Memory Evolution: Multi-Functioning Unified-Random Access Memory (URAM)

Yang-Kyu Choi and Jin-Woo Han

School of EECS, KAIST, 373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
E-mail: ykchoi@ee.kaist.ac.kr, Phone: +82-42-869-3477, Fax: +82-42-869-8565

Abstract

A new paradigm for silicon memory technology is proposed. A technological breakthrough that will overcome the saturation in revenue obtained from ‘scaling’, a novel type of fusion memory is presented. A high-speed DRAM and non-volatile flash memory are integrated in a single memory transistor. The memory cell is named Unified-Random Access Memory (URAM), as multi-functional operation is processed in a single memory cell. The paradigm shift from ‘scaling’ to ‘multi-function’ will create new value and continue the evolution of silicon memory technology.

1. Introduction

For over the past three decades, ‘scaling’ has been an important growth factor in semiconductor technology. In silicon memory technology, scaling has allowed an increase of memory density and a decrease of cost per bit. However, on the basis of traditional scaling, silicon memory is approaching physical and technical limits. For example, the critical dimension of the state-of-the-art SONOS memory has approached the fundamental limit in currently developed charge storage materials [1]. This implies that the revenue from scaling will decrease as scaling slows. Therefore, an entirely new concept is required in order for silicon memory technology to remain competitive. According to this stringent requirement, a number of studies have reported emerging memory technologies, but none can viably replace the Si-based memory framework in the near future. In addition, a chip-based fusion memory package that contains various functional memory blocks has been reported. For example, DRAM, SRAM, and flash memory functions can be operated in a single chip. However, because these memory chips involve system or package level fusion, hybrid integration of these functional memories is obstructed by factors related to performance, process, and cost.

In this paper, a novel fusion memory concept is proposed. Unlike the traditional chip-based fusion memory, the single memory transistor reported here can operate high-speed DRAM and non-volatile flash memory function. By adopting a SONOS and a floating body structure in a FinFET, a charge trapped memory for non-volatile function and a floating body memory for high-speed operation are realized. The floating body is formed on a SOI substrate as well as a bulk wafer by adopting a buried n-well or a buried $Si_{1-y}C_y$ substrate.

2. Traditional Silicon Based Memory to Emerging Memory

An ideal memory device should satisfy three requirements: high speed, high density, and non-volatility. Unfortunately, a memory that can satisfy all these requirements has yet to be developed. Consequently, memory devices have been advanced by pursuing one among these requirements. Fig. 1 illustrates a representative memory and the aforementioned key domains. During the evolution of memory devices, ‘scaling’ has played a core role in the exponential growth of the memory industry. Memories have reached 2Gb density with a 50nm node for DRAM and 32Gb density with a 40nm node for NAND flash. As silicon technology enters the deep nanoscale dimensions, physical and technical limits are expected to be encountered in the near future. As such, growth in silicon memory technology advanced by scaling cannot be sustained indefinitely. Thus, a technological breakthrough is necessary in order to continue memory evolution.

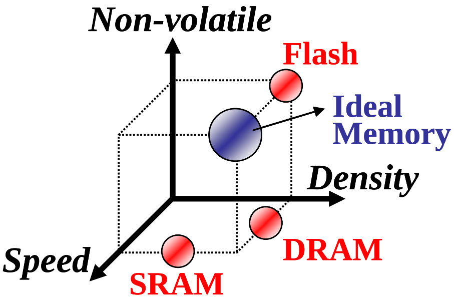

Figure 1. A schematic view of the domains of various memory devices. Ideal memory should satisfy non-volatility, high speed, and high density.

Numerous exotic solutions have been investigated for beyond silicon based memory technology. These alternative memories include ferroelectric RAM (FRAM), magnetic RAM (MRAM), phase-change RAM (PRAM), polymer RAM (PoRAM), resistive-switching (ReRAM), and mechanical memory. Even though each memory type presents unique advantages, all have weaknesses in terms of replacing the current silicon based memory in the near future. The major difficulties of these memories lie in material stability, variability, and integration with a Si platform. Even though the emerging memories are fantastic, it would not easy to assure their reliability characteristics.

978-1-4244-2186-2/08/$25.00 ©2008 IEEE

3. Fusion Memory

Fusion memory has been developed to meet the demands of multi-functions and high-performance digital applications. A fusion memory chip is combines various types of memory such as DRAM and flash memory. The first type of fusion memory is physical combination of chips, which are combined by a multi-chip package (MCP) or system-in-package (SIP). However, in MCP/SIP, the time delay caused by wiring can degrade the data transmission speed. This implies that effective communication between the CPU and memory will be difficult in a multimedia chip. In a notable breakthough, two or more kinds of memory chips have been combined on one chip, which is implemented by system-on-chip (SOC) technology. However, the hybrid integration of these functional memories is obstructed by process and cost. Furthermore, the fundamental limitation of aforementioned fusion memories is that the portion of the different memory blocks is fixed, and hence users cannot adjust the strength of each memory's density and speed. But if a single memory transistor can process different memory functions, users can optimize specifications of the memory to fit customers' demands. A prototype of a unified-RAM (URAM) that provides this capacity is proposed in this work.

