13.2 A 1Tb 4b/Cell 96-Stacked-WL 3D NAND Flash Memory with 30MB/s Program Throughput Using Peripheral Circuit Under Memory Cell Array Technique

Hwang Huh, Wanik Cho, Jinhaeng Lee, Yujong Noh, Yongsoon Park, Sunghwa Ok, Jongwoo Kim, Kayoung Cho, Hyunchul Lee, Geonu Kim, Kangwoo Park, Kwanho Kim, Heejoo Lee, Sooyeol Chai, Chankeun Kwon, Hanna Cho, Chanhui Jeong, Yujin Yang, Jayoon Goo, Jangwon Park, Juhyeong Lee, Heonki Kim, Kangwook Jo, Cheoljoong Park, Hyeonsu Nam, Hyunseok Song, Sangkyu Lee, Woopyo Jeong, Kun-Ok Ahn, Tae-Sung Jung

SK hynix, Icheon, Korea

Ever since a 3b/cell (TLC) NAND Flash memory became the mainstream in non-volatile memory market, a new demand for a 4b/cell (QLC) NAND flash memory has been emerging for low-cost applications. However, QLC has inherently much longer page program time than TLC because of 16-state programming within a limited program and erase (PE) window, as well as narrower V_{th} distributions. The longer page-program time, subsequently, degrades sequential write performance. Thus it is not possible to meet the required sequential-write performance in applications such as mobile devices and solid state drives (SSDs).

We present a 1Tb QLC 3D NAND Flash memory with 30MB/s program throughput, featured as a 4-plane architecture, using a 16kB-page size, and a page-program time of 2.15ms. Its density reaches 8.4Gb/mm^2 with a 4-plane architecture. For reducing page-program time, we use two kind of the V_{th}-distribution improvement techniques: a high-gain bandgap reference (BGR) circuit, and an accurate sensing scheme. As a result, a 2.15ms page-program time is achieved using a two-step programming method: coarse 16-level in the 1st step, and a fine 16-level in the 2nd step.

Figure 13.2.1 shows the concept of peripheral circuit under memory cell array (PUC). A 16kB page buffer (PB), the largest area in peripheral logic circuitry, has been arranged below the plane of cell array. In previous 2-plane structured 3D QLC NAND memories [1,2], it is hard to add more planes alongside the memory array within a given chip area because PB arrays are placed in the same X/Y direction to the memory cells. However, in this work, PUC enables 16kB/plane and 4-plane architecture in a small die area by putting PB arrays underneath the memory cell in Z direction. These doubled planes make the device 2× faster in sequential write throughput with the same program time. To enable more efficient peripheral-under-cell structure, a slit metal contact is introduced to connect page buffers with cell bit lines. In the middle of vertical 3D cell arrays, the eight open spaces per plane are regularly patterned for the pathways of contacts. The contacts pass through the opening way vertically in order to link above-cell metal lines to below-cell metal lines. Also, the size reduction in peripheral circuitry is performed to fit them underneath cell arrays as much as possible. Most of peripheral circuits are settled under array, and only 10% of the entire chip is organized, outside of cell array, for peripheral circuits, power metal lines, row decoders, and pads. The height of peripheral circuits outside cell array reaches about 2 times the height of the pad as shown in Fig. 13.2.6. This is distinguished from other similar work [3]. In addition to small 16KB 4-plane PUC, for faster write performance, the proposed following V_{th} distribution improvements are applied, and then we leverage these techniques to reduce page program time.

In developing narrow V_{th} distributions of 4b/cell 3D NAND, it is important that various bias voltage generators used in read and program operations have to precisely generate their target levels within small variations responding to temperatures. Figure 13.2.2 shows the schematic and simulation result of high-gain BGR that we implement in this device. In the scheme, diode M1 creates a strong complementary to absolute temperature (CTAT) current, I_{CTS}, corresponding to temperatures. The current generated by BJT pairs is a proportional to absolute temperature (PTAT) current, I_{PTAT}. The summation of I_{CTS} and I_{PTAT} is a weak CTAT current, I_{CTW}, and the reference voltage is generated by flowing I_{CTW} through R5 which is a diffusion resistor with the PTAT characteristic. The main negative factors determining the temperature variation are not only component mismatches but also nonlinearities of error-amplifier. The former is minimized by common centroid matching in component layouts while the latter is reduced by upgrading an error amplifier. The error amplifier uses NMOS input pairs for low input offset to reduce a random dopant fluctuation. In order to reduce an input offset, the error amplifier designs that DC gain is above 74dB with all PVT variations. The Monte-Carlo simulation shows the variation is decreased by 46.9%, comparing to conventional one.

