13.4 A 512Gb 3-bit/Cell 3D 6th-Generation V-NAND Flash Memory with 82MB/s Write Throughput and 1.2Gb/s Interface

Dongku Kang, Minsu Kim, Su Chang Jeon, Wontaeck Jung, Jooyong Park, Gyosoo Choo, Dong-kyo Shim, Anil Kavala, Seung-Bum Kim, Kyung-Min Kang, Jiyoung Lee, Kuihan Ko, Hyun-Wook Park, Byung-Jun Min, Changyeon Yu, Sewon Yun, Nahyun Kim, Yeonwook Jung, Sungwhan Seo, Sunghoon Kim, Moo Kyung Lee, Joo-Yong Park, James C. Kim, Young San Cha, Kwangwon Kim, Youngmin Jo, Hyunjin Kim, Youngdon Choi, Jindo Byun, Ji-hyun Park, Kiwon Kim, Tae-Hong Kwon, Youngsun Min, Chiweon Yoon, Youngcho Kim, Dong-Hun Kwak, Eungsuk Lee, Wook-ghee Hahn, Ki-sung Kim, Kyungmin Kim, Euisang Yoon, Won-Tae Kim, Inryoul Lee, Seung hyun Moon, Jeongdon Ihm, Dae Seok Byeon, Ki-Whan Song, Sangjoon Hwang, Kye Hyun Kyung

Samsung Electronics, Hwasung, Korea

Data storage is one of the hottest discussion topics in today's connected world. The amount of data growth is expected to be exponential, while budget and space remain constricted. Since the transformation of storage device from planar NAND to 3D V-NAND [1], the areal density of semiconductor storage devices has continuously evolved and has surpassed the density of magnetic hard drives. By providing the largest storage capacity in the smallest footprint, 3D V-NAND has been leading the data center revolution in recent years. However, 3D-technology scaling faces several technical challenges [2]. (1) As the number of WL stacks increases the channel-hole etch process becomes a limit, since the total WL-mold height increases. (2) Interference between cells increases since the distance between WLs becomes smaller. (3) Faster data transfer speeds are required to support higher IO bandwidth.

This paper presents a 512Gb, 3b/cell, flash memory featuring the 6th generation V-NAND technology, which integrates more than 120-WL layers. An 16kB two-plane cell array with 8kB sub-plane architecture is implemented. Read and write performance is improved to 820 and 82MB/s. To achieve this it was critical to minimize the WL and BL setup time for read and verify (during programming) operations. Also, due to the increased number of WL layers, variation in program and erase characteristics, as well as power dissipation became more serious issues. Therefore, a customized V_{th} window management scheme and a low-power sensing scheme are implemented. Finally, to meet the 1.2Gb/s IO performance target, a channel training scheme and an enhanced on-die termination schemes were developed. Figure 13.4.1 summarizes the key features of this work.

To improve read and write throughput the BL precharge time should be minimized. As shown in Fig. 13.4.2(a) (1) V_1 is applied to BLSHF to force V_{BL1} to BL as a first step; (2) V_2 is applied to BLSHF to force V_{BL2} to BL as shown in Fig 13.4.2(b). To accelerate BL precharge, we need to ensure that $V_1 > V_2$ to keep $V_{BL1} > V_{BL2}$. However, when a V_1 to V_2 transition occurs, the sudden disconnect from the page buffer to BL across BLSHF can be observed when the BL voltage near the page buffer is higher than V_2-V_t(BLSHF). This degrades the overall precharge performance because the BL is not precharged during that period. To solve this problem, enhanced BL precharge scheme is proposed as shown in Fig. 13.4.2(c). In this scheme, rather than applying V_2 as a step function from V_1, V_2 is applied gradually to promote a continuous current flow in any situation across BLSHF, thereby minimizing the BL precharge time.

