micromachines

Article

Self-Adaption of the GIDL Erase Promotes Stacking More Layers in 3D NAND Flash

Tao Yang [1,2], Bao Zhang [1], Qi Wang [1], Lei Jin [2] and Zhiliang Xia [2,*]

[1] Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: xiazhiliang@ime.ac.cn

Abstract: The bit density is generally increased by stacking more layers in 3D NAND Flash. Gate-induced drain leakage (GIDL) erase is a critical enabler in the future development of 3D NAND Flash. The relationship between the drain-to-body potential (V_{db}) of GIDL transistors and the increasing number of layers was studied to explain the reason for the self-adaption of the GIDL erase. The dynamics controlled by the drain-to-body and drain-to-gate potential contribute to the self-adaption of the GIDL erase. Increasing the number of layers leads to a longer duration of the maximum value of V_{db} (V_{db_max}), combined with the increased drain-to-gate potential, which enhances the GIDL current and further boosts channel potential to reach the same value at different positions of the NAND string. We proposed a method based on the correlation between the duration of V_{db_max} and the number of layers to obtain the limited layers of the GIDL erase. The limited layers allowed are more than four times the number of layers used in the current simulation. Combining the novel method of dividing the channel into multi-regions with the asynchronous GIDL erase method will be useful for further stacking more layers in 3D NAND Flash.

Keywords: 3D NAND Flash; GIDL erase; self-adaption; drain-to-body potential

Citation: Yang, T.; Zhang, B.; Wang, Q.; Jin, L.; Xia, Z. Self-Adaption of the GIDL Erase Promotes Stacking More Layers in 3D NAND Flash. *Micromachines* **2023**, *14*, 686. https://doi.org/10.3390/mi14030686

Academic Editor: Zhongrui Wang

Received: 9 February 2023
Revised: 7 March 2023
Accepted: 17 March 2023
Published: 20 March 2023

1. Introduction

With the continuous development of smartphones, 5G, artificial intelligence, and cloud computing, the demand for higher bit density in the market has grown rapidly. The bit density is generally increased by stacking more layers in 3D NAND Flash [1–9]. However, the connection scheme of using a body contact spacer (BCS) between channel polysilicon and the array common source line faces challenges during the process of removing the bottom of the gate stack, especially when stacking more than two stacks due to the overlap problem [10]. The novel connection scheme of the channel-hole sidewall ONO butting (CSOB) scheme was proposed to boost the storage density by using the sidewall connection scheme between channel polysilicon and the array common source line [10]. The erase method was changed from body erase in the BCS scheme to GIDL erase in the CSOB scheme. Therefore, holes needed to be generated by gate-induced drain leakage (GIDL) to increase the channel potential and achieve the erase operation. The first challenge for GIDL erase is the control of the doping profile in GIDL transistors, especially when it is first used in Bit Cost Scalable (BiCS). Since the manufacturing process of the bottom select gate is formed prior to the memory cell, the memory cell requires processes with a high thermal budget to achieve a large memory window and better performance. However, this leads to the problem of doping in the bottom select gate, as it can spread from the heavily doped source diffusion during the high thermal budget process [11]. To solve this problem, the manufacturing process of the bottom select gate was changed in Pipe-BiCS, placing it after the memory cell and adopting a shallower and steeper select gate channel doping profile with the low thermal budget [12]. Since then, the GIDL erase has faced challenges related to the variability and uniformity of the erase potential due to the

indirect bias of GIDL erase caused by the potential drop for band-to-band tunneling. Caillat et al. proposed an optimization method for GIDL erase with dual-side (bottom select gate side and top select gate side) GIDL injection, which led to better erase effectiveness and more effective tail control of the erase threshold voltage distribution for variability improvement [13]. Malavena et al. studied the GIDL erase dynamics process in the vertical channel NAND Flash, focusing on the increase in channel potential, and proposed a compact model for the dynamic process of the GIDL erase to explore the directions for continuous optimization [14,15]. Nowadays, most of the advanced manufactured 3D NAND Flash structures in production use the GIDL erase method. The industry expects that the cost-effective manufacturing of 3D NAND Flash can achieve over 1000 stack layers via innovations in processes and novel materials [16]. Meanwhile, combining these innovative solutions with file-system and firmware-level optimizations can greatly improve memory performance and expand the application market of 3D NAND Flash memory [17]. GIDL erase is a critical enabler for the future development of 3D NAND Flash. However, the challenge of GIDL erase is the limitation layer, as more holes are required to boost the channel potential with stacking more layers. This has become a potential risk for the realization of thousands of layers in 3D NAND Flash.

