Scaling Challenges of NAND Flash Memory and Hybrid Memory System with Storage Class Memory & NAND flash memory

Ken Takeuchi
Department of Electrical, Electronic and Communication Engineering, Chuo University
Tokyo, Japan, E-mail: takeuchi@takeuchi-lab.org

Abstract- This paper summarizes the scaling challenges of the conventional 2D floating-gate cell NAND flash memories [1, 2]. The scaling trends and limits of the bulk and SOI NAND flash memories are investigated in terms of short channel effects and channel boosting leakage from 20nm to below 10nm generation using 3D-device simulation. In the bulk NAND cell, 13nm generation is the scaling limit for realizing both channel boosting during program-inhibit and SCE suppression. The SOI NAND cell scaling limit is decreased to 8nm generation. Then, scaling problems and device design for 3D-stackable NAND flash memory are investigated [3]. Control gate length \( L_g \) and spacing \( L_{spac} \) are paid attention since they can be separately varied in 3D NAND and significantly affect the cell area of the 3D NAND as well as the electrical characteristics. \( L_g \) and \( L_{spac} \) should be the same to cope with the tradeoff between memory window and disturbance. If the number of stacked layers is 18 with the layer pitch of 40nm, the effective cell size of the 3D NAND corresponds to that of 15nm planar NAND technology. Then, this paper discusses an error prediction (EP) low density parity check (LDPC) error correcting code (ECC) which realizes an over 10-times extended lifetime [4, 5]. As the design rule shrinks, the floating gate (FG)-FG capacitive coupling among neighboring memory cells seriously degrades the memory cell reliability. The EP-LDPC ECC calibrates the inter-cell coupling without access time penalty. Finally, this paper overviews a state-of-the-art hybrid memory solution with storage class memory (SCM) and NAND flash memory for the big data solid-state storage system [5, 6]. Data fragmentation of MLC NAND flash memory is suppressed and efficient MLC NAND flash usage is realized by storing small hot data to SCM. The 3D TSV hybrid SSD realizes 11 times performance increase, 6.9 times endurance enhancement and 93% write energy reduction.

I. INTRODUCTION

SSDs and emerging storage class non-volatile memories such as PCRAM, ReRAM and MRAM have enabled innovations in various nano-scale VLSI memory systems for personal computers, smart phones, tablets and enterprise servers. There is a strong demand for continuous scaling in floating-gate (FG) type NAND flash memories below 20nm generation. However, serious program disturb errors due to the interferences from neighboring cells [7-10] become prominent. On the other hand, channel engineering and its scaling limitation in NAND flash memory cells are also concerns [11, 12]. Due to the large EOT (16~20nm) in NAND cells, short channel effect (SCE) is degraded in scaled generations. As a result, DIBL induced program disturb has been reported [13]. Also, operation margins may decrease in MLC technologies since large S-factor worsens on/off current ratio [14]. However, suppressing SCE by high channel doping concentration leads to junction leakage during program-inhibit in bulk NAND flash memories [11].

A thin body, thin BOX fully depleted SOI NAND flash memory is one of the candidates for the future scaled NAND flash memory for the excellent SCE controllability [13, 15]. Moreover, the drawbacks of thin body SOI transistors such as high parasitic resistance, low \( V_{TH} \) controllability by channel doping and \( V_{TH} \) increase by quantum confinement are less critical in the NAND flash memory than in logic device because the NAND flash memory does not require high cell current during the read and precise initial \( V_{TH} \) control (\( V_{TH} \) is controlled by the amount of electrons in the FG).

II. SCALING CHALLENGES OF 2D-NAND FLASH MEMORIES [1, 2]

In this section, the scaling trends of bulk and SOI NAND cells are newly investigated in terms of SCE and leakage during channel boosting from 20nm to below 10nm generation using 3D-device simulation [1, 2]. The SOI NAND pushes the scaling limit of the short channel effect (SCE) and channel boosting leakage from 13nm to 8nm generation.

Fig. 1 shows the program operation of a NAND flash memory. In a program-inhibit bit-line, channel voltage of the NAND string is boosted up to more than 8V to avoid \( V_{PDM} \) cell disturb. Thus, the junctions of the NAND cells must withstand high voltage stress. Otherwise, junction leakage occurs and program-inhibit fails because the channel voltage does not sufficiently increase. Hence, channel doping concentration cannot be increased in a bulk NAND cell for suppressing SCE. In SOI NAND flash cell, thinner BOX has better SCE characteristics while too thin BOX may cause BOX leakage during channel boosting. If there is no BOX leakage, channel leakage is greatly reduced because the junction area is very small.

