Extended Abstracts of the 2010 International Conference on Solid State Devices and Materials, Tokyo, 2010, pp748-749

E-3-1(Invited)

Current Development Status and Future Challenges of Charge-Trapping NAND Flash

Hang-Ting Lue, Kuang-Yeu Hsieh, and Chih-Yuan Lu

Emerging Central Lab., Macronix International Co., Ltd., 16 Li-Hsin Road, Hsinchu Science Park, Hsinchu, Taiwan (e-mail: htlue@mxic.com.tw)

Abstract- Although conventional floating gate (FG) Flash memory has already gone into the 2Xnm node, the technology challenges are formidable beyond 20nm. The fundamental challenges include FG-FG interference, few-electron storage caused statistical fluctuation, poor short-channel effect, WL-WL breakdown, poor reliability, and edge effect sensitivity. Although charge-trapping (CT) devices have been proposed very early and studied for many years, these devices have not prevailed over FG Flash in the planar 2D NAND. However, beyond 20nm the advantage of CT devices may become more significant. Especially, due to the simpler structure and no need for charge storage isolation, CT is much more desirable than FG in 3D stackable Flash memory. Optimistically, 3D CT Flash memory may allow the density increase to continue for at least another decade. In this paper, we review the operation principles of CT devices and several variations such as MANOS and BE-SONOS. We will then discuss various 3D memory architectures. Technology challenges and the poly-silicon thin film transistor (TFT) issues will be addressed in detail.

I. Introduction

The scaling limitation of FG NAND and its potential replacement have become the focus of research in recent years [1,2]. Among the many emerging memory technologies, charge-trapping (CT) device is the most mature with high volume production of the physically 2b/c NOR Flash using local channel hot electron injection [3]. Meanwhile, a similarly structured devices using uniform FN programming have become promising solutions to continue NAND Flash scaling. **Figure 1** briefly compares the FG and CT NAND. CT NAND has several advantages: (1) planar structure and simple geometry that allow better scaling capability. (2) immunity to FG-FG interference, and (3) immunity to SILC due to discrete traps.

In this work, the current development status and future challenges of CT NAND are discussed.

II. Review of Gate Stack Engineering of CT NAND Device

Figure 2 illustrates several CT NAND devices [4]. The conventional SONOS and MONOS NAND afford no memory window. Two extensively studied improvements are briefly described below:

(1) Using high-K top dielectric and a higher work function metal gate (TANOS) [5]: Since the electrons are deeply trapped in the SiN, very high electric field (> 15 MV/cm) is required to de-trap them at reasonable erasing speed. The combination of high-K top dielectric and metal gate helps reduce the gate injection during erasing [6] and thus the use of high erase field. The high field stressing, however, degrades the tunnel oxide reliability, and thus further innovation is still needed to achieve fast erase and good data retention.

(2) Using barrier engineer (BE) tunneling barrier (BE-SONOS) [7]: Substrate hole injection was originally proposed to erase SONOS but this requires very thin tunnel oxide, which cannot stop hole injection from the substrate during retention. Barrier engineering provides a "variable thickness" tunneling barrier that both allow hole injection at high field and good data retention in low field during retention. P^+-poly gate is used to reduce gate injection during erase operation, but P^+-poly cannot completely eliminate gate injection and erase saturation may occur at high field and longer erasing time.

It is possible to combine the merits of these two approaches into a BE-MANOS device [6]. Since high-K is not a good insulator, to improve the reliability a buffer oxide [8] between high-K and SiN is introduced to improve the retention.

III. (2D bulk) BE-SONOS NAND Flash Development

Figure 3 shows the dumb-mode programmed state Vt distribution of CT BE-SONOS NAND. The distribution of BE-SONOS NAND with only one programming pulse without any verification and subsequent shots is already tight, thanks to the near-planar and much simpler geometry that minimize the process variation. This gives superior programming throughput because fewer program and verify (PV) sequences are needed.

Figure 4 illustrates the typical MLC Vt distribution of 75nm NAND BE-SONOS NAND test chip. With optimized interface engineering of Si/O1 we can achieve high endurance of P/E >100K for SLC and > 3K for MLC within one block [9] without any help of error correction or system-level help. **Figure 5** shows the intrinsic retention of the BE-SONOS NAND test chip under a low P/E cycling stress. We prove that with an optimized SiN trapping layer the CT device can achieve excellent retention without any issue of charge lateral spreading, contrary to the common misperception. In fact, electrons are deeply trapped in SiN (with a good SiN, not detuned) and unlike SONOS NOR Flash for which a local internal dipole field from CHE electrons and BBHH holes exists at the junction edge [3], CT NAND has no lateral electric field.

