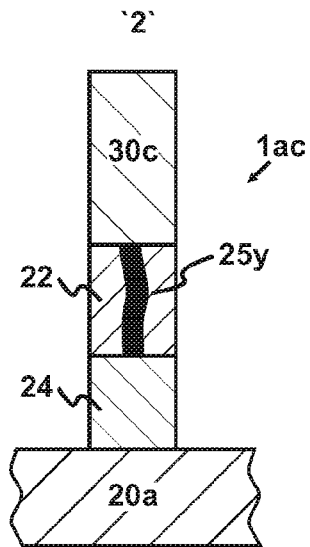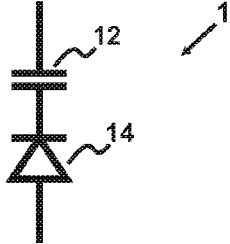
Fig. 5C

Fig. 6A

Fig. 6B

Fig. 7A

Fig. 7B

Fig. 7C

Fig. 8A

Fig. 8B

Fig. 9

MULTI-BIT-PER-CELL THREE-DIMENSIONAL ONE-TIME-PROGRAMMABLE MEMORY

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority from Chinese Patent Application 201610238012.7, filed on Apr. 14, 2016, in the State Intellectual Property Office of the People's Republic of China (CN), the disclosure of which is incorporated herein by reference in its entirety.

BACKGROUND

1. Technical Field of the Invention

[0002] The present invention relates to the field of integrated circuit, and more particularly to one-time-programmable memory (OTP).

2. Prior Art

[0003] Three-dimensional one-time-programmable memory (3D-OTP) is a monolithic semiconductor memory. It comprises a plurality of vertically stacked OTP cells. In a conventional OTP, the memory cells are formed on a two-dimensional (2-D) plane (i.e. on a semiconductor substrate). In contrast, the memory cells of the 3D-OTP are formed in a three-dimensional (3-D) space. The 3D-OTP has a large storage density and a low storage cost. Because the 3D-OTP has a long data retention (>100 years), it is suitable for long-term data storage.

[0004] U.S. Pat. No. 5,838,396 issued to Zhang on Nov. 10, 1998 discloses a 3D-OTP 00. It comprises a semiconductor substrate 0 and a plurality of OTP memory levels 100, 200 stacked above the semiconductor substrate 0. Among them, the memory level 200 is stacked above the memory level 100. Transistors in the substrate 0 and interconnects thereof form a substrate circuit (including the peripheral circuit of the OTP memory levels 100, 200). Each OTP memory level (e.g. 100) comprises a plurality of address lines (e.g. word lines 20a, 20b . . . , and bit lines 30a, 30b . . .) and memory cells (e.g. 1aa-1bb . . .). Each OTP memory level 100 further comprises a plurality of OTP arrays. Each OTP array is a collection of all OTP cells which share at least one address line. Contact vias 20av, 30av couple the address lines 20a, 30a with the substrate 0.

[0005] The 3D-OTP in Zhang is a single-bit-per-cell 3D-OTP, wherein each 3D-OTP cell stores a single bit. Namely, each OTP cell has two states '1' and '0': the '1' OTP cell is in a low-resistance state, whereas the '0' cell is in a high-resistance state. To further improve the storage density and lower the storage cost, it is desired to store more bits in each 3D-OTP cell.

OBJECTS AND ADVANTAGES

[0006] It is a principle object of the present invention to provide a 3D-OTP with a large storage capacity.

[0007] It is a further object of the present invention to provide a 3D-OTP with a low storage cost.

[0008] It is a further object of the present invention to provide a properly working 3D-OTP even with leaky OTP cells.

[0009] It is a further object of the present invention to provide a properly working 3D-OTP even under external interferences.

[0010] In accordance with these and other objects of the present invention, the present invention discloses a multi-bit-per-cell 3D-OTP.

