A 128Gb (MLC)/192Gb (TLC) Single-Gate Vertical Channel (SGVC) Architecture 3D NAND using only 16 Layers with Robust Read Disturb, Long-Retention and Excellent Scaling Capability

Hang-Ting Lue, Pei-Ying Du, Wei-Chen Chen, Yung-Chun Lee, Tzu-Hsuan Hsu, Teng-Hao Yeh, Kuo-Pin Chang, Chih-Chang Hsieh, Chiatze Huang, Guan-Ru Lee, Chih-Ping Chen, Chieh-Fang Chen, Chia-Jung Chiu, Y. J. Chen$^+$, W. P. Lu$^+$, Tahone Yang$^+$, Kuang-Chao Chen$^+$, Chun-Hsiung Hung$^\#$, Keh-Chung Wang, and Chih-Yuan Lu

Macronix International Co., Ltd
Emerging Central Lab., Technology Development Center$^+$, and Design Center$^\#$
16 Li-Hsin Road, Hsinchu Science Park, Hsinchu, Taiwan. E-mail: htlue@mxic.com.tw

Abstract: We have successfully developed a 128Gb MLC (or 192Gb TLC) 3D NAND Flash using 16-layer SGVC architecture. The produced memory density is 1.6 Gb/mm^2 for MLC or 2.4 Gb/mm^2 for TLC (including CMOS peripheral area, spared BL's and blocks). Such memory density is comparable to 48-layer 3D NAND using the popular gate-all-around (GAA) structures. SGVC has the important advantage of much smaller cell size and pitch scaling capability which allows very high-density memory at much lower stacking layer number. SGVC possesses very robust read disturb immunity (>120M read) and long-retention (> 40 years at room temperature) at fresh state that can suppress the very frequent wear-leveling and refresh operations needed for other 3D NAND Flash devices and is very suitable for read-intensive memory. With further stacking/scaling, it is possible to realize low-cost 1Tb single-chip solution at merely 48 layers.

I. Introduction

An SGVC 3D NAND Flash architecture was proposed [1]. In this work, we have successfully developed a 128Gb MLC 3D NAND Flash product using 16 effective-layer SGVC. The memory features and device performances are illustrated in detail.

II. Architecture Features and Memory Design

Figure 1(a) illustrates the 16-layer SGVC architecture. It is a flat-channel thin-body poly-silicon TFT device, arranged in a U-turn NAND string architecture, where both bitline (BL) and source contacts are connected at top. The BL contact is arranged in a twisted layout to allow double-density M2 BL's to enhance the bandwidth performances [1].

For each block, there are plural string-select transistors (SSL), where every wordline (WL) shares the plural SSL (pages) in order to reduce the WL decoder area. The ground-select transistor (GSL) is also shared for the same block. The bottom inversion gate (IG) serves as the dummy WL that turn-on the bottom U-turn region.

Figure 1(b) shows that the 16 effective WL structure has actual 21 oxide-poly (OP) layers. There are 3 dummy WL's at the top to protect the edge WL0/WL31 to relieve the hot-carrier disturb during programming. The top SSL/GSL has long channel length (>0.1um) like 2D NAND to avoid punch-through leakage during self-boosting programming inhibit. The bottom IG also has pretty long-channel length (>0.2um), simply because of the process need to well control the bottom etch recess depth.

Figure 1(c) shows the staircase formation. It is arranged in a matrix-like layout to save the area. Such staircase process is formed at the WL pad which is periodically carried out per ~200um-length WL. We used the metal-strapped WL's to connect the low-resistance M3 (Cu) to the bottom high-resistance P$^+$ poly gate per ~200um in order to reduce the WL RC delay. In the product, the WL pulse waveform can be controlled within 10usec only. The short WL programming pulse not only improves the performances but also improves the program-disturb window.

There is no complex metal gate replacement process needed, contrary to VNAND [2]. Experimental data shows that we have excellent WL bridge yield with very low bad-block (BB) rate of ~1% for most chips. This is great for product yield. The layout penalty of using the metal-strapped WL is less than 5% overhead.