We have successfully scaled the BE-SONOS CT NAND toward 38nm half pitch [9]. In summary, the BE-SONOS CT NAND has excellent reliability that can surpass the consumer applications' requirement.

Moreover, we have not observed any tail bit during retention baking or read disturb stress, and this is a fundamental advantage over the SILC issue of FG. On the other hand, CT NAND without metal-gate/high-K has a high erased-verified Vt (EV) and slower erase speed due to the erase saturation. Also, the total self-boosting disturb-free window is limited by the smaller ISPP slope (<1) so that TLC (3b/c) is more challenging.

Further scaling of CT NAND becomes more critical and faces the same fundamental physical limit of FG NAND. In **Fig. 6**, the stored electrons would be only a few tens at 25nm node. This not only threatens the retention, but also the programming statistics. Poisson statistics ($Sigma \propto \sqrt{N}$) [10] naturally come into play such that the program distribution becomes broader, as shown in **Figs. 7 and 8**. This limits the ultimate programming accuracy and impacts the multi-level operation.

In our understanding CT NAND probably has only limited capability to extend the current 2D NAND. Few-electron storage issue and the large fringing field effect at sub-20nm node do not spare CT NAND.

IV. Outlook of 3D CT NAND Flash

3D NAND Flash provides a very promising path to continue the NAND Flash [11-12]. The most important breakthrough is the introduction of the bit-cost scalable (BiCS) [11] approach that uses only one critical process (a drilled hole) to define many memory layers simultaneously, offering a possibility of ultra low cost memory. CT devices are naturally best suited for 3D because there is no need to isolate nitride in 3D structures thus it provides a low-cost processing approach.

Figure 9 briefly compares various 3D architectures [12]. Most 3D NAND have lateral scaling limitation at F>50nm and larger cell size of $>6F^2$. The larger cell size is difficult to compete with 20nm FG NAND. Moreover, when more memory layers are stacked, the processing cost inevitably increases and the array efficiency is generally lost such that the bit cost saturates after more than 30 memory layers.

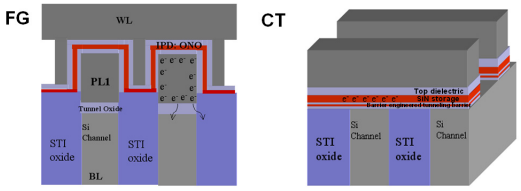
In **Fig. 10**, vertical gate (VG) shows the best horizontal scaling capability and no read current degradation when adding more layers. We consider VG as the most promising path to scale below 30nm cell size [12-13]. **Figure 11** illustrates the device characteristics of a 75nm half-pitch, 8-layer memory stack of 3D VG TFT BE-SONOS NAND [13]. It shows very successful device performances and further encourages the future 3D NAND Flash development.

One important issue of 3D NAND is that it is inevitable to use the TFT device, which raises uniformity issues. One interesting thing is that as we scale TFT device to sub-30nm node [14], it generally behaves excellently and approaches the bulk device's performance. However, some tail bits often happen mainly because of the grain boundary. Fortunately, with suitable P/E algorithm a tight Vt distribution can still be managed.

In summary, we suggest that reducing the cell size and improving the array efficiency by optimized decoding method are key factors to reduce the bit cost for 3D memory. Further more, significant repair methods and design efforts are also required to solve the yield loss and correct the device-level issues. With optimized CT device reliability is guaranteed.

References:

[1] K. Kim, et al, IEDM 2007, pp.27-30. [2] R. Liu, VLSI 2010 short course. [3] B. Eitan, et al, IEEE EDL, 2000, pp. 543-545. [4] H. T. Lue, IMW 2010 short course. [5] C. H. Lee, et al, IEDM 2003, 26.5.1-26.5.4. [6] S. C. Lai. et al, NVSMW 2007, pp. 88-89. [7] H. T. Lue, et al, IEDM 2005, pp. 555-558. [8] S. C. Lai. et al, NVSMW 2008, pp. 101-102. [9] C. C. Hsieh, et al, IEDM 2010, in submission. [10] H. T. Lue, et al, IMW 2010, pp. 92-95. [11] H. Tanaka, etl al, VLSI 2007, pp. 14-15. [12] Y. H. Hsiao, et al, IMW 2010, pp. 142-145. [13] H. T. Lue, et al, VLSI 2010, pp.131-132. [14] T. H. Hsu, et al, IEDM 2009, pp. 629-632.



