13.1 A 1Tb 4b/cell NAND Flash Memory with t_{PROG}=2ms, t_R=110µs and 1.2Gb/s High-Speed IO Rate

Doo-hyun Kim, Hyunggon Kim, Sungwon Yun, Youngsun Song, Jisu Kim , Sung-Min Joe, Kyung-Hwa Kang, Joonsuc Jang, Hyun-Jun Yoon, Kangbin Lee, Minseok Kim , Joonsoo Kwon, Jonghoo Jo, Sehwan Park, Jiyoon Park, Jisoo Cho, Sohyun Park, Garam Kim , Jinbae Bang, Heejin Kim, Jongeun Park , Deokwoo Lee, Seonyong Lee, Hwajun Jang, Han-Jun Lee, Donghyun Shin, Jungmin Park, Jungkwan Kim , Jongmin Kim, Kichang Jang, Il Han Park, Seung Hyun Moon, Myung-Hoon Choi, Pansuk Kwak, Joo-Yong Park, Youngdon Choi, Sang-Lok Kim, Seungjae Lee, Dongku Kang, Jeong-Don Lim, Dae-Seok Byeon, Kiwhan Song, Junghwan Choi, Sang Joon Hwang, Jaeheon Jeong

Samsung Electronics, Hwaseong, Korea

3D NAND flash memory has enhanced its areal density by more than 50% per year by virtue of the aggressive development of 3D WL stacking technology for the recent three consecutive years [1-3]. Also storage market still requires more bits for diverse digital applications. [4]

Since the announcement of a 4b/cell topology, based on three-dimensional (3D) stacked-word-line NAND Flash memory in previous work [5], solid state drives (SSDs) using 4b/cell technology are emerging in the market and have received a lot of attention. The 4b/cell technology is one promising solution that can fulfill the demand for explosive data growth. As the mainstream NAND Flash memory market has been rapidly replaced by 3b/cell technology, the 4b/cell technology is also expected to become the mainstream in this area. However, the technology is inherently plagued with performance and reliability issues, because the programming of the required 16 states has to be performed within a limited cell threshold voltage (V_{th}) window.

This paper presents a 1Tb 4b/cell NAND flash memory, which is successfully developed and manufactured using 92 stacked WLs. The chip achieves 7.53Gb/mm^2 areal density with a 18MB/s program throughput, t_R=110µs, and 1.2Gb/s IO speeds. Chip size is reduced by 25% over previous work [5]. The program and read performance are enhanced by 33% and 24%, respectively by extremely improving cell characteristics with while maintaining reliability criteria. In the following sections the schemes for reducing program time as well as improving reliability, by more than 2×, are presented. Improvements in all key performance parameters are achieved by mitigating the disadvantage of the multi-stack by the application of reliability and performance enhancing schemes.

The proposed predictive program scheme is shown in Fig. 13.1.1. If the cell, whose target state is P_{n+1}, is determined to be off-cell in the P_n verify operation, it is inhibited from program after applying 1 pulse program without P_{n+1} verify operation. In general, this degrades cell V_{th} distribution. However, the degradation can be minimized if programming step voltage that means increment of program pulse per one loop and verify voltage difference between adjacent states are not significantly different. The re-program scheme consisting of coarse program and fine program is used to make the cell threshold voltage distribution becomes narrower. [5] The predictive program scheme can be applied to coarse program, not fine program because the coarse program uses a larger programming step voltage than that of fine program and it is similar to V_{th} difference between adjacent states. As coarse program consists of half of the total programming time and also, as verification operation consists of 70% of coarse program, we can reduce total programming time by about 16.5% and verify count by 47% with this novel verification technique.