SUMMARY OF THE INVENTION

[0011] The present invention discloses a multi-bit-per-cell three-dimensional one-time-programmable memory (3D-OTPMB). It comprises a plurality of OTP cells stacked above a semiconductor substrate. Each OTP cell comprises an antifuse layer, which is irreversibly switched from a high-resistance state to a low-resistance state during programming. By adjusting the magnitude of the programming currents, 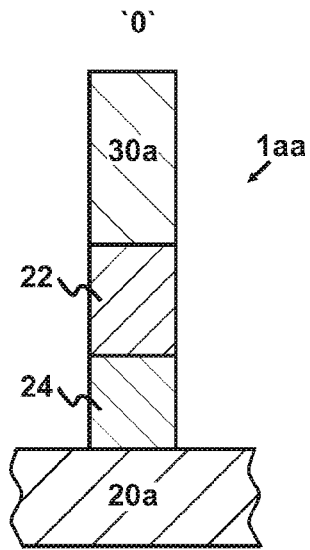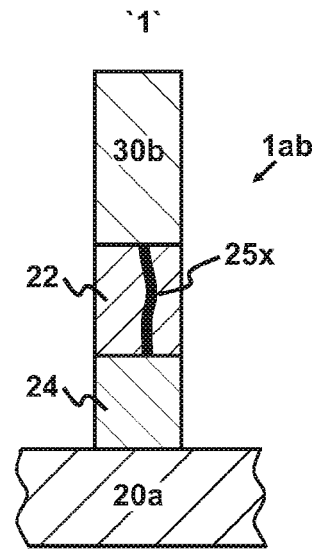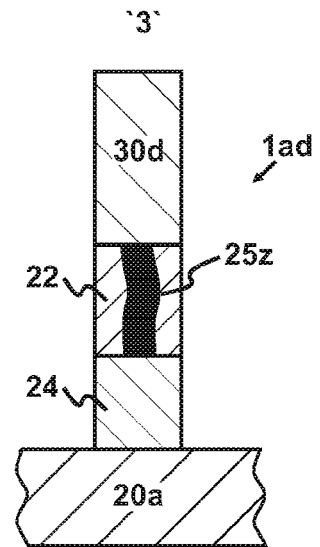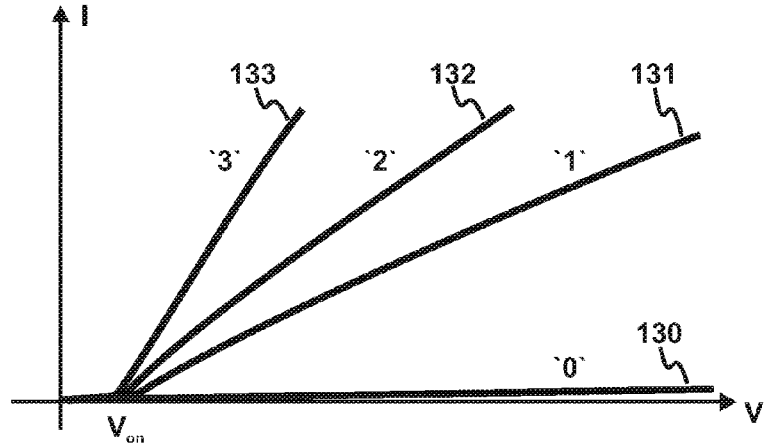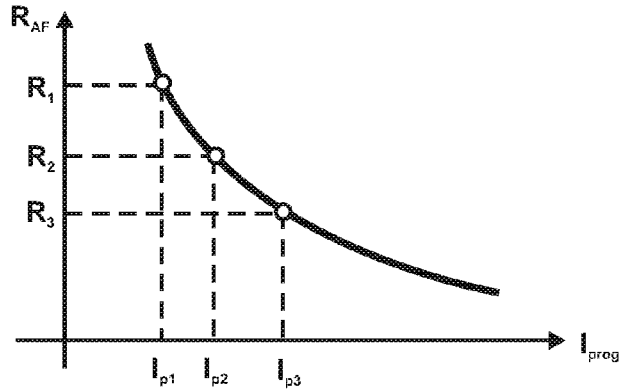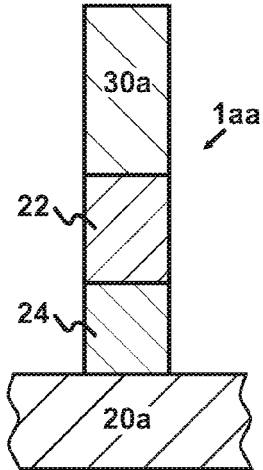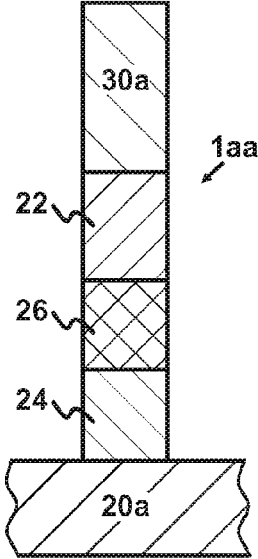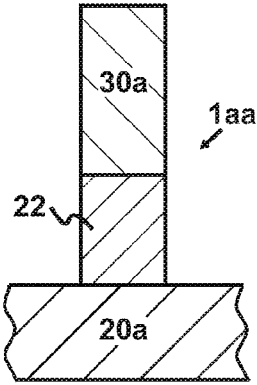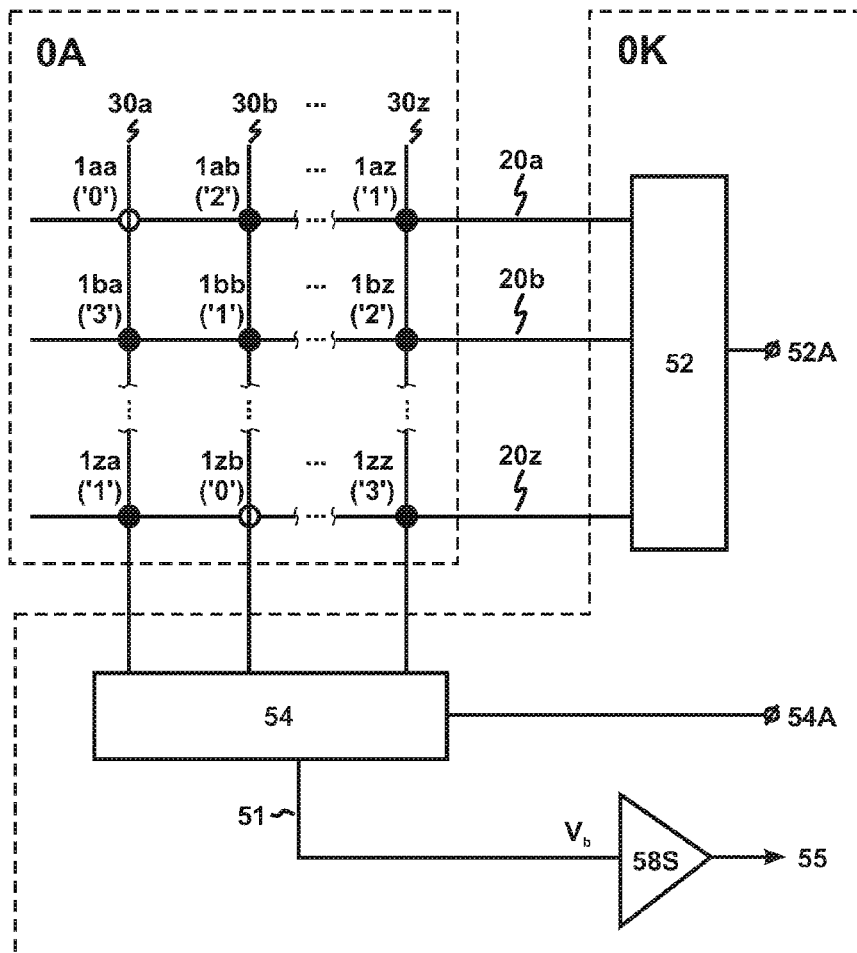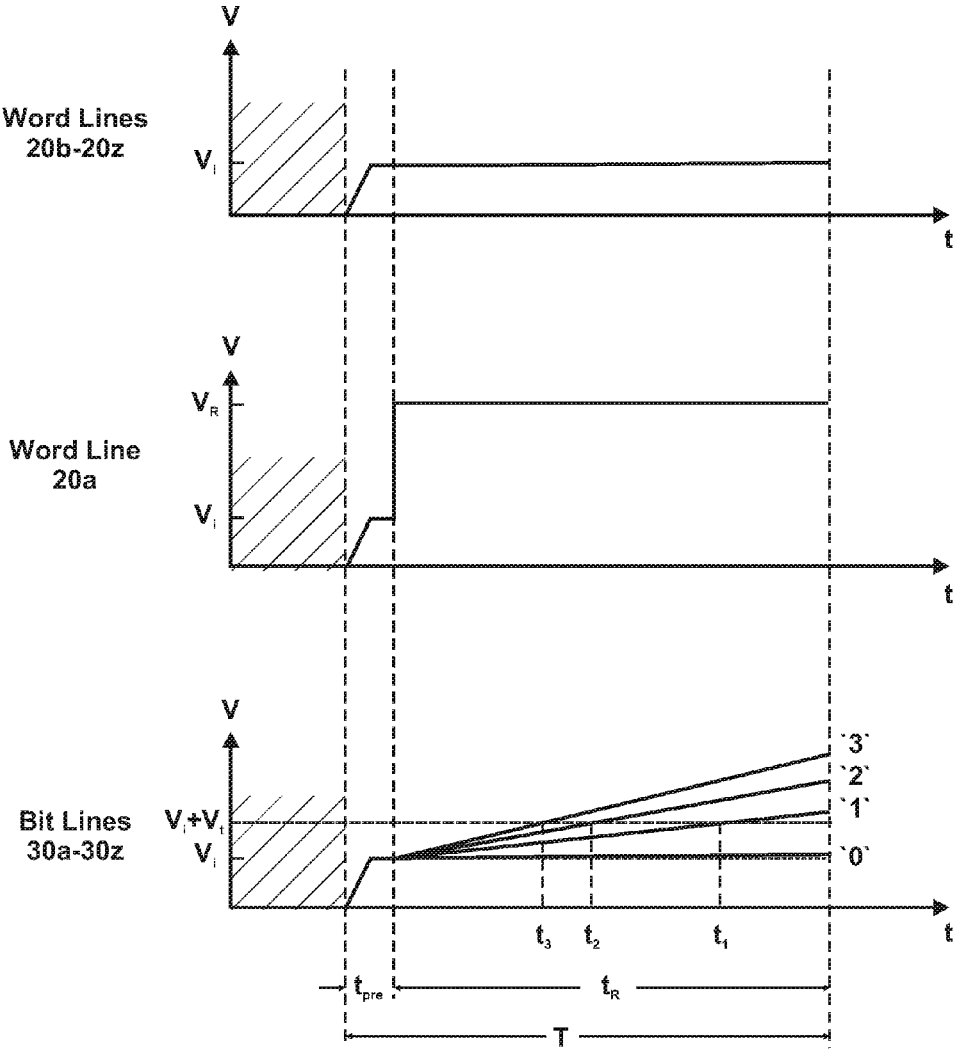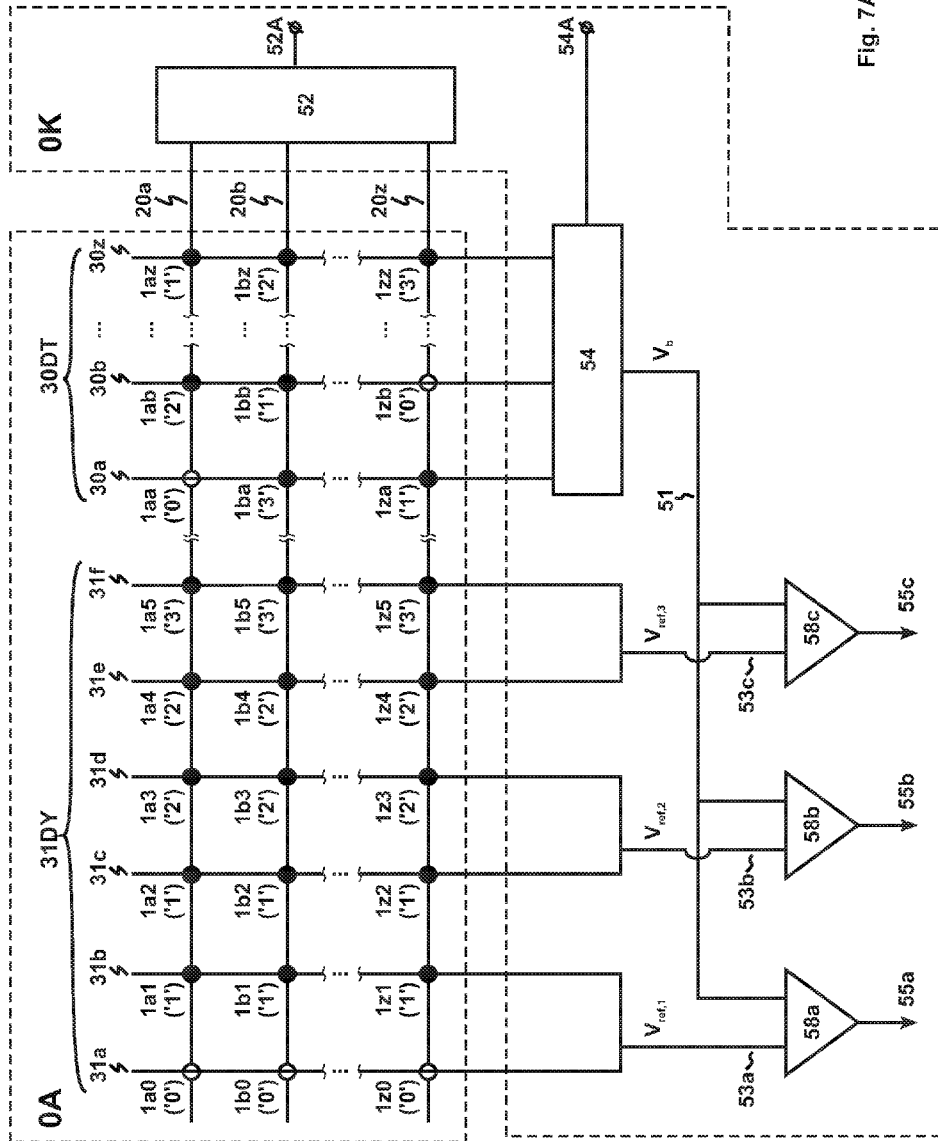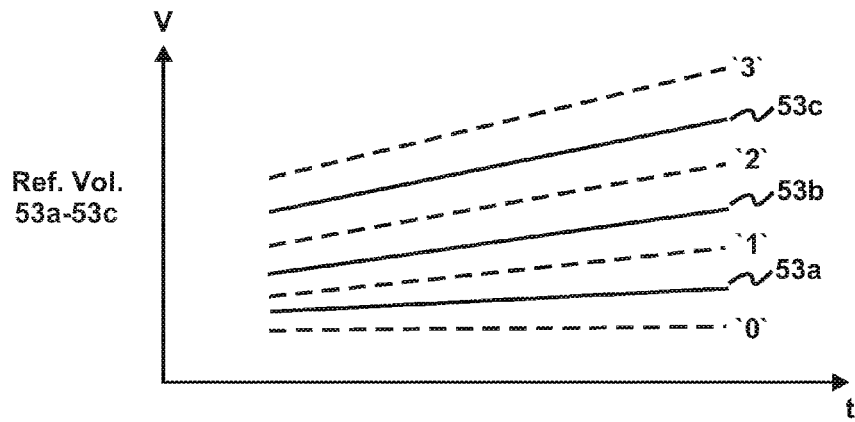