1. Gap filling difficulty and complicated topology → 1. Near-Planar structure
2. FG-FG interference and disturbs → 2. Discrete traps, no FG
3. Few-electron retention and sensitivity to oxide defect (SILC) → 3. Deep traps (High E_A). Immune to point defect in tunnel oxide

Fig. 1 Comparison of FG and CT NAND. CT NAND has the advantages of simpler planar structure, deep traps that are immune to SILC and no FG coupling interference issues.

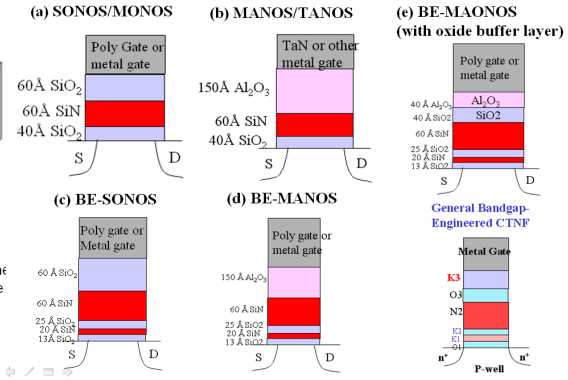


Fig. 2 Schematic diagram of several important CT NAND devices with barrier engineering (BE) and high-K/metal gate [4].

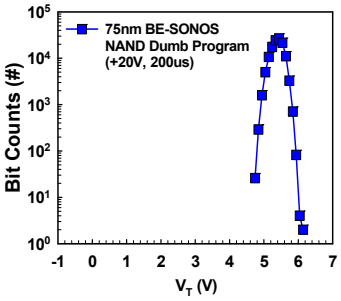


Fig. 3 Dumb program (without verify and one shot program only) Vt distribution of 75nm BE-SONOS. BE-SONOS has narrower Vt distribution due to its simple geometry.

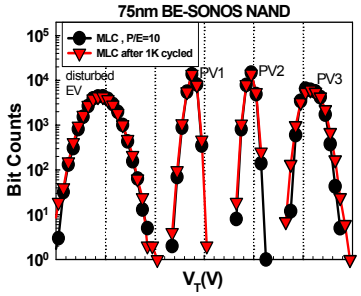


Fig. 4 Typical MLC Vt distribution of BE-SONOS NAND Flash test chip. Gap between each state is >200mV within 1 block without any ECC help.

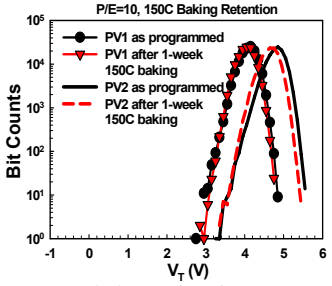


Fig. 5 Intrinsic retention of BE-SONOS at P/E=10. It shows perfect retention at Vt=4V. Higher Vt shows slightly larger charge loss because of the larger built-in retention field that induces larger vertical charge loss.

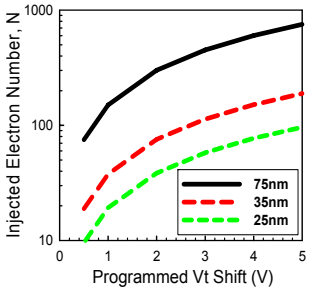


Fig. 6 Stored electron number for CT NAND for various technology nodes and programmed Vt.

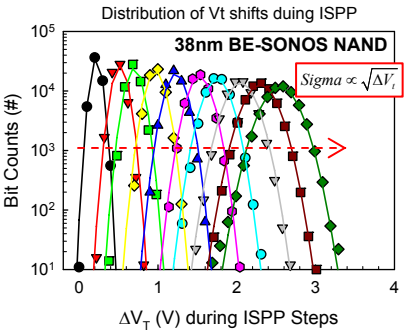


Fig. 7 Distribution of delta Vt shift during ISPP for the 38nm BE-SONOS NAND. Higher programmed Vt shows larger standard deviation.

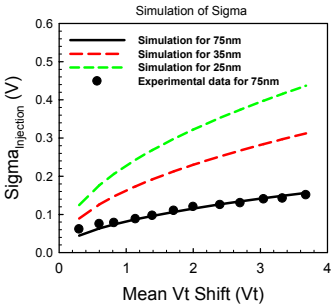


Fig. 8 Model of standard deviation (Sigma) of programming statistics for various technology nodes. It well follows the Poisson statistics [10].

