Review

Recent Progress on 3D NAND Flash Technologies

Akira Goda

Micron Memory Japan, Tokyo 144-0052, Japan; agoda@micron.com

Abstract: Since 3D NAND was introduced to the industry with 24 layers, the areal density has been successfully increased more than ten times, and has exceeded 10 Gb/mm^2 with 176 layers. The physical scaling of XYZ dimensions including layer stacking and footprint scaling enabled the density scaling. Logical scaling has been successfully realized, too. TLC (triple-level cell, 3 bits per cell) is now the mainstream in 3D NAND, while QLC (quad-level cell, 4 bits per cell) is increasing the presence. Several attempts and partial demonstrations were made for PLC (penta-level cell, 5 bits per cell). CMOS under array (CuA) enabled the die size reduction and performance improvements. Program and erase schemes to address the technology challenges such as short-term data retention of the charge-trap cell and the large block size are being investigated.

Keywords: 3D NAND; floating gate cell; charge-trap cell; CMOS under array

1. Introduction

After 2D NAND reached the scaling limit around 15 nm in process node, 3D NAND was proposed as a solution for the continuous NAND scaling [1]. The 3D NAND was introduced into production with 24 layers and MLC technology [2]. The scaling trend of the areal density of 2D NAND and 3D NAND is summarized based on the NAND publications in IEEE ISSCC conferences, (Figure 1). As seen in Figure 1, 3D NAND successfully replaced 2D NAND and has achieved more than 10 Gb/mm^2 areal density [2–31].

At the early stage of the 3D NAND development, various types of 3D NAND technologies were proposed. A comprehensive survey can be found in [32]. After extensive research on various technology options, vertical NAND string architecture with gate-all-around (GAA) cells was introduced to the industry for floating gate (FG) cells and charge-trap cells [2,33].

In this paper, the status of 3D NAND scaling is reviewed and discussed by using the ISSCC conference publications as a reference. This review mainly focuses on the vertical 3D NAND technology adopted in the industry. The physical and logical scaling of 3D NAND will be discussed. The recent progress and topics since the prior review of 3D NAND [34] will be reviewed, including progress on the QLC technology and beyond.

Figure 1. Areal density trend of NAND flash chip presented in ISSCC conferences since 2008 [2–31]. Adapted with permission from ref. [34], Copyright 2020 IEEE.

2. 3D NAND Architecture and Operations

Figure 2 shows the 3D NAND cell array architectures. The strings are placed in the vertical direction. Word lines (WLs) have a plate-like shape and are stacked vertically for 3D cell stacking. There are multiple select gates at the drain side (SGDs) in a block. The channel of the NAND string has a cylinder shape.

(**a**)   (**b**)
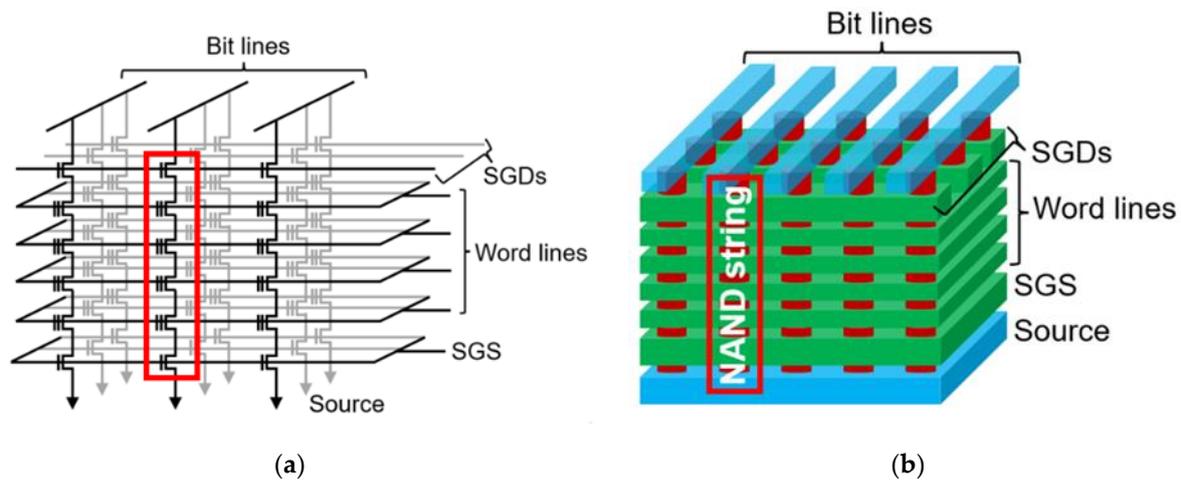
Figure 2. 3D NAND array architecture [35]. (**a**) Schematics of 3 × 3 NAND strings and (**b**) bird's eye view of 3 × 5 NAND strings. Only 4 word lines (WLs) in a string are shown for visibility. The actual 3D NAND has more than 128 WLs in a string.

A block is a unit of the erase operation. As shown in Figure 3, there are two types of erase methods in 3D NAND—the body erase (Figure 3a) and the GIDL erase (Figure 3b) [1,36]. In the body erase, NAND strings are connected to the Si-substrate, and holes are supplied to the NAND string from the Si-substrate, enabling the positive body potential required for erase. In the GIDL erase, the NAND strings are de-decoupled from the Si-substrate and formed on the N+ source layer instead. During erase, the electron-hole pairs are generated at source and drain N+ junctions by GIDL mechanism to supply holes to the NAND strings. The GIDL erase is used for the CMOS under Array (CuA) technology, which will be discussed later [33].

(**a**)   (**b**)
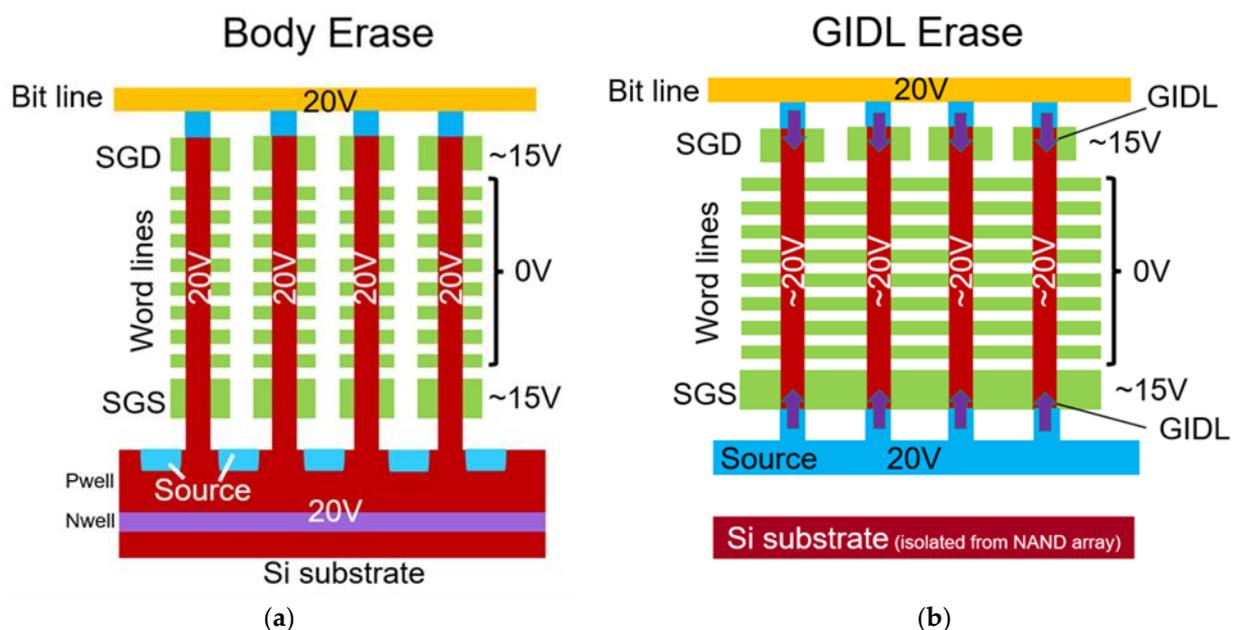
Figure 3. Erase schemes of 3D NAND [35]. (**a**) Body erase scheme directly biases the channels at erase voltage. Holes are supplied from the Si-substrate. (**b**) GIDL erase scheme generates electron-hole pairs at the source and drain junctions. The generated holes are supplied to the channels to boost the potential to around the erase voltage.

During program and read, one of the SGDs is selected in a block so that one NAND string can be selected per bit line (Figure 4).
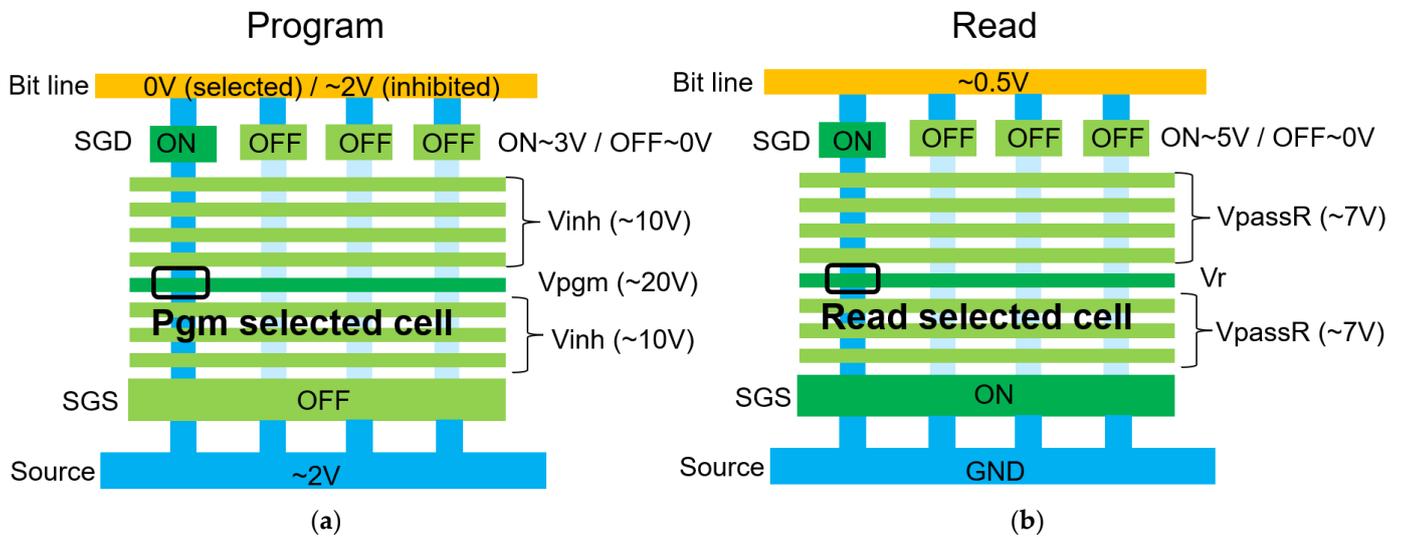
Figure 4. (**a**) Program and (**b**) read operations of 3D NAND. One cell per bit line is selected by selecting one SGD and one WL [35].