Retention issue owing to charge loss with CTF (charge trap flash) geometry is one of the critical issues that impact the reliability of 4b/cell NAND Flash memory. Moreover, as increasing the number of stacked WL layers, the gate length and space lengthbetween WLs should be scaled down due to the difficulty of etching process during the channel hole formation, which leads to change of retention characteristics in accordance with states of adjacent WL cells, so called 'lateral spreading effect' [6]. When adjacent WL's state is in a deep erased state, deterioration of retention characteristics due to the lateral spreading effect increases. To reduce the lateral spreading effect, a deep erase compensation (DEC) scheme is proposed as shown in Fig. 13.1.2. First, deep erased cells are verified and then a program of the verified cells is performed with optimal program voltage (V_{PGM}) in order to move deep erased cells towards higher V_{th}. By applying this scheme it is possible to eliminate those deep erased cells of adjacent WLs. As a result, the V_{th} distribution skew of WL_n cells due to the lateral spreading effect decreases after a retention time. Therefore, the retention characteristic of QLC increases more than 10%.

As QLC has double the states than TLC, it is more difficult to acquire read window margin within QLC cell V_{th} distribution. Furthermore, considering cell Vth distribution shift due to a retention, it makes even more difficult to satisfy reliability criteria. To overcome such obstacle, we developed an adaptive read scheme using cell count (ARC), in which read levels can be adapted based on the cell V_{th} distribution shift of the highest states due to retention. As suggested in this paper, the concept of ARC is as follows: first, measure the cell-count of the highest state which most likely to shift with the greatest distance. Second, based on the cell-count information carried out with the highest state corresponding page mapping, other lower states' read level is varied as shown in Fig. 13.1.3(a). As the number of cells in each 16 states are well managed by randomizer, the information of the highest states' off-cell numbers represent actual cell V_{th} distribution shifts well. As Information of cell count can also reflect retention speed due to temperature and process variations, it can predict read levels more accurately compared to the method where it varies read levels depending only on the time passed. When the page read command is invoked, the chip starts reading from the highest state to the lowest state in order. From the first read, off-cell count information of the highest state can be used to estimate retention amount is acquired by the on-chip logical cell counter module. This value acquired from the first read of highest state is compared with the pre-set reference value and then in accordance with the pre-set cases, read level for next read can be adjusted. As shown in Fig. 13.1.3(b), it is shown in a chip test, by using ARC method retention reliability characteristic has increased about twice.

In order to support the next-generation host IO interface such as PCIe Gen4.0 and UFS 3.0, the NAND IO speed is required over 1.2Gb/s. In order to meet the speed requirement, the Toggle 4.0 specification is just established. The presented device fully supports the Toggle 4.0 specification including not only 1.2Gb/s operation but also the read/write training, the read DCC (duty-cycle corrector), the ZQ calibration [3], and the multi-purpose ODT (on-die termination) using the ODT pin. The DCC calibration loop consists of the coarse tuning and the fine tuning as illustrated in Fig. 13.1.4(a). The coarse tuning loop employs SAR (successive approximation algorithm) logic for a fast locking time while the fine tuning loop utilizes a linear counter to achieve the target accuracy. Since the proposed calibration loop is performed only during the power-up period, it is difficult to fully keep up with voltage and temperature variations in real time because the toggle NAND device usually does not have a free running clock. In order to overcome this weakness, we additionally suggest the run-time duty correction sequence to compensate such environmental changes, as shown in Fig. 13.1.4(b), 13.1.5(a) and 13.1.5(b) verify that the proposed DCC scheme can improve eye-open windows by up to 8% in mono chips and properly adjust the duty distortion in the read clock path, thus successfully achieving high-speed 1.2Gb/s NAND IO Rate. Moreover, since maintaining the eye-open window gets more severe as the number of dies in a package increases, the proposed DCC scheme can be an effective way of preserving the required signal integrity in multi-die environments.

Figure 13.1.6 shows the fabricated die photograph, with an area of 136mm^2. Key parameters are summarized in Fig. 13.1.7.

References:

[1] K.-T. Park et al., "Three-Dimensional 128Gb MLC Vertical NAND Flash-Memory with 24-WL Stacked Layers and 50MB/s High-Speed Programming," *ISSCC*, pp. 334-335, Feb. 2014.

[2] J.-W. Im et al., "A 128Gb 3b/Cell V-NAND Flash Memory with 1Gb/s I/O Rate," *ISSCC* pp. 130-131, Feb. 2015.