Figure 1(d) shows the die photo graph of the 128Gb MLC product. The dies size is 76.5mm^2, including CMOS peripheral. In the array, it includes the necessary spared BL's (~12%) for error correction code (ECC), and also some spared blocks for bad-block repair and other necessary blocks for controller operations. So far we haven't developed CMOS under Array (CuA) [3], so that the CMOS peripheral circuits are designed in a conventional way. The integration method follows our previous work [4] which makes the array first in the deep trench so that the top surface of both array and CMOS are flattened before the contact and BEOL processing.

For the peripheral design, the page buffer is designed using the 25nm half pitch metal 2 (M2) BL design rules. Many peripheral logic circuits and internal charge pumping circuits are designed to support both MLC and TLC operations.

The memory density (including the CMOS peripheral and spared BL's and blocks) is 1.6 Gb/mm^2 for MLC or 2.4Gb/mm^2 for TLC. Excluding CMOS area, the memory density is 2.2Gb/mm^2 and 3.4 Gb/mm^2 for MLC and TLC, respectively (still including spared BL's and blocks). Such memory density is already comparable to the 48-layer GAA 3D NAND devices.

Figure 2(a) and 2(b) shows the cross-sectional view of channel-length and channel-width direction, respectively. The SGVC is a flat-channel, thin-body poly-silicon TFT device. The charge-trapping device employs the bandgap engineered tunnel barrier (BE-SONOS) [5] together with certain top oxide engineering in order to optimize the device P/E memory window. Ultra-thin poly silicon channel (Tsi<=6nm) is developed to improve the TFT device performance. In the channel-width direction, a special etching technique is developed to isolate the channel poly vertically by a hole-type etching, as shown in Fig. 2(b). The device channel width edge needs certain optimized corner engineering to optimize the device window and reliability.

The advantage of SGVC device is illustrated in **Fig. 3**. The ONO and channel poly keep flat even for the non-ideal deep etching. Experimental results show that the FN tunneling keep similar behavior with the same ISPP slope for all WL's. Thus the advantage of SGVC is less critical to the etching angle that makes the processing friendly for further stacking.

III. Device Characteristics and Memory Window

Figure 4(a) shows the IdVg curves of a typical SGVC device during ISPP programming. The IdVg curves keep parallel shifted with large memory window. The average subthreshold slope (SS) is excellent (~250mV/dec) as shown in **Fig. 4(b)**. The superior SS is attributed to the ultra-thin body TFT that improves

978-1-5386-3559-9/17/$31.00 ©2017 IEEE

the short-channel effect, and it also suppresses the impact of random grain boundary traps.

Figure 5(a) shows the ISPP programming of product chip. The peak Vt reaches 6V at 24V programming bias, with average ISPP slope ~0.8. Without the field enhancement effect caused by curvature of GAA device, the ISPP programming of our flat-cell SGVC device creates >8V peak P/E window.

The onset of FN tunneling bias starts at ~14V. This indicates that at lower voltages, both Vpass (<11V) program disturb and read disturb (<7V) are negligible.

The erasing transient is shown in **Fig. 5(b)**. The erased Vt high bound can be lowered below -2V after sufficient erasing time. The overall P/E memory window is already enough to enable MLC and TLC operations.

Figure 6 shows the MLC (2bits/cell) Vt distribution. It should be mentioned that there are still some finite interferences from WL-WL (Z-directional) and back-gate (Y-directional) effects. But these Y/Z interferences are not severe and can be well controlled. There are two programming methods. We can use the WL iteration method like the "LM" method [6] commonly used in 2D FG NAND to iterate the Vt distribution between various WL's. The resultant program-verify (PV) distribution is tighter. The other way is to use the simple full-sequence programming (FSP) to produce MLC (low/high pages are programmed together) in one ISPP sequence without WL iteration. FSP is faster in programming but the PV distribution is wider. Both methods are feasible to produce MLC with moderate ECC usage. The initial raw-bit error rate (RBER) can be controlled around 1E-4 for MLC distribution, which is comparable to the standard FG NAND Flash.

Figure 7 illustrates the impact of shared page number (SSL's) per WL. In 3D NAND, to share more SSL's in a block can reduce the WL decoder area and is cost saving. However, we found that it will generate hot-carrier disturbance after many number of programming (NOP) disturb by the many SSL's operation, especially when we adopt the WL iteration method (LM-like). It results in some tail bit behaviors in Vt distribution and increased fail-bit count (FBC). To control the FBC tail bits when using the WL iteration method, it is suggested to minimize the SSL number to <=8 per block to avoid strong NOP disturb. FSP method has more tolerance on multi-SSL's operation.