4. Unified-RAM (URAM)

The concept of URAM is illustrated in **Fig. 2**. In URAM technology, a single memory transistor can process non-volatile memory and high speed capacitorless 1T-DRAM. These operations are identified by the bias conditions of V_g and V_d. By implementation of O/N/O as an electron trapped zone for non-volatile memory and a floating body as a hole storage zone for capacitorless 1T-DRAM, URAM is realized. For the first prototype of URAM, a FinFET SONOS structure on a SOI substrate was presented in [2], and the same functional structure fabricated on a bulk substrate was subsequently reported in [3].

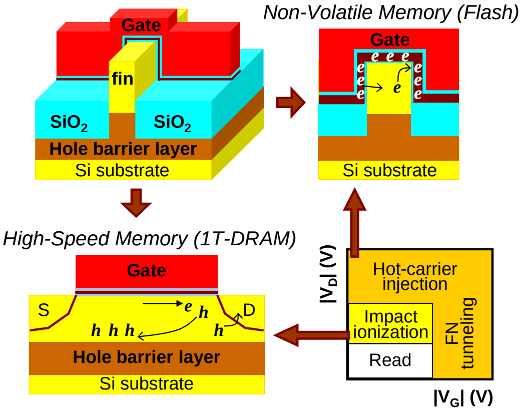

Figure 2. Schematic view of URAM concept. SONOS memory and 1T-DRAM operation are implemented in a memory cell.

4-1. Device fabrication process

Because the device processed on the SOI substrate inherently includes a floating body, the operation of 1T-DRAM is easily achieved. However, the SOI substrate can suffer from heat dissipation and cost problem. In order to form a floating body on a bulk substrate, an energy-band engineered substrate formed by a buried n-well or hetero-epitaxially grown $Si/S_{i-y}C_y$ was used. If the buried n-well is formed by high energy n-type implantation carried out on the p-type substrate, the built-in potential of the PN junction makes a floating surface channel [3]. Also, hetero-epitaxially grown $Si/S_{i-y}C_y$ builds a band offset in the valence band. The energy band lineup of SOI and a buried $Si_{1-y}C_y$ substrate are similar except that the hole energy barrier of SOI is higher than that of the buried $Si_{1-y}C_y$ substrate. This implies that a buried $Si_{1-y}C_y$ substrate can also be implemented for floating body applications.

After preparation of the substrates, the subsequent process sequence of URAM is identical to that of a conventional FinFET SONOS, as presented in [2] for SOI and in [3] for a bulk substrate. After Si_3N_4 deposition for fin hard mask, a fin is patterned. High density plasma (HDP) SiO_2 is deposited, planarized by CMP, and partially recessed by HF until the Si fin is exposed. As a gate dielectric, O/N/O is stacked sequentially and n+ *in-situ* poly-Si is deposited and patterned. Finally, S/D implantation and an activation process are conducted.

The fabricated URAM is shown in **Fig. 3**. All types of URAM are based on a FinFET SONOS structure.

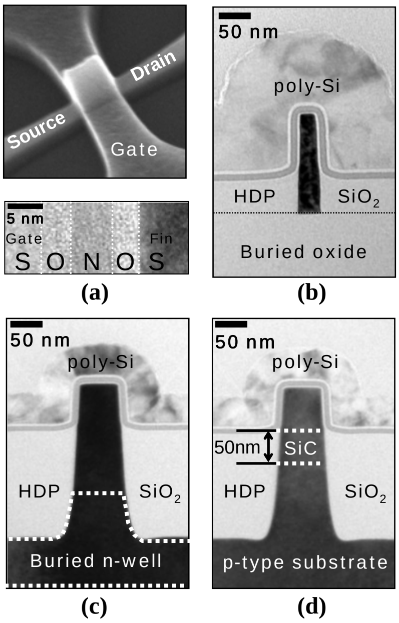

Figure 3. TEM images of URAM. (a) Bird eye's veiw of URAM and gate dielectric O/N/O. (b) URAM on SOI substrate. (c) URAM on buried n-well substrate. (d) URAM on buried $Si_{1-y}C_y$ substrate.