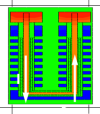
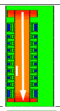
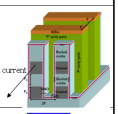
	P-BiCS	TCAT	VSAT	VG
String				
Cell Shape				
Cell Size in X, Y direction	$6F^2$ (3F*2F)	$6F^2$ (3F*2F)	$6F^2$(3F*2F)	$4F^2$ (2F*2F)
Gate Process	Gate first	Gate last	Gate First	Gate Last
Current Flow direction	U-turn	Vertical	Multi-U-turn	Horizontal
Device Structure	GAA	GAA	Planar	Double Gate
Possible minimal F	~50 nm	~50 nm	~50 nm	~2X nm

Fig. 9 A brief comparison of various 3D NAND Flash architectures. 3D Vertical gate (VG) NAND may be the best scalable device in horizontal pitch [12].

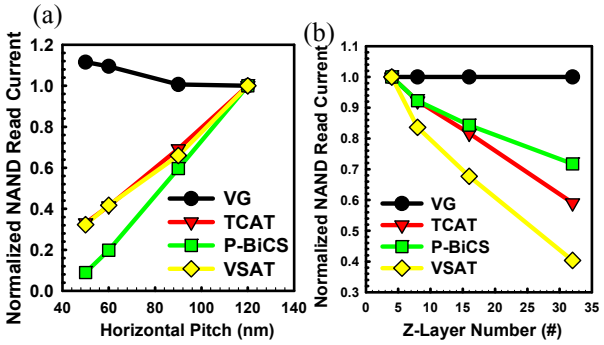
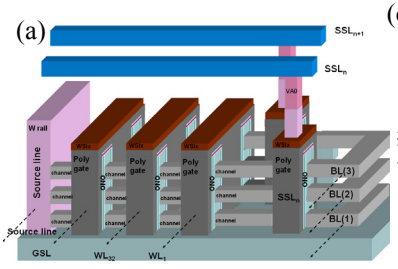
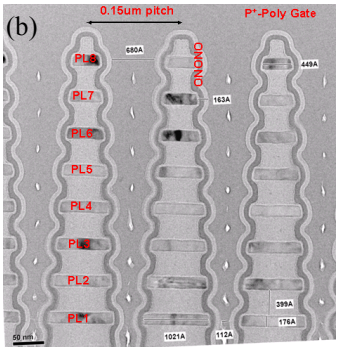


Fig. 10 **(a)** Comparison of read current for various horizontal pitch (2F). **(b)** Comparison of read current of various Z layer number. VG NAND shows the best lateral scaling and no degradation when more memory layers are added.





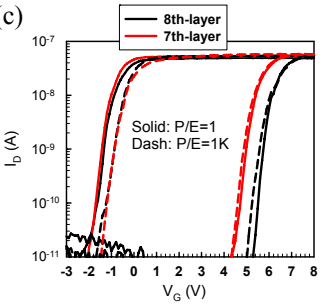
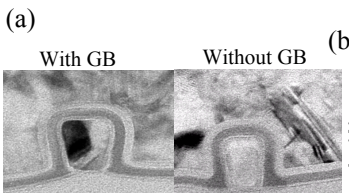
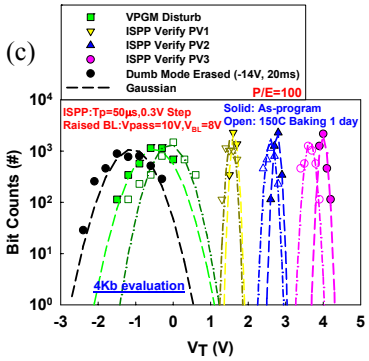


Fig. 11 **(a)** Proposed 3D VG NAND architecture and decoding method. **(b)** The 8-layer VG TFT BE-SONOS structure [13] shown in the X direction. The deep trench with a high aspect ratio is successfully fabricated. Both ONONO and poly gate gap fill-in are successful. **(c)** The corresponding IV characteristics during P/E cycling. It shows well-behaved performances.





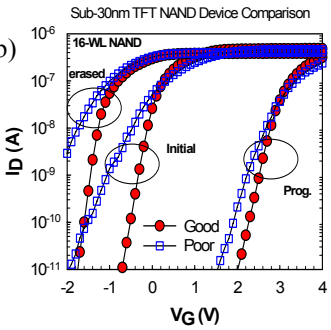


Fig. 12 **(a)** The TEM cross-sectional view of the sub-30nm TFT BE-SONOS device with/without grain boundary (GB) [14]. **(b)** The *IV* curve of the device with GB has worse IV characteristic such as larger S.S.. On the other hand, the device without GB has excellent *IV* characteristics that are close to the bulk device. **(c)** Even with tail distribution, we can still obtain a MLC tight Vt distribution by suitable P/E algorithm.