Figure 8 shows the read disturb test. The product chip is continuously read for a single page in a selected WL, while the neighbor WL's are applied a pass-gate voltage (Vpass,read=7V). The Vpass,read will continuously stress the entire blocks and will gradually shift the other WL's Vt. Usually both the standard FG NAND and 3D GAA VC NAND can only survive read disturb stress at ~100K to 500K cycling. Beyond the read disturb cycling when controller detects higher FBC it will move the data (refresh).

For our SGVC device, it can surpass the strong 120M read disturb test without the need to refresh the data. Such superior read disturb is due to the intrinsic advantage of flat-cell device that has much smaller low-voltage disturb.

Figure 9 shows the retention test. The high-temperature baking broadens the PV distribution due to the charge relaxation effects in the nitride-trapping device. Hole lateral migration [7] is indeed a challenge that degrades the retention when the device is erased by hole injection. For SGVC device, we have optimized the device to ensure the fresh state retention. In Fig. 9, the post 85C 300-hour baking still keeps the FBC ~30 extrapolated to 1E-7 chunk probability. The good news is that our retention loss becomes saturated after long-term baking. Using simplified Arrehenius model with Ea=1.1eV, the retention already surpass 40 years at room-temperature storage, and more than 5 years at 40 degree. The RBER increase slowly at 85C baking, and it is expected to be only ~1E-3 at 85C 10 year, which is correctable by BCH method. We believe that this retention performance is much better than ordinary TLC 3D NAND Flash.

The robust read disturb and long retention at fresh state makes the SGVC device very suitable for the highly read-intensive memory applications such as game-grade memory. Without the need to refresh and wear leveling, it is very advantageous for system design.

Figure 10 shows the MLC PE cycling test. It shows the feasibility of >1K PE cycling endurance. The PV distribution keeps tight after 1K cycling. There are some issues of erase degradation, SSL disturb, and worse retention after higher PE cycling stress. We will improve the device to provide higher-endurance performances in the future. It should be mentioned that the much fewer need of wear leveling by SGVC requires much lower PE cycling to fulfill ordinary applications.

Figure 11 shows the TLC (3bit/cell) distribution after completing the programming of all WL's. The WL iteration method is conducted to cancel the WL interferences to produce tighter PV distributions. The memory window between each state is tight. The advanced error correction method of low-density parity check (LDPC) [8] can be applied to correct the data. The TLC window management strategy is to suppress the sensing noise as much as we can, and enlarge the peak P/E window. To make better correction capability by LDPC, the FBC for all pages are suggested to be uniformly and randomly distributed so that LDPC can maximize its efficiency in correcting higher RBER. Usually the elder-programmed WL's have tendency of higher RBER due to the array effect. It should be minimized by optimizing the operation conditions.