4-2. Memory Characteristics

A. Non-Volatile Memory Characteristics

Data states in non-volatile memory operation are distinguished by the existence of trapped charges in the nitride layer of O/N/O. Program/erase can be carried out by Fowler-Nordheim (FN) tunneling or channel hot-electron injection (CHE), and the range of the program/erase voltage is relatively high. The I_D-V_G characteristics of non-volatile memory operation at the SOI substrate are shown in **Fig. 4**. Despite the thick O/N/O stack, superior short-channel characteristics are obtained due to a narrow fin effect. In program/erase, a 80 μsec pulse is required to obtain a 3V threshold voltage window with 11V programming (P) and -11V erasing (E). Excellent endurance and reliability characteristics are also obtained. The non-volatile memory characteristics in the case of other substrates can be found in [3].

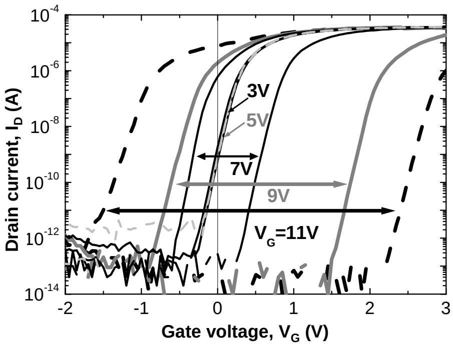

Figure 4. Measured programming and erasing characteristics of I_d-V_g. P/E is carried out by a Fowler-Nordheim mechanism.

B. 1T-DRAM Characteristics

Floating body MOSFETs can play a role of a capacitorless 1T-DRAM. Data states of 1T-DRAM are stored in the partially depleted zone of the floating-body. The generated holes are isolated by the valence band barrier in the buried layer: BOX, n-well, or SiC. In a bulk platform, a PN built-in potential formed by a buried n-well or valence-band-offset formed by a hetero-epitaxially grown layer can confine charges in the channel. In programming, holes created by impact ionization are accumulated in the floating body and thereafter the channel potential is lowered. In erasing, accumulated holes are eliminated toward the drain by the negative drain voltage, and the channel potential is raised again. Therefore, the data state is identified by the presence of excessive holes in the floating body. Current sensing at the source then allows the detection of the data states. **Fig. 5** shows the energy band lineup for devices on various substrates. The position of the buried n-well should be carefully set in order to avoid a junction short to the n+ source/drain.

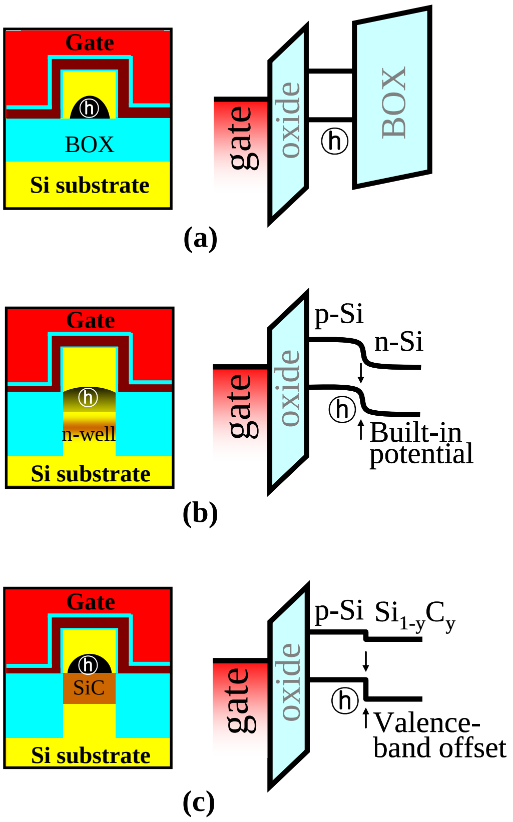

Figure 5. Cross-sectional schematics and energy band lineup for various substrates. (a) SOI substrate, (b) buried n-well, and (c) buried $Si_{1-y}C_y$.

Most holes at different energy distributions cannot surmount the buried oxide barrier. However, a fraction of holes at the high energy tail can escape over the barrier. Therefore, it is important to evaluate the minimum barrier height for effective holes storage. According to the simulation results, an energy barrier of 0.1eV results in the loss of only ~20% holes and thus retains ~80%. A built-in potential of ~0.6 eV between the p-channel/n-well is obtained by calculation with the doping profile. In a Si/$Si_{1-y}C_y$ system, the valence band offset can be evaluated via a photo-luminescent experiment [4]. However, examination of the p-type capacitance-voltage characteristics provides a simpler approach [5]. Comparing and fitting the simulation data to the measured curve, the energy barrier was estimated as 0.1 eV. An advantage of a bulk over a SOI substrate is that a positive substrate (back gate) voltage can effectively increase the barrier height. Therefore, even though the intrinsic barrier height of the bulk is lower than that of SOI, the back gate voltage enhances the charge storage performance. But back gate voltage is restricted below ~0.6V, because the source/drain to body (PN junction) diode can turn on if the back gate voltage is higher than ~0.6V.