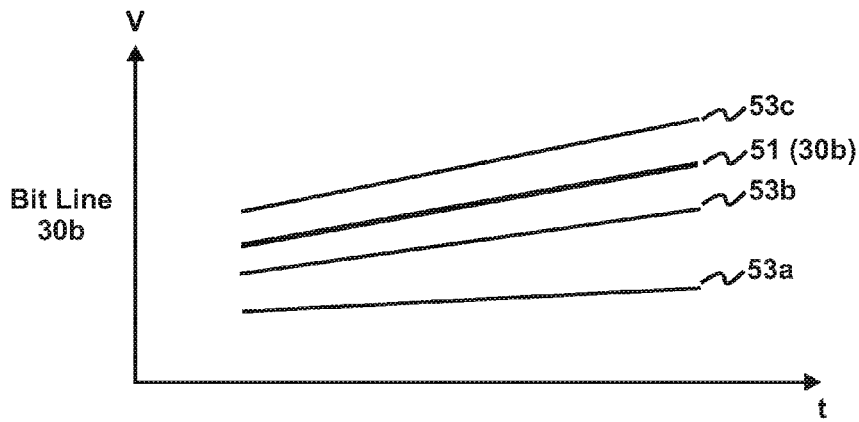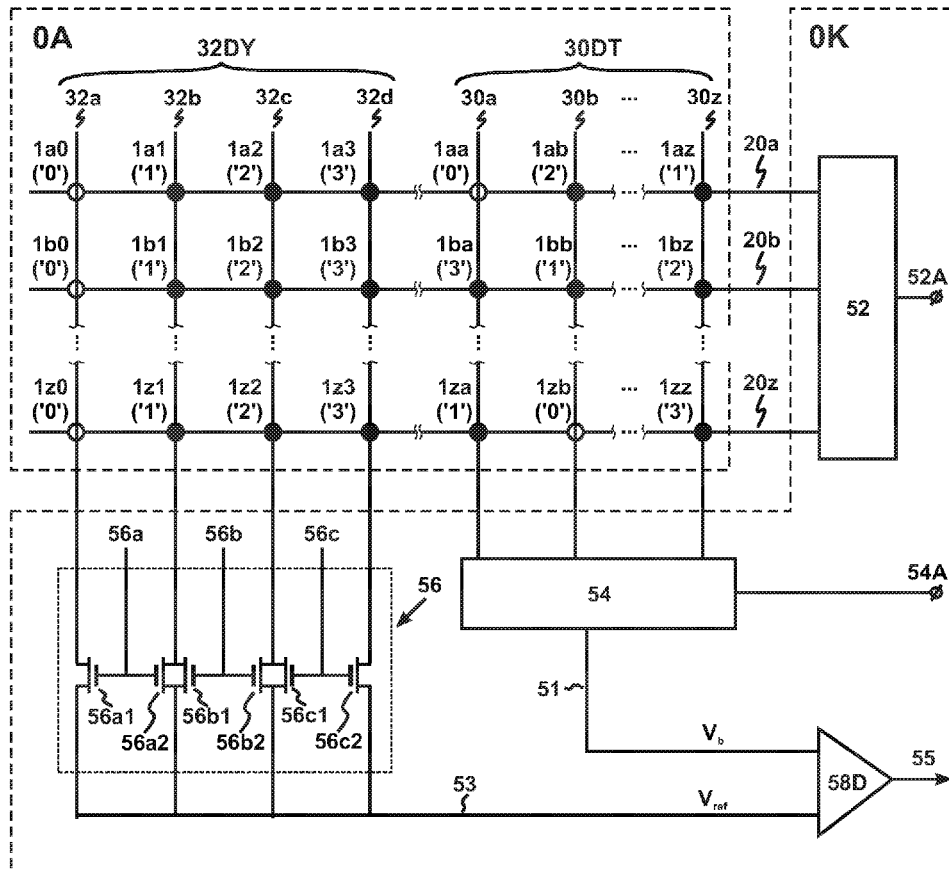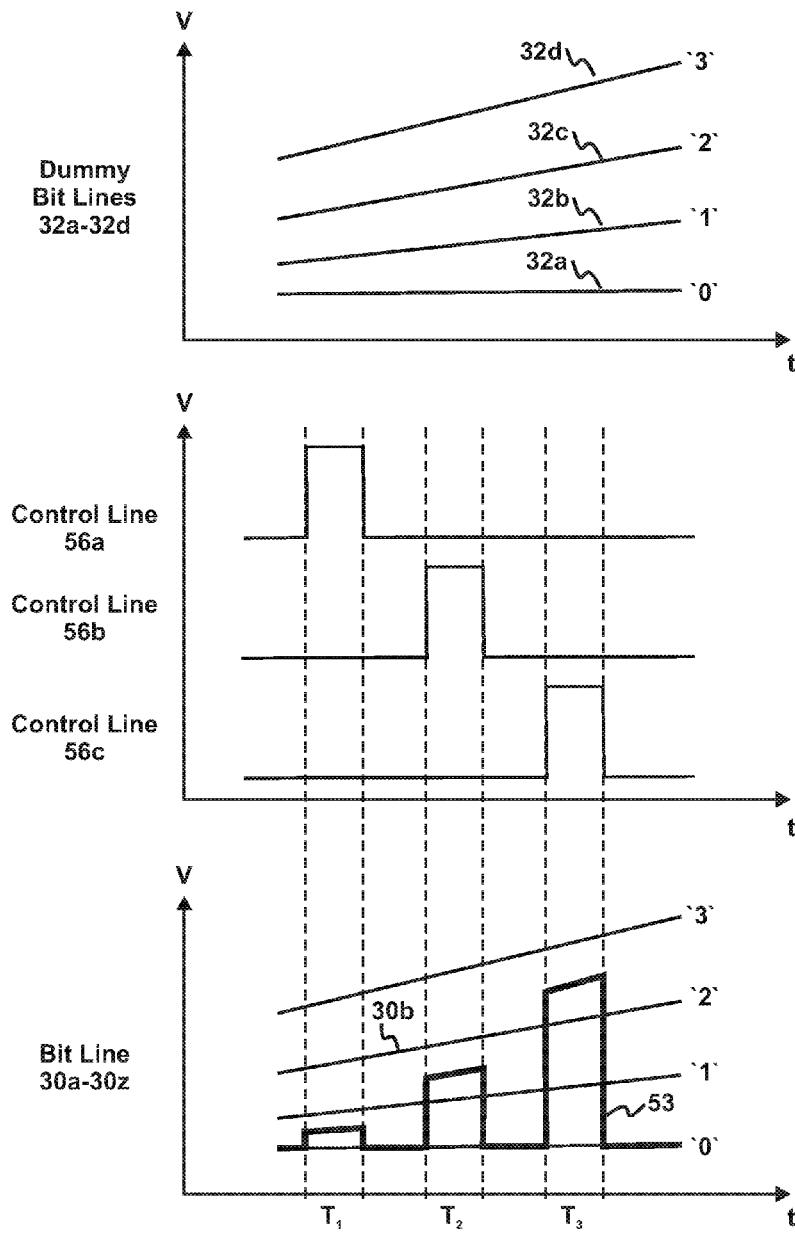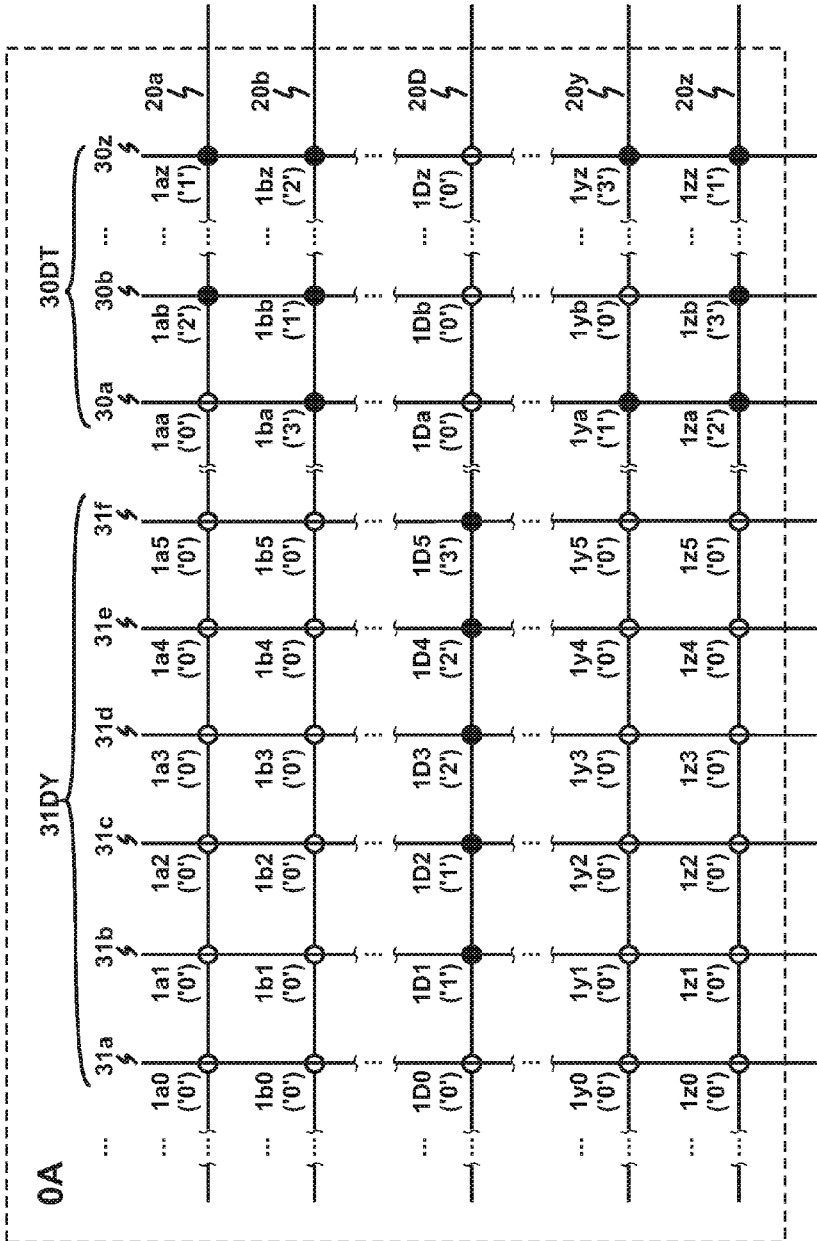the programmed antifuses have different resistances. Using the resistance to represent the digital states, the OTP cells have N (N>2) states: 0, 1, . . . N−1, whose respective resistances are R_O, R_1, . . . R_{N-1}, with $R_0 > R_1 > . . . > R_{N-1}$. Having N states, each OTP cell stores more than one bit.