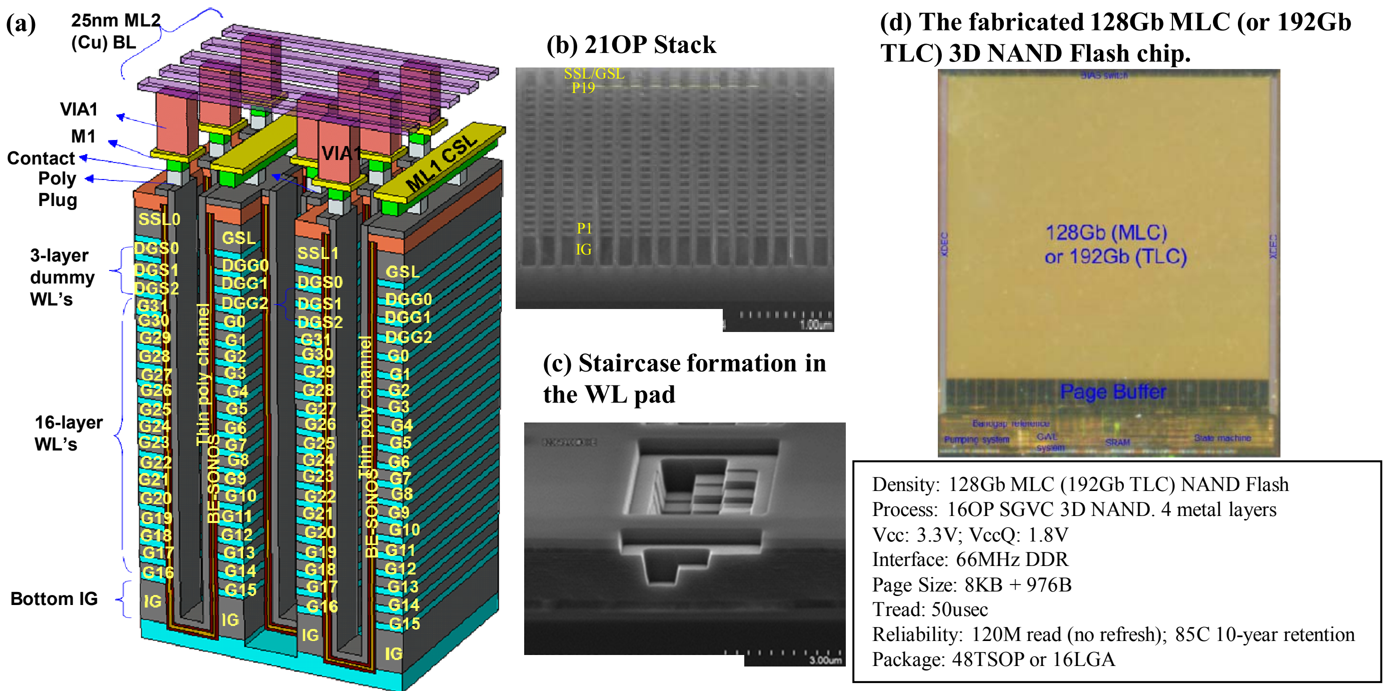
III. Memory Density Roadmap and Summary

Figure 12 shows the predicted memory density (Gb/mm^2, considered for the whole chip size) for various stacked layer number, assuming the same pitch and design rule. SGVC will produce >6Gb/mm^2 (TLC) at 48OP that can achieve the single-chip 1Tb solution with bit cost of <0.04USD/GByte.

We are continuously carrying out both processing and device improvements to enhance the non-volatile memory performances of SGVC 3D NAND Flash. With further stacking and pitch scaling, it is possible to get even lower "bit cost" than the trend estimated in [9]. In Fig. 2(b), it is estimated that there are rooms for further pitch scaling whenever the etching capability is enhanced. The flat-cell property allows similar FN tunneling and read disturb property after dimension scaling, in contrary to the GAA device. Pitch scaling get better cost down than high stacking and will be considered in the SGVC scaling.

References:

[1] H. T. Lue, et al, pp. 44-47, IEDM 2015. [2] J. Lee, et al, pp. 284-287, IEDM 2016. [3] K. Parat, et al ,pp. 48-51, IEDM 2015. [4] C. H. Hung, et al, pp. 227-230, IEDM 2012. [5] H. T. Lue, et al, pp. 555-558, IEDM 2005. [6] H. Kim et al., JSSC, Vol. 43, No. 4, pp. 919-928, 2008. [7] H. J. Kang, et al, pp. 182-183, VLSI 2015. [8] K. C. Ho, et al, IEEE transactions on very large scale integration (VLSI) systems, vol. 24, NO. 4, pp. 1293-1304. [9] EE Times report: "64 Layers is 3D NAND's Sweet Spot", in http://www.eetimes.com/document.asp?doc_id=1331911



Density: 128Gb MLC (192Gb TLC) NAND Flash
Process: 16OP SGVC 3D NAND. 4 metal layers
Vcc: 3.3V; VccQ: 1.8V
Interface: 66MHz DDR
Page Size: 8KB + 976B
Tread: 50usec
Reliability: 120M read (no refresh); 85C 10-year retention
Package: 48TSOP or 16LGA

Fig.1 **(a)** The schematics of 16-layer SGVC architecture. **(b)** The SEM cross-sectional view. A total of 21 OP layers, including 19 period OP layers, one bottom IG, one top SSL/GSL. The IG and SSL/GSL have much longer channel length than center OP layers. **(c)** The staircase formation fabricated at WL pad to reduce the WL RC delay. **(d)** The designed 128Gb MLC (or 192Gb TLC) memory product chip. The die size is **76.5mm^2.**

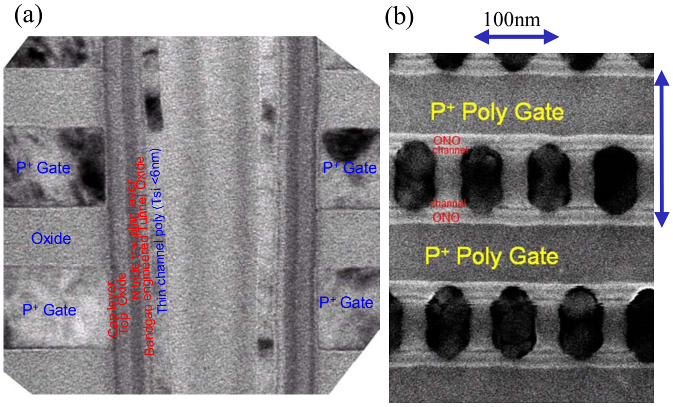


Fig.2 **(a)** The TEM cross-sectional view of the channel length-direction. A little taper angle of etching is observed, but this is tolerable for SGVC device. **(b)** The plane-view (cross-sectional view of one center layer) of the channel width direction. WL (Y) and BL (X) pitch are 220 and 100nm, respectively.

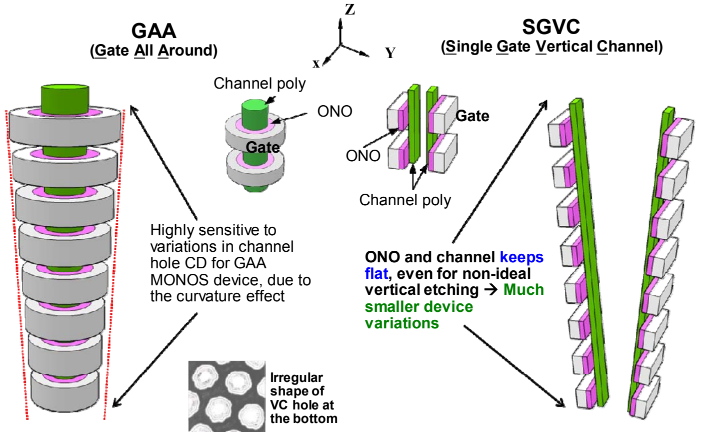

GAA
(Gate All Around)
Z
X
Y
SGVC
(Single Gate Vertical Channel)
Channel poly
ONO
Gate
ONO
Gate
Channel poly
Highly sensitive to variations in channel hole CD for GAA MONOS device, due to the curvature effect
ONO and channel keeps flat, even for non-ideal vertical etching → Much smaller device variations
Irregular shape of VC hole at the bottom

Fig.3 Schematic diagram to illustrate the advantage of SGVC flat-channel device. ONO and channel poly keep flat even for a non-ideal vertical etching. This gives much tolerance to 3D etching performance.

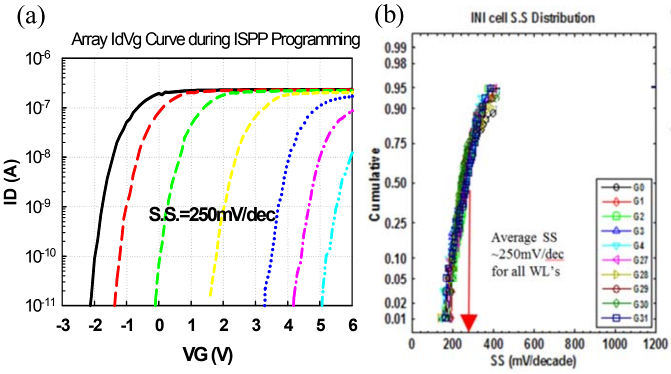


Fig.4 **(a)** The IdVg curves during ISPP programming for array cell. **(b)** The average subthreshold slope (SS) is only 250mV/dec, thanks to the thin-body TFT device.

