

TCAD-Based Word Line Driver Optimization for NAND-type 3D Flash Memory Scaling

Sudarshan Narayanan
Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
sudarshan.narayanan@sandisk.com

Qinghua Zhao
Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
qinghua.zhao@sandisk.com

Satoru Mayuzumi
Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
satoru.mayuzumi@sandisk.com

Derek Dou
Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
derek.dou@sandisk.com

Chao Qin
Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
chao.qin2@sandisk.com

Mohan Dunga
Technology Development Engineering
Sandisk Technologies Inc.
Milpitas, CA, USA
mohan.dunga@sandisk.com

Abstract— Word Line Switch (WLSW) transistors play a critical role in NAND-type 3D Flash memory by transferring high Program and Erase voltages (25V) to the gates of memory cells. As memory capacity increases via word-line (WL) stacking, the number of WLSW transistors grows proportionally, causing substantial chip area expansion. However, shrinking WLSW width is constrained by maximum transferable Program voltage (VPGM) limiting the memory array performance. In this paper, the cause of WLSW width scaling limitation is identified through calibrated 3D TCAD simulations. Furthermore, various factors adversely impacting WLSW scaling are systematically addressed and new process and architectural solutions to mitigate them are presented.

Keywords— WLSW, CMOS Scaling, CBA structure.

I. INTRODUCTION

In CMOS-bonded-array (CBA) architecture for NAND-type 3D Flash [1,2], CMOS die dimensions must track array die size for cost efficiency. While vertical stacking of word lines (WLs) has enabled aggressive density scaling in NAND-type 3D Flash, it also introduces a proportional increase in

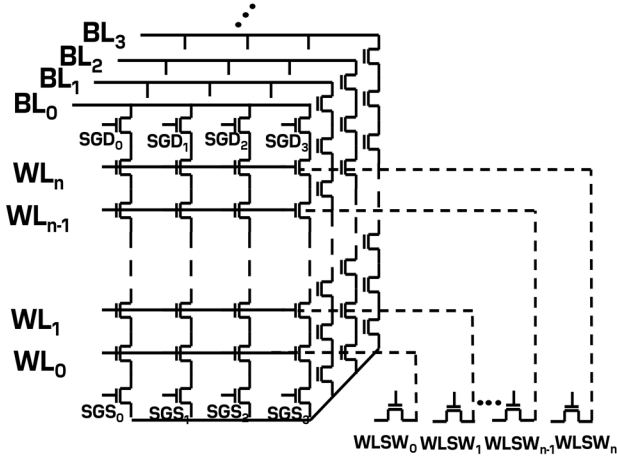


Fig. 1: Illustration of NAND-type 3D Flash architecture - every WL in each block is connected to a dedicated WLSW.

peripheral circuitry required for array operation. Among these, Word Line Switch (WLSW) transistors represent a dominant contributor to area, since each WL requires a dedicated WLSW as illustrated in Fig. 1. As a result, WLSW X-pitch scaling becomes increasingly critical with rising WL layer count and directly impacts overall die size and cost.

In addition to area and cost considerations, WLSWs are responsible for passing and withstanding high Program and Erase voltages (~25V). This high voltage operation imposes

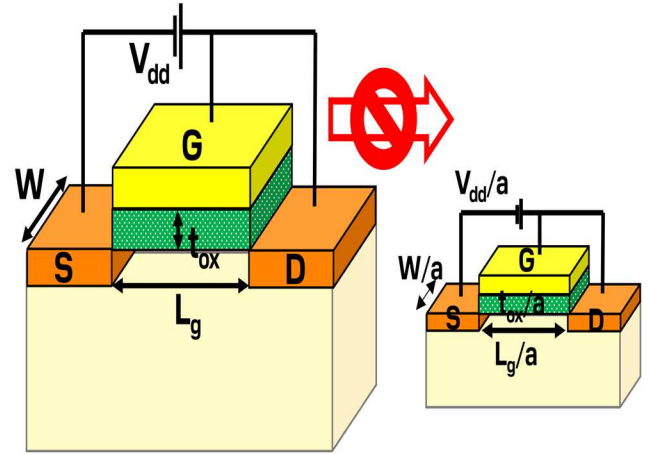


Fig. 2: Dennard scaling is not applicable to HV WLSW transistors as supply voltage does not scale.