[0012] To minimize read error due to leaky OTP cells, the present invention discloses a full-read mode. For the full-read mode, the states of all OTP cells on a selected word line are read out during a read cycle. The read cycle includes two read phases: a pre-charge phase and a read-out phase. During the pre-charge phase, all address lines (including all word and all bit lines) in an OTP array are charged to an input bias voltage of an amplifier associated with the OTP array. During the read-out phase, after its voltage is raised to the read voltage VR, a selected word line starts to charge all bit lines through the associated OTP cells. By measuring the voltage change on the bit lines, the states of the associated OTP cells can be determined.

[0013] To minimize read error due to external interferences, the present invention further discloses a differential amplifier for measuring the states of the OTP cells. One input of the differential amplifier is the bit-line voltage V_b from a data OTP cell (i.e. the OTP cell that stores data), while the other input is a reference voltage V_{ref} from a dummy OTP cell (i.e. the OTP cell that provides V_{ref} for the differential amplifier). Like the data OTP cells, the dummy OTP cells have N states. The value of the reference voltage (e.g. $V_{ref,1}$) is between the voltages (e.g. $V_{\cdot 0}$, $V_{\cdot 1}$) on the bit lines associated with the dummy OTP cells in adjacent states (e.g. '0', '1'), preferably equal to the average of the two. To determine the state of a selected data OTP cell, N−1 measurements are taken. The data OTP cell is in the state 'k' if $V_{ref,k-1} < V_b < V_{ref,k}$ (k=1, 2, . . . N−1).

[0014] Accordingly, the present invention discloses a multi-bit-per-cell 3D-OTP (3D-OTPMB), comprising: a semiconductor substrate including transistors thereon; a plurality of OTP cells stacked above said semiconductor substrate, each of said OTP cells comprising an antifuse layer, where said antifuse layer is irreversibly switched from a high-resistance state to a low-resistance state during programming; a plurality of contact vias coupling said OTP cells to said semiconductor substrate; wherein said OTP cells have more than two states, the OTP cell in different states being programmed by different programming currents.

[0015] The present invention further discloses a multi-bit-per-cell 3D-OTP (3D-OTP_{MB}), comprising: a semiconductor substrate including transistors thereon; a plurality of OTP

cells stacked above said semiconductor substrate each of said OTP cells comprising an antifuse layer, where said antifuse layer is irreversibly switched from a high-resistance state to a low-resistance state during programming; a plurality of contact vias coupling said OTP cells to said semiconductor substrate; wherein the resistance of said antifuse layer is determined by a programming current, said OTP cells being programmed by at least two programming currents.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] FIG. 1 is a perspective view of a 3D-OTP.

[0017] FIGS. 2A-2D are cross-sectional views of four preferred 3D-OTP cells in different states.

[0018] FIG. 3 is a symbol of an OTP cell.

[0019] FIG. 4A shows a current-voltage (I-V) characteristics of four preferred OTP cells in different states; FIG. 4B shows a relationship between the resistance of the programmed antifuse and the programming current (R_{AF}-I_P).

[0020] FIGS. 5A-5C are cross-sectional views of three preferred 3D-OTP cells;

[0021] FIG. 6A is a circuit block diagram of a preferred OTP array in the full-read mode; FIG. 6B is its signal timing diagram.

[0022] FIG. 7A is a circuit block diagram of a first preferred OTP array comprising differential amplifiers; FIGS. 7B-7C are the signal timing diagrams for the reference voltages and the bit-line voltages, respectively.

[0023] FIG. 8A is a circuit block diagram of a second preferred OTP array comprising a differential amplifier; FIG. 8B is its signal timing diagram.

[0024] FIG. 9 is a preferred 3D-OTP comprising a dummy word line.

[0025] It should be noted that all the drawings are schematic and not drawn to scale. Relative dimensions and proportions of parts of the device structures in the figures have been shown exaggerated or reduced in size for the sake of clarity and convenience in the drawings. The same reference symbols are generally used to refer to corresponding or similar features in the different embodiments. In FIG. 6A, FIG. 7A, FIG. 8A and FIG. 9, solid dots represent programmed OTP cells, while open dots represent unprogrammed OTP cells. The symbol "/" means a relationship of "and" or "or".

[0026] Throughout the present invention, the phrase "on the substrate" means the active elements of a circuit are formed on the surface of the substrate, although the interconnects between these active elements are formed above the substrate and do not touch the substrate; the phrase "above the substrate" means the active elements are formed above the substrate and do not touch the substrate.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0027] Those of ordinary skills in the art will realize that the following description of the present invention is illustrative only and is not intended to be in any way limiting. Other embodiments of the invention will readily suggest themselves to such skilled persons from an examination of the within disclosure.

[0028] Referring now to FIG. 2A-2D, four preferred OTP cells 1*aa*-1*ad* are disclosed. They are in the states: '0', '1', '2', '3', respectively. Each OTP cell (e.g. 1*aa*) comprises a

top electrode **30***a*, a bottom electrode **20***a*, an antifuse layer **22** and a quasi-conductive layer **24**. The antifuse layer **22** has high resistance before programming (FIG. **2A**), and is irreversibly switched to low resistance after programming (FIGS. **2B-2D**). The quasi-conductive has the following properties: its resistance at the read voltage (i.e. the read resistance) is substantially lower than when the applied voltage has a magnitude smaller than or polarity opposite to that of the read voltage.

[0029] Because the OTP cell 1*aa* is unprogrammed, no conductive filament is formed in its antifuse layer **22**. On the other hand, because the OTP cells 1*ab*-1*ad* are programmed, conductive filaments **25***x*-**25***z* of different sizes are formed therein. Among them, the conductive filament **25***x* of the OTP cell 1*ab* is thinnest and has the largest resistance; the conductive filament **25***z* of the OTP cell 1*ad* is thickest and has the lowest resistance; the conductive filaments **25***y* of the OTP cell 1*ac* has an intermediate size and therefore, has an intermediate resistance.

[0030] FIG. **3** is a symbol of an OTP cell **1**. It comprises an antifuse **12** and a diode **14**. The antifuse **12** comprises the antifuse layer **22** and its resistance is irreversibly switched from high to low during programming. The diode **14** comprises the quasi-conductive layer **24** and is broadly interpreted as any two-terminal device whose resistance at the read voltage is substantially lower than when the applied voltage has a magnitude smaller than or polarity opposite to that of the read voltage.

[0031] Referring now to FIGS. **4A-4B**, the electrical characteristics of the '0'-'3' OTP cells are disclosed. FIG. **4A** shows current-voltage (I-V) characteristics of the '0'-'3' OTP cells. The I-V curve **130** corresponds to a '0' OTP cell 1*aa*; the I-V curve **131** corresponds to a '1' OTP cell 1 *ab*; the I-V curve **132** corresponds to a '2' OTP cell 1 *ac*; the I-V curve **133** corresponds to a '3' OTP cell 1 *ad*. The diode **14** has a turn-on voltage V_{on}. Once the applied voltage is larger than V_{on}, the resistance of the diode **14** drops rapidly. At this time, the resistance of the OTP cell **1** primarily comes from the antifuse layer **12**.

[0032] FIG. **4B** shows a relationship between the resistance of the programmed antifuse (RAF) and the programming current (I_P). As R_{AF} is inversely proportional to I_P, the programmed antifuse would have different R_{AF} by adjusting I_P. For the state '1', the programming current I_{P1} is relatively small, thus the antifuse resistance R_1 is relatively large. For the state '3', the programming current I_{P3} is relatively large, thus the antifuse resistance R_3 is relatively small. The state '2' is between the states '1' and '3'. Overall, $I_{P1} < I_{P2} < I_{P3}$ leads to $R_1 > R_2 > R_3$.

[0033] Referring now to FIGS. **5A-5C**, three preferred 3D-OTP cells 1 *aa* are shown. In the preferred embodiment of FIG. **5A**, the bottom electrode (word line) **20***a* comprises a metallic or highly doped semiconductor material. The top electrode (bit line) **30***a* comprises a metallic or highly doped semiconductor material. The antifuse layer **22** is a layer of insulating dielectric (e.g. silicon oxide, silicon nitride). The quasi-conductive layer **24** is used to form a diode **14**. For a semiconductor diode1 **4**, the bottom electrode **20***a* comprises a P+ semiconductor material, the quasi-conductive layer **24** comprises an N− semiconductor material, while the top electrode **30***a* comprises an N+ semiconductor material. Alternatively, the bottom electrode **20***a* comprises a metallic material, the quasi-conductive layer **24** comprises a P+/N−/ N+ diode, while the top electrode **30***a* comprises another

3

metallic material. For a Schottky diode **14**, the bottom electrode **30***a* comprises a metallic material, the quasi-conductive layer **24** comprises an N– semiconductor material, while the top electrode **30***a* comprises an N+ semiconductor material. For a ceramic diode **14**, the bottom electrode **30***a* comprises a metallic material, the quasi-conductive layer **24** comprises a ceramic material (e.g. a layer of metal oxide), while the top electrode **30***a* comprises another metallic material.

[0034] The preferred embodiment of FIG. **5B** is similar to that of FIG. **5A**, except that a conductive layer **26** separates the antifuse layer **22** from the quasi-conductive layer **24**. The conductive layer **26** preferably comprises at least a metallic material, which minimizes heat damage to the quasi-conductive layer **24** during programming. The preferred embodiment of FIG. **5C** is even simpler than those of FIGS. **5A-5B** in that it does not comprise a separate quasi-conductive layer. After the antifuse layer **22** is ruptured, a diode is naturally formed at the junction of the top electrode **30***a* and the bottom electrode **20***a*. As an example, the bottom electrode **20***a* comprises a highly doped P+ semiconductor material and the top electrode **20***a* comprises a highly doped N+ semiconductor material. For those skilled in the art, the OTP cell **1***aa* could take other forms.

[0035] To minimize read error due to leaky OTP cells, the present invention discloses a full-read mode. For the full-read mode, all OTP cells on a selected word line are read out during a read cycle T. FIGS. **6A-6B** discloses a preferred OTP array OA in the full-read mode. The OTP array OA comprises word lines **20***a*-**20***z*, bit lines **30***a*-**30***z* and OTP cells **1** *aa*-**1** *zz* (the numbers in the parenthesis represent the states of the OTP cells) (FIG. **6A**). The preferred embodiment further comprises a single-ended amplifier **58S**. The read cycle T includes two read phases: a pre-charge phase t_{pre} and a read-out phase t_R (FIG. **6B**). During the pre-charge phase t_{pre}, all address lines **20***a*-**20***z*, **30***a*-**30***z* in the OTP array OA are charged to an input bias voltage V_i of the amplifier **58S**.

[0036] During the read-out phase t_R, all bit lines **30***a*-**30***z* are floating. Based on the row address **52A**, the row decoder **52** raises the voltage on a selected word line **20***a* to the read voltage V_R, while voltage on unselected word lines **20***b*-**20***z* remains at the input bias voltage V_i. After this, the selected word line **20***a* starts to charge the bit lines **30***a*-**30***z* through the OTP cells **1** *aa*-**1** *az* and the voltages on the bit lines **30***a*-**30***z* begin to rise. At this time, the voltage on each bit line is sent to the amplifier **58S** by rotating the column address **54A**. For each column address **54A**, the column decoder **54** selects a bit line (e.g. **30***b*) and sends its voltage V_b to the input **51** of the amplifier **58S**. When the value of the voltage V_b exceeds the threshold voltage V_T of the amplifier **58S**, the output **55** is toggled. By measuring the toggling time, the state of each OTP cell (e.g. the OTP cell **1***ab* at the intersection of the selected word line **20***a* and the selected bit line **30***b*) can be determined.

[0037] During the above measurement, because the V_T of the amplifier **58S** is relatively small (~0.1V or smaller), the voltage changes delta(V) on the bit lines **30***a*-**30***z* are small. The largest voltage change $delta(V)_{max} \approx N*V_T$ is far less than the read voltage V_R. As long as the I-V characteristics of the OTP cell satisfies $I(V_R) >> I(-N*V_T)$, the 3D-OTP$_{MB}$ would work properly even with leaky OTP cells.

[0038] To minimize read error due to external interferences, the present invention further discloses differential

amplifiers for measuring the states of the OTP cells. FIGS. **7A-C** disclose a first preferred OTP array OA with differential amplifiers. In addition to the regular bit line **30***a*-**30***z* and OTP cells **1** *aa*-**1** *zz* (FIG. **6A**), the preferred OTP array OA further comprises dummy bit lines **31** *a*-**31** *f* and dummy OTP cells **1** *a*0-**1** *z*5 (FIG. **7A**). For clarity, the regular bit line **30***a*-**30***z* are referred to as data bit lines, while the regular OTP cells **1***aa*-**1***zz* are referred to as data OTP cells. The dummy bit lines **31***a*-**31***f* form a dummy bit-line set **30DY**, while the data bit lines **30***a*-**30***z* form a data bit-line set **30DT**.

[0039] This preferred embodiment further comprises N–1 (in this case, =3) differential amplifiers **58***a*-**58***c* (FIG. **7A**). All differential amplifiers **58***a*-**58***c* comprise two inputs: first inputs are the bit-line voltage V_b from a data OTP cell (i.e. the OTP cell stores data), while second inputs are the reference voltages $V_{ref,1}$–$V_{ref,3}$. For example, the reference voltage for the differential amplifier **58***a* is $V_{ref,1} \neq (V_{\cdot 0} + V_{\cdot 1})/2$; the reference voltage for the differential amplifier **58***b* is $V_{ref,2} \neq (V_{\cdot 1} + V_{\cdot 2})/2$; and, the reference voltage for the differential amplifier **58***c* is $V_{ref,3} \sim (V_{\cdot 2} + V_{\cdot 3})/2$. As used herein, $V_{\cdot i}$ (i=0, 1, 2, or 3) is the voltage on the associated bit line when reading out a state 'I' OTP cell.

[0040] To generate these reference voltages $V_{ref,1}$-$V_{ref,3}$, the OTP array OA uses 2N-2 (in this case, =6) dummy bit lines **31***a*-**31***f*. Each word line (e.g. **20***a*) is associated with 2N-2 (in this case, =6) dummy OTP cells (e.g. **1***a*0-**1***a*5). Like the data OTP cells **1** *aa*-**1** *az*, the dummy OTP cells **1***a*0-**1***a*5 have N states. For example, the dummy OTP cells **1** *aa*0-**1** *a*5 on the word line **20***a* are in the states '0', '1', '1', '2', '2', '3', '3', respectively (FIG. **7A**). The reference voltage $V_{ref,1}$ on the second input **53***a* of the amplifier **58***a* is generated by shorting the dummy bit lines **31** *a* (coupled with a '0' dummy OTP cell **1** *a*0) and **31***b* (coupled with a '1' dummy OTP cell **1***a*1). Namely, $V_{ref,1} = (V_{31a} - V_{31b})2 = (V_{\cdot 0} + V_{\cdot 1})/2$(FIG. **7B**). The reference voltage $V_{ref,2}$ on the second input **53***b* of the amplifier **58***b* is generated by shorting the dummy bit lines **31***c* (coupled with a '1' dummy OTP cell **1** *a*2) and **31***d* (coupled with a '2' dummy OTP cell **1** *a*3). Namely, $V_{ref,2} = (V_{31c} + V_{31d})/2 = (V'1' + V'2')/2$. The reference voltage $V_{ref,3}$ on the second input **53***c* of the amplifier **58***c* is generated by shorting the dummy bit lines **31***e* (coupled with a '2' dummy OTP cell **1** *a*4) and **31***f* (coupled with a '3' dummy OTP cell **1** *a*5). Namely, $V_{ref,3} = (V_{31e} + V_{31f})/2 = (V'2' + V'3')/2$.

[0041] To determine the state of a selected data OTP cell, N–1 measurements are taken concurrently at the N–1 amplifiers **58***a*-**58***c*. The data OTP cell is in the state 'k' if $V_{ref,k-1} < V_b < V_{ref,k}$(k=1, 2, . . . N–1). For example, to measure the state of the data OTP cell **1** ab, the column decoder **54** sends the voltage on the bit line **30***b* to the first inputs of all amplifiers **58***a*-**58***c*. The amplifiers **58***a*-**58***c* make three measurements concurrently (FIG. **7C**). As the data OTP cell **1** ab is in the state '2', the bit-line voltage V_b on the first input **51** is larger than the reference voltages $V_{ref,1}$, $V_{ref,2}$ on the second inputs **53***a*, **53***b* of the amplifiers **58***a*, **58***b*, respectively. However, V_b is smaller than $V_{ref,3}$ on the second input **53***c* of the amplifier **58***c*. Thus, the outputs **55***a*-**55***c* of the amplifiers **58***a*-**58***c* are high, high, low. Accordingly, the state of the data OTP cell **1** *ab* can be determined.

[0042] FIGS. **8A-8B** discloses a second preferred OTP array **0A** comprising an amplifier. Instead of three amplifiers **58***a*-**58***c* (FIG. **7A**), it uses a single amplifier **58D** (FIG. **8A**).

The first input of the amplifier **58**D is the bit-line voltage V_b of a selected data OTP cell, while the second input is a selected reference voltage V_{ref}. This preferred embodiment uses fewer dummy bit lines and dummy OTP cells than FIG. **7**A. The OTP array **0**A comprises only N (in this case, =4) dummy bit lines **32**a-**32**d and each word line (e.g. **20**a) is coupled with N dummy OTP cells (e.g. **1**a0-**1**a3). Like the data OTP cells **1**aa-**1**az, the dummy OTP cells **1**a0-**1**a3 have N states. For example, the dummy OTP cells **1**aa0-**1**a3 on the word line **20**a are in the states 0 , 1 , 2 , 3 , respectively (FIG. **8**A). Adjacent dummy bit lines (e.g. **32**a, **32**b) are coupled through a pair of pass transistors (e.g. **56**a1, **56**a2), which is controlled by a control signal **56**a. An appropriate reference voltage Vref is generated by asserting the corresponding control signals. For example, by asserting the control signal **56**a, the passing transistors **56**a1, **56**a2 is turned on and the reference voltage Vref is equal to the average of the voltages on the dummy bit lines **32**a, **32**b. Namely, $V_{ref} = V_{32a} + V_{32b} = (V_{\cdot 0} + V_{\cdot 1\cdot})/2$.

[0043]   To determine the state of a selected data OTP cell, N−1 measurements are taken sequentially at the amplifier **58**D (FIG. **8**B). The OTP cell is in the state 'k' if $V_{ref,k-1} < V_b < V_{ref,k}$ (k=1, 2, . . .N−1). For example, to determine the state of the data OTP cell **1**ab, the column decoder **54** sends the voltage $V_b (= V_{\cdot 2\cdot})$ on the bit line **30**b to the first input **51** of the amplifier **58**D. Three sequential measurements are taken. During the first measurement T_1, the control signal **56**a is asserted to turn on the pass transistors **56**a1, **56**a2. The reference voltage Vref on the second input **53** of the amplifier **58**D is $V_{ref} = (V_{\cdot 0} + V_{\cdot 1\cdot})/2$. As $V_{ref} < V_b$, the output **55** of the amplifier **58**D is high. Similarly, during the second measurement T_2, by turning on the pass transistors **56**b1, **56**b2, the output **55** of the amplifier **58**D stays high as $V_{ref} = (V_{\cdot 1} + V_{\cdot 2\cdot})/2 < V_b$; during the third measurement T_3, by turning on the pass transistors **56**c1, **56**c2, the output **55** of the amplifier **58**D turns low as $V_{ref} = (V_{\cdot 2\cdot} + v_{\cdot 3\cdot})/2 > V_b$. By analyzing the results from three measurements T_1-T_3, the state of the data OTP cell **1**ab can be determined.

[0044]   In the preferred embodiments of FIGS. **6**A-**8**B, the row decoder **52**, the column decoder **54** and the amplifier **58**S, **58**a-**58**c, **58**D are formed on the substrate **0** and are a portion of the substrate circuit OK. The OTP array **0**A is stacked above the substrate circuit OK and covers at least a portion thereof. The 3D-OTP$_{MB}$ has a small die size and a low die cost.

[0045]   All dummy OTP cells need to be pre-programmed before shipping. During pre-programming, the resistances of the dummy OTP cells need to be adjusted precisely. For the preferred embodiment of FIG. **7**A (or, FIG. **8**A), each word line **20**a is associated with 2N−2 (or, N) dummy OTP cells. The total number of the dummy OTP cells in the OTP array OA (with M data word lines and N data bit lines) is (2N−2) *M (or, N*M). Pre-programming all of them is time consuming and costly. To reduce the pre-programming time, the present invention further discloses a dummy word line. FIG. **9** discloses a preferred OTP array OA comprising a dummy word line **20**D. In addition to the data word lines **20**a, **20**b, . . . **20**y, **20**z, the preferred OTP array OA comprises a dummy word line **20**D. The pre-programming is only carried out to the dummy OTP cells **1**D1-**1**D5 located at the intersections of the dummy word line **20**D and the dummy bit lines **31**b-**31**f. Because majority of the dummy OTP cells (e.g. **1**a0-**1**b5, **1**y0-**1**z5) do not need pre-programming, the pre-programming time is significantly reduced.

[0046]   During read, both voltages on the selected data word line (e.g. **20**a) and the dummy word line **20**D are raised to V_R. Because the dummy OTP cells **1**Da-**1**Dz at the intersections of the dummy word line **20**D and the data bit lines **30**a-**30**z are un-programmed, the voltage rise on the dummy word line **20**D would not affect the signals on the data bit lines **30**a-**30**z. Moreover, because the dummy OTP cells **1**a0-**1**a5 at the intersections of the data word line **20**a and the dummy bit lines **31**a-**31**f are unprogrammed, the voltage rise on the data word line **20**a would not affect the signals on the dummy bit lines **31**a-**31**f, either. Accordingly, the operation of this preferred embodiment is similar to those in FIGS. **7**A-**8**B.