![Fig. 1 Schematic of the program operation in a NAND flash memory. Channel voltage is boosted in program-inhibit bit-line and high drain-substrate voltage is applied to the cells](image-url)
Considering above, the device design is discussed. Figs. 2(a) and 2(b) show the device structure of the bulk and SOI NAND flash memory cells at 15nm generation used in this simulation. Bulk source/drain junction depth \( X_J \) and SOI thickness \( T_{soi} \) are fixed to 6nm. Punch-through stopper (PTS) layer, where the doping concentration is higher than the other channel region, is added to the bulk cell. \( T_S \) (distance between PTS layer and source/drain junction) is changed to control the SCE. Although SCE is better in smaller \( T_S \), the junction leakage increases. For the SOI NAND cell, \( T_{BOX} \) (BOX thickness) is varied.

Channel leakage during channel boosting of the bulk and SOI NAND cells having the same SCE characteristics are compared. Junction electric field \( E_{junction} \) and BOX electric field \( E_{BOX} \) are evaluated for the channel leakage. Although the bulk and SOI NAND cells have the same \( V_{TH} \) roll-off characteristics, S-factor and DIBL in the SOI NAND cell are better.

Figs. 3(a) and 3(b) show the potential profile during the channel boosting simulation at 11nm generation for bulk and SOI NAND cells, respectively. SCE in both bulk and SOI NAND cells are made the same as the case in 15nm generation. Source and drain voltages are set to 8V. In the bulk NAND cell (Fig. 3(a)), \( E_{junction} \) is found to be higher than 1MV/cm, which is the critical value for the junction leakage. Thus, channel boost in the bulk cell fails at 11nm generation. On the other hand, \( E_{BOX} \) in the SOI NAND cell is found to be below 10MV/cm where FN tunneling starts to occur. Note that BOX is made of thermal oxide and its quality is higher than that of the oxide (typically TEOS) filled between the FGs and CGs. Therefore, no BOX leakage occurs even at 11nm generation in the SOI cell.

The scaling trends for the \( E_{junction} \) and \( E_{BOX} \) in the SOI NAND cell as a function of \( T_{BOX} \) are shown in Figs. 5(a) and 5(b). \( E_{BOX} \) is still below 10MV/cm at 11nm generation for all \( T_{BOX} \). S-factor for 11nm generation is also worse than 13nm generation.
III. 3D-STACKABLE NAND FLASH MEMORY DESIGN [3]

3D-stackable NAND flash memory (3D NAND) [15-18] has been attracting much attention to overcome the scaling limit of the planar NAND flash memory. In these devices, the number of stacked layers Nlayer is increased because they are completely different from the planar NAND. Fig. 8(a) shows the simplified cross sectional view of the bit-cost scalable (BiCS) [16, 17] type 3D NAND. One of the problems of the 3D NAND is the decrease of cell density in the planar direction. The BiCS hole must be filled with O/N/O film (~20nm) and silicon channel. Since the O/N/O film is not aggressively scaled to maintain memory window and reliability, the diameter of the BiCS hole is not so scalable. Therefore, Nlayer should be increased to compensate this drawback under finite taper angle θ in the BiCS hole (Fig. 8(a)). On the other hand, since the minimal line and space lithography pattern is not required for the control gate (CG) formation in 3D NAND, CG length Lg and spacing Lspace can be separately chosen. This design flexibility is only allowed for 3D NAND. Thus suitable device design for 3D NAND can be explored in terms of Lg and Lspace.

3D NAND scaling and design methodologies are investigated [17]. Comparing with the planar NAND, the requirements for the Lg and Lspace of the 3D NAND are comprehensively studied from the cell size and electric characteristics.

The effective cell area (Aeff) of the 3D NAND is discussed. Aeff is A/Nlayer where A is the cell area in each layer [19]. The Aeff of the 3D NAND in Fig. 8(a) is approximately given as (square layout is assumed [17]),

\[ A_{eff} = \frac{2R_{B} + 2L_{layer}(L_{g} + L_{space})\tan\theta + 2t_{ONO} + F}{2/N_{layer}} \]

where R_B, t_{ONO} and F are the bottom radius of the BiCS hole, total thickness of O/N/O layer and feature size (spacing between BiCS hole), respectively. Fig. 8(b) shows the effective cell area as a function of Nlayer with various layer pitches (Lg + Lspace). Aeff for the planar NAND is also shown in the figure.