strong electrostatics constraints that fundamentally distinguish WLSW scaling from conventional low voltage CMOS scaling. Classical Dennard [3] scaling as shown in Fig. 2, is inapplicable in this regime as supply voltage cannot be

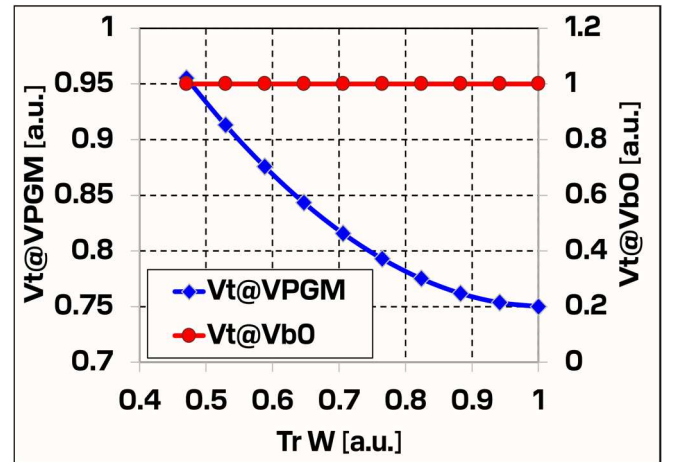


Fig. 3: Experimental data and TCAD trend shows non-linear increase in $V_t@V_{PGM}$ with transistor width reduction, while $V_t@V_{b0}$ has no width dependence.

proportionally reduced. As observed in Fig. 3, the dominant WLSW-X pitch scaling barrier is the non-linear increase in back-bias threshold voltage at Program voltage ($V_t@V_{PGM}$) as width shrinks. While threshold voltage at zero back-bias ($V_t@V_{b0}$) remains largely invariant with width, the strong dependence of $V_t@V_{PGM}$ on transistor width highlights the role of high back-bias electrostatics in limiting device scalability. This threshold voltage increase from scaled

WLSW width limits the ability of WLSW to transfer VPGM, thereby impacting memory array performance.

In subsequent sections of the manuscript, we investigate the physical origin of this High Back-bias Narrow Width Effect (HiB-NWE) using calibrated 3D TCAD simulations. By analyzing electrostatic behavior across device geometries and bias conditions, we identify the dominant mechanism contributing to this scaling limitation. Based on this understanding, we further explore process, architectural, and biasing strategies to mitigate HiB-NWE. Finally, we quantify the cumulative benefit of the proposed solutions on WLSW width scaling potential for future NAND-type 3D Flash technologies.

II. RESULTS AND DISCUSSION

A. TCAD Simulation Setup:

Three-dimensional TCAD using Synopsys Sentaurus (SProcess/SDevice) [4] was employed to replicate the WLSW layout and bias environment as shown in Fig. 4. The electrostatics in the TCAD deck was calibrated against silicon data by reproducing the CV/IV curves. Threshold voltage was

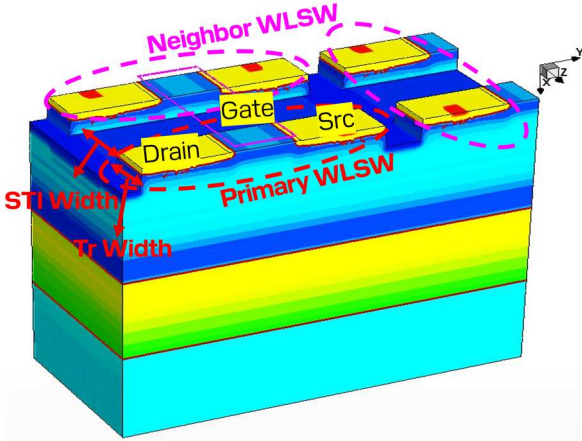


Fig. 4: TCAD simulation domain indicating primary and neighboring WLSW.