Fig. 6 shows the program/erase characteristics for 1T-DRAM. For programming, drain voltages for impact ionization were V_D=1.5V for SOI and V_D=2V for bulk. Since a fraction of charges surmounts the barrier height in bulk whereas most holes are stored in SOI, programming voltage for the bulk is higher than that for the SOI.

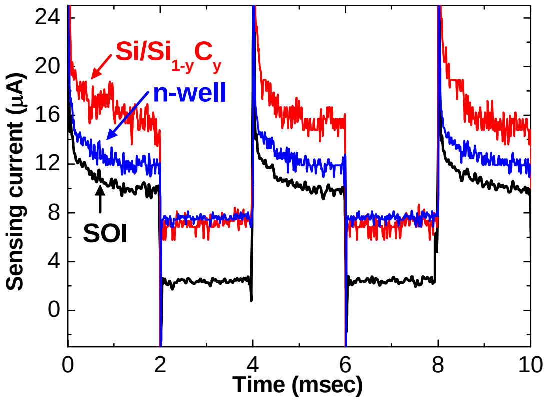

Figure 6. Measured programming and erasing characteristics of 1T-DRAM. For programming, impact ionization voltages of V_D=1.5Vand V_G=1V for SOI, and V_D=2Vand V_G=1V for bulk are used. For erasing, forward junction voltage of V_D=-1V for all devices is used. During all operations, the substrate voltages are V_{Sub}=0V for SOI, and V_{Sub}=0.3V for n-well and $Si_{1-y}C_y$.

In the case of SOI substrate, a 8μA sensing current window is achieved. However, for the bulk substrate, in order to obtain a larger sensing window, the substrate voltage should be set to 0.3V. This implies that even though the intrinsic 1T-DRAM performance of SOI is superior to that of bulk, the performance of bulk with the aid of higher programming voltage and positive substrate voltage can be comparable to that of SOI.

The operational voltage between the FN tunneling regime for the non-volatile memory and impact ionization regime for 1T-DRAM should be clarified. Unfortunately, hot electrons generated by impact ionization during programming of the 1T-DRAM can charge the nitride trap, which can cause undesired soft programming for non-volatile memory. Therefore, the programming voltage for the 1T-DRAM is higher than the impact ionization voltage. This parameter should be sufficiently small so as to avoid soft programming in charge trapping non-voltage memory. In the erase condition, the negative drain voltage does not disturb the charge trapped state in the non-volatile memory. The absence of interference between the two operational modes is verified in **Fig. 7**. A threshold voltage shift is not observed before and after 10^4 sec of 1T-DRAM operation. This means that charge trapping is negligible during 1T-DRAM operation if the operation voltage domains are carefully set. This verifies distinctive operation for both memory modes.

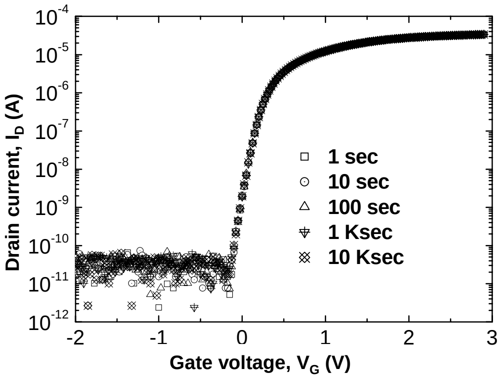

Figure 7. I_D-V_D curves after 1T-DRAM operation. The programming disturbance between impact ionization for 1T-DRAM and charge trapping for non-volatile memory is found to be negligible.

Conclusions

A Unified-RAM (URAM) for multi-functioning of non-volatile memory and 1T-DRAM is demonstrated with a band-offset technology. By combining an O/N/O gate dielectric and a floating body into a FinFET structure, charge-trapping memory and capacitorless 1T-DRAM functions are realized in a single transistor. This fusion allows users to adjust the memory capacity and speed; they can optimize specifications of memory to fit their demands. With multi-functioning URAM technology, memory vendors can dynamically and flexibly respond to the shifting demands of customers. This paradigm shift from a multi-bit cell to a multi-function cell can sustain Moore's law and continue the evolution of silicon technology.

Acknowledgments

This work was supported by the National Research Program for the 0.1-Terabit Nonvolatile Memory Development, sponsored by the Ministry of Knowledge Economy. The authors wish to thank all staff of the National Nanofab Center for their unwavering support of device fabrication.

References

[1] H. Lee et. al., *VLSI*, p.114 (2007).
[2] J.-W. Han et. al., *IEDM*, p.929 (2007).
[3] J.-W. Han et. al., *VLSI*, (2008).
[4] K. Brunner et. al., *APL*, 2, p.303 (1996).
[5] K. Rim et. al., *APL*, 18, p.32286 (1998).