It is known that BL-coupling noise between neighboring BLs diminishes the accuracy in a sensing operation, and thus it broadens V_{th} windows. Figure 13.2.3(d) displays the curve of I-trip, cell current to discriminate whether the cell is turned on or off, responding to both adjacent BL patterns and the threshold voltage of the cell. When adjacent BLs (aggressors) are grounded, I-trip of victim BL holds its value without coupling noise in a sensing operation. However, I-trip of victim BL fluctuates due to coupling noise when the adjacent BLs are precharged. If the states of adjacent BLs are controlled to ground levels during a sensing operation, the BL-to-BL capacitive coupling noise can be eliminated, and the result of sensing becomes highly confident. The random data is generally programmed by a user for a better reliability. An individual BL is programmed to one threshold voltage out of 16 V_{th} levels, and it is likely that the cell V_{th} of desired BL is different from neighboring BLs after programming. We invent the 4-step sequence for suppressing BL-to-BL coupling noise in Fig. 13.2.4: (1) Sensing upper V_{th} level; (2) Sensing lower V_{th} level; (3) Discharging off-cell BL in (1) and on-cell BL in (2) to ground for shielding; and (4) sensing target V_{th} level. In previous method [4], sensing lower V_{th} level and discharging on-cell BL were employed while both off-cell and on-cell BL are discharged with higher and lower V_{th} sensing respectively in this scheme. These both-sided ground BLs can refine the BL-to-BL coupling noise effectively. Thanks to adjacent BLs which are kept at ground, target level sensing is performed with removing BL-to-BL coupling noise. This sequence consequently leads to increase page read time (t_R) by reason of additional sensing operations. A sensing evaluation time is controlled to minimize the impact on increasing $t_{R,}$ instead of changing directly wordline (WL) bias voltages at every sense. In comparison with conventional scheme, the almost equivalent read time is achieved, and, as a result, this technique reduces the raw bit error rate by 46% in measured data.

Figure 13.2.6 shows the microscope image of this QLC NAND and its key parameters are summarized in Fig. 13.2.5.

References:

[1] Seungjae 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.
[2] N. Shibata et al., "A 1.33Tb 4-bit/Cell 3D-Flash Memory on a 96-Word-Line-Layer Technology," *ISSCC*, pp. 210-212, Feb. 2019.
[3] Chang Siau et al., "A 512Gb 3-bit/Cell 3D Flash Memory on 128-Wordline-Layer with 132MB/s Write Performance Featuring Circuit-Under-Array Technology," *ISSCC*, pp.218-219, Feb. 2019.
[4] Sungdae Choi et al., "A 93.4mm^2 64Gb MLC NAND-Flash Memory with 16nm CMOS Technology," *ISSCC*, pp. 328-329, Feb. 2014.

978-1-7281-3205-1/20/$31.00 ©2020 IEEE

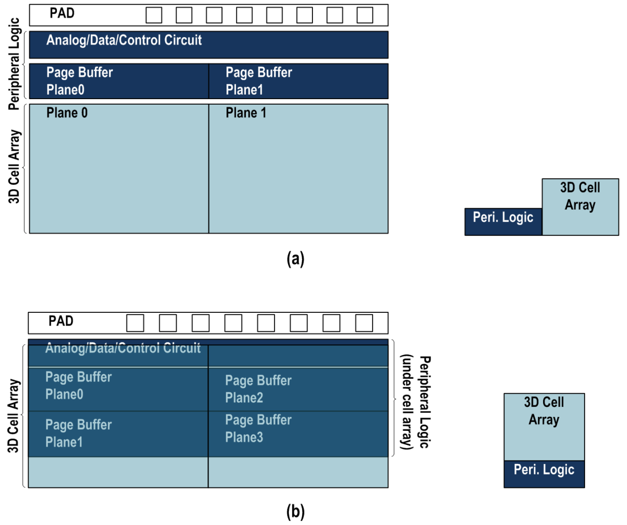


Figure 13.2.1: Comparison between (a) a two-plane non-PUC and (b) a four-plane PUC 3D NAND.

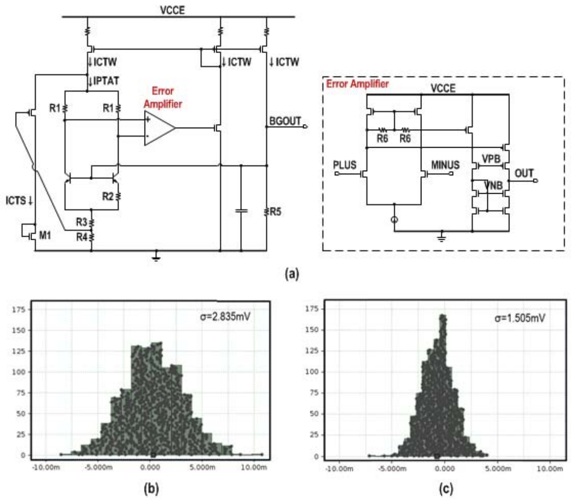


Figure 13.2.2: (a) High-gain bandgap reference (error-amplifier) scheme. Monte-Carlo simulation results for (b) the conventional and (c) the proposed BGR.