3. Floating Gate NAND and Replacement Gate NAND

3.1. Architectures of FG NAND and RG NAND

The floating gate (FG) cell technology was used in 2D NAND. In 3D NAND, in addition to the FG technology (FG NAND), replacement gate cell technology (RG NAND) is also utilized [33,36]. Figure 5 compares the cross-sections of the NAND strings for FG NAND and RG NAND.

In FG NAND, the FG storages are separated between cells because the FG is made of the conductive polysilicon material. The advanced cell integration scheme is adopted to realize the FG separation [33]. In this scheme, the cells are formed outside the pillar holes. Therefore, the diameter of the pillar etching needs to be aligned with the final channel diameter of the NAND string.
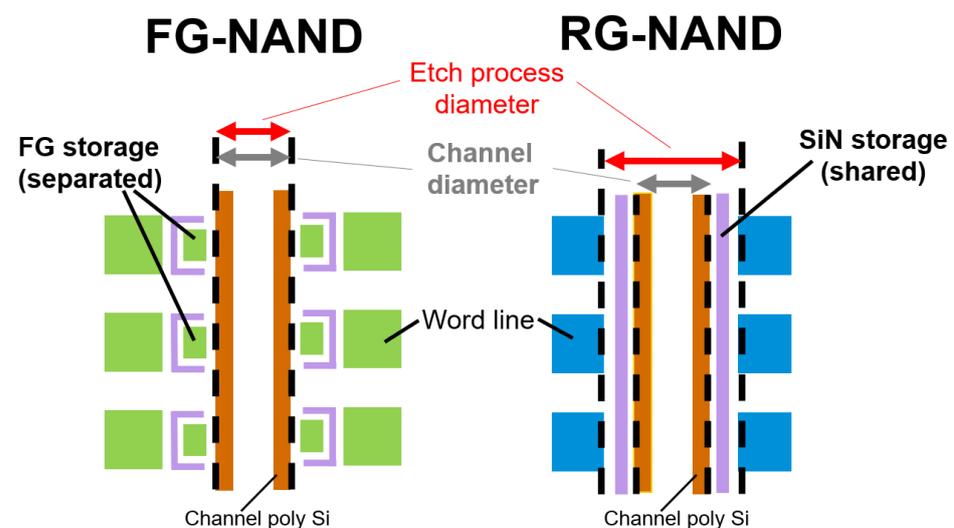
Figure 5. Cross-section comparison of NAND strings between floating gate (FG) NAND and replacement gate (RG) NAND. RG NAND has a larger diameter for the memory hole etching which is advantageous for the pillar height scaling.

In RG NAND, charge-trap cells with a silicon-nitride (SiN) storage layer are employed [2]. Because the SiN storage is a dielectric material which can trap charges, the storage layer can be shared and continuous among the cells. The SiN storage film and other films composing the charge-trap cells are formed after the dry etching of the pillar holes. Therefore, the diameter of the pillar holes for etching can be larger than that of the final channel diameter by the thicknesses of the cell dielectrics. This is advantageous for the pillar etching, especially for the very tall pillar when many cells are stacked vertically. The word line is formed by tungsten metal by replacing the SiN films stacked originally [36]. Therefore, the technology is called replacement gate NAND. RG NAND is the combination of the replacement gate technology and the charge-trap cell technology.

3.2. Band-Engineered Tunneling Dielectrics of the Charge-Trap Cell

In the FG cell, electrons are injected to or emitted from the FG by Fowler–Nordheim (FN) Tunneling (Figure 6a). In the charge-trap cells, the programming is similar to FG cells, where the electrons are injected to the SiN storage by modified FN tunneling. For erase, holes are injected to the SiN storage by direct tunneling (DT) (DT erase) (Figure 6b). In the charge-trap cell, in order to enhance erase efficiently, the cell stack is engineered as shown in Figure 7a. First, the band engineered (BE) tunnel directrices has been introduced [37]. As BE-tunnel dielectrics, an ONO stacked film or a film with an engineered nitrogen profile can be used instead of the SiO_2 tunnel layer. With BE-tunnel dielectrics, holes can tunnel only the thin oxide layer during the erase, while the full ONO thickness can be utilized during the retention. Second, the High-k/Metal gate is used to reduce the unwanted electron injection from the control gate [38]. By combining the BE-tunnel layer and the High-k/Metal gate, good erase characteristics can be achieved with the charge-trap cell (Figure 7b).
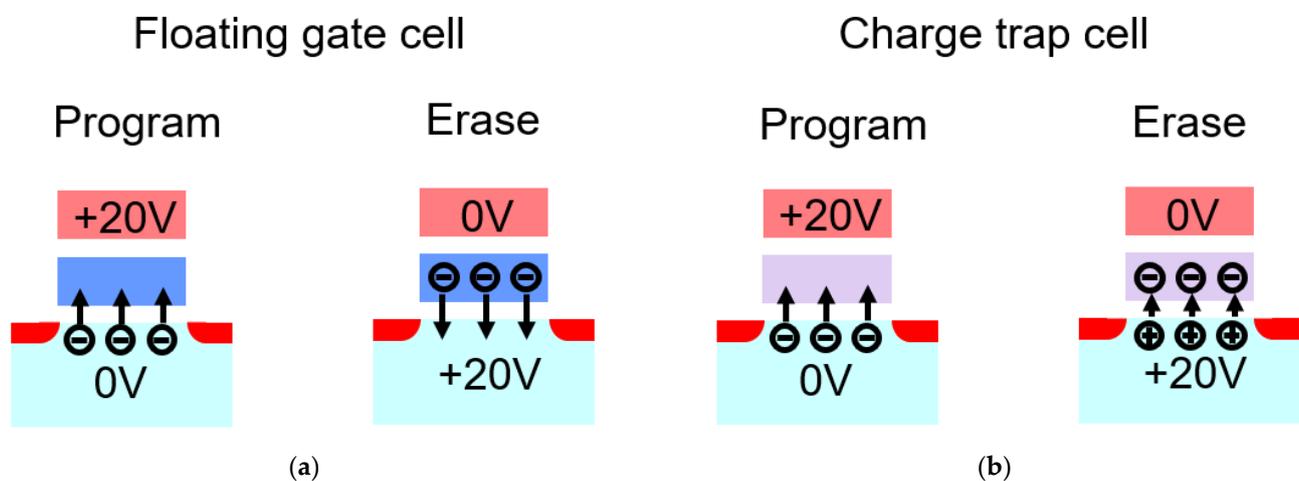
Figure 6. Program and erase operations for (**a**) floating gate cell and (**b**) charge-trap cell. In the charge-trap cell, holes are injected to the storage layer by direct tunneling.

3.3. Data Retention Mechanisms of the Charge-Trap Cell

The short-term data retention has been a challenge for the charge-trap cell [38]. Figure 8 shows various mechanisms potentially causing the short-term data retention [39]. The first mechanism is charge migration and relaxation. After the programming, the trapped charge in SiN storage could move both laterally (lateral migration, LM) or vertically (vertical relaxation, VR). The second mechanism is detrapping from the SiN storage layer by trap-assisted tunneling (TAT). On top of these, there is trapping at the BE-tunnel oxide as it includes SiN or a nitrogen-rich layer. Due to the very short distance between the trapping sites at BE and the poly Si channel, the detrapping can occur in a very short time.