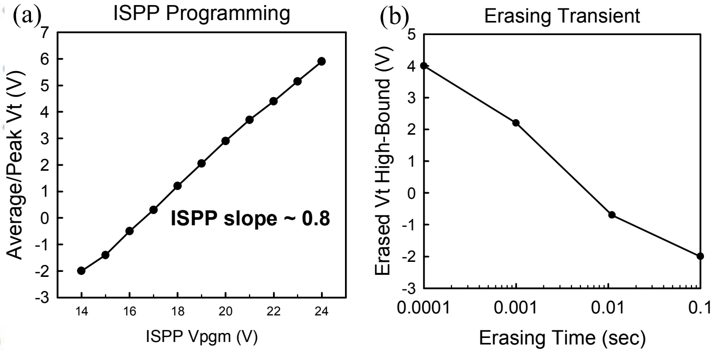


Fig.5 **(a)** The average (mean) Vt of product chip during ISPP programming. The ISPP slope ~0.8. **(b)** The erasing transient (20V) of erased Vt high bound. High-bound Vt can be lower than -2V.

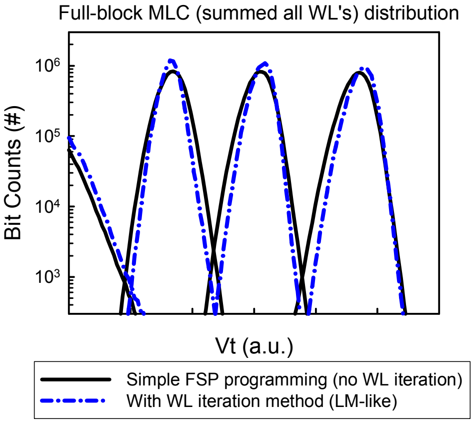


Fig.6 MLC Vt distribution. Two methods are compared. 1st one is the simplest full-sequence programming (FSP) without any WL iteration. The 2nd one is with WL iteration technique (LM-like) that can cancel the interferences effects.

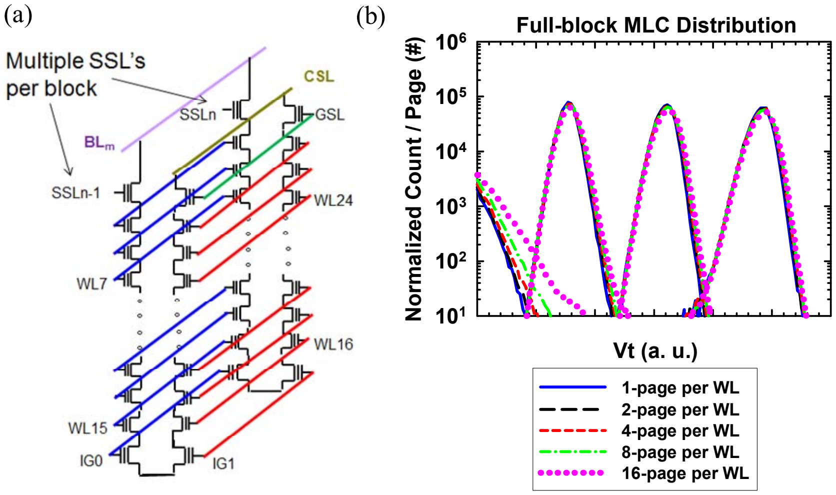


Fig.7 **(a)** Schematic to illustrate the SGVC 3D NAND architecture. Every BL is connected to multiple NAND strings in a block, where each string has its own string-select transistor (SSL) for the control. **(b)** The MLC memory window for various number of strings (SSL) per block. When SSL number is 16 per WL, it shows increased program disturb and tail bits.

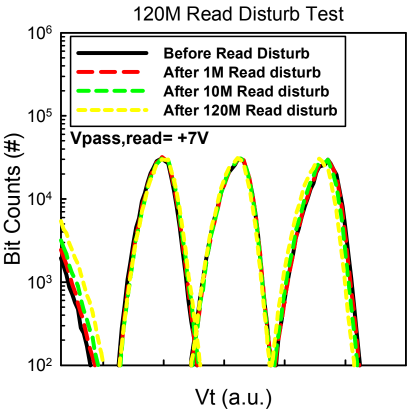


Fig.8 The120M read-disturb test results. The center WL10 is continuously read after 120M read cycling. The far-side WL's will suffer a continuous Vpass,read (=7V) stress. The device is robust against read disturb stress.

