2024 Design, Automation & Test in Europe Conference (DATE 2024)

# A DTCO Framework for 3D NAND Flash Readout

Mattia Gerardi, Arvind Sharma, Yang Xiang, Jakub Kaczmarek,
Fernando García Redondo, Maarten Rosmeulen, Marie García Bardon
*imec*, Leuven, Belgium

***Abstract*—To continue increasing the storage density of 3D NAND flash memories, new technology options need to be evaluated early on. This work presents a unique predictive parametric framework for Multi-Level Cell 3D NAND Flash read operation at the array level. This framework is used to explore the read sensitivity to multiple parameters and technology options. We identify the trade-offs between number of layers, read-current and read time to be the most determinant factors to ensure the array readability while enabling stacks of more than 300 layers and maximizing the memory density.**

***Keywords*—Non-Volatile Memory, 3D NAND Flash, DTCO Memory Readout, Framework.**

## I. Introduction

The advances in storage density, power consumption, speed, reliability, and cost have positioned NAND Flash as today's storage standard [1]. This study addresses the existing gap in predictive Multi-Level Cell (MLC) 3D NAND Flash compact modeling for read operations, offering a swift and adaptable Design-Technology Co-Optimization (DTCO) tool for the optimization and design of future 3D NAND structures, complementing Technology Computer-Aided Design (TCAD) simulations. We present a parametric model to forecast the effects of design choices on system-level read operations. This model includes a wide range of parameters, from transistor to array level, capturing geometrical attributes, array dimensions, and material selection.

## II. Background and Related Work

***Background*** – 3D NAND Flash memories are highly dense, Non-Volatile Memory (NVM) arrays in which transistors are serially and vertically stacked [1]. Improving the bit density today relies on increasing the number of stacked cells beyond 200 layers, reducing Bit Line (BL) pitches below 100 nm [2], and storing multiple bits per cell [1]. Multi-, Triple- and Quad-Level Cells (MLC, TLC, QLC) provide 2, 3, and 4 bits per cell, respectively. As density is increased, readability is more challenging: read-current ($I_{READ}$) lowers, RC parasitics increase, *Vth* distributions overlap. High mobility materials and periphery design for smaller sense margins are studied to overcome these limits.

***Related Work and Motivation*** - Cell-level TCAD simulations, analytical and compact models are accessible in literature, and they have been considered for the current project development. Previous work related to planar NAND Flash compact modelling exists [3], as well as a 3D NAND Flash behavioral compact model is available [4]. Nevertheless, a comprehensive 3D NAND Flash framework accounting for bit cell and array level characteristics in an MLC technology is missing. Addressing these needs, this study introduces a tool that allows memory designers to predict system-level outcomes during the read phase when adjusting design parameters. The current framework aims to implement a quick and complete tool, through which multiple parametric sweeps can be performed and evaluated. Model parameters include array size (number of layers), geometry (BL spacing), and transistors characteristics (*Vth* variability, channel material).

## III. Methodology and Framework

As described in Fig. 1, the framework has been developed in Python and orchestrates the circuit simulations running on Cadence Spectre. A netlist generator receives user-defined parameters describing the memory, and creates netlist and measurement files based on the unit subcircuits, considering a current-to-voltage conversion sensing scheme. Circuit testbench simulations are then parallelized by a management module and processed by a processing module.

***Memory Architecture*** - The tool modularly creates the target netlists by first defining the required subcircuits that are repeatedly placed, following the 3D array structure. Each subcircuit has cell parameters through which the cell behavior is described. The basic subcircuits conforming the 3D array are a unit RC-cell and a bit cell. Bit states are randomly or deterministically assigned to the cells, according to the needs.

***Subcircuits Definition*** - An RC T-model is defined to consider resistive contributions and capacitive coupling (parallel plate and fringe) between adjacent BLs and Word Lines (WLs). Conformal mapping has been employed to evaluate fringe capacitances [5]. A 5-parameter transistor compact model is used for the memory cell description [6]. Such parameters can be employed to fit the model I-V curves to any measurements or TCAD simulation. In this work, I-V results from [7] have been taken as reference. MLC modeling is rendered with the definition of nominal writable cell *Vth* in a given window. The model can receive distributions describing the *Vth* statistics to start Montecarlo simulations, while deterministic *Vth* deviations are here presented enabling error-rate analysis.