For a 3b/cell V-NAND, a program pulse is followed by seven times of verify operations to form the seven distinctive V_{th} states. The verify operations are a large overhead for a program operation, since the program pulse time is typically far less than the verify time for a programming cycle. As shown in Fig. 13.4.3(a), when a verify (or target) voltage is applied to WL_N, WL_{N+1} and WL_{N-1} are subjected to the V_{READ} voltage just like the other unselected WLs. Therefore, from WL_N's perspective, a couple of capacitances exist looking towards the upper and lower WL directions. In general, WL_{N+1} always maintains an erase state when WL_N is programmed and verified. By utilizing this principle, we can apply a lower voltage than V_{READ} to WL_{N+1} without affecting the sensing current for the verify operation. In this work, to minimize the capacitive coupling between WL_N and WL_{N+1}, we apply a copy of the selected WL voltage to WL_{N+1}, as shown in Fig. 13.4.3(b). This technique improves the programming throughput by reducing the WL setup time for the verify operations.

An increase in the number of WL stacked layers results in degraded channel-hole variation [3], as shown in Fig. 13.4.4(a & b). Hence, to minimize this effect, an adaptive erase scheme is used to minimize erase V_{th} variation, as shown in Fig. 13.3.4(c). Generally, a cell's reliability to repetitive erase-and-program operations is inversely proportional to a cell's V_{th} window. Hence, the V_{th} distance between an erase state and the highest programed state is required to be constant, since all WLs should be able to handle the same amount of erase-and-program cycles. Hence, to achieve balanced reliability characteristic for all WLs, we apply a progressive V_{th} window scheme as shown in Fig. 13.4.4(d).

Power consumption is another imminent issue for increasing WL stacks. As shown in Fig. 13.4.5(a), pre-pulse operation is required for every sensing operation to initialize a channel-hole potential. Otherwise, the channel voltage might fluctuate, depending on the situation, due to the accumulated voltage when consecutive sensing operations are performed. When the accumulated channel voltage exceeds a certain level, it might also introduce V_{th} disturbance for cells. However, during a pre-pulse operation, the WL to channel-hole capacitance is instantaneously inflated multiple times, since the WL-channel capacitances for all SSLs (shown in Fig. 13.4.5(b)) become effective. This increases power consumption. (Fig. 5(b)) To overcome this excessive power consumption, we developed a random pre-pulse sensing scheme, which is shown in Fig. 13.4.5(c). This scheme reduces the occurrence of pre-pulse operations to lower the average current consumption. Also, sporadic execution of pre-pulse operations prevents channel voltage accumulation.

To support a 1.2Gb/s toggle 4.0 double-data rate, it was necessary to overcome pin-to-pin load mismatch and read-clock duty-cycle error. Hence, the input and output interface is implemented with read and write DQ training and duty-cycle correction (DCC) training circuits. For instance, write DQ training improves t_{DSH} (data setup and hold time margin) by delaying each pin at the host controller according to its load mismatch, as shown in Fig. 13.4.6(a). Similarly, read training compensates t_{DVW} (data valid window) by adjusting the input strobe for each pin in the controller, as shown in Fig. 13.4.6(b). When training the controller uses a designated data pattern for V-NAND and reads this pattern from the V-NAND to perform validation. In addition to the read and write training, an explicit and implicit DCC improves t_{DVW} by correcting the read clock's duty cycle error, which is caused by channel losses, as shown in Fig. 13.4.6(c). Explicit DCC training is performed with a specific command sequence, whereas implicit DCC training is performed during normal read operations. Finally, to enhance signal integrity (SI) for a multi-die-stack package, two types of on-die-termination (ODT) schemes are supported, as shown in Fig. 13.4.6(d). In the first (designated) ODT scheme, designated chips (CHIP2 and CHIP4) of different packages are used for termination when the nODTx signal is enabled. However, in the second (hybrid-CE) scheme, the termination is enabled with the chip enable signal (nCEx). These termination techniques provide additional options to achieve better SI with lower termination power. The termination modes (read or write) are detected by reading the read clock (nREx) with nODTx signal.