The impacts on the Vth distribution and programming algorithms solutions for the short-term data retention will be discussed in a later section.
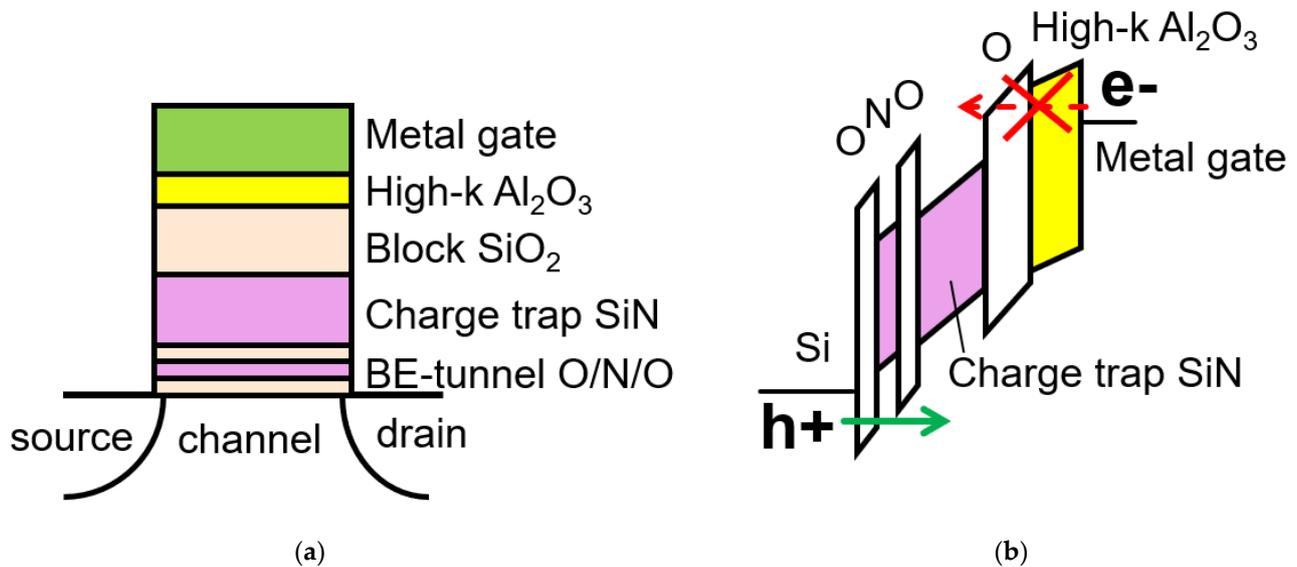
Figure 7. (**a**) Cell film stack of the charge-trap cell. (**b**) Band diagram during erase of charge-trap cell. Band-engineered (BE) tunnel dielectrics enhances hole injections while High-k/Metal gate prevents electron back tunneling [35].
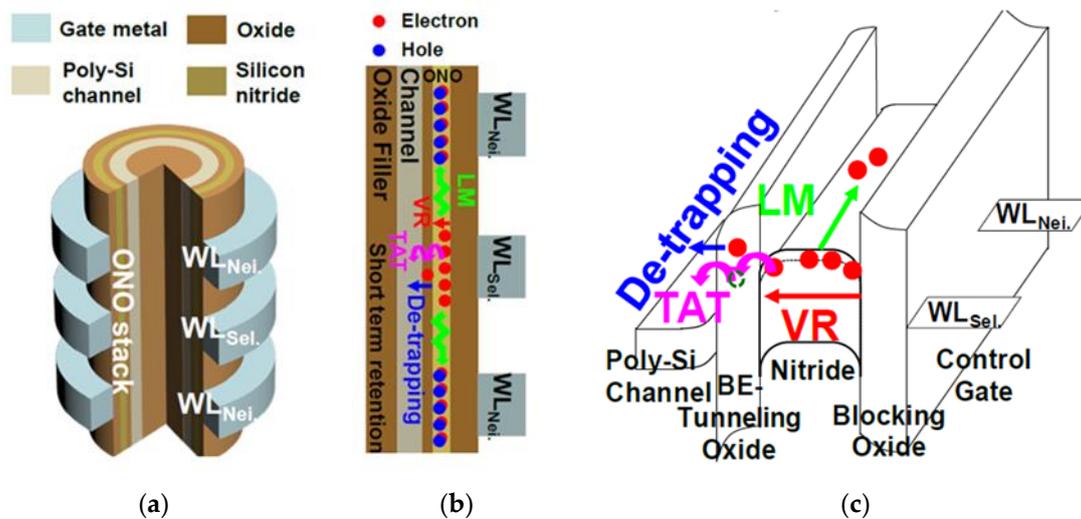
Figure 8. (**a**) The bird's eye view and (**b**) cross-sectional view of 2D NAND with charge-trap cells. (**c**) Energy band diagram schematic for short-term retention and possible retention mechanisms. Reprinted with permission from ref. [40], copyright 2020 IEEE.

4. 3D NAND Array Physical Scaling by Process Integration

4.1. Conventional Scaling

The physical scaling of the 3D NAND array can be described as XYZ scaling, as shown in Figure 9. XY scaling means reducing the cell footprint. Z scaling means stacking more layers, which is often done together with layer pitch shrink (Z shrink) in order to minimize the increase in the physical height. Z scaling (stacking) has been the main scaling enabler so far. In the ISSCC publications, the steady progress of the layer stacking has been shown, except for in 2020, when the focus was on circuit design technologies for single-level-cells (SLC) and quad-level-cells (QLC) rather than the physical array scaling. In the latest achievement, 176-layer stacked 3D NAND has been demonstrated in both publication and mass production (Figure 10).
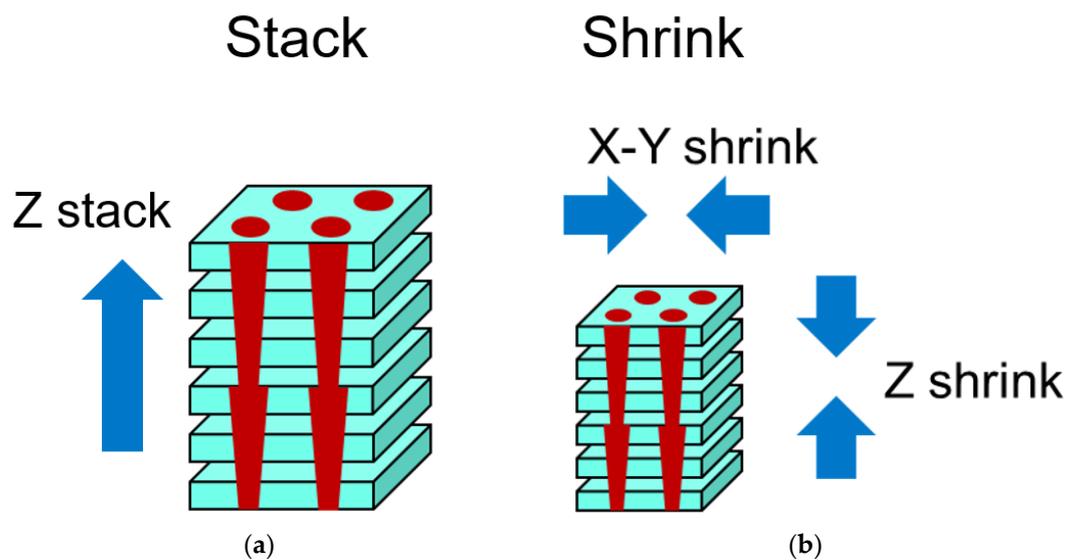
Figure 9. XYZ scaling of 3D NAND. (**a**) Layer stacking in Z-direction. (**b**) Footprint shrink of XY dimensions and layer pitch reduction as Z-shrink. Z-shrink is often combined with Z-stack to minimize the increase in the pillar height.
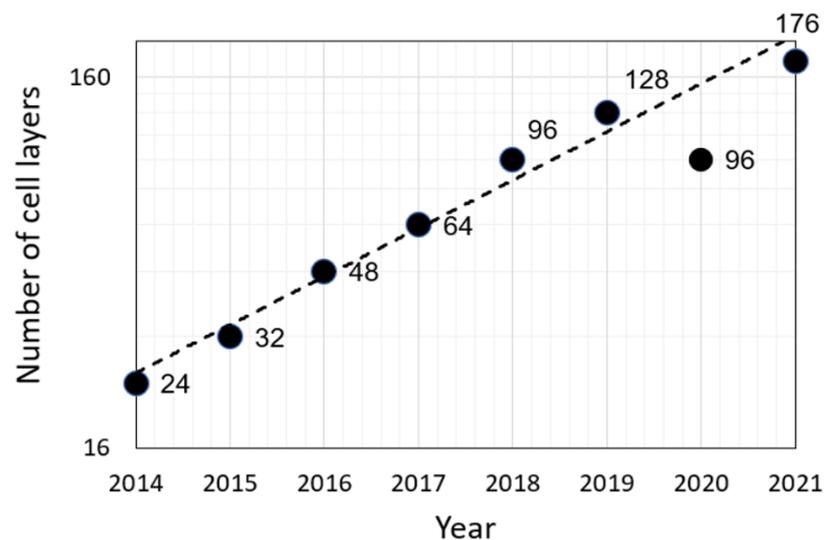
Figure 10. The year-by-year trend of cell layer stacking in 3D NAND published in ISSCC conferences. The reverse trend in 2020 is because the publications are about SLC and QLC technologies on the 96-layer NAND baseline.

The effort for XYZ scaling has been focused on efficiency improvement of the pillar (memory hole) layout [41]. Due to the various layout space requirements such as source line contacts, SGD-to-SGD separations, block-to-block separations, the pillar layout is far from the ideal hexagonal close-packed (HCP) layout. There have been continuous improvements in the array layout so that the pillar arrangement has become closer to HCP, enabling XY scaling (Figure 11).
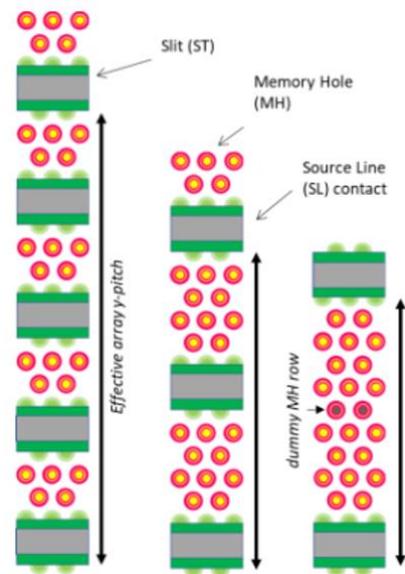
Figure 11. XY scaling by pillar (memory hole) layout efficiency improvement. The pillar layout approaches HPC configuration. Reprinted with permission from ref. [41], Copyright 2020 IEEE.