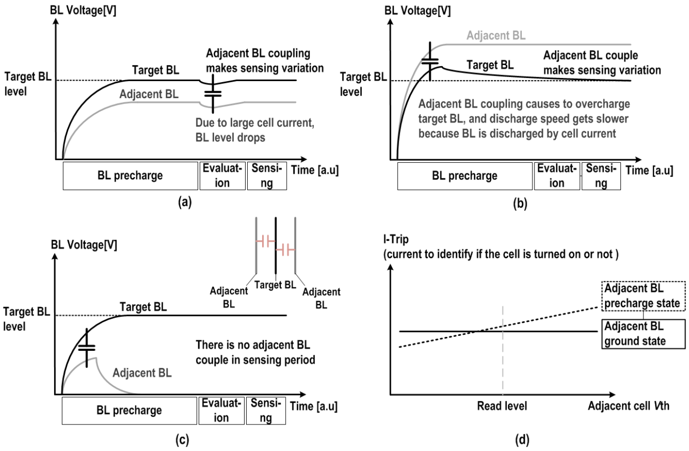


Figure 13.2.3: Capacitive coupling noise to desired BL under different scenarios. (a) Adjacent BL precharge & lower cell V_{th}, (b) Adjacent BL precharge & higher cell V_{th}, (c) Adjacent BL to ground. (d) I-trip graph, corresponding to adjacent BL patterns & cell V_{th}.

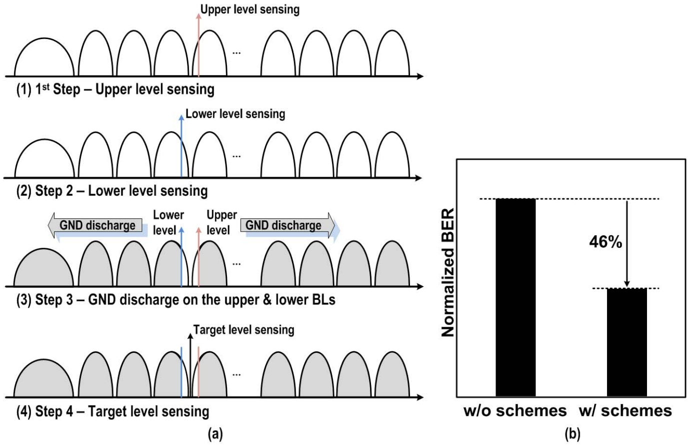
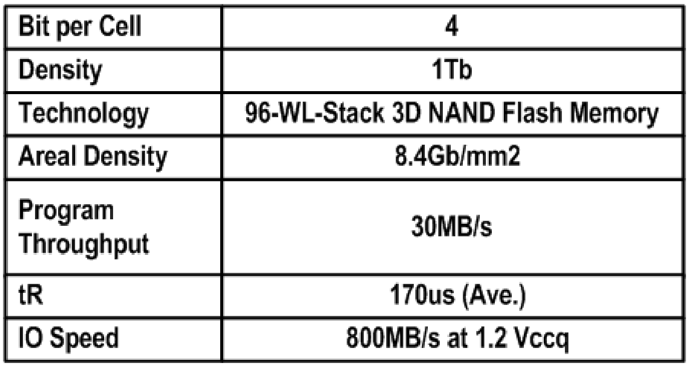

Upper level sensing
(1) 1st Step – Upper level sensing
Lower level sensing
(2) Step 2 – Lower level sensing
GND discharge
Lower level
Upper level
GND discharge
(3) Step 3 – GND discharge on the upper & lower BLs
Target level sensing
(4) Step 4 – Target level sensing
(a)
Normalized BER
46%
w/o schemes
w/ schemes
(b)

Figure 13.2.4: (a) Proposed sequence for accurate sensing (b) A 46% improvement in BER is observed.



Bit per Cell	4
Density	1Tb
Technology	96-WL-Stack 3D NAND Flash Memory
Areal Density	8.4Gb/mm2
Program Throughput	30MB/s
tR	170us (Ave.)
IO Speed	800MB/s at 1.2 Vccq

Figure 13.2.5: Table summarizing key test-chip parameters.

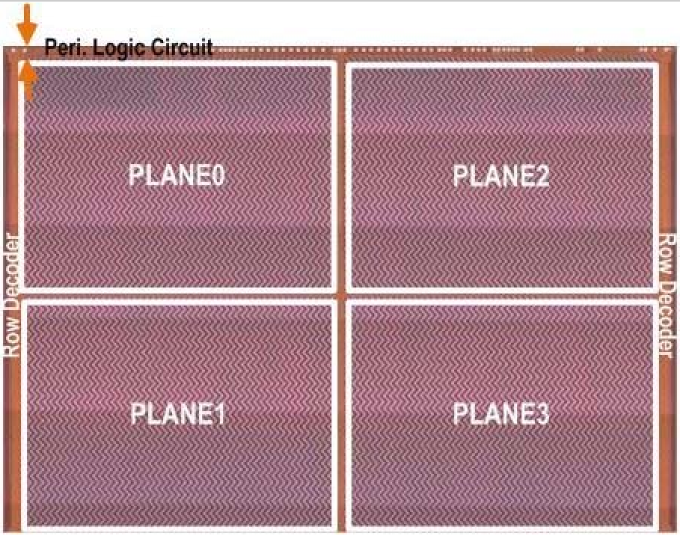


Figure 13.2.6: Die microphotograph: outside-of-cell-array circuits within 10 percent of the total chip.