References:

[1] J. Jang, et al., "Vertical Cell Array using TCAT (Terabit Cell Array Transistor) Technology for Ultra High Density NAND Flash Memory," *IEEE VLSI*, pp. 192-193, 2009.

[2] D. Kang, et al., "256 Gb 3 b/Cell V-nand Flash Memory With 48 Stacked WL Layers," *JSSC*, vol. 52, no. 1, pp. 210–217, Jan. 2017.

[3] H. Kim, et al., "Evolution of NAND Flash Memory: From 2D to 3D as a Storage Market Leader," *IEEE Intl. Memory Workshop*, 2017.

[4] C. Kim, et al., "A 512-Gb 3-b/Cell 64-Stacked WL 3-D-NAND Flash Memory," *JSSC*, vol. 53, no. 1, pp. 124-133, Jan. 2018.

[5] H. Maejima, et al., "A 512Gb 3b/Cell 3D Flash Memory on a 96-Word-Line-Layer Technology," *ISSCC*, pp. 336-337, 2018.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

	ISSCC 2018 [5]	ISSCC 2019
Bits per cell	3	3
Density	512Gb	512Gb
Pagesize	16 KB/Page	16 KB/Page
I/O Bandwidth	Max. 1Gb/s	Max. 1.2Gb/s
tBERS	3.5ms (Typ.)	3.5ms (Typ.)
tR	58us	45us
Program Throughput	57MB/s	82MB/s
Vcc	2.30V to 3.6V	2.35V to 3.6V
Vccq	1.8V	1.2V

Figure 13.4.1: Feature Summary.

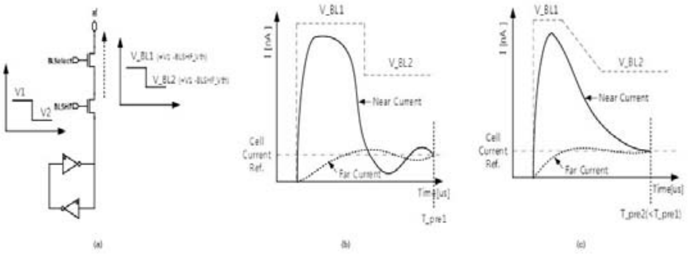
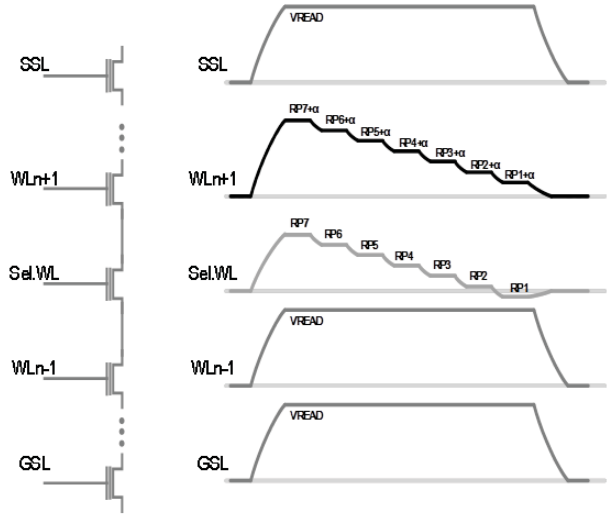


Figure 13.4.2: Enhanced BL precharge scheme.



< Fig. a >   < Fig. b >

Figure 13.4.3: Couple-capacitance-minimizing technique.