***Simulation Read Sequence*** - Read ($V_{READ}$) and pass ($V_{PASS}$) voltages are applied to the WLs. The generated string current depends on the programmed *Vth*, and pre-charged capacitive sense nodes connected to each BL discharge accordingly [1]. The full MLC 3D NAND Flash readout sequence is here implemented. The MLC readout takes place in 3 subsequent phases during which 3 different $V_{READ}$ are applied to the WL connected to those cells to be read, while

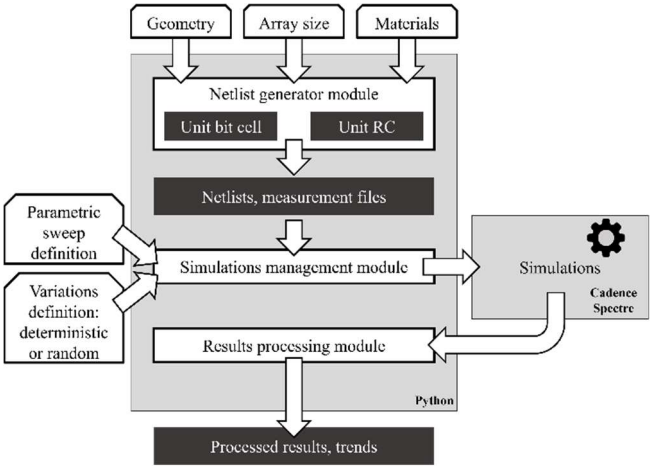


Fig. 1. 3D NAND DTCO Framework, composed of the three key modules: *netlist/measurements generator*, *simulations management module* and *results processing module*.

979-8-3503-4859-0/DATE24/© 2024 EDAA

$V_{PASS}$ biases the WLs controlling the other cells placed in series. The pre-charged sense node ideally discharges only if the cell of interest has a *Vth* lower than the applied $V_{READ}$. Logic operations are performed on the results of each read phase, so that the bit content can be completely read by associating the discharge sequence to the assumed encoding. Current and voltages are evaluated for 3 different timestamps in each phase: 75 ns, 145 ns and 215 ns after the start.

***Design Space Exploration and Figures of Merit -*** Netlists and measurements files are flexibly generated, and sweep simulations are run (Table I). Worst-case bit assignment is assumed: the highest possible *Vth* has been written in every cell, apart from the cells being read. The latter is written with the second-highest *Vth*, to which the lowest $I_{READ}$ discharge at the sense node is associated. The worst-case has been considered in terms of cell position too: the furthest cell from the WL driver and from the BL node is the one being read. Off-current ($I_{OFF}$) is the subthreshold current flowing when the $V_{READ}$ is lower than the *Vth*, and it undesirably discharges the pre-charged sense node. Sense node sensitivity is computed as the voltage difference between the sense node discharged by the $I_{OFF}$ and the one discharged by the $I_{READ}$. The highest $V_{READ}$ phase is of interest, being associated to the highest $I_{OFF}$.

TABLE I. SIMULATIONS SETTINGS

| **Parametric sweeps** | | | | | |
|---|---|---|---|---|---|
| | *N_z* [*units*] | *s_BL* [*nm*] | *SS* [*mv/dec*] | *Vth* variability [*mV*] | *Vth* windows [*V*] |
| **Nominal** | *NA* | 40 | 200 | 0 | 1.5(S) |
| **Sweep Values** | 1 - 800 | 20 - 60 | 150 - 300 | ±10, ±50, ±100 | 1.5(S), 2.25(M), 3.75(L) |

## IV. RESULTS AND DISCUSSION

Parametric sweeps (Table I) results are here presented.

***Number of Layers*** – Sweeps of the number of layers (*N_z*) for different *Vth* windows (S, M, L) have been performed. For each *N_z*, the best sensitivity among the 3 timestamps has been considered in Fig. 2. The sensitivity decreases with *N_z* for any *Vth* window: with a fixed read time, the maximum sensitivity is set by the maximum achievable discharge driven by the $I_{READ}$, which decreases as *N_z* rises. Larger voltage windows (M, L) bring the advantage of higher reliability. On the other hand, the increase of the $I_{OFF}$ in the case of M and L windows leads to a decrease in the sensitivity below the smaller S window for $N\_z > 200$. If *N_z* is fixed, an optimum read time exists, trading off the sensitivity with the read speed.

***Bit Line Spacing -*** The effects of BLs spacing (*s_BL*) on the sensitivity are shown in Fig. 3. The sensitivity falls at smaller *s_BL*: larger capacitive coupling is introduced by closely spaced lines. A larger discharge on the sense node corresponding to the highest *Vth* cell is due to discharging adjacent lines causing larger crosstalk on the victim line.