extracted for WLSW at zero back-bias and under high back-bias when WLSW transfers VPGM i.e. back-bias voltage of VPGM. Calibrated TCAD simulations reproduced the interesting silicon trend wherein threshold voltage at zero back-bias, $V_t@V_{b0}$ showed no width dependence, while threshold voltage during VPGM transfer, $V_t@V_{PGM}$ showed non-linear increase for narrower widths. This divergence between low and high back-bias behavior indicates that the scaling limitation is not driven by intrinsic channel electrostatics alone, but rather by additional depletion mechanisms that become prominent under high-voltage biasing. Therefore, isolating the origin of this effect is essential for identifying effective mitigation strategies.

B. Rootcause of HiB-NWE:

TCAD simulations of devices with varying transistor widths revealed that the lateral depletion extension beneath STI was nearly invariant to transistor width as shown in Fig. 5. The extent of this edge depletion region is set primarily by the back-bias, STI sidewall geometry and the doping under the STI. Under high back-bias conditions, the depletion region expands laterally underneath the STI, creating a parasitic charge component that is independent of transistor width. Hence, the edge depletion charge is a constant overhead that does not scale with transistor width.

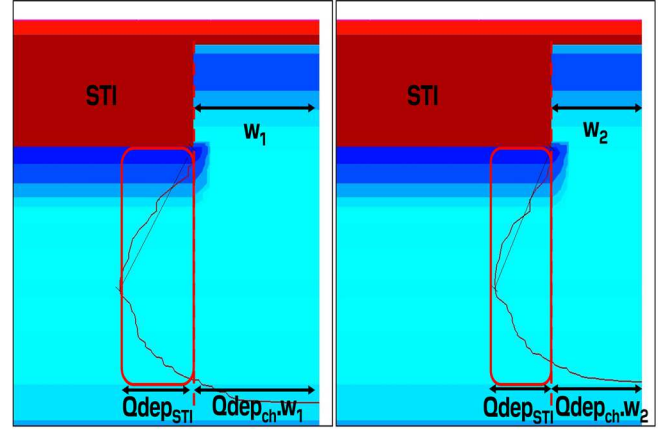


Fig. 5: Illustration of HiB-NWE: Cross-section of two transistors with widths $w_1 > w_2$ show nearly width independent depletion extension under STI represented by edge depletion Q_{dep_STI} and depletion in channel represented by Q_{dep_ch} . Q_{dep_STI} dominates at narrower widths leading to HiB-NWE.

As illustrated in Fig. 5, the cross-section of two transistors with different widths w_1 and w_2 wherein $w_1 > w_2$, shows similar depletion under the STI. The total depletion charge Q_{dep_Tot} in the structure is given by

$$Q_{dep_Tot} = Q_{dep_STI} + w \cdot Q_{dep_ch} \quad (1)$$

where, Q_{dep_STI} is the depletion charge under the STI, Q_{dep_ch} is the depletion charge per unit width under the channel region and w is the width of the transistor. From (1), it is evident that as the transistor width narrows, the edge depletion charge Q_{dep_STI} makes up a larger fraction of the total depletion charge Q_{dep_Tot} , thereby increasing the average depletion charge per unit width that the gate must neutralize. Consequently, a higher gate bias is needed to reach the same surface potential for inversion at high back-bias — manifesting as a higher $V_t@V_{PGM}$ and a stronger HiB-NWE.

C. Solutions to Mitigate HiB-NWE:

Guided by the physical insight into the rootcause of HiB-NWE, various process and architectural strategies to mitigate it were explored extensively through TCAD simulations. These strategies are designed to either reduce the fixed depletion component or improve gate control over the channel