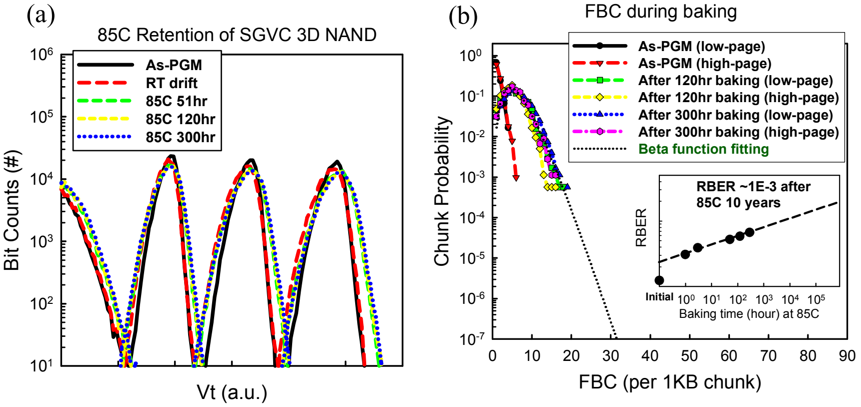


Fig.9 **(a)** The 85C long-term retention (fresh, P/E=1). The PV distribution is broadened after baking by the charge relaxation effect. **(b)** The FBC (per 1KB chunk size) approximately follows the beta function. Extrapolated to 1E-7 chunk probability, the FBC is still controlled within 35 after baking. The inset shows the RBER increase to only ~1E-3 at 85C 10 year.

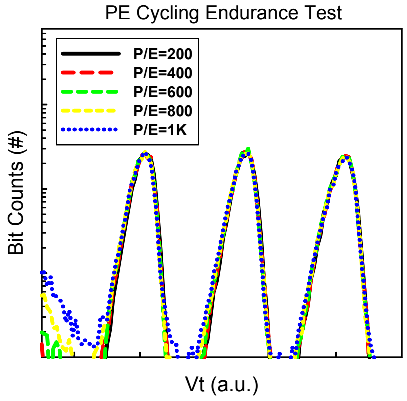


Fig.10 The PE cycling capability test. The device shows feasibility to support >1K cycling stress. The PV distribution keeps tight. Further improvements will be continued to achieve higher endurance performances.

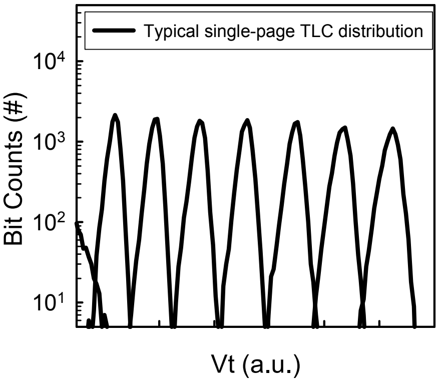


Fig.11 The typical single-page TLC window after completing all other WL's programming. WL iteration method is carried out to cancel the Y/Z interferences. Extensive ECC method using LDPC can correct the data.

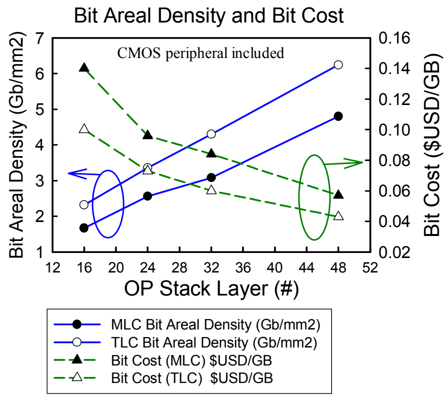


Fig.12 Memory density and bit cost (ballpark only) analysis of SGVC 3D NAND. 48OP layers can produce 1Tb single-chip solution that achieves density >6Gb/mm^2, with bit cost < 0.04 USD/GB. (PS: Assume wafer cost ranges from 1700 to 2000 USD, with 80% yield)