![Fig. 8](image-url)
Fig. 9 $L_g$ and $L_{space}$ design window for (a) 3D NAND and (b) planar NAND structure.

Table 1 summarizes the comparison of the 3D and planar NAND. 3D NAND achieves very good $I_{on}$, S.S. and low $V_{pgm}$ compared with planar NAND. Slight degradations in $V_{on}$ roll-off and $V_{th}$ shift by the stored electrons in the neighboring cell are observed only at the small $L_g$ and $L_{space}$ region.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>3D NAND</th>
<th>Planar NAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{on}$</td>
<td>Very good</td>
<td>Poor</td>
</tr>
<tr>
<td>$V_{on}$</td>
<td>Poor</td>
<td>Fair</td>
</tr>
<tr>
<td>S.S.</td>
<td>Very good</td>
<td>Fair</td>
</tr>
<tr>
<td>$V_{th}$</td>
<td>Good</td>
<td>Fair</td>
</tr>
<tr>
<td>$V_{pgm}$</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>$L_{space}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$L_g$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$V_{th}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$V_{pgm}$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1 Summary of 3D and 2D NAND

IV. ERROR PREDICTING (EP) LDPC ECC [4, 5]

As the design rule shrinks, the floating gate (FG)-FG capacitive coupling among neighboring memory cells seriously degrades the memory cell reliability. To enhance the error correction capability, an LDPC ECC is proposed for 1Xnm flash memories instead of the Bose-Chaudhuri-Hocquenghem (BCH) ECC. In the 2 bit/cell, 3 reference voltages ($V_{ref}$) are needed for the BCH. The conventional LDPC requires many, e.g. 21, $V_{ref}$ to get accurate $V_{TH}$ information. The inter-cell coupling is also considered to calibrate the interference. However, the increase in $V_{ref}$ number requires more sequential read cycles. Assuming 50µs cell read time and 21 $V_{ref}$ the read access time is as much as 1050µs. In case of the 3bit/cell or 4bit/cell, the read access time increases by twice or five-times, which is unacceptably long. To realize both fast read and high reliability, the error prediction LDPC (EP-LDPC) utilizing only 3 $V_{ref}$ is proposed as show in Fig. 10 [4, 5].

![Fig. 10 Comparison of the conventional and the EP-LDPC ECC [4, 5.]](image)

The read is 7-times faster than the conventional LDPC. The EP-LDPC corrects errors most effectively because in addition to the $V_{TH}$ and the inter-cell coupling, the write/erase cycles and the retention time are considered for the calibration. As a result, over 10-times extended lifetime is realized.

Fig. 11 shows the hardware architecture of the SSD with ED-LDPC. The error prediction sequence is realized with the simple logic gates in the NAND controller. The additional NAND controller circuit area to the conventional LDPC is negligibly small.

![Fig. 11 Hardware architecture of SSD with EP-LDPC ECC [4, 5.]](image)
V. HYBRID MEMORY SOLUTION WITH STORAGE CLASS MEMORY & NAND FLASH MEMORY [5, 6]

There is a growing demand for a high performance, highly reliable and low power SSD. A 3D TSV-integrated SSD with hybrid memory configuration which uses storage class memories (SCMs) and NAND flash memories is a promising solution. Among various SCMs, ReRAM is the best candidate due to its high speed, low power operation and potentially high scalability [20, 21]. In [5, 6], the detailed specifications for the ReRAM and architecture for the hybrid SSD are proposed.

The block diagram of the hybrid SSD is shown in Fig. 12(a). The ReRAM uses NAND-like I/F. The polling (Ready/Busy status), which is used in NAND I/F, allows a variable access time. Fig. 12(b) shows the physical image of the proposed SSD with TSVs.

![Block diagram of the proposed 3D TSV-integrated hybrid ReRAM/MLC NAND SSD](image)

Fig. 12 (a) Block diagram of the proposed 3D TSV-integrated hybrid ReRAM/MLC NAND SSD. Proposed ReRAM uses NAND-like I/F. (b) Physical image of the proposed SSD.