In this work, we have investigated the correlation between GIDL erase and the number of layers (N_{Layers}) based on TCAD simulation. The relationship between the drain-to-body potential (V_{db}) and N_{Layers} was studied. We found self-adaption of the GIDL current (I_{gidl}) with an increase in the N_{Layers}. As the N_{Layers} increased, the duration of the maximum value of V_{db} ($V_{\mathrm{db\_max}}$) also increased, resulting in a longer duration of maximum GIDL current. The self-adaption of the GIDL erase is achieved through the dynamic control by the drain-to-body and drain-to-gate potential. At the end of the duration of $V_{\mathrm{db\_max}}$, which leads to maximum I_{gidl}, the same potential can be reached at different positions of the NAND string. The proposed method enables the determination of the limited number of layers for the GIDL erase based on the correlation between the duration of $V_{\mathrm{db\_max}}$ and the number of layers. Finally, we propose to divide the channel into multi-regions and implement an asynchronous GIDL erase method to overcome the limitation of GIDL erase and enable the stacking of more layers in 3D NAND Flash.

2. Simulations

Figure 1a shows the schematic of the 3D NAND Flash structure. The Synopsys Sentaurus Sdevice simulator (Synopsys, Mountain View, CA, USA) was used for the simulation. The simulation structure of this device is shown in Figure 1b, with one top select gate (TSG), one bottom select gate (BSG), and more than two dummy cells (DMY) between the select gate and memory cells, and the number of memory cells will be dynamically adjusted according to the N_{Layers}. The TSG and BSG function as GIDL transistors. The parameters of the simulation structure are based on the current technology products. Furthermore, the drift–diffusion model, Coulomb scattering mobility model, thermionic emission, Shockley–Read–Hall recombination, and band-to-band-tunneling (BTBT) model were introduced in the polysilicon channel, which was well-calibrated based on our previous works [18] and GIDL current experiments. The inset shows the detailed structure of the GIDL transistors. The BTBT current occurs in the virtual PN junction around the GIDL transistors, and it follows the channel conduction direction through the BTBT path. Meanwhile, the electric field in the virtual PN junction is controlled by the drain-to-gate potential and the drain-to-body potential. GIDL transistors at the top and bottom of the strings can be applied as a negative bias to the gate relative to the drain. The erase voltage can be transferred to the channel via hole injection generated by the BTBT during the erase operation.

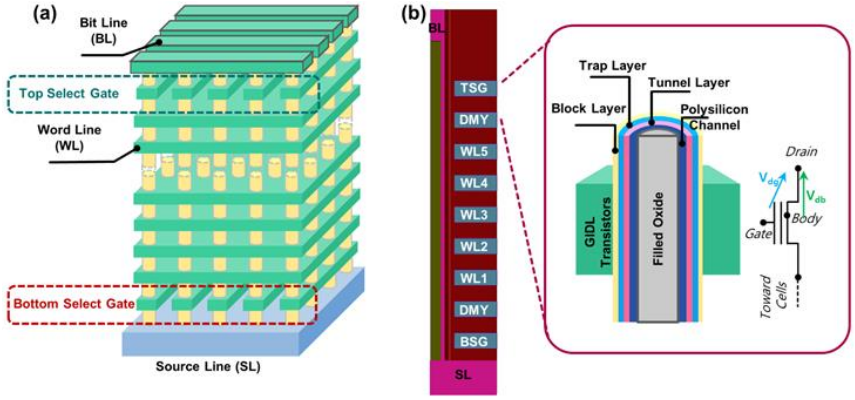

Figure 1. (**a**) Schematic diagram of 3D NAND Flash structure; (**b**) schematic diagram of TCAD simulation structure: the inset shows the detailed structure of the GIDL transistor, and the gate stack consists of the tunnel layer, trap layer, and block layer, respectively.