***Threshold Voltage Variability*** – Custom sensitivity analysis as function of *Vth* variability can be performed. Progressively larger percentage variation of the $I_{READ}$ can be observed as *N_z* increases, when deterministic *Vth* variation is introduced to account for the worst-case sensitivity. Smaller voltage overdrives are biasing the cells in the case of larger *N_z*, and these overdrives are reduced proportionally in larger measure by the same absolute *Vth* variation. Introducing process-aware *Vth* variations enables error-rates analysis.

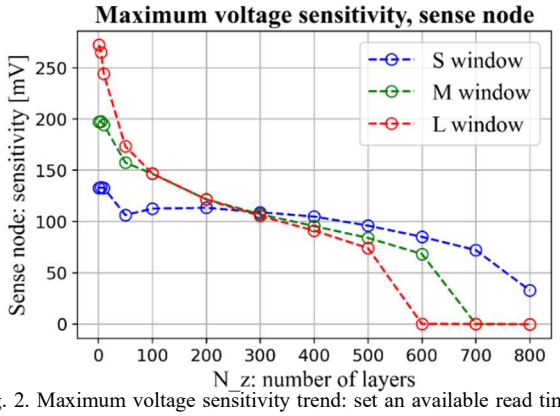


Fig. 2. Maximum voltage sensitivity trend: set an available read time, the sensitivity decreases with *N_z* for all voltage windows.

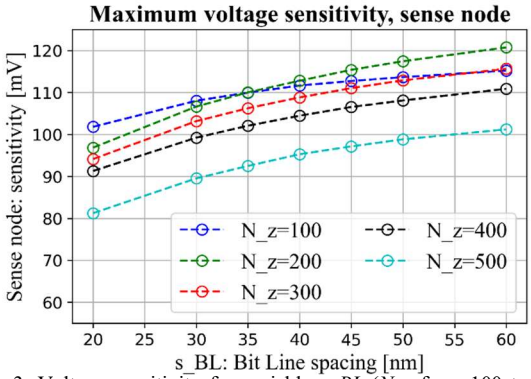


Fig. 3. Voltage sensitivity for variable *s_BL* (*N_z* from 100 to 500). Increased crosstalk with closely spaced BLs reduces the sensitivity.

***Transistor parameters*** – $I_{OFF}$ and $I_{READ}$ values are directly influenced by transistor parameters like the Sub-threshold slope (*SS*) and carrier mobility ($\mu$). The increase of the $I_{OFF}$ at larger *SS* is counteracted by the reduction of gate-to-source and source-to-drain voltages for larger *N_z*. Given a fixed reading time, $V_{PASS}$ and $V_{READ}$, smaller arrays show a sensitivity decrease with $I_{READ}$ improvement due to higher gate-to- and drain-to-source biasing in subthreshold region leading to higher $I_{OFF}$ values too.

## V. CONCLUSION

This work has introduced a predictive parametric framework for MLC 3D NAND Flash readout at the array level. We discerned that number of layers, read-current and read time stand out as pivotal factors in ensuring readability while maximizing memory density beyond the 300-stack.

## REFERENCES


[1] C. Monzio Compagnoni, *et al.*, "Reviewing the Evolution of the NAND Flash Technology," *Proc. IEEE*, vol. 105, no. 9, pp. 1609–1633, Sep. 2017.

[2] S. Rachidi, *et al.*, "Enabling 3D NAND Trench Cells for Scaled Flash Memories," in *2023 IEEE IMW*, Monterey, CA, USA: IEEE, May 2023, pp. 1–4.

[3] L. Larcher, *et al.*, "Modeling nand Flash Memories for IC Design," *IEEE EDL*, vol. 29, no. 10, pp. 1152–1154, Oct. 2008.

[4] S. Sahay and D. Strukov, "A Behavioral Compact Model for Static Characteristics of 3D NAND Flash Memory," *IEEE EDL*, vol. 40, no. 4, pp. 558–561, Apr. 2019.

[5] A. Bansal, *et al.*, "An Analytical Fringe Capacitance Model for Interconnects Using Conformal Mapping," *IEEE TCAD*, vol. 25, no. 12, pp. 2765–2774, Dec. 2006.

[6] D. G. A. Neto, *et al.*, "A 5-DC-parameter MOSFET model for circuit simulation in QucsStudio and SPECTRE," in *2023 IEEE NEWCAS*, Edinburgh, United Kingdom: IEEE, Jun. 2023, pp. 1–5.

[7] D. Verreck, *et al.*, "Quantitative 3-D Model to Explain Large Single Trap Charge Variability in Vertical NAND Memory," in *2019 IEEE IEDM*, San Francisco, CA, USA: IEEE, Dec. 2019, p. 32.1.1-32.1.4.