In [5, 6], three data management algorithms are proposed for the 3D hybrid SSD. The key idea is to store hot fragmented data less than the page size to ReRAM and use MLC NAND for sequential data. To evaluate the hybrid 3D hybrid NAND SSD a TLM (transaction level modeling) -based SSD emulator that can comprehensively simulate performance, energy consumption and P/E cycles has been developed. The results for the write performance, write energy and average P/E cycles are shown in Fig. 13.

![Normalized write data size](image)

Fig. 13 (a) Write performance, (b) write energy and (c) average P/E cycles of the conventional and hybrid SSDs. The horizontal axis for (a) and (c) is the data size written to the SSD normalized by the SSD MLC NAND total capacity. 100ns/sector is assumed for the ReRAM write and read latency.

Compared with the conventional MLC NAND SSD, the hybrid SSD shows 11 times higher performance and 79% lower write energy (Figs. 13(a) and 13(b)). By using 3D TSV interconnects, the I/O energy is reduced by 27 times because the huge capacitance of the wire bonding is almost eliminated. As a result, the total SSD energy reduction reaches 93%. Furthermore, the slope of the average MLC NAND P/E cycles is decreased by 6.9 times in Fig. 13(c) by the hybrid SSD. This directly corresponds to a reduction in the replacement cost of a SSD storage system because the slope determines the aging speed of the SSD. In ReRAM, a data fragmentation does not occur because the partial overwrite is possible. As a result, the slope of the ReRAM P/E cycles is limited to 28 times of that of the MLC NAND in the hybrid SSD. Assuming MLC NAND endurance of \(3 \times 10^9\), the required P/E cycles for ReRAM is less than \(10^7\), which is acceptable for the ReRAM device characteristics. Fig. 14 shows the valid page map of the conventional and hybrid SSD. The valid pages are scattered in the conventional SSD indicating that frequent overwrites have occurred to the MLC NAND. On the other hand, the hybrid SSD efficiently uses ReRAM and shows less fragmentation of MLC NAND because overwrites to MLC NAND are suppressed.
The required ReRAM latency to obtain sufficient improvements is also investigated in [5, 6]. Fig. 15 shows the SSD write performance and energy as a function of the ReRAM write latency. ReRAM read latency is also varied. From the figures, both ReRAM write and read latency should be less than 3μs to maintain high performance and low power operation. Considering 50ns write pulse, the 3μs access is achievable for ReRAM in write verify operation.

![Comparison of the SSD valid page location.](image)

**Fig. 14** Comparison of the SSD valid page location.

VI. CONCLUSION

Scaling challenges of 2D and 3D NAND flash memory and solution for the scaling blockage are discussed. The scaling trends and limits of the bulk and SOI NAND flash memories are investigated in terms of short channel effects and channel boosting leakage from 20nm to below 10nm generation using 3D-device simulation. In the bulk NAND cell, 13nm generation is the scaling limit for realizing both channel boosting during program-inhibit and SCE suppression. The SOI NAND cell scaling limit is decreased to 8nm generation. Then, scaling problems and device design for 3D-stackable NAND flash memory are investigated. If the number of stacked layers is 18 with the layer pitch of 40nm, the effective cell size of the 3D NAND corresponds to that of 15nm planar NAND technology.

An error prediction (EP) low density parity check (LDPC) error correcting code (ECC) which realizes an over 10-times extended lifetime is discussed. As the design rule shrinks, the floating gate (FG)-FG capacitive coupling among neighboring memory cells seriously degrades the memory cell reliability. The EP-LDPC ECC calibrates the inter-cell coupling without access time penalty.

Finally, the hybrid memory solution with storage class memory (SCM) and NAND flash memory are reviewed for the big data solid-state storage system. Data fragmentation of MLC NAND flash memory is suppressed and efficient MLC NAND flash usage is realized by storing small hot data to the big data solid memory (SCM) and NAND flash memory are reviewed. If the number of stacked layers is decreased to 8nm generation, boosting during program

ACKNOWLEDGMENTS

The author sincerely appreciates Kousuke Miyaji, Shuhei Tanakamaru, Hiroki Fujii, Koh Johguchi, Kazuhide Higuchi, Chao Sun, Yuki Yanagihara, Masafumi Doi and Chinglin Hung for their continuous hard work and efforts. This paper is the summary of their originally reported work about the NAND flash memory, SSD and hybrid memory systems [1-4, 6].

REFERENCES

学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，
提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。
图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具