3. Results and Discussion

Figure 2 shows the GIDL current during the GIDL erase operation for various N_{Layers}. There is a significant dependence of the I_{gidl} on the N_{Layers}, which can be divided into four regions. Region I is before I_{gidl} reaches the maximum value. Region II is before the voltage of GIDL transistors starts to increase (T_{rise}), Region III is before the end time of the ramp-up of the drain voltage, and Region IV is the main erase region with constant erase voltage. In Region I, the maximum I_{gidl} value for larger N_{Layers} occurs later than the maximum value for small N_{Layers}. As the N_{Layers} increase, there is a time delay in the appearance of the maximum value of the I_{gidl}. Once the I_{gidl} reaches the maximum value, it rapidly begins to decrease. In Region II, the larger N_{Layers} exhibit a steeper decrease in the I_{gidl} from the maximum value. In Region III, the I_{gidl} remains relatively constant after the voltage of the GIDL transistors starts to increase. The larger N_{Layers} result in a higher I_{gidl}. In Region IV, the I_{gidl} also decreases rapidly, which shows an obvious dependence on the N_{Layers}. Finally, the GIDL current tends to be the same value regardless of the N_{Layers} in the main erase region. The reason why the I_{gidl} depends on the number of layers is that the capacitance of the channel increases as more layers are stacked, requiring more holes to be injected to boost the channel potential for GIDL erase. Then, we extracted the maximum value of the I_{gidl} for each of the N_{Layers}. As shown in Figure 3, there is a clear correlation between the maximum value of I_{gidl} and the N_{Layers} in a log scale. The I_{gidl} generated by the BTBT has increasing self-adaption with an increase in the N_{Layers}.

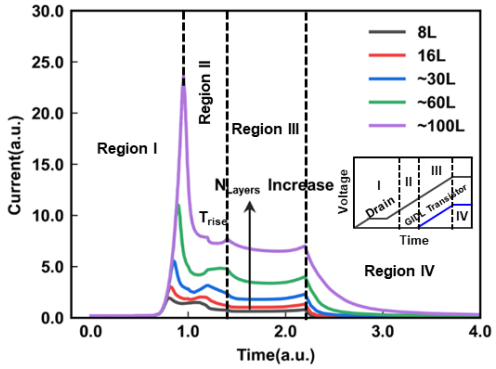

Figure 2. The I_{gidl} with the increasing N_{Layers}: the inset shows the waveform of the GIDL erase operation.

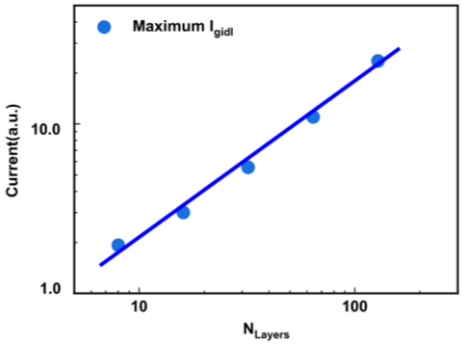

Figure 3. The correlation between the maximum value of the I_{gidl} and the N_{Layers} extracted using Sentaurus TCAD.

The electric field that drives the BTBT in the junction is controlled both by the drain-to-gate potential (V_{dg}) and the drain-to-body potential (V_{db}). In the experimental test, the V_{dg} can be easily monitored. However, monitoring the changes in the drain-to-body potential is difficult. To investigate the reason for the self-adaption of I_{gidl} with increasing N_{Layers}, we used Sentaurus TCAD simulation tools to study the relationship between V_{db} and the N_{Layers}. Figure 4 shows the dependence of V_{db} on the N_{Layers}. Before V_{db} reaches the maximum value (V_{db_max}), the trend of V_{db} is independent of the N_{Layers}. This is due to the small I_{gidl}, resulting in a slower rise in the channel potential in the body area after the hole injection compared to the ramp speed of the drain voltage. However, once V_{db} reaches V_{db_max}, the duration of V_{db_max} exhibits a clear dependence on the N_{Layers}. The increase in the N_{Layers} leads to a longer duration of V_{db_max}. The appearance of V_{db_max} is attributed to the equal rise rate of the GIDL transistors' body potential and the external drain voltage ramp speed (shown in Figure 5), which allows V_{db} to remain in the V_{db_max} state. This can be explained by the fact that the capacitance of the channel increases with an increase in the N_{Layers}, which requires a larger I_{gidl} to affect the far-end channel potential and needs more time to meet the demand of I_{gidl}. Therefore, keeping V_{db} at V_{db_max} between the drain-to-body and V_{dg} will create a further increase, promoting a significant increase in the GIDL current to meet the demand for higher layers.