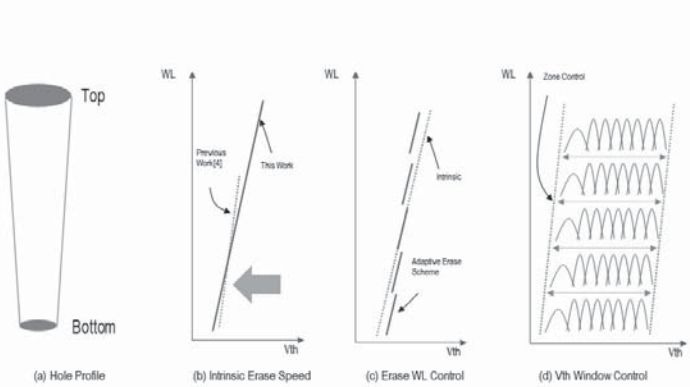
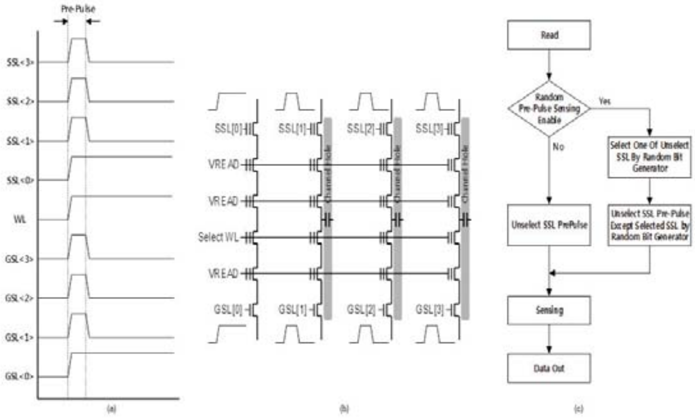


Figure 13.4.4: Progressive V_{th} window scheme.



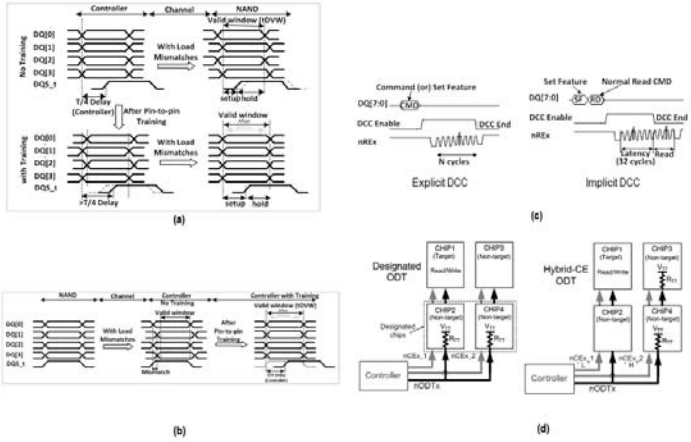


Figure 13.4.5: Random pre-pulse sensing scheme.

Figure 13.4.6: Toggle 4.0 (a) write training (b) write training (c) Explicit & Implicit DCC Training (d) designated and hybrid-CE ODT.

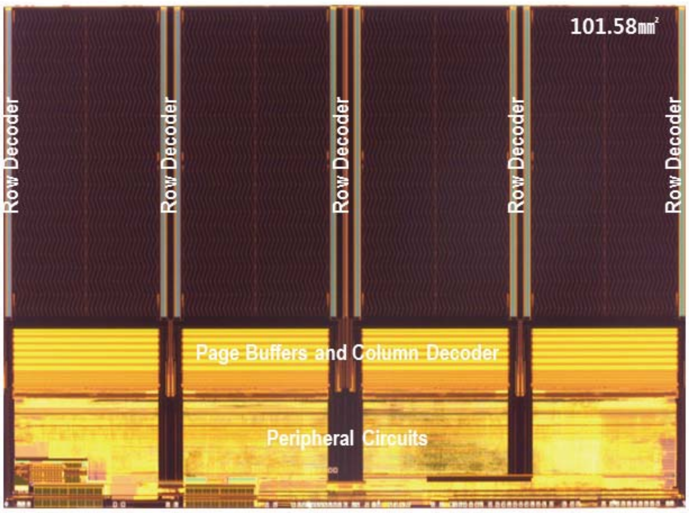


Figure 13.4.7: Die photograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE