Understanding the Floating-Body Effect Simulation and Optimization in 3D-DRAMs

Salvatore Maria Amoroso
Synopsys NE, Ltd
Glasgow, G3 8HB UK
samoroso@synopsys.com

Geert Eneman
Imec
Leuven B-3001, Belgium
geert.eneman@imec.be

Plamen Asenov
Synopsys NE, Ltd
Glasgow, G3 8HB UK
pasenov@synopsys.com

Meng-Hsuan Ke
Synopsys Taiwan Co., Ltd.
Hsinchu Hsien, Taiwan
mke@synopsys.com

Nouredine Rassoul
Imec
Leuven B-3001, Belgium
nouredine.rassoul@imec.b

Inhee Lee
Imec
Leuven B-3001, Belgium
inhee.lee@imec.be

Attilio Belmonte
Imec
Leuven B-3001, Belgium
attilio.belmonte@imec.be

Ko-Hsin Lee
Synopsys Taiwan Co., Ltd.
Hsinchu Hsien, Taiwan
kohsin@synopsys.com

Xi-Wei Lin
Synopsys, Inc
Sunnyvale, CA, USA
xiwei@synopsys.com

Victor Moroz
Synopsys, Inc
Sunnyvale, CA, USA
victor.moroz@synopsys.com

In this paper we study the physics and simulation of the floating-body effect (FBE) in advanced DRAM technologies, such as 4F2 and 3D DRAMs. Our results highlight that this effect is modulated by the bitline voltage transition speed and, therefore, proper transient TCAD simulations in the nanoseconds regime must be carried out to accurately describe the FBE. We also show how the DRAM transistor can be optimized, by means of TCAD-driven Multi-Objective Optimization, for write/read speed and retention including FBE minimization. Finally, we highlight the role of statistical variability, induced by random dopants and discrete traps, in enhancing leakage and FBE and how this can affect the transistor optimization directions.

Keywords—Floating Body Effect, TCAD, 3D-DRAM, 4F2 DRAM, memory, variability, reliability.

I. Introduction

Planar DRAM technology is facing challenges that may become unsurmountable beyond the 10nm (half-pitch, F) feature size **[1]**. A way to circumvent the scaling roadblocks is to achieve density increase by means of vertical stacking instead of feature size shrinkage. A first step towards this approach is the shift from a buried channel array transistor (BCAT) $6F^2$ DRAM layout **[2,3]** to a $4F^2$ layout featuring a (Gate-All-Around or Double-Gate) vertical channel transistor (VCT) with a self-aligned capacitor stacked on top **[4]**. The next consequential step in this bit density race consists of the adoption of 3D stacked DRAMs **[1,5]**, where the VCT-based vertical cell is now flipped on the side and many of these cells can then be vertically stacked: this relieves some of the lithography burdens, but introduces new challenges related to topography, etching uniformity, mechanical stress, variability and reliability **[1,6]**. Moreover, the adoption of a Gate-All-Around (GAA) or Double-Gate (DG) transistor introduces the risk of floating body effects, which may affect data retention during bitline (BL) switching – known as "dynamic retention" **[4,7]**. This paper focuses on the physics and simulation of the floating-body effect (**FBE**) in $4F^2$ and 3D-DRAMs, highlighting how this effect is modulated by the BL transition speed and how to properly account for this in transient TCAD simulation. We will then show how the DRAM transistor can be optimized, by means of TCAD-driven Multi-Objective Optimization (MOO), for write/read speed and retention including FBE minimization. Finally, we will highlight the role of statistical variability, induced by random dopants and discrete traps, in enhancing leakage and FBE and in affecting the transistor optimization directions.

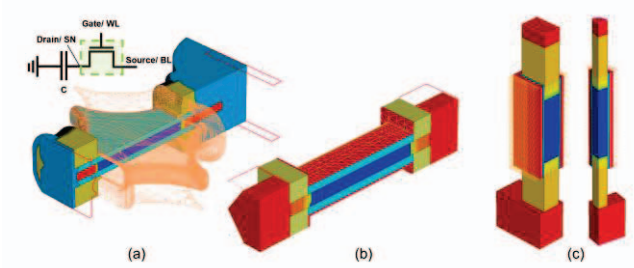

Figure 1: *(a) Realistic 3D-DRAM structure from Process Explorer emulation and (b) simplified version for 3D TCAD simulation; (c) Double-Gate and Gate-All-Around $4F^2$ DRAMs.*

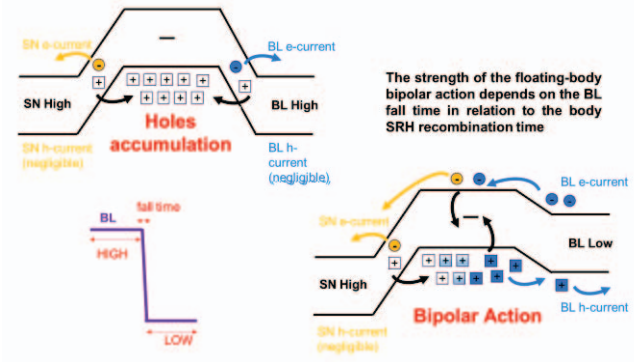

Figure 2: *Illustrative band diagram to explain the FBE physics during a BL switch (including the body SRH recombination)*

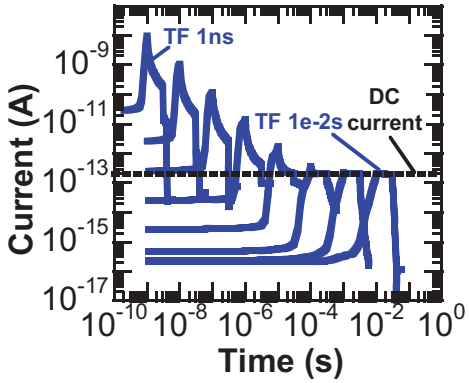

Figure 3: *Transient TCAD simulation of the drain current for BL fall times (TF) varying from 1ns to 1e-2s.*

II. Simulation Methodology

The process and device simulation is carried out by means of Synopsys Sentaurus Process and Sentaurus Device **[8,9]** tools. Although a state-of-the-art process emulation flow has been implemented in Synopsys Sentaurus Process Explorer **[10]** to obtain realistic hypocycloid channel shapes (**Fig.1a**), for this study we are adopting a simplified GAA structure (**Fig.1b**) , which offers similar electric characteristics, high simulation throughput and tunability for the optimization phase. To evaluate the retention and FBE we perform transient simulations featuring band-to-band (BTB) and trap-assisted-tunneling (TAT) **[9]**, whilst program and read performances are evaluated with a set of orientation/stress-dependent mobility models (including Philips, Lombardi, Canali) **[9]**. The transistor optimization is carried out using the Synopsys Sentaurus Calibration Workbench **[11]**, where the non-dominated sorting genetic algorithm NSGA-II **[12]** has been implemented to perform Multi-Objective-Optimization (MOO) tