Vertical stack array of one-time programmable nonvolatile memory based on pn-junction diode and its operation scheme for faster access

Seongjae Cho¹, Sunghun Jung², Sungjun Kim², and Byung-Gook Park²a)

¹ Department of Electronic Engineering, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do 461–701, Republic of Korea
² Inter-university Semiconductor Research Center (ISRC) and Department of Electrical and Computer Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151–744, Republic of Korea
a) bgpark@smu.ac.kr

Abstract: In this work, a three-dimensional (3-D) architecture of one-time programmable (OTP) nonvolatile memory (NVM) arrays is introduced and its viable process integration and operation method are schemed. Vertical stack architecture is highly pursued for higher-level integration and NVMs based on devices free from transistors and charge trapping layers would be one of the candidates. In this work, in an effort for the NVM technology trend, architecture, fabrication process, and operation scheme for faster data access are studied in depth. Silicon (Si) pn-junction diode is focused by its virtues of cost-effectiveness, process maturity, and compatibility to peripheral Si CMOS circuits.

Keywords: vertical stack, one-time programmable (OTP) memory, nonvolatile memory (NVM), pn-junction diode, three-dimensional (3-D) architecture, metal-insulator-semiconductor (MIS)

Classification: Integrated circuits

References

1 Introduction

In the history of electronic engineering, one-time programmable (OTP) memories were mainly used for read-only-memory (ROM) applications where the central processing unit (CPU) instructions were stored to operate a whole computer system [1, 2, 3]. By grafting a novel structuring with the concept of OTP memory and the nanoscale silicon (Si) processes, innovative high-density nonvolatile memory (NVM) applications can be implemented. Although there might be various kinds of OTP NVM devices and array architectures, a novel technology based on simple materials and fabrication was previously proposed [4, 5]. In this work, it is evolved into three-dimensional (3-D) architecture for achieving very-large-scale integration (VLSI) and accompanying process integration and operation scheme for faster access to the stored data are developed. The program operation is performed by breaking down the thin oxide layer between vertical Si channel and bitline (BL). Thus, the program state can be simply determined by whether the read current through the oxide layer is high or low (programmed or not). In spite of the high-level scalability of charge
trap flash (CTF) NVM, reliability issues in terms of retention and endurance are remaining to be resolved [6, 7]. They are usually originated from the quality of CT layer and it is hard to completely eliminate them as long as the NVM device is operated by the charge-based mechanisms. In this regard, it is understood that the diode-based OTP NVM is one of the plausible candidates that are to a large degree free from the reliability issues of CTF NVMs. Furthermore, it is more advantageous for higher integration and simpler processes compared with the metal-oxide-semiconductor field-effect transistor (MOSFET) based NVMs. Recently, for even higher-level integration of NVMs, design and process developments for stack arrays have been intensively researched [8, 9, 10]. In this paper, along with the introduction of pn-junction diode OTP device, viable process integration and operating schemes for the diode-based 3-D OTP NVM with full Si CMOS compatibility are developed.

2 Device structure and program operation physics

Figure 1 shows a schematic cross-sectional view of the diode-based OTP NVM device of metal-insulator-semiconductor (MIS) vertical pillar structure.

![Fig. 1. Schematic view of the proposed OTP NVM device [5].](image)

Although silicon (Si) has been chosen in this study as the channel material for full Si complementary metal-oxide-semiconductor (CMOS) compatibility, other possible options such as Schottky barrier or resistive switching materials can be applied [11, 12, 13]. The semiconductor regions of each type can be formed through ion implantation or low-pressure chemical vapor deposition (LPCVD) with in situ doping. The thin oxide before a program operation, fuse oxide (FO), can be thermally grown or deposited. The common signal line with n+ doping connected to the n-type pillar bottoms acts as the deposited wordline (DWL). In order to perform a program operation, a positive program voltage \( V_{PRG} \) is applied to a BL. The high electric field between BL and p-type pillar breaks down the FO and the program/erase (P/E) states of a cell are determined by its leakage current level conducting through the FO: program operation breaks down the FO in a specific cell, which increases the leakage current prominently and this state can be identified as State 0. At this point, after the breakdown by a program operation, the thin oxide no longer can be called...
as an FO but needs to be termed as an anti-fuse oxide (AFO). On the other hand, the erase state at which the FO is still conserved can be read as State 1 (numbering each state can be reversed by cases). The FO is broken down by the electron-hole pairs generated near the FO-Si pillar interface and the positive feedback in the electron tunneling due to the holes trapped in the FO [14, 15, 16]. Figs. 2 (a) and 2 (b) demonstrate the electrostatic potential and electric fields across a cell in the vertical direction under forward bias conditions [5], which was performed by device simulations [17].

![Image](image_url)

**Fig. 2.** Simulation analyses on (a) electrostatic potential and (b) electric field as a function of location from the BL top (Oxide thickness = 30 Å, Si width/height = 20 nm/100 nm, and $N_{WL}=10^{21} \text{cm}^{-3}$).

### 3 A viable process integration

The stacked OTP NVM array is made to have two layers that can be integrated laterally or vertically sharing a common platform. Since the memory cells on the lower and upper layers (it is assumed that the two layers are integrated in a vertical manner) should have the same electrical characteristics, it would be desirable to construct the Si pillars by deposition of polycrystalline Si (poly-Si). However, the poly-Si deposition needs to be performed on a thick bottom oxide for electrical isolation from...
the Si substrate since it would be complicated to carry out anisotropic etching processes for BLs and DWLs crossing perpendicularly one another by a top-down fabrication method. Fig. 3 (a) through 3 (j) suggests the fabrication processes for the stacked OTP NVM array. Isolation bottom oxide (IBO) is deposited for the isolation between the cell arrays and the Si substrate (Fig. 3 (a)). Since the IBO should be thick enough for effective electrical isolation, plasma-enhanced CVD (PECVD) or high-density plasma CVD (HDPCVD) process with a high deposition rate will be required [18]. p⁺ poly-Si is deposited for the lower-layer bitlines (LBLs) and the thin FO is thermally grown or deposited with low enough deposition rate (Fig. 3 (b)). LPCVD with in situ doping would be the most efficient way to form the LBLs. p- and n-type regions for the vertical diode can be deposited and doped by an LPCVD with in situ doping as shown in Fig. 3 (c). The pn-junction diode is constructed by simple controls over the doping gases and switching time. Patterning and anisotropic etch processes in the direction A down to the IBO follow as shown in Fig. 3 (d). Then, patterning and etch will be performed along the direction B which is perpendicular to direction A. The anisotropic etch should be continued until the p- and n-type Si layers on the FO are completely removed (Fig. 3 (e)). Subsequently, isolation between BLs and chemical mechanical polishing (CMP) are conducted, which reveals the n-type ends of diode structures as shown in Fig. 3 (f). Following processes are the same with those for the single-layer array [5]. Poly-Si with in situ n⁺ doping is deposited for the DWL. p- and n-type regions for the reversed vertical diode are formed by LPCVD with in situ doping in sequence (Fig. 3 (g)). Patterning and anisotropic etch processes in the direction B down to the isolation oxide on the first layer and in the direction down to the DWL are performed continuously. Isolation oxide deposition and CMP are conducted subsequently as shown in Fig. 3 (h). The second FO layer and p⁺ poly-Si for the upper-layer bitlines (UBLs) are deposited (Fig. 3 (i)). Patterning and dry etch in the direction of A will be following and p- and n-type Si layers are completely removed until the p-type ends of the vertical diodes are exposed (Fig. 3 (j)). Stacked two-layer array having pn-junction diode channels in vertically reverse-aligned positions is finally constructed through the suggested process integration. This strategy can be further extended to achieve 3-D memory architecture with more number of stacks employing not only Si pn-junction diodes but also other types of rectifying components and resistive switching materials [11].

4 Operation methods for the stacked OTP NVM array

Figure 4 demonstrates a simplified array of the 3-D OTP NVM. The basic operation schemes are based on those already confirmed in the previous works [5]. It is revealed that a program operation over the whole page at once is not possible. In Fig. 4, only the program operation over the upper layer is concerned and the cells with rigid boxes are to be programmed. “P” and “I” on the cells indicate that the intended operations are “possible” and “impossible”, respectively. For example, the cell with P notation and without a rigid box is possible to remain program-inhibited while it is impossible to program one with I identification and with a rigid box. Other possibilities can be checked in the same manner. In order to program
Fig. 3. Fabrication processes for the stacked array of OTP NVM based on pn-diode pillars. From (h) to (j), the upper isolation oxide layer is made to be transparent to look inside the array more clearly.
CellU3W1 and CellU2W2 (where CellUiWj denotes the cell on the cross point of UBLi and WLj), WL1 and WL2 should be grounded and program voltages ($V_{PRG}$'s) should be applied to UBL2 and UBL3. Although it is originally intended that CellU3W2 should be program-inhibited, it is programmed inevitably under these operation conditions. Other cells with “I” notations shown in Fig. 4 will undergo either unwanted program or program-inhibition operations. In this regard, the program operation cannot be performed by page but only by line in sequence. Fig. 5 (a) shows the WL-by-WL sequential row program operation assuming that only the upper layer is considered. If the rigid-boxed cells are to be programmed, the operations are performed from the top DWL, WL1, down to the bottom DWL, WL4. The UBLs are assigned with serial signals for the program operations onto the specific cells: 

