Ultimately Scaled 20nm Unified-RAM

Dong-II Moon, Sung-Jin Choi, Chung-Jin Kim, Jee-Yeon Kim, Jin-Seong Lee, Jae-Sub Oh*, Gi-Sung Lee*
Yun-Chang Park*, Dae-Won Hong*, Dong-Wook Lee*, Young-Su Kim*, Jeoung-Woo Kim*, Jin-Woo Han, and Yang-Kyu Choi

Department of EE, KAIST, *National Nanofab Center, Daejeon 305-701, Korea
Email: ykchoi@ee.kaist.ac.kr, Phone: +82-42-350-5477, Fax: +82-42-350-8565

Abstract
A 20 nm all-around-gate (AAG) FET is demonstrated for unified-RAM by the use of an 8 nm Si-nanowire on a bulk-substrate. High performances of NVM and 1T-DRAM are presented.

Introduction
Integration of various types of memories in one cell has become important with respect to device level fusion. Unified-RAM (URAM), which combines non-volatile memory (NVM) and capacitor-less 1T-DRAM in a single transistor, was proposed for multi-functional memory applications [1]. In this work, a gate length of 20 nm with an 8 nm diameter Si-nanowire (SiNW), which is fully surrounded by the gate, is demonstrated. A suspended SiNW, which is fully depleted (FD), is fabricated onto the bulk-Si substrate by employment of deep reactive-ion-etching (RIE). The concept of the proposed URAM, which is utilized by implementation of an O/N/O layer as an electron trapped zone for NVM and a SiNW as a hole storage zone for 1T-DRAM, is shown in Fig. 1. For 1T-DRAM programming, the parasitic BJT method (Gen2) [2] is used to minimize the cross-talk between the 1T-DRAM and the NVM, as shown in Fig. 2. The distinctive functions of NVM and 1T-DRAM are demonstrated, and there is no mutual disturbance between operation modes in a single transistor. Additionally, the BJT method enables 1T-DRAM operation without a partially depleted (PD) body, which inhibits ultimate scaling of 1T-DRAM.

Device Fabrication
The process flow of the AAG URAM is illustrated in Fig. 4. A photolithography process with the 0.18 μm technology with a photo-resist ashing and an oxide hard mask trimming process was used to realize small size features. One of the notable processes in the overall process is patterning of the SiNW and isolation of adjacent FETs. An 8 nm SiNW was formed by deep RIE, otherwise known as the Bosch process [3], as shown in Figs. 3 and 5. In a previous study, SiNW was separated from the bulk-Si by employment of a partial isotropic etching and a subsequent sacrificial oxidation process [4]. In this work, however, the SiNW was completely separated from the bulk-Si by the Bosch process, and sacrificial oxidation was only applied for further reduction of the SiNW size. Therefore, reproducibility could be improved and the profile of the SiNW can be accurately controlled.

Device-to-device isolation was realized by shallow trench isolation (STI) oxide deposition and subsequent chemical-mechanical polishing (CMP). The remaining thickness of the STI oxide is approximately 70 nm. Afterwards, an O/N/O layer of 3 nm/6 nm/8 nm, serving as a charge trapping layer in the NVM mode, and an n+ in-situ doped poly-Si layer were sequentially stacked. An AAG with a gate length (LG) of 20 nm was then patterned, as shown in Fig. 5. After gate patterning, oxide spacers were formed. The S/D was doped with arsenic, and the dopant was activated by a rapid thermal annealing (RTA) process at 1000 °C for 3 sec. LG is in a range from 20 to 150 nm; the width of the SiNW (WNW) is scaled down to 8 nm, as shown in Fig. 6.

Results and Discussion
The current was normalized by the perimeter of the SiNW with consideration of its shape. Measured transfer characteristics (ID, VG) are shown in Fig. 7. The threshold voltage (VT) roll-off, drain-induced barrier lowering, and subthreshold swing (SS) for various diameters of the SiNW (WNW) are summarized in Fig. 8. Although the device characteristics are clearly improved upon reduction of the WNW dimension, short-channel effects were not sufficiently suppressed due to the thick gate dielectric (EOT = 14 nm). However, the URAM functioned well.

NVM characteristics
NVM functionality is shown with programming/erasing (P/E) transient behaviors in Figs. 9 and 10. P/E operation is performed with FN tunneling. A VT window of 3 V is achieved with a 100 μsec pulse at programming bias (VPGM) of 10 V and erasing bias (VER) of -10 V. Fast erasing speed, in particular, is attributed to the concentrated electric-field at angular corners. A VT shift (ΔVT), i.e., programming window, is also investigated for various LG and WNW, as shown in Fig. 11. The amount of ΔVT increases with smaller WNW, which arises from the intensified vertical electric-field [5].
**IT-DRAM characteristics**