4.2. Disruptive Scaling

As a disruptive XY scaling, the split cells have been investigated (Figure 12). In the conventional 3D NAND, the cell has a cylinder shape. In the split cell, the cell is split into two parts so that the cell density increases. There are two different types of the split cell proposals. One is a planar-like split cell [42] and the other is a half-cylindrical cell [43,44]. The planar-like cell is similar to the 2D NAND cell, while the half-cylindrical cell can be seen as an evolution of the 3D NAND cylindrical cell. For both split cells, the challenges are the process integration of the cell split, the increased cell-to-cell interference and the reduced gate-coupling ratio.

(**a**)   (**b**)
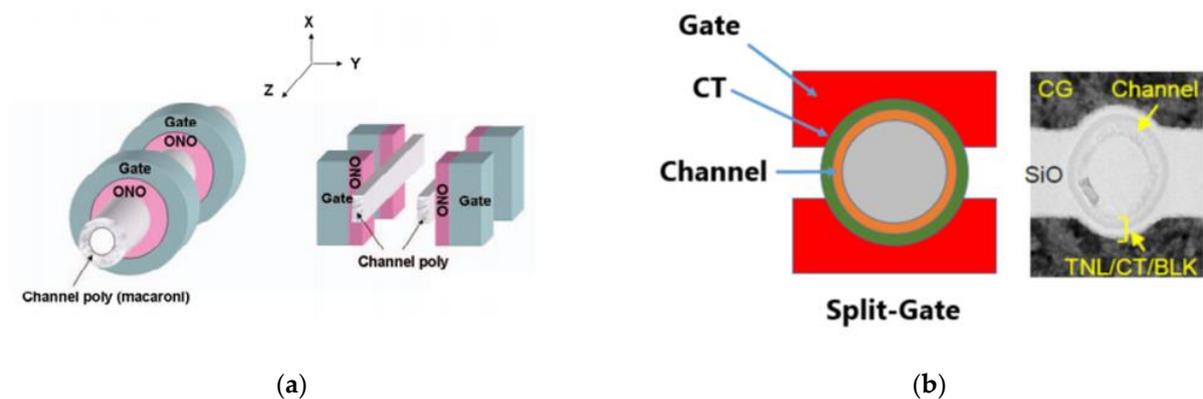
Figure 12. Split cells with (**a**) planar cell and (**b**) semicylindrical cell. Both of these achieve physical density scaling. The extra cell-to-cell interference needs to be managed to achieve good cell characteristics and high reliability. Reprinted with permission from refs. [42,44], Copyright 2015 and 2019 IEEE, respectively.

As discussed earlier, the pitch shrink of the WL layer is critical to manage the pillar height with layer stacking. In the charge-trap cell, the SiN storage layer is continuous between neighboring cells. With the WL pitch scaling, the trapped charge migration between the neighboring cells would raise the reliability concern. In order to overcome this issue, the SiN storage separation has been proposed (Figure 13) [45]. The process flow is similar to that of 3D FG NAND where the diameter of the pillar etching is smaller

compared to the conventional 3D RG NAND. Therefore, the process integration needs to be well-managed for the successful implementation of the SiN storage separation.
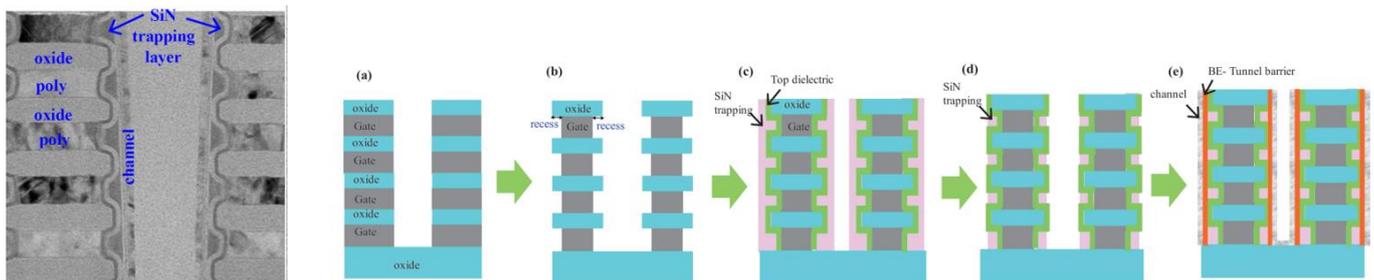
Figure 13. Charge-trap cell with the confined SiN storage layer and the process flow. The SiN storage is successfully separated between cells. The pillar etching diameter needs to be smaller compared to the conventional charge-trap cell. Reprinted with permission from ref. [45], Copyright 2019 IEEE.

5. 3D NAND Array Logical Scaling by Cell Device Engineering

In addition to the physical cell density scaling, the logical density scaling (i.e., more bits per cell scaling) has been actively pursued in 3D NAND.

Figure 14 shows the number of 3D NAND publications in ISSCC conferences. 3D NAND started as MLC technology and rapidly transitioned to TLC, owing to the excellent cell characteristics and reliability. Recently, QLC technology has been introduced to mass production and the presence of the QLC technology has also been increasing in the publications. Currently, TLC is the mainstream for high-performance and high-endurance usages. QLC is becoming mainstream for the high-density and low-cost usages.
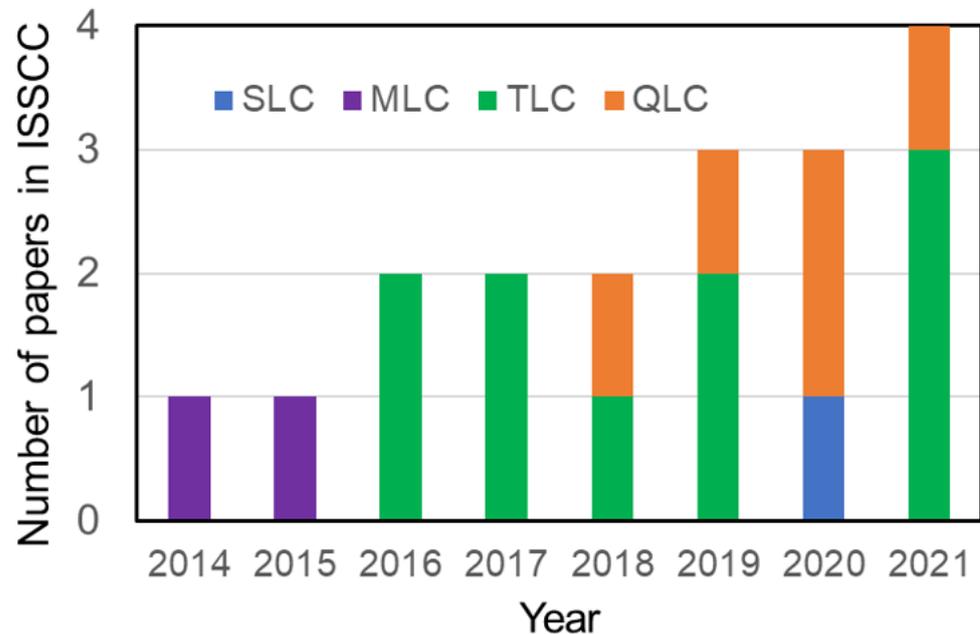
Figure 14. 3D NAND publications in ISSCC for SLC, MLC, TLC and QLC technologies. TLC has been the mainstream for 3D NAND. The interest in QLC is rapidly increasing.

Recently, "beyond QLC" efforts have been reported for both FG and charge-trap cells. PLC (5 bits per cell) distributions were shown with FG cells (Figure 15a) [46]. In this work, the excellent data-retention properties of the FG cell were identified as critical to enable PLC. HLC (6 bits per cell) Vth distributions were experimentally shown at a cryogenic temperature of 77 K for charge-trap cells (Figure 15b) [47]. In that work, it was shown that random telegraph noise (RTN) improves at the 77 K while it degrades at 300 K relative

to 358 K. Performance and reliability characteristics as well as the operational conditions warrant further study to enable further bit-per-cell scaling beyond QLC.

(**a**)　　　　　　　　　　　　　　　　　　　　　　(**b**)
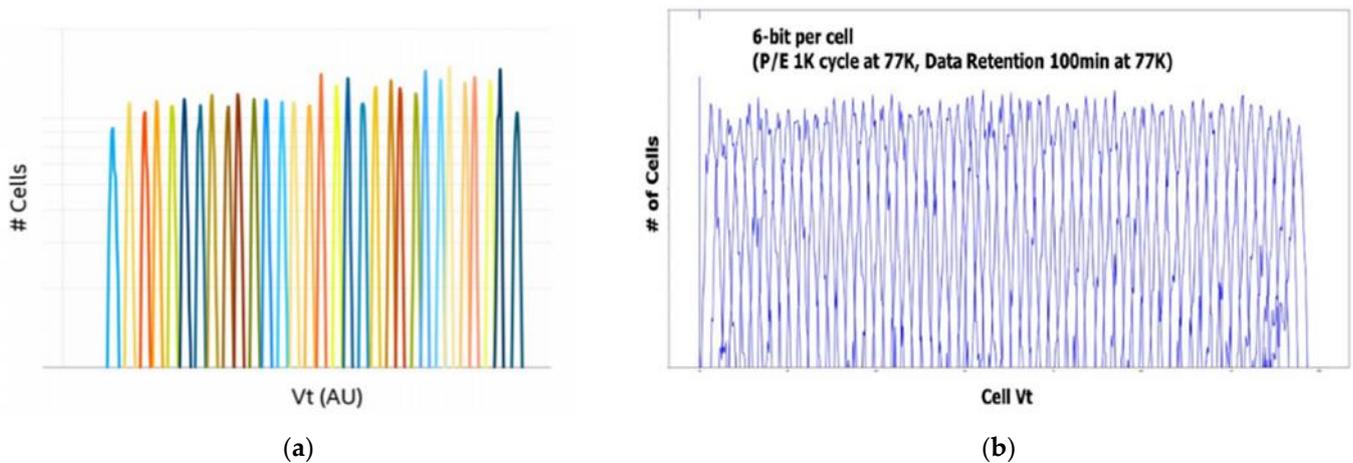
Figure 15. (**a**) PLC (5 bits per cell) Vth distribution with FG cell. (**b**) HLC (6 bits per cell) Vth placement with charge-trap cell at cryogenic temperature of 77 K. Reprinted with permission from refs. [46,47], Copyright 2020 and 2021 IEEE, respectively.

6. 3D NAND Performance Scaling by Design and Algorithms

6.1. Write Bandwidth for TLC

Write bandwidth is provided by

$$Page\ size \times number\ of\ planes/tProg \tag{1}$$

Therefore, the large write parallelism (= *page size × number of planes*) and short *tProg* are critical to achieve high write bandwidth. Figure 16 shows the trend of TLC write bandwidth from the ISSCC publications. There has been incremental improvement over the years and a steep increase in recent years.
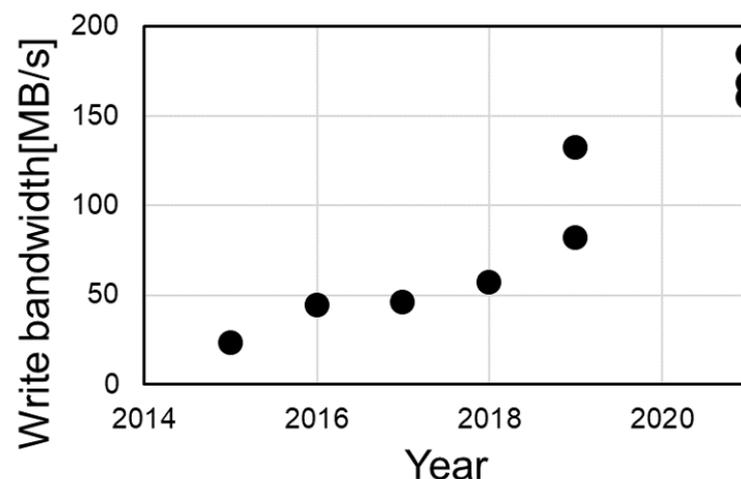
Figure 16. TLC write bandwidth trend reported in ISSCC conferences.

The former is due to the continuous improvement of TLC tProg. Figure 17 is the TLC tProg trend published or estimated from ISSCC publications. In TLC RG NAND, all seven programmed states are programmed in a single programming pass. The improvements in tProg are realized by combinations of WL and BL bias time reduction, fine tuning of the programming voltage compensating cell characteristics' variability across the pillar and the reduction in the program verify operations. Figure 18 shows the center XDEC (WL-driver) architecture which shortens the time for WL loading [31].
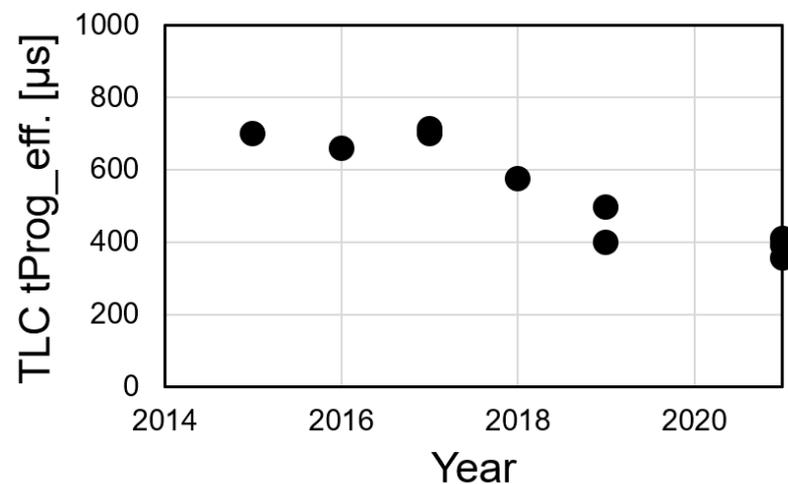
Figure 17. TLC effective tProg reported or calculated from ISSCC publications.

(a)　　　　　　　**(b)**
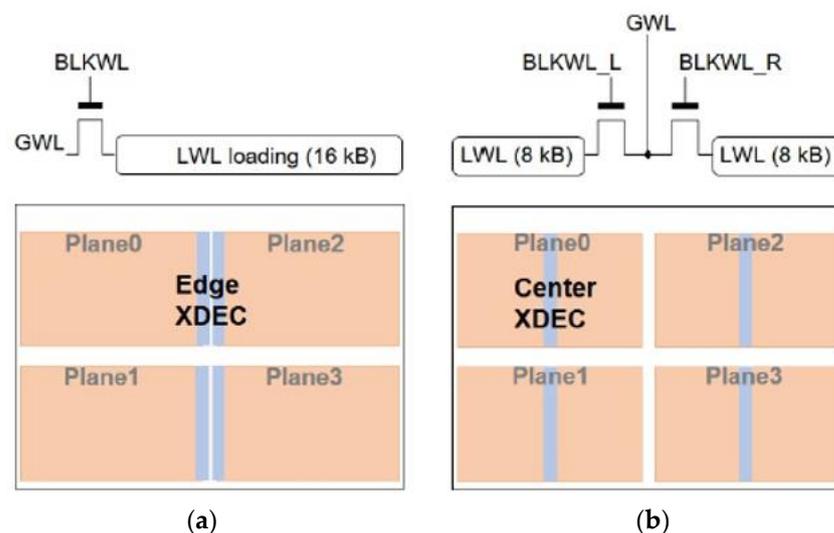
Figure 18. XDEC (WL driver) architecture. (**a**) XDEC is placed at the edge of the local WL. (**b**) XDEC is placed at the center of the local WL, reducing the LWL load to a half for faster operations. Reprinted with permission from ref. [31], Copyright 2021 IEEE.

The latter (the steep increase in the write bandwidth) is realized by the increased number of planes owing to the CMOS under Array (CuA) architecture. Figure 19 shows three different CMOS architectures utilized in 3D NAND. In CMOS outside array (CoA), the CMOS circuits are placed next to the array. Therefore, the die size increases if the CMOS area increases. To increase the parallelism, a larger amount of CMOS circuits such as page buffer circuits need to be placed, which results in the increase in the die size. In the CMOS under Array (CuA), the CMOS circuits are placed under the array. More parallelism (more planes) can be realized with CuA because the larger area is available for CMOS circuits. Another variation is to place the CMOS over array by using wafer bonding technologies (Wafer on wafer, WoW) [48]. With this architecture, CMOS is processed separately from the array processing by using a dedicated wafer for CMOS. Therefore, the process flow can be optimized for CMOS devices and interconnect.
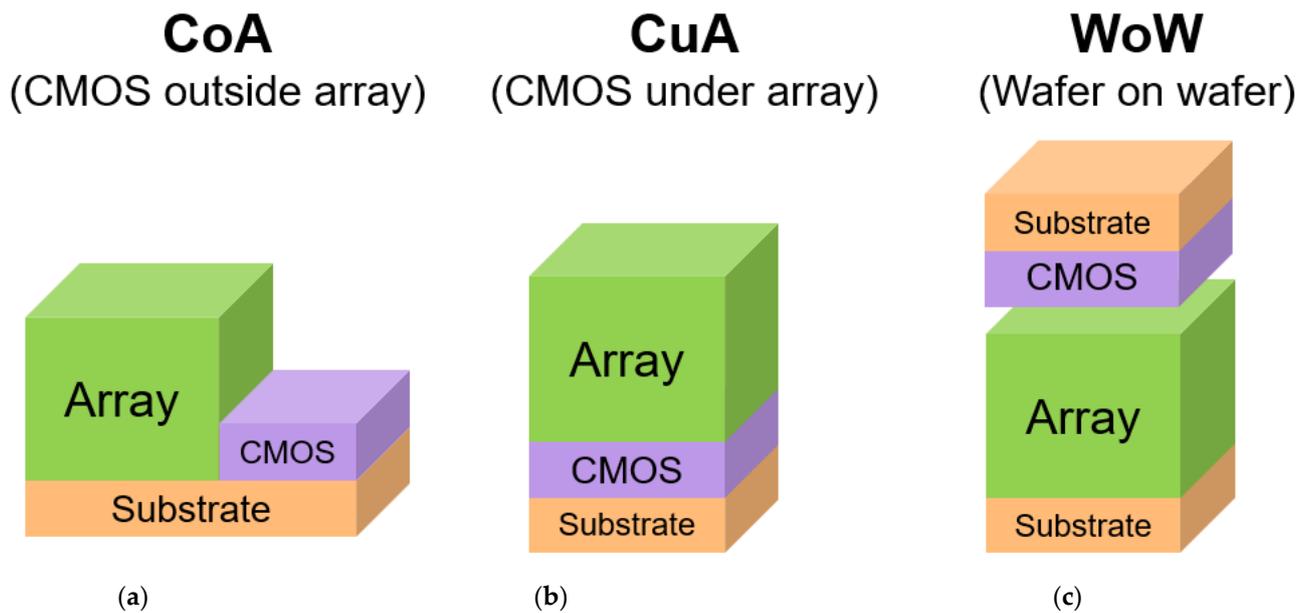
Figure 19. 3D NAND architectures for CMOS placement. (**a**) CMOS outside array (CoA); (**b**) CMOS under Array (CuA), (**c**) Wafer on wafer bonding (WoW).

Figure 20 shows the number of publications in ISSCC for CNA and CuA. After CuA was presented in 2016 for the first time in an ISSCC conference, publications with CuA have been increasing, and most of the recent 3D NAND publications are for CuA. CuA technology is now the mainstream for 3D NAND.
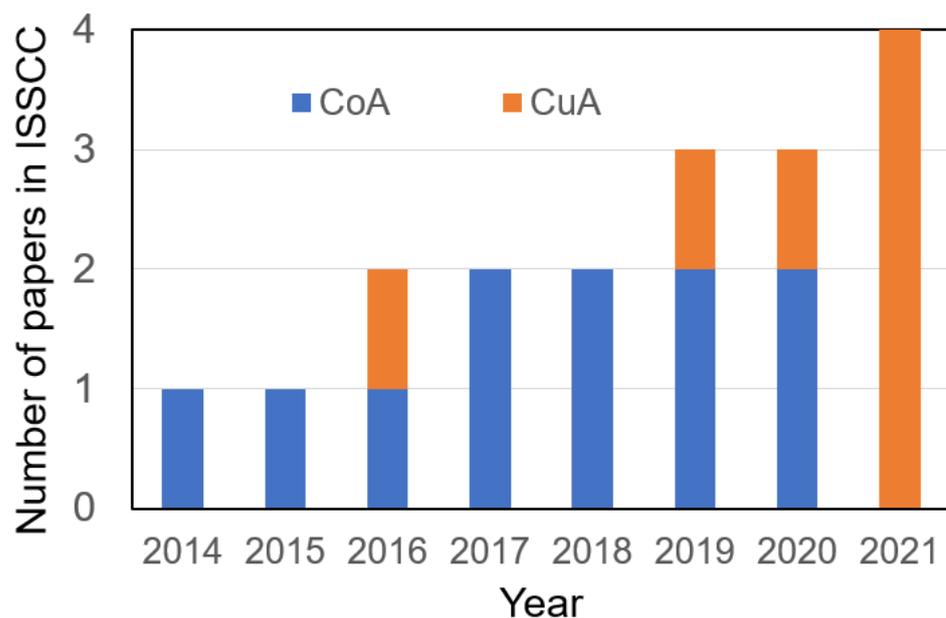
Figure 20. 3D NAND publications in ISSCC for CMOS outside array (CoA) and CMOS under array (CuA) architectures. All publications in 2021 are for CuA.

6.2. QLC Program Schemes

QLC tProg is much longer than TLC tProg (as shown in Figure 21), in order to realize the tight Vth distributions. As discussed earlier, in RG NAND with the charge-trap cells, there is a phenomenon known as short-term data retention. This causes shift and widening of Vth distributions right after programming. To realize the tight Vth distributions for QLC, it is important to manage the short-term data retention effects.
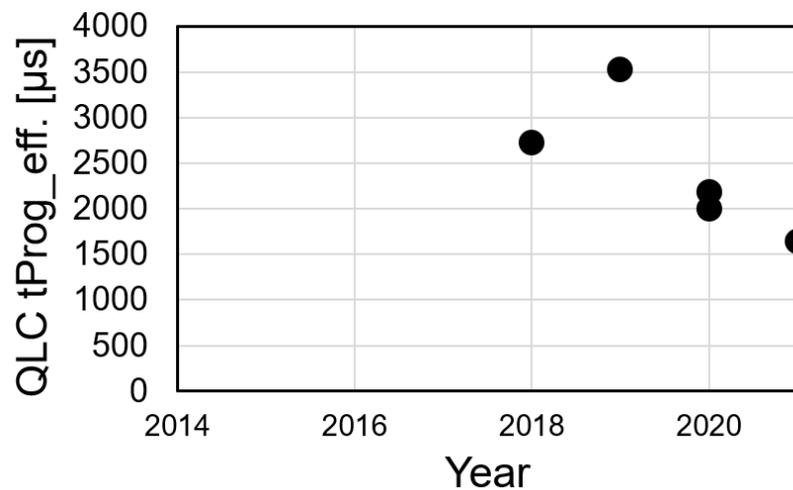
Figure 21. QLC effective tProg reported or calculated from ISSCC publications.

The coarse–fine programming scheme has been introduced for QLC RG NAND [26]. As shown in Figure 22a, all 16 levels are programmed in the first pass with relatively wide distribution widths. Due to the short-term data retention, the Vth distributions shift and widen right after the coarse programming. The fine programming is performed as the second programming path, which tightens the Vth distributions by touching up the distribution tails caused by the short-term data retention. This scheme can be called the 16-16 scheme as the 16-level programming is performed twice. To reduce QLC tProg, an 8-16 scheme was proposed (Figure 22b). With the 8-16 scheme, eight levels are programmed at the coarse programming pass, contributing to the tProg reduction.

| (**a**) | (**b**) |
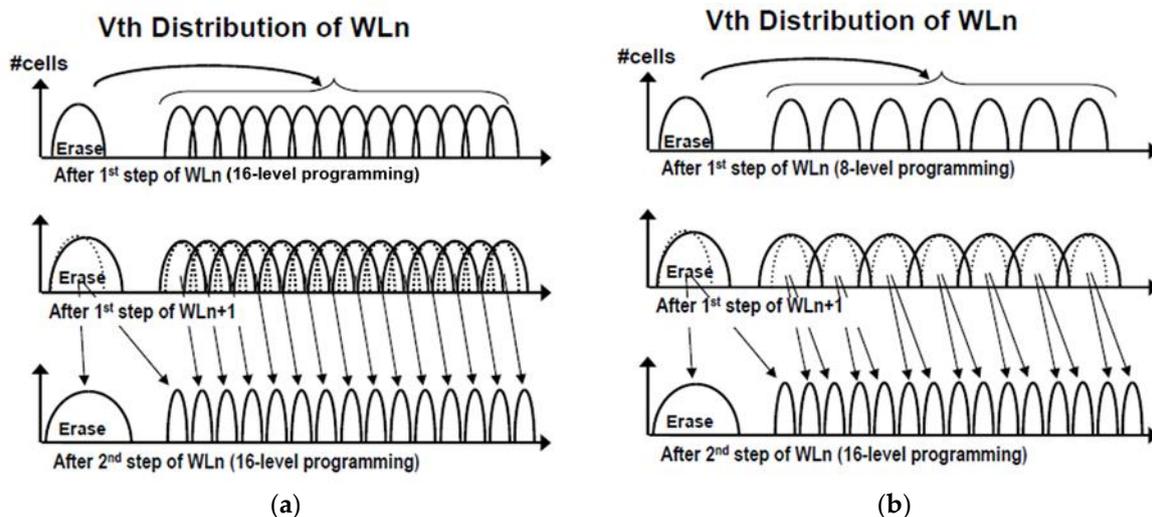
Figure 22. QLC coarse–fine programming schemes with (**a**) 16-16 scheme and (**b**) 8-16 scheme. 16-16 is more efficient to compensate the short-term retention while the 8-16 scheme would have tProg advantage. Reprinted with permission from ref. [26], Copyright 2019 IEEE.

In the FG NAND, the first programming pass only has four levels. The 16 levels are then completed at the second pass [30] because the short-term retention is much smaller in the FG cell and does not require the touch-up operation.

So far, 1.6–2 msec tProg has been reported for QLC and ~0.4 msec tProg has been reported for TLC. For the further enhancement of QLC tProg, solving the short-term data retention by programming scheme and cell improvement is critical.

6.3. Block Size Scaling

With the WL stacking increasing, the block size increases. Figure 23a shows the TLC block size as a function of number of WL stacking. The number of SGDs per block is chosen as a parameter. Typically, FG NAND runs 12–16 SGDs/block, while RG NAND has 4–8 SGDs/blocks. Given that the block size is the minimum granularity of erase, the increase in the block size could increase the system burden of the data management and would degrade system performance. In order to mitigate this problem, the block-by-deck scheme was proposed (Figure 23b) [30]. In this scheme, the NAND string is divided into multiple segments (three segments in this example) and each segment is treated as a different block. During erase operation, the entire pillar is biased to the erase voltage. The WLs of the selected deck block are grounded while the WLs of the unselected deck blocks are ramped to the channel potential. In this way, the cells in the unselected deck blocks can be inhibited for erase.

(**a**)   (**b**)
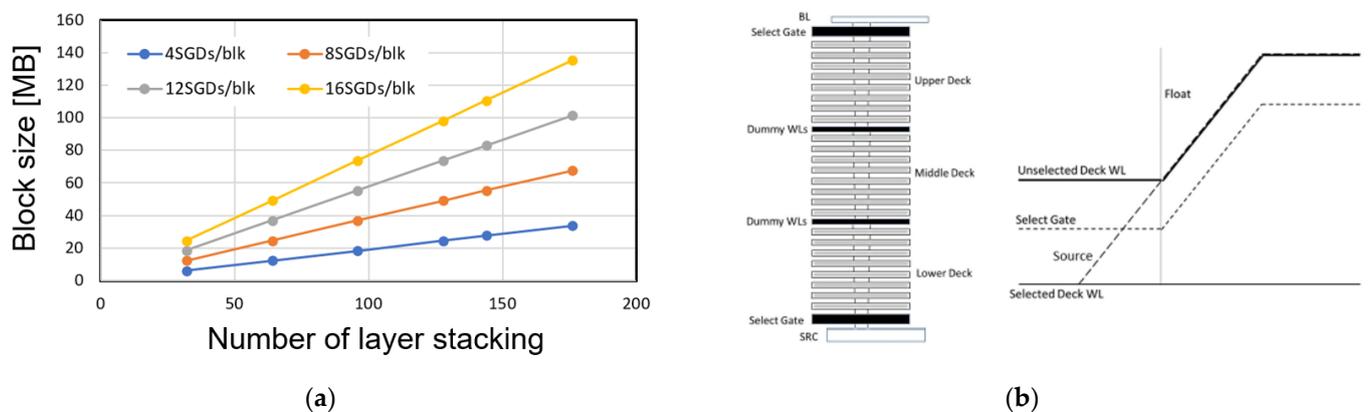
Figure 23. (**a**) Block size scaling as a function of number of layer stacking. (**b**) Block-by-deck erase scheme. The conventional physical block is divided into three logical blocks [30]. Reprinted with permission from ref. [23], Copyright 2021 IEEE.

7. Conclusions

3D NAND scaling has been successfully achieved. The layer stacking has reached 176 layers. QLC has been introduced in both FG NAND and RG NAND. Research on PLC is active, with partial demonstrations for cell capability. CMOS under array (CuA) has been widely adopted and enables performance enhancement by increasing the number of planes. For future scaling, on top of the continuous XYZ physical scaling, disruptive technologies such as split cells are proposed. Program and erase schemes are being developed further to solve cell reliability challenges and block size scaling challenges.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Tanaka, H.; Kido, M.; Yahashi, K.; Oomura, M.; Katsumata, R.; Kito, M.; Fukuzumi, Y.; Sato, M.; Nagata, Y.; Matsuoka, Y.; et al. Bit Cost Scalable technology with Punch and plug process for ultra high density flash memory. In Proceedings of the 2007 IEEE Symposium on VLSI Technology, Kyoto, Japan, 12–14 June 2007; pp. 14–15. [CrossRef]
2. Park, K.T.; Han, J.M.; Kim, D.; Nam, S.; Choi, K.; Kim, M.S.; Kwak, P.; Lee, D.; Choi, Y.H.; Kang, K.M.; et al. Three-dimensional 128 Gb MLC vertical NAND Flash-memory with 24-WL stacked layers and 50 MB/s high-speed programming. In Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, 9–13 February 2014; Volume 57, pp. 334–335. [CrossRef]
3. Cernea, R.; Pham, L.; Moogat, F.; Chan, S.; Le, B.; Li, Y.; Tsao, S.; Tseng, T.Y.; Nguyen, K.; Li, J.; et al. A 34 MB/s-program-throughput 16 Gb MLC NAND with all-bitline architecture in 56 nm. In Proceedings of the 2008 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 3–7 February 2008; Volume 51, pp. 420–624. [CrossRef]

4. Li, Y.; Leo, S.; Fong, Y.; Pan, F.; Kuo, T.C.; Park, J.; Samaddar, T.; Nguyen, H.; Mui, M.; Htoo, K.; et al. A 16 Gb 3 b/Cell NAND flash memory in 56 nm with 8 MB/s write rate. In Proceedings of the 2008 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 3–7 February 2008; Volume 51, pp. 506–508. [CrossRef]

5. Kanda, K.; Koyanagi, M.; Yamamura, T.; Hosono, K.; Yoshihara, M.; Miwa, T.; Kato, Y.; Mak, A.; Chan, S.L.; Tsai, F.; et al. A 120 mm^2 16 Gb 4-MLC NAND flash memory with 43 nm CMOS technology. In Proceedings of the 2008 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 3–7 February 2008; Volume 51, pp. 430–625. [CrossRef]

6. Zeng, R.; Chalagalla, N.; Chu, D.; Elmhurst, D.; Goldman, M.; Haid, C.; Huq, A.; Ichikawa, T.; Jorgensen, J.; Jungroth, O.; et al. A 172 mm^2 32 Gb MLC NAND flash memory in 34 nm CMOS. In Proceedings of the 2009 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 8–12 February 2009; pp. 236–237. [CrossRef]

7. Trinh, C.; Shibata, N.; Nakano, T.; Ogawa, M.; Sato, J.; Takeyama, Y.; Isobe, K.; Le, B.; Moogat, F.; Mokhlesi, N.; et al. A 5.6 MB/s 64 Gb 4 b/Cell NAND flash memory in 43 nm CMOS. In Proceedings of the 2009 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 8–12 February 2009; pp. 246–248. [CrossRef]

8. Futatsuyama, T.; Fujita, N.; Tokiwa, N.; Shindo, Y.; Edahiro, T.; Kamei, T.; Nasu, H.; Iwai, M.; Kato, K.; Fukuda, Y.; et al. A 113 mm^2 32 Gb 3 b/cell NAND flash memory. In Proceedings of the 2009 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 8–12 February 2009; Volume 1, pp. 242–243. [CrossRef]

9. Lee, C.; Lee, S.K.; Ahn, S.; Lee, J.; Park, W.; Cho, Y.; Jang, C.; Yang, C.; Chung, S.; Yun, I.S.; et al. A 32 Gb MLC NAND-flash memory with Vth-endurance-enhancing schemes in 32 nm CMOS. In Proceedings of the 2010 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 7–11 February 2010; Volume 53, pp. 446–447. [CrossRef]

10. Marotta, G.G.; Macerola, A.; D'Alessandro, A.; Torsi, A.; Cerafogli, C.; Lattaro, C.; Musilli, C.; Rivers, D.; Sirizotti, E.; Paolini, F.; et al. A 3 bit/Cell 32 Gb NAND flash memory at 34 nm with 6 Mb/s program throughput and with dynamic 2 b/cell blocks configuration mode for a program throughput increase up to 13 MB/s. In Proceedings of the 2010 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 7–11 February 2010; Volume 53, pp. 444–445. [CrossRef]

11. Kim, H.; Park, J.H.; Park, K.T.; Kwak, P.; Kwon, O.; Kim, C.; Lee, Y.; Park, S.; Kim, K.; Cho, D.; et al. A 159 mm^2 32 nm 32 Gb MLC NAND-flash memory with 200 MB/s asynchronous DDR interface. In Proceedings of the 2010 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 7–11 February 2010; Volume 53, pp. 442–443. [CrossRef]

12. Kim, T.Y.; Lee, S.D.; Park, J.S.; Cho, H.Y.; You, B.S.; Baek, K.H.; Lee, J.H.; Yang, C.W.; Yun, M.; Kim, M.S.; et al. A 32 Gb MLC NAND flash memory with Vth margin-expanding schemes in 26 nm CMOS. In Proceedings of the 2011 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 20–24 February 2011; pp. 202–203. [CrossRef]

13. Park, K.T.; Kwon, O.; Yoon, S.; Choi, M.H.; Kim, I.M.; Kim, B.G.; Kim, M.S.; Choi, Y.H.; Shin, S.H.; Song, Y.; et al. A 7 MB/s 64 Gb 3-bit/cell DDR NAND flash memory in 20 nm-node technology. In Proceedings of the 2011 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 20–24 February 2011; Volume 43, pp. 212–213. [CrossRef]

14. Fukuda, K.; Watanabe, Y.; Makino, E.; Kawakami, K.; Sato, J.; Takagiwa, T.; Kanagawa, N.; Shiga, H.; Tokiwa, N.; Shindo, Y.; et al. A 151 mm^2 64 Gb MLC NAND flash memory in 24 nm CMOS technology. In Proceedings of the 2011 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 20–24 February 2011; pp. 198–199. [CrossRef]

15. Lee, D.; Chang, I.J.; Yoon, S.Y.; Jang, J.; Jang, D.S.; Hahn, W.G.; Park, J.Y.; Kim, D.G.; Yoon, C.; Lim, B.S.; et al. A 64 Gb 533 Mb/s DDR interface MLC NAND Flash in sub-20 nm technology. In Proceedings of the 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 19–23 February 2012; Volume 55, pp. 430–431. [CrossRef]

16. Naso, G.; Botticchio, L.; Castelli, M.; Cerafogli, C.; Cichocki, M.; Conenna, P.; D'Alessandro, A.; De Santis, L.; Di Cicco, D.; Di Francesco, W.; et al. A 128 Gb 3 b/cell NAND flash design using 20 nm planar-cell technology. In Proceedings of the 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 17–21 February 2013; Volume 56, pp. 218–219. [CrossRef]

17. Helm, M.; Park, J.K.; Ghalam, A.; Guo, J.; Ha, C.W.; Hu, C.; Kim, H.; Kavalipurapu, K.; Lee, E.; Mohammadzadeh, A.; et al. A 128 Gb MLC NAND-Flash device using 16 nm planar cell. In Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 9–13 February 2014; Volume 57, pp. 326–327. [CrossRef]

18. Choi, S.; Kim, D.; Choi, S.; Kim, B.; Jung, S.; Chun, K.; Kim, N.; Lee, W.; Shin, T.; Jin, H.; et al. A 93.4 mm^2 64 Gb MLC NAND-flash memory with 16 nm CMOS technology. In Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 9–13 February 2014; Volume 57, pp. 328–329. [CrossRef]

19. Sako, M.; Watanabe, Y.; Nakajima, T.; Sato, J.; Muraoka, K.; Fujiu, M.; Kouno, F.; Nakagawa, M.; Masuda, M.; Kato, K.; et al. A low-power 64 Gb MLC NAND-flash memory in 15 nm CMOS technology. In Proceedings of the 2015 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 22–26 February 2015; Volume 58, pp. 128–129. [CrossRef]

20. Im, J.W.; Jeong, W.P.; Kim, D.H.; Nam, S.W.; Shim, D.K.; Choi, M.H.; Yoon, H.J.; Kim, D.H.; Kim, Y.S.; Park, H.W.; et al. A 128 Gb 3 b/cell V-NAND flash memory with 1 Gb/s I/O rate. In Proceedings of the 2015 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 22–26 February 2015; Volume 58, pp. 130–131. [CrossRef]

21. Tanaka, T.; Helm, M.; Vali, T.; Ghodsi, R.; Kawai, K.; Park, J.K.; Yamada, S.; Pan, F.; Einaga, Y.; Ghalam, A.; et al. A 768 Gb 3 b/cell 3D-floating-gate NAND flash memory. In Proceedings of the 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 31 January–4 February 2016; Volume 59, pp. 142–144. [CrossRef]

22. Yamashita, R.; Magia, S.; Higuchi, T.; Yoneya, K.; Yamamura, T.; Mizukoshi, H.; Zaitsu, S.; Yamashita, M.; Toyama, S.; Kamae, N.; et al. A 51 2Gb 3 b/cell flash memory on 64-word-line-layer BiCS technology. In Proceedings of the 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 5–9 February 2017; Volume 60, pp. 196–197. [CrossRef]

23. Kim, C.; Cho, J.H.; Jeong, W.; Park, I.H.; Park, H.W.; Kim, D.H.; Kang, D.; Lee, S.; Lee, J.S.; Kim, W.; et al. A 512 Gb 3 b/cell 64-stacked WL 3D V-NAND flash memory. In Proceedings of the 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 5–9 February 2017; Volume 60, pp. 202–203. [CrossRef]

24. Lee, S.; Kim, C.; Kim, M.; Joe, S.M.; Jang, J.; Kim, S.; Lee, K.; Kim, J.; Park, J.; Lee, H.J.; et al. A 1 Tb 4 b/cell 64-stacked-WL 3D NAND flash memory with 12 MB/s program throughput. In Proceedings of the 2018 IEEE International Solid—State Circuits Conference—(ISSCC), San Francisco, CA, USA, 11–15 February 2018; Volume 61, pp. 340–342. [CrossRef]

25. Maejima, H.; Kanda, K.; Fujimura, S.; Takagiwa, T.; Ozawa, S.; Sato, J.; Shindo, Y.; Sato, M.; Kanagawa, N.; Musha, J.; et al. A 512 Gb 3 b/Cell 3D flash memory on a 96-word-line-layer technology. In Proceedings of the 2018 IEEE International Solid—State Circuits Conference—(ISSCC), San Francisco, CA, USA, 11–15 February 2018; Volume 61, pp. 336–338. [CrossRef]

26. Sugawara, H.; Hosono, K.; Hisada, T.; Kaneko, T.; Nakamura, H. A 1.33 Tb 4-bit/Cell 3D-Flash Memory on a 96-Word-Line-Layer Technology. In Proceedings of the 2019 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 17–21 February 2019; Volume 43, pp. 210–212.

27. Siau, C.; Kim, K.H.; Lee, S.; Isobe, K.; Shibata, N.; Verma, K.; Ariki, T.; Li, J.; Yuh, J.; Amarnath, A.; et al. A 512 Gb 3-bit/Cell 3D Flash Memory on 128-Wordline-Layer with 132 MB/s Write Performance Featuring Circuit-Under-Array Technology. In Proceedings of the 2019 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 17–21 February 2019; pp. 218–220. [CrossRef]

28. Higuchi, T.; Kodama, T.; Kato, K.; Fukuda, R.; Tokiwa, N.; Abe, M.; Takagiwa, T.; Shimizu, Y. A 1 Tb 3 b/Cell 3D-Flash Memory in a 170+ Word-Line-Layer Technology. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; pp. 428–430.

29. Huh, H.; Cho, W.; Lee, J.; Noh, Y.; Park, Y.; Ok, S.; Kim, J.; Cho, K.; Lee, H.; Kim, G.; et al. A 1 Tb 4 b/Cell 96-Stacked-WL 3D NAND Flash Memory with 30 MB/s Program Throughput Using Peripheral Circuit under Memory Cell Array Technique. In Proceedings of the 2020 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 16–20 February 2021; pp. 220–221. [CrossRef]

30. Khakifirooz, A.; Balasubrahmanyam, S.; Fastow, R.; Gaewsky, K.H.; Ha, C.W.; Haque, R.; Jungroth, O.W.; Law, S.; Madraswala, A.S.; Ngo, B.; et al. A 1 Tb 4 b/Cell 144-Tier Floating-Gate 3D-NAND Flash Memory with 40 MB/s Program Throughput and 13.8 Gb/mm^2 Bit Density. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 16–20 February 2021; Volume 64, pp. 424–426. [CrossRef]

31. Park, J.; Kim, D.; Ok, S.; Park, J.; Kwon, T.; Lee, H.; Lim, S.; Jung, S.; Choi, H.; Kang, T.; et al. A 176-Stacked 512 Gb 3 b/Cell 3D-NAND Flash with 10.8 Gb/mm^2 Density with a Peripheral Circuit Under Cell Array Architecture. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; pp. 422–423.

32. Lee, G.H.; Hwang, S.; Yu, J.; Kim, H. Architecture and process integration overview of 3d nand flash technologies. *Appl. Sci.* **2021**, *11*, 6073. [CrossRef]

33. Parat, K.; Dennison, C. A floating gate based 3D NAND technology with CMOS under array. In Proceedings of the 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, USA, 7–9 December 2015; pp. 3.3.1–3.3.4. [CrossRef]

34. Goda, A. 3-D NAND Technology Achievements and Future Scaling Perspectives. *IEEE Trans. Electron Devices* **2020**, *67*, 1373–1381. [CrossRef]

35. Hemink, G.; Goda, A. NAND technology status and perspectives. In *Semiconductor Memories and Systems*; Redaelli, A., Pellizzer, F., Eds.; Elsevier: Amsterdam, The Netherlands, 2022; in press.

36. Jang, J.; Kim, H.S.; Cho, W.; Cho, H.; Kim, J.; Sun, I.S.; Jang, Y.; Jeong, J.H.; Son, B.K.; Dong, W.K.; et al. Vertical cell array using TCAT(terabit cell array transistor) technology for ultra high density NAND flash memory. In Proceedings of the 2009 Symposium on VLSI Technology, Kyoto, Japan, 15–17 June 2019; pp. 192–193.

37. Lue, H.T.; Wang, S.Y.; Lai, E.K.; Shih, Y.H.; Lai, S.C.; Yang, L.W.; Chen, K.C.; Ku, J.; Hsieh, K.Y.; Liu, R.; et al. BE-SONOS: A bandgap engineered SONOS with excellent performance and reliability. In Proceedings of the IEEE International Electron Devices Meeting, Washington, DC, USA, 5 December 2005; pp. 547–550. [CrossRef]

38. Park, Y.; Choi, J.; Kang, C.; Lee, C.; Shin, Y.; Choi, B.; Kim, J.; Jeon, S.; Sel, J.; Park, J.; et al. Highly manufacturable 32 Gb multi—Level NAND flash memory with 0.0098 μm^2 cell size using TANOS(Si—Oxide—Al$_2$O$_3$—TaN) cell technology. In Proceedings of the 2006 International Electron Devices Meeting, San Francisco, CA, USA, 11–13 December 2006; Volume 2, pp. 5–8. [CrossRef]

39. Chen, C.P.; Lue, H.T.; Hsieh, C.C.; Chang, K.P.; Hsieh, K.Y.; Lu, C.Y. Study of fast initial charge loss and it's impact on the programmed states Vt distribution of charge-trapping NAND flash. In Proceedings of the 2010 International Electron Devices Meeting, San Francisco, CA, USA, 6–8 December 2010; pp. 118–121. [CrossRef]

40. Woo, C.; Kim, S.; Park, J.; Shin, H.; Kim, H.; Choi, G.B.; Seo, M.S.; Noh, K.H. Modeling of Charge Failure Mechanisms during the Short Term Retention Depending on Program/Erase Cycle Counts in 3-D NAND Flash Memories. In Proceedings of the 2020 IEEE International Reliability Physics Symposium (IRPS), Dallas, TX, USA, 28 April–30 May 2020; pp. 3–8. [CrossRef]

41. Alsmeier, J.; Higashitani, M.; Paak, S.S.; Sivaram, S. Past and future of 3D flash. In Proceedings of the 2020 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 12–18 December 2020; pp. 6.1.1–6.1.4. [CrossRef]

42. Lue, H.T.; Hsu, T.H.; Wu, C.J.; Chen, W.C.; Yeh, T.H.; Chang, K.P.; Hsieh, C.C.; Du, P.Y.; Hsiao, Y.H.; Jiang, Y.W.; et al. A novel double-density, single-gate vertical channel (SGVC) 3D NAND Flash that is tolerant to deep vertical etching CD variation and possesses robust read-disturb immunity. In Proceedings of the 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, USA, 7–9 December 2015; pp. 3.2.1–3.2.4. [CrossRef]

43. Fujiwara, M.; Ishikawa, T.; Arayashiki, Y.; Hirayama, K.; Koyama, Y.; Kashiyama, S.; Cai, W.; Goki, Y.; Sawa, K.; Ikeno, D.; et al. 3D Semicircular Flash Memory Cell: Novel Split-Gate Technology to Boost Bit Density. In Proceedings of the 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 7–12 December 2019; pp. 642–645. [CrossRef]

44. Ishimaru, K. Future of Non-Volatile Memory -From Storage to Computing. Procedding of the 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 7–12 December 2019; pp. 12–17. [CrossRef]

45. Fu, C.H.; Lue, H.T.; Hsu, T.H.; Chen, W.C.; Lee, G.R.; Chiu, C.J.; Wang, K.C.; Lu, C.Y. A Novel Confined Nitride-Trapping Layer Device for 3D NAND Flash with Robust Retention Performances. In Proceedings of the 2019 Symposium on VLSI Technology, Kyoto, Japan, 9–14 June 2019. [CrossRef]

46. Kalavade, P. 4 bits/cell 96 Layer Floating Gate 3D NAND with CMOS under Array Technology and SSDs. In Proceedings of the 2020 IEEE International Memory Workshop (IMW), Dresden, Germany, 17–20 May 2020; pp. 14–17. [CrossRef]

47. Aiba, Y.; Tanaka, H.; Maeda, T.; Sawa, K.; Kikushima, F.; Miura, M.; Fujisawa, T.; Matsuo, M.; Sanuki, T. Cryogenic Operation of 3D Flash Memory for New Applications and Bit Cost Scaling with 6-Bit per Cell (HLC) and beyond. In Proceedings of the 2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), Chengdu, China, 8–11 April 2021; pp. 28–30. [CrossRef]

48. Yang, S. Unleashing 3D NAND's Potential with an Innovative Architecture. In Proceedings of the Flash Memory Summit, Santa Clara, CA, USA, 7 August 2018.