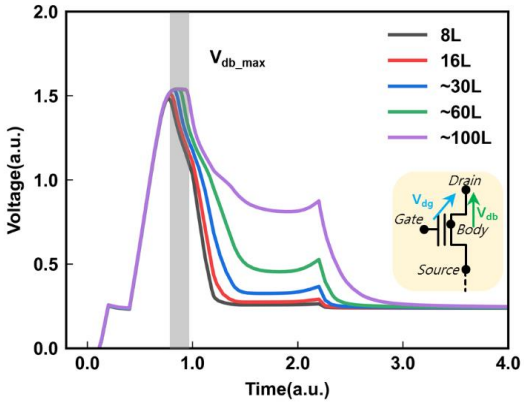

Figure 4. The drain-to-body potential with increasing N_{Layers}, and large N_{Layers} with long V_{db_max} time.

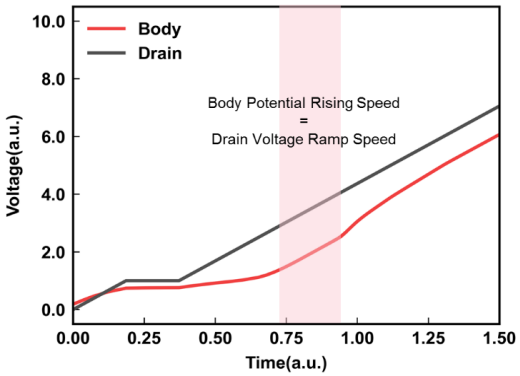

Figure 5. The correlation between drain and body potential.

Figure 6 illustrates the band diagram at various times with the V_{db_max}. With an increase in time, a larger V_{dg} reduces the BTBT distance, leading to a higher I_{gidl}. At the final time of V_{db_max}, I_{gidl} reaches its maximum value. After that, V_{db} starts to decrease sharply from the maximum value, and this region is closely related to the N_{Layers}. The final V_{db} increases with the number of layers. At this time, the V_{dg} remains constant, and the junction is affected by V_{db} to dynamically generate I_{gidl} to meet the demand for a higher number of layers. The dynamics controlled by the drain-to-body and drain-to-gate potential contribute to the self-adaption of the GIDL erase, which enables one to stack more layers in 3D NAND Flash.

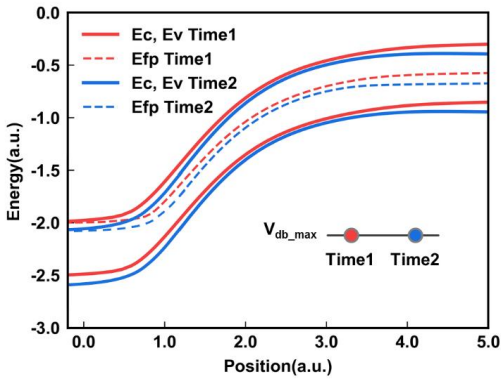

Figure 6. The band diagram of GIDL transistors at various times with the V_{db_max}, Time2 > Time1.

Next, the channel potential at different positions of the string during the GIDL erase operation was studied, as shown in Figure 7. Malavena et al. assumed that the channel potential remains as a constant channel potential due to uniform hole accumulation [15]. To investigate the channel potential position dependence of the GIDL erase, we used a verified channel length longer than 5 um. However, our results show an interesting phenomenon. There is an apparent channel position dependence on the channel potential before the GIDL current reaches its maximum value. When the GIDL current reaches the maximum value, the channel potential starts to reach the same potential, and I_{gidl} becomes large enough to boost the total channel potential. Then, the channel potential and the position potential of the GIDL transistors become the same. It is clear that the I_{gidl} reaches the maximum value during the GIDL erase process, and the same potential can be reached at different positions along the NAND string.

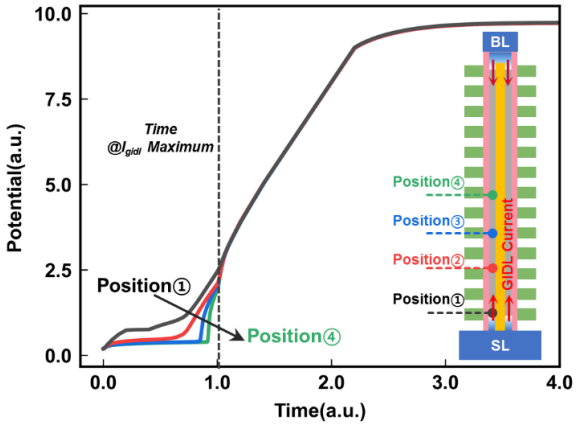

Figure 7. The channel potential at different positions of string during the GIDL erase operation.

To explore the limitation of the maximum stacking layers along the NAND string of the GIDL erase operation, it is essential to determine the time when maximum I_{gidl} occurs to achieve a consistent channel potential at different positions. Based on the relationship between V_{db} and the N_{Layers}, as the N_{Layers} increase, the duration of V_{db_max} also increases, and the subsequent I_{gidl} is determined by V_{dg}. The maximum V_{dg} value is determined by the ramp of the GIDL transistor voltage, which affects the V_{dg} voltage jointly with the drain or source of the GIDL transistors. Therefore, the time when the GIDL transistor voltage increases determines the longest duration of V_{db_max} during the GIDL erase operation. By extracting the duration of V_{db_max} for lower layers and the longest duration of V_{db_max}, a linear correlation between the duration of V_{db_max} and the number of layers is obtained. In order to verify the extensibility of the correlation, we simulated the GIDL erase by stacking more layers, and extracted the duration of V_{db_max}, finding that it still follows this correlation (red square). Finally, the limited layers of the GIDL erase were obtained using the proposed method, as shown in Figure 8. The limited layers allowed for more than four times the number of layers used in the current simulation. To achieve stacking more layers in 3D NAND Flash, the GIDL erase method needs further optimization. The current GIDL erase method erases the entire channel simultaneously, that is, the synchronous GIDL erase operation. Therefore, reducing the capacitance of each erase will effectively break the limitation of the synchronous GIDL erase method. By dividing the channel into multiple regions and applying the asynchronous GIDL erase method, the stacking of more layers in 3D NAND Flash can be achieved.

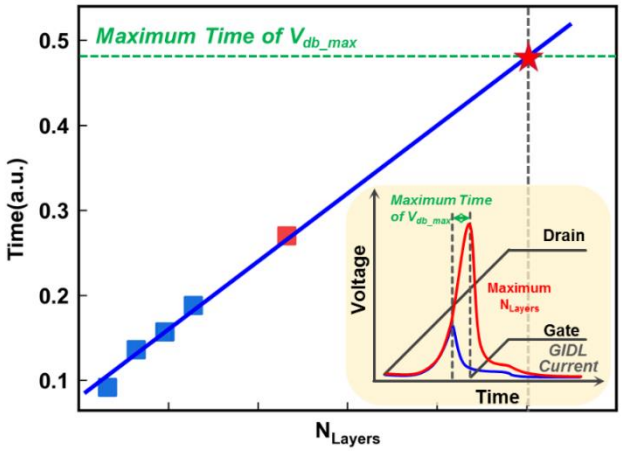

Figure 8. The correlation between the duration of V_{db_max} and the N_{Layers}, whereby the maximum time of V_{db_max} determines the maximum stacking layers.

4. Conclusions

The bit density is generally increased by stacking more layers in 3D NAND Flash. Gate-induced drain leakage (GIDL) erase is a critical enabler for the future development of 3D NAND Flash. We have investigated the correlation between GIDL erase and the number of layers based on TCAD simulation. By studying the relationship between the drain-to-body potential of GIDL transistors and increasing the number of layers, we have explained the reason for the self-adaption of the GIDL erase. The self-adaption of the GIDL erase is achieved through the dynamic control via the drain-to-body and drain-to-gate potential. As the number of layers increases, the duration of $V_{\mathrm{db\_max}}$ also increases, which, in combination with increased drain-to-gate potential, enhances the GIDL current dynamically. This leads to the same potential being reached at different positions of the NAND string. Based on the correlation between the duration of $V_{\mathrm{db\_max}}$ and the number of layers, we obtained the limited layers of the GIDL erase using the proposed method. The limited layers allowed are more than four times the number of layers used in the current simulation. Furthermore, by combining the novel method of dividing the channel into multi-regions using the asynchronous GIDL erase method, 3D NAND Flash can stack more layers. We expect that the GIDL erase will contribute to the further development of 3D NAND Flash.

Author Contributions: Writing—original draft preparation, T.Y.; writing—review and editing, B.Z., Q.W., L.J. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1107700, and in part by the National Science and Technology Major Project of China under Grant No. 21-02.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to confidentiality request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Kim, M.; Yun, S.W.; Park, J.; Park, H.K.; Lee, J.; Kim, K.S.; Na, D.; Choi, S.; Song, Y.; Lee, J.; et al. A 1Tb 3b/Cell 8th-Generation 3D-NAND Flash Memory with 164MB/s Write Throughput and a 2.4 Gb/s Interface. In Proceedings of the 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 20–26 February 2022; Volume 65, pp. 136–137.
7. Huo, Z.; Cheng, W.; Yang, S. Unleash Scaling Potential of 3D NAND with Innovative Xtacking®Architecture. In Proceedings of the 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 12–17 June 2022; pp. 254–255.
8. Lim, J.; Chung, S.; Ko, D.H.; Sung, S.-K.; Noh, K.; Ahn, S.J.; Song, J.H. Development of 7th generation 3D VNAND Flash Product with COP structure for Growing Demand in Storage Market. In Proceedings of the 2022 International Conference on Electronics, Information, and Communication (ICEIC), Jeju, Republic of Korea, 6–9 February 2022; pp. 1–4.

9. Heineck, L.; Liu, J. 3D NAND Flash Status and Trends. In Proceedings of the 2022 IEEE International Memory Workshop (IMW), Dresden, Germany, 15–18 May 2022; pp. 1–4.
10. Kang, J.; Lee, J.; Yim, Y.; Park, S.; Kim, H.S.; Cho, E.S.; Kim, T.; Lee, J.H.; Kim, J.; Lee, R.; et al. Highly Reliable Cell Characteristics with CSOB (Channel-hole Sidewall ONO Butting) Scheme for 7th Generation 3D-NAND. In Proceedings of the 2021 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 11–16 December 2021; pp. 10.1.1–10.1.4.
11. 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.
12. Katsumata, R.; Kito, M.; Fukuzumi, Y.; Kido, M.; Tanaka, H.; Komori, Y.; Ishiduki, M.; Matsunami, J.; Fujiwara, T.; Nagata, Y.; et al. Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices. In Proceedings of the 2009 Symposium on VLSI Technology, Kyoto, Japan, 15–17 June 2009; pp. 136–137.
13. Caillat, C.; Beaman, K.; Bicksler, A.; Camozzi, E.; Ghilardi, T.; Huang, G.; Liu, H.; Liu, Y.; Mao, D.; Mujumdar, S.; et al. 3DNAND GIDL-assisted body biasing for erase enabling CMOS under array (CUA) architecture. In Proceedings of the 2017 IEEE International Memory Workshop (IMW), Monterey, CA, USA, 14–17 May 2017; pp. 1–4.
14. Malavena, G.; Lacaita, A.L.; Spinelli, A.S.; Compagnoni, C.M. Investigation and compact modeling of the time dynamics of the GIDL-assisted increase of the string potential in 3-D NAND flash arrays. *IEEE Trans. Electron. Devices* **2018**, *65*, 2804–2811. [CrossRef]
15. Malavena, G.; Mannara, A.; Lacaita, A.L.; Spinelli, A.S.; Compagnoni, C.M. Compact modeling of GIDL-assisted erase in 3-D NAND Flash strings. *J. Comput. Electron.* **2019**, *18*, 561–568. [CrossRef]
16. Kim, K. The Smallest Engine Transforming Humanity: The Past, Present, and Future. In Proceedings of the 2021 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 11–16 December 2021; pp. 1.1.1–1.1.8.
17. Wang, Y.; Huang, J.; Chen, J.; Mao, R. PVSensing: A Process-Variation-Aware Space Allocation Strategy for 3D NAND Flash Memory. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2021**, *41*, 1302–1315. [CrossRef]
18. Yang, T.; Xia, Z.; Shi, D.; Ouyang, Y.; Huo, Z. Analysis and optimization of threshold voltage variability by polysilicon grain size simulation in 3D NAND flash memory. *IEEE J. Electron. Devices Soc.* **2020**, *8*, 140–144. [CrossRef]