$0110 \rightarrow 1100 \rightarrow 0011 \rightarrow 0101$. Signals 1 and 0 on a BL indicate $V_{PRG}$ and pass voltage ($V_{PASS}$), respectively. Since the corresponding WL is grounded on each step, $(V_{BL}, V_{WL}) = (V_{PRG}, GND)$ programs a cell (signal 1) and $(V_{BL}, V_{WL}) = (V_{PASS}, GND)$ inhibits the program operation on a cell (signal 0). Here, $V_{PASS}$ should be an intermediate positive voltage between $V_{PRG}$ and 0 V so that both $|V_{PRG} - V_{PASS}|/T_{FO}$ and $|V_{PASS}|/T_{FO}$ are smaller than the breakdown electric field ($T_{FO}$: fuse oxide thickness). The other WLs not participating in the program operation over a specific WL are applied by $V_{CC}$ which is also a positive bias to screen the electric field induced by $V_{PRG}$ over the cell to be program-inhibited, which is applied to the management of $V_{PASS}$. Fig. 5 (b) tabulates the bias conditions for the program operations on the upper layer with the signal sequencing. If only the upper layer is to be programmed intentionally as mentioned above, all the LBLs should be fed up with $V_{PASS}$ to avoid unwanted program operations regardless of $V_{WL}$’s. However, the cells on the lower level can be programmed simultaneously; it
is a genuine advantage of the stack OPT NVM array that the cells on the double layers having a WL in common can be programmed without any time loss.

Fig. 6 (a) shows the cells on both layers to be programmed. The rigid and dotted boxes indicate the cells to be programmed on the upper and the lower layers, respectively. The program operations can be conducted by supplying voltage signals in sequence, \([0011/0001] \rightarrow [0100/1000] \rightarrow [1000/1001] \rightarrow [1001/0110]\). The pair of 4-digit numbers in a bracket indicates the program signals from the UBLs and LBLs, respectively, and the signal set

Fig. 5. Program operation by row sequence. (a) Cell to be programmed on the upper-layer array. (b) Program operation schemes tabulated by UBL signals with pass-signaled LBLs.
in the bracket is fed at one time so that there is no redundant time consumption even though the number of the cells is doubled. Fig. 6 (b) summarizes the bias conditions for the double-layer program keeping the total number of operations unchanged at all. Read operations can be also carried out without any additional time consumption compared with the case of single-layered array. The memory states of the two cells under cross-points made by a specific WL and both upper and lower BLs in the same position (LBLj and UBLk where j = k) can be sensed simultaneously as shown in Fig. 7.

Fig. 6. Simultaneous program operation by row sequence on stacked arrays. (a) Cells to be programmed on the double-layer arrays. (b) Program operation schemes tabulated by U/LBL signals.
5 Conclusions

In this work, we have demonstrated the viable process integration and operation methods for the stacked novel OTP NVM array. The fabrication process is very simple and fully compatible with the conventional CMOS processing. Also, it was confirmed that the proposed operation schemes on the double-layered stack could be performed without any loss in time even though the integration density was doubled by making simultaneous operations on both layers. With the aids by the genuine features of the unit cell as well as its array stackability, the proposed OTP NVM array and accompanying operation schemes would be highly practical in the 3-D OTP and other 3-D cross-bar-array-based NVM technologies.

Acknowledgments

This work was supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as Global Frontier Project (CISS-2012M3A6A6054186). Also, it was in part supported by the IT R&D program of the Korean Ministry of Knowledge Economy (MKE/KEIT, Project No. 10035320), “Development of Novel 3-D Stacked Devices and Core Materials for the Next Generation Flash Memory”.

Fig. 7. Cross-sectional view of the stacked array illustrating the simultaneous read operations.