The single transistor latch (STL) characteristics triggered by the parasitic BJT method are depicted in Fig. 12. At low $V_D$, the device shows normal transfer behaviors. If the value of $V_D$ is further increased, $SS$ becomes very steep, i.e., $SS < 10 \text{ mV/dec.}$ $I_D$ was abruptly increased by the activation of the parasitic BJT in the floating SiNW; its unique hysteresis loop between the forward and reverse scan was observed via a positive feedback mechanism. The minimum $V_D$ to trigger the STL, namely latch voltage ($V_{latch}$), was analyzed for different device parameters in order to investigate the relationship between the device scaling and the STL conditions and to establish proper operational conditions in the BJT-based IT-DRAM, as shown in Fig. 13. It should be noted that a shortened $L_G$ reduces $V_{latch}$ by virtue of enhanced gain of the parasitic BJT. In contrast, narrowed $W_{NW}$ acts adversely on the decrement of $V_{latch}$. A numerical simulation was performed to verify the effects on the STL of $W_{NW}$ reduction, as shown in Fig. 14. The simulation results support the idea that an aggressively scaled $W_{NW}$ is not desirable for low operational voltage, because the narrowed width exacerbates the non-local effect [6-7]. Fig. 15 shows the typical BJT-based IT-DRAM characteristics under the operational timing diagram with a 5 nsec programming and erasing pulse. The binary states are clearly distinguished in a scaled device with a 20 nm $L_G$ and an 8 nm $W_{NW}$. The available $V_D$ bias range for reliable memory operation was measured and the results are shown in Fig. 16. If the value of $V_D$ is less than $V_{latch}$ the parasitic BJT is not activated. On the other hand, if $V_D$ is too high, the current sensing margin of the BJT-based IT-DRAM can approach zero due to high leakage current. The current sensing margin is increased up to 190 $\mu$A/μm at $V_D = 3.2 \text{ V}$, which is the best result on a bulk-substrate reported to date. Fig. 17 shows the read retention characteristics. The read states are sustained without loss of data over a few seconds. One of the important advantages of the BJT-based IT-DRAM is non-destructive readout. The conventional read method destroys data during the read operation; thus, a refresh step is periodically required, which degrades the operation speed and causes increased power consumption. However, the BJT read method automatically charges the body during the read operation, which greatly improves the operation speed. Fig. 18 shows the cumulative curve of each data state in terms of the sensing window. The sensing current window is high enough that a sense amplifier is not necessary to identify the data states; this can remarkably simplify the periphery circuit architecture. To evaluate the data hold characteristics, an optimal bias was exploited and the results are plotted in Fig. 19. After the hold period, each memory state is successfully read out. Moreover, multiple read operations are allowed without loss of the data state, as shown in Fig. 20. Once the parasitic BJT is triggered and holes are charged in the floating SiNW, pre-stored data are sustained in a stand-by (hold) state by $V_D = -2.5 \text{ V}$ and $V_D = 0 \text{ V}$. Previously reported IT-DRAM characteristics were compared with those in this work and the results are summarized in Table 1.

**Conclusions**

An AAG URAM cell with a 20 nm $L_G$ and 8 nm $W_{NW}$, constituting the smallest IT-DRAM realized thus far, has been demonstrated. Moreover, this device showed the highest sensing margin among devices implemented to date on bulk-substrates. Finally, URAM operation was presented with NVM characteristics. URAM with low cost and a CMOS compatible process is a promising structure for next generation fusion memory and high performance SoC applications.

**Acknowledgement**

The authors would like to thank Dr. Hee Chul Lee for his managerial support from NNFC. This research was partly supported by Nano R&D program through the NRF of Korea funded by the MEST (2009-0082583) and the IT R&D program of MKE/KEIT (10029953, Terabit Nonvolatile Memory Development).

**References**

Fig. 1 Schematic of the SiNW AAG URAM operation. A 1T-DRAM is enabled by stored holes in a floating SiNW and a NVM is activated by trapped electrons in an O/N/O layer.

Fig. 2 Bias map for the BJT-based 1T-DRAM (Gen2) and the NVM implemented on URAM.

Fig. 3 (a) The formation of the suspended SiNW by the Bosch process. (b) A tilted SEM image after patterning of the SiNW.

Fig. 4 Process flow of the AAG URAM on a bulk substrate. After SiNW formation, boron ions with a dose of $10^{14}$/cm$^2$ were implanted on a bulk substrate to eliminate the leakage path. A sacrificial oxide was grown and removed to alleviate etching damage. 1000 °C spike annealing was performed.

Fig. 5 (a) A TEM image along the x-x’ direction of Fig. 3 (b). The distance of the suspended SiNW from the bulk substrate is approximately 150 nm. (b) In-line CD-SEM image after poly-Si gate patterning. (c) A bird’s eye view of the fabricated device after spacer formation.

Fig. 6 (a) A TEM image across the a-a’ of Fig. 5 (c). $L_G$ is 20 nm and the length of the spacer is 17 nm. (b) Cross-sectional TEM image of the SiNW inside the poly-Si gate along the b-b’ direction of Fig. 5(c). An AAG structure was fabricated on a bulk substrate. (c) Close-up image of Fig. 6(b). $W_{NW}$ is defined as the averaged width of the SiNW. (d) A TEM image of O/N/O (3/6/8nm) layers for flash memory application.

Fig. 7 Transfer characteristics of an 8 nm $W_{NW}$ device. The current is normalized by the perimeter of the SiNW.

Fig. 8 Short-channel effects of the fabricated devices with various $L_G$ and $W_{NW}$. (a) $V_T$ roll-off. (b) Subthreshold swing (SS). (c) Drain induced barrier lowering (DIBL).

Fig. 9 Program transient characteristics under various bias conditions.
Hold characteristic of 1T-DRAM.

The STIL is deactivated below 2.7V and the highest sensing margin reported to date. Failure of read ‘0’ is mainly caused by leakage current.

**Table 1.** Comparison table between the proposed 1T-DRAM and previously announced 1T-DRAM. This work shows the smallest device size and the highest sensing margin reported to date.