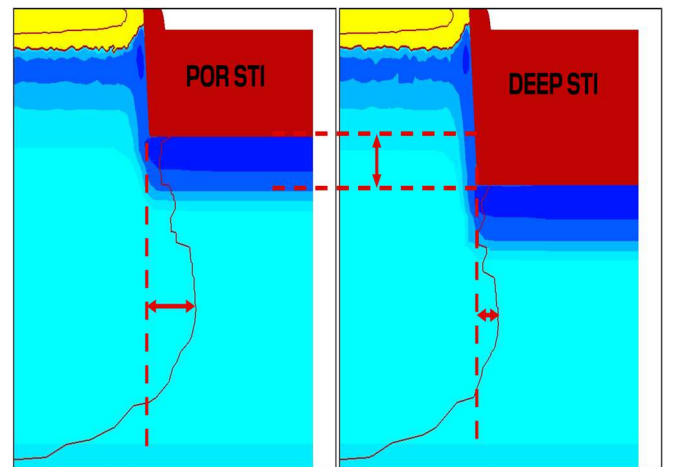


Fig. 6: TCAD simulations for WLSW with POR STI on the left panel and Deep STI on the right panel show edge depletion width extension significantly reduced with deeper STI, thereby lowering HiB-NWE.

under high back-bias conditions. The main solutions identified to lower this phenomenon are outlined here:

(1) **STI depth increase:** As shown in Fig. 6, TCAD simulations demonstrate that deepening STI curtails edge depletion spread under STI and hence reducing the width-independent depletion component Q_{depSTI} .

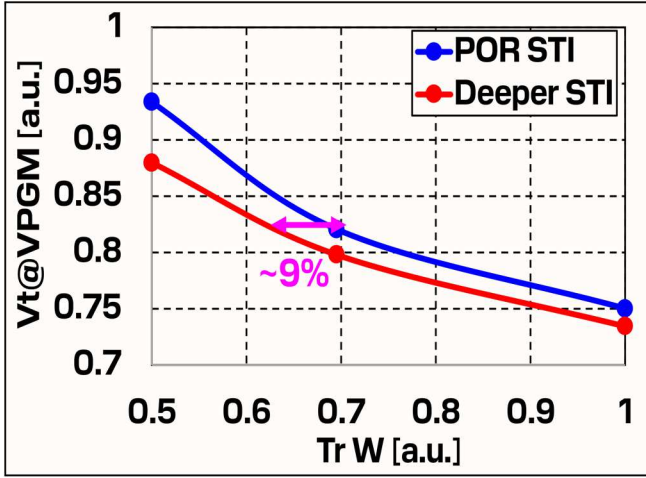


Fig. 7: TCAD simulation demonstrates the $V_t@VPGM$ reduction due to deeper STI, enabling ~9% width reduction at given $V_t@VPGM$.

Hence, the threshold voltage simulations as seen in Fig. 7 show lower $V_t@VPGM$ due to deeper STI. Consequently, for a given $V_t@VPGM$, large transistor width reduction of ~9% can be achieved. This approach represents a process-level solution that modifies device geometry to mitigate the fundamental electrostatic limitation.

(2) **Shared-gate architecture:** Beyond process modifications, architectural approaches can further mitigate HiB-NWE. To support the depletion charge under the STI, it

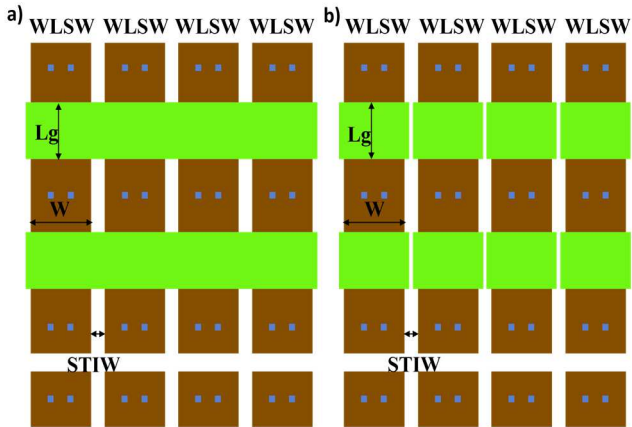


Fig. 8: Illustration showing the WLSW layout with a) Shared Gate and b) Separate Gate

was proposed to extend the active gate over the STI and realize a shared gate between neighboring WLSWs as seen in Fig. 8.