[0047]   In the preferred embodiments of FIGS. **7**A-**9**, to manufacture high-quality dummy OTP cells, all dummy word line, dummy bit lines and dummy OTP cells are preferably formed in the middle of the OTP array **0**A. For example, the dummy word line **20**D is formed in the middle of the OTP array **0**A, so are the dummy bit lines **31**a-**31**f.

[0048]   Although examples disclosed in these figures are horizontal 3D-OTP (i.e. the OTP memory levels **100**, **200** are horizontal), the inventive spirit can be extended to vertical 3D-OTP (i.e. the OTP memory strings are vertical to the substrate). More details of the vertical 3D-OTP are disclosed in Chinese Patent Application No. 201610234999. 5, filed on Apr. 16, 2017.

[0049]   While illustrative embodiments have been shown and described, it would be apparent to those skilled in the art that many more modifications than that have been mentioned above are possible without departing from the inventive concepts set forth therein. For example, beside N=4 (i.e. each OTP cell stores two bits), the present invention can be extended to N=8 (i.e. each OTP cell stores three bits) or more. The invention, therefore, is not to be limited except in the spirit of the appended claims.

What is claimed is:

1. A multi-bit-per-cell three-dimensional read-only memory (3D-OTP$_{MB}$), comprising:

a semiconductor substrate including transistors thereon;

a plurality of OTP cells stacked above said semiconductor substrate, each of said OTP cells comprising an antifuse layer, where said antifuse layer is irreversibly switched from a high-resistance state to a low-resistance state during programming;

a plurality of contact vias coupling said OTP cells to said semiconductor substrate;

wherein said OTP cells have more than two states, the OTP cell in different states being programmed by different programming currents.

2. The 3D-OTP$_{MB}$ according to claim **1**, further comprising:

an OTP array comprising said plurality of OTP cells including a data OTP cell, a plurality of word lines including a data word line, and a plurality of bit lines including a data bit line;

an amplifier coupled with said data OTP cell through said data bit line.

3. The 3D-OTP$_{MB}$ according to claim **2**, wherein all OTP cells coupled with said data word line in said OTP array are read out during a read cycle.

4. The 3D-OTP$_{MB}$ according to claim **3**, wherein said word lines and said bit lines are charged to an input bias voltage of said amplifier during a pre-charge phase of said read cycle.

5. The 3D-OTP$_{MB}$ according to claim **3**, wherein a read voltage is applied to said data word line during a read phase of said read cycle.

6. The 3D-OTP$_{MB}$ according to claim **2**, wherein said OTP array further comprises:

a first dummy bit line associated with a first unprogrammed dummy OTP cell;

a second dummy bit line associated with a second programmed dummy OTP cell;

a first differential amplifier comprising a first input, wherein said first input is coupled with said first and second dummy bit lines.

7. The 3D-OTP$_{MB}$ according to claim **6**, wherein said OTP array further comprises:

a third dummy bit line associated with a third programmed dummy OTP cell, said second and third dummy OTP cells having a same state;

a fourth dummy bit line associated with a fourth programmed dummy OTP cell, said second and fourth dummy OTP cells having different states;

a second differential amplifier comprising a second input, wherein said second input is coupled with said third and fourth dummy bit lines.

8. The 3D-OTP$_{MB}$ according to claim **2**, wherein said OTP array further comprises:

a first dummy bit line associated with a first unprogrammed dummy OTP cell;

a second dummy bit line associated with a second programmed dummy OTP cell;

a third dummy bit line associated with a third programmed dummy OTP cell, said second and third dummy OTP cells having different states;

a differential amplifier comprising an input, wherein said input is coupled with said first and second dummy bit lines during a first measurement, and said input is coupled with said second and third dummy bit lines during a second measurement.

9. The 3D-OTP$_{MB}$ according to claim **2**, wherein said OTP array further comprises:

a dummy word line in parallel with said data word line;

a dummy bit line in parallel with said data bit line;

a first dummy OTP cell formed at the intersection of said dummy word line said dummy bit line, wherein said first OTP dummy cell is programmed;

a second dummy OTP cell formed at the intersection of said data word line and said dummy bit line, wherein said second dummy cell is unprogrammed.

10. The 3D-OTP$_{MB}$ according to claim **1**, further comprising a quasi-conductive layer, wherein the resistance of said antifuse layer is larger than the read resistance of said quasi-conductive layer.

11. A multi-bit-per-cell three-dimensional read-only memory (3D-OTP$_{MB}$), comprising:

a semiconductor substrate including transistors thereon;

a plurality of OTP cells stacked above said semiconductor substrate each of said OTP cells comprising an antifuse layer, where said antifuse layer is irreversibly switched from a high-resistance state to a low-resistance state during programming;

a plurality of contact vias coupling said OTP cells to said semiconductor substrate;

wherein the resistance of said antifuse layer is determined by a programming current, said OTP cells being programmed by at least two programming currents.

12. The 3D-OTP$_{MB}$ according to claim **11**, further comprising:

an OTP array comprising said plurality of OTP cells including a data OTP cell, a plurality of word lines including a data word line, and a plurality of bit lines including a data bit line;

an amplifier coupled with said data OTP cell through said data bit line.

13. The 3D-OTP$_{MB}$ according to claim **12**, wherein all OTP cells coupled with said data word line in said OTP array are read out during a read cycle.

14. The 3D-OTP$_{MB}$ according to claim **13**, wherein said word lines and said bit lines are charged to an input bias voltage of said amplifier during a pre-charge phase of said read cycle.

15. The 3D-OTP$_{MB}$ according to claim **13**, wherein a read voltage is applied to said data word line during a read phase of said read cycle.

16. The 3D-OTP$_{MB}$ according to claim **12**, wherein said OTP array further comprises:

a first dummy bit line associated with a first unprogrammed dummy OTP cell;

a second dummy bit line associated with a second programmed dummy OTP cell;

a first differential amplifier comprising a first input, wherein said first input is coupled with said first and second dummy bit lines.

17. The 3D-OTP$_{MB}$ according to claim **16**, wherein said OTP array further comprises:

a third dummy bit line associated with a third programmed dummy OTP cell, said second and third dummy cells having a same state;

a fourth dummy bit line associated with a fourth programmed dummy OTP cell, said second and fourth dummy cells having different states;

a second differential amplifier comprising a second input, wherein said second input is coupled with said third and fourth dummy bit lines.

18. The 3D-OTP$_{MB}$ according to claim **12**, wherein said OTP array further comprises:

a first dummy bit line associated with a first unprogrammed dummy OTP cell;

a second dummy bit line associated with a second programmed dummy OTP cell;

a third dummy bit line associated with a third programmed dummy OTP cell, said second and third dummy OTP cells having different states;

a differential amplifier comprising an input, wherein said input is coupled with said first and second dummy bit lines during a first measurement, and said input is coupled with said second and third dummy bit lines during a second measurement.

19. The 3D-OTP$_{MB}$ according to claim **2**, wherein said OTP array further comprises:

a dummy word line in parallel with said data word line;

a dummy bit line in parallel with said data bit line;

a first dummy OTP cell formed at the intersection of said dummy word line said dummy bit line, said first dummy OTP cell is programmed;

a second dummy OTP cell formed at the intersection of said data word line and said dummy bit line, said second dummy OTP cell is unprogrammed.

20. The 3D-OTP$_{MB}$ according to claim **1**, further comprising a quasi-conductive layer, wherein the resistance of said antifuse layer is larger than the read resistance of said quasi-conductive layer.

* * * * *