[3] D. Kang et al., "256Gb 3b/Cell V-NAND Flash Memory with 48 Stacked WL Layers," *ISSCC*, pp. 130-131, Feb. 2016.

[4] C. Kim et al., "A 512Gb 3b/cell 64-Stacked WL 3D-NAND flash memory," *ISSCC*, pp. 202-203, Feb. 2017.

[5] S. Lee et al., "A 1Tb 4b/Cell 64 Stacked WL 3D NAND Flash Memory with 12MB/s program throughput," *ISSCC*, pp. 340-342, Feb. 2018.

[6] H. Kang et al., "Space Program Scheme for 3-D NAND Flash Memory Specialized for the TLC Design," *VLSI*, pp. 201-202, 2018.

 978-1-7281-3205-1/20/$31.00

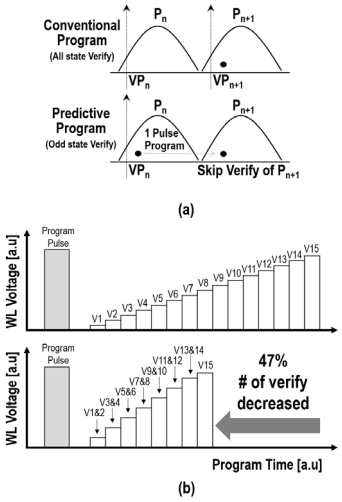

Figure 13.1.1: (a) Proposed program scheme using predictive verify concept. (b) The number of verifies is decreased by 47% using the predictive program scheme.

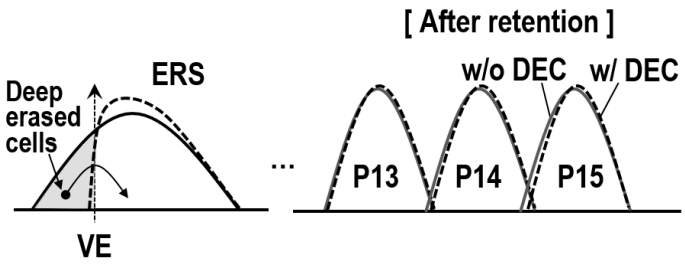

Figure 13.1.2: Illustration of deep-erase compensation (DEC) scheme.

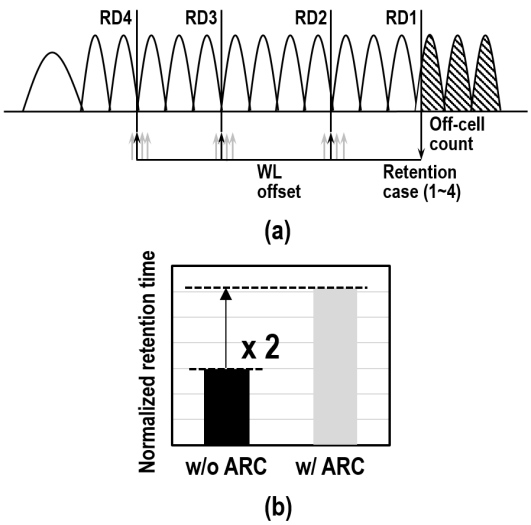

Figure 13.1.3: (a) Adaptive-read scheme using cell count (ARC) scheme. (b) Reliability is shown to be improved by 2×.