Fig. 9 illustrates the WLSW width-direction 2D cross-section for shared- and separated-gate designs. Compared with the separated-gate layout, the shared-gate architecture increases the effective gate area and strengthens electrostatic control. By enabling adjacent gates to help balance the STI depletion charge, it reduces the charge burden on individual transistor. Consistent with this mechanism, TCAD simulations in Fig. 10 show a significant reduction in

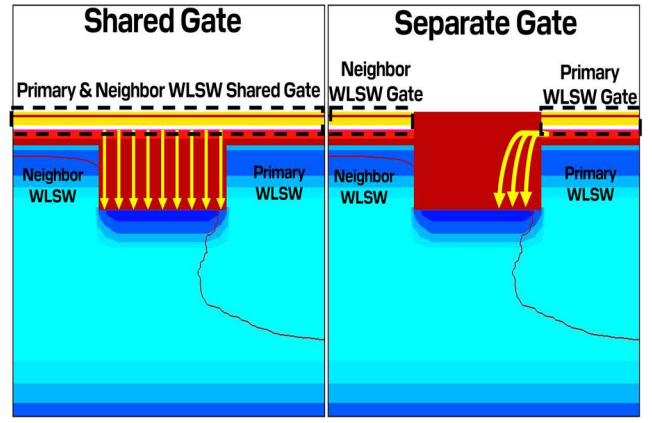


Fig. 9: 2D cross-section along width shows the electric field lines from the gate in Shared Gate on the left panel and Separate Gate on the right panel. Additional gate area in Shared Gate layout supports depletion charge under STI and lowers HiB-NWE.

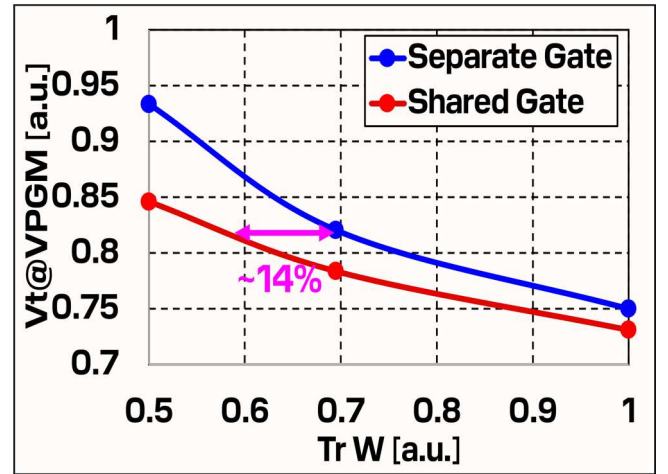


Fig. 10: TCAD simulations show significant reduction in $V_t@VPGM$ for Shared Gate architecture, enabling ~14% WLSW width reduction at a given $V_t@VPGM$.

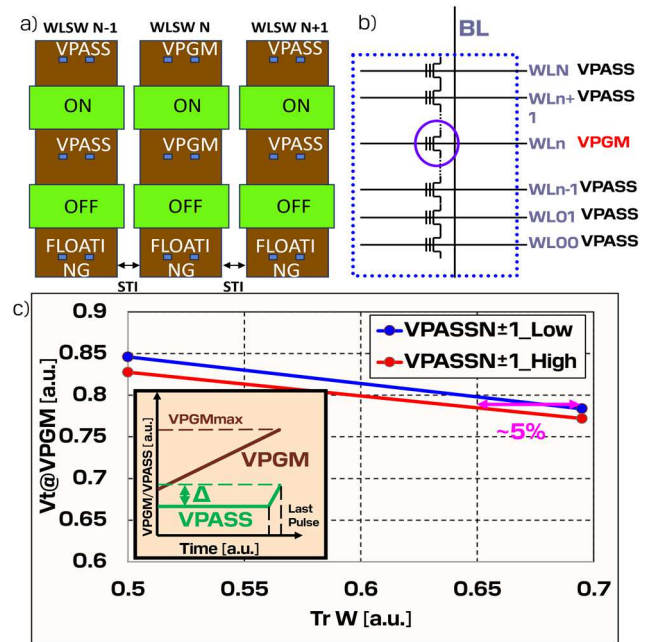


Fig. 11: a) WLSW layout with bias on adjacent transistors. b) Corresponding gate voltages on memory string. c) $V_t@VPGM$ benefit with novel bias scheme - inset shows VPASS step-up at last pulse.

$V_t@V_{PGM}$, enabling up to ~14% WLSW width reduction with shared-gate layouts.

(3) Novel bias scheme optimization: In addition to spatial redistribution of electrostatic control through innovative layout, the bias conditions on WLSW can be leveraged to dynamically control depletion charge under the STI. Fig. 11a) illustrates the WLSW layout with bias on adjacent transistors, while Fig. 11b) denotes the corresponding gate voltages on the memory string. During normal Program operation, when the selected WLSW N transfers VPGM, the adjacent WLSW N $\pm$ 1 transfers lower voltage VPASS. By appropriately modulating the bias on neighboring word lines, it is possible to engineer the interaction of depletion regions. Specifically, a higher bias on adjacent WLSW N $\pm$ 1 merges the depletion regions, thereby allowing two gates to share the depletion charge and lower the $V_t@V_{PGM}$ for WLSW N. We propose a novel biasing scheme to mitigate HiB-NWE as shown in the inset of Fig. 11c), wherein VPASS on the neighbor WLSW N $\pm$ 1 is stepped up just as Program voltage VPGM reaches target threshold value of VPGMmax. This timing-dependent biasing enhances electrostatic sharing during critical operation to lower $V_t@V_{PGM}$ and enables 5% width reduction benefit as observed from TCAD simulations in Fig. 11c).

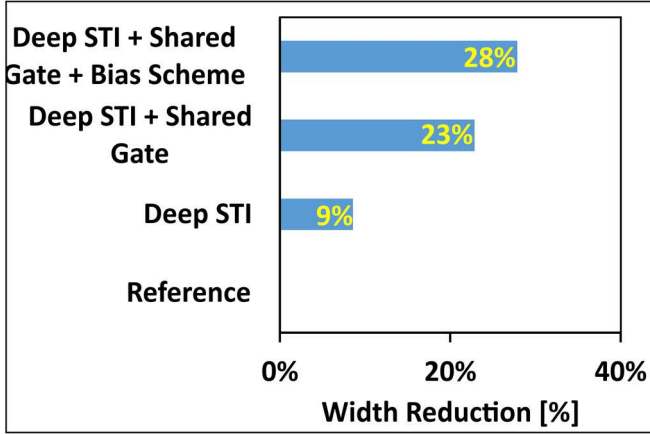


Fig. 12: Benefit of HiB-NWE lowering Process/Architectural solutions on WLSW width.

Collectively, the combination of STI engineering, architectural innovation, and bias optimization effectively suppresses HiB-NWE by targeting both reduction and sharing of depletion charge under the STI. As summarized in Fig. 12, these approaches enable a cumulative WLSW width reduction of ~28%, demonstrating that multi-dimensional co-optimization is essential for overcoming high-voltage scaling limitations in WLSW transistors.

III. CONCLUSION

In this paper, the need to shrink WLSW footprint was highlighted and HiB-NWE was identified as a key impediment to shrink WLSW width. Through TCAD simulations, the root cause of HiB-NWE was identified as the spatial extension of depletion charge under the STI and beyond the active gate area while transferring high VPGM. We proposed optimized STI shape, shared WLSW gate architecture and novel WLSW biasing scheme as the key strategies to lower HiB-NWE and successfully scale WLSW width by ~28%. These improvements will contribute towards significant cost benefit due to overall die size reduction for future NAND-type 3D Flash technologies.

REFERENCES

- [1] R. Katsumata, 2024 IEEE International Memory Workshop (IMW), pp. 1–4, 2024.
- [2] Y. Li, IEEE Solid-State Circuits Magazine, 12(4):56–65, 2020.
- [3] R. Dennard et al., IEEE J. Solid-State Circuits, 9(5):256–268, Oct. 1974.
- [4] Synopsys Inc., Sentaurus Device User Guide, 2024.