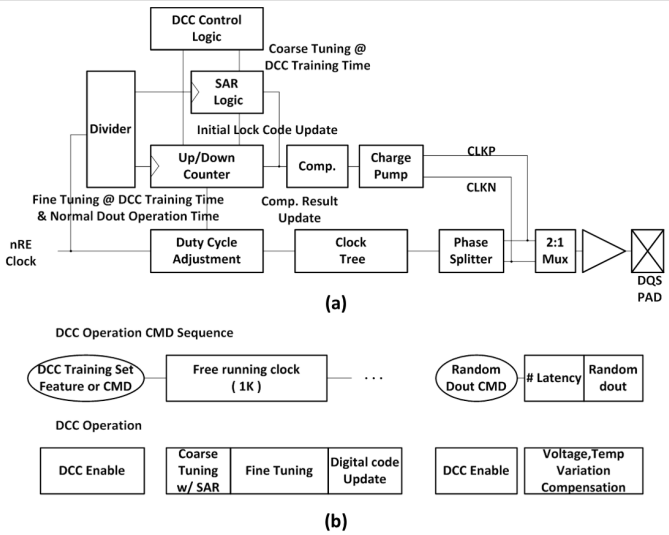

Figure 13.1.4: (a) Overall block diagram of the proposed duty cycle correction (DCC) scheme, (b) an example DCC operation sequence.

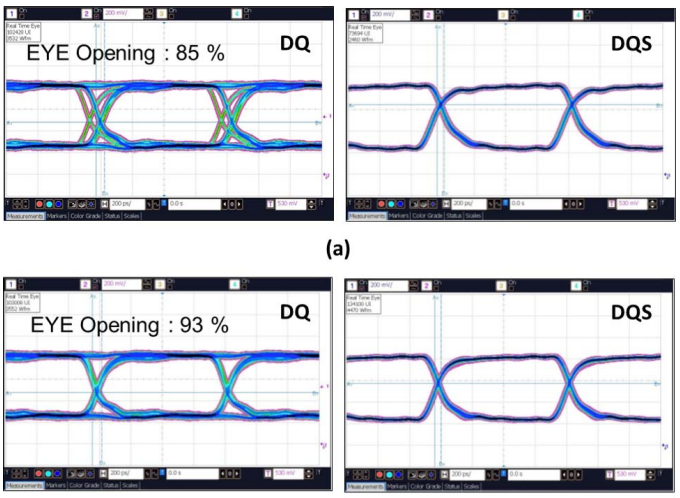

Figure 13.1.5: Measured read eye diagrams for DQ and DQS (a) before, and (b) after using DCC at 1.2Gb/s IO speed.

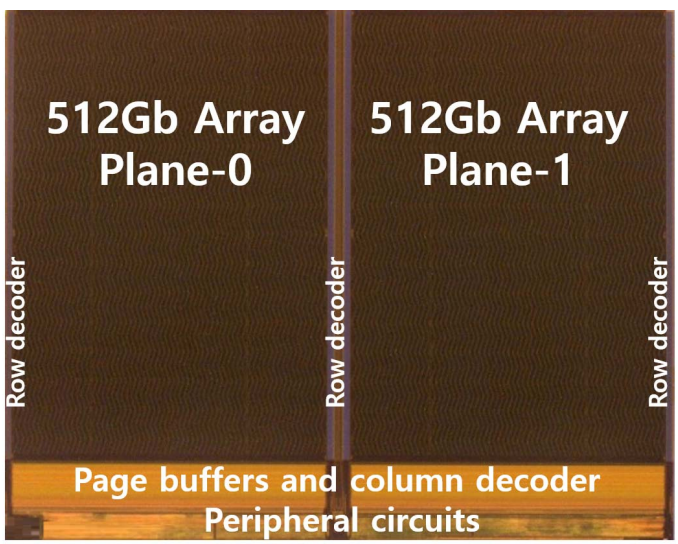

Figure 13.1.6: Die photograph of the 92-stacked 1Tb 4b/cell 3D V-NAND Flash.

	Previous Work [5]	This Work
Bits per cell	4	4
Density	1Tb	1Tb
Chip Size	182mm^2	136mm^2
Technology	4th 3D V-NAND with 64 stacked WL layer	5th 3D V-NAND with 92 stacked WL layer
tBERS	3.5ms (Typ.)	3.5ms (Typ.)
tPROG	3ms	2ms
tR (4K)	145µs	110µs
IO Rate	Max. 1.0Gb/s	Max. 1.2Gb/s

Figure 13.1.7: Table listing key parameters for the 1Tb 4b/cell 3D NAND Flash, with a comparison to previous work.

978-1-7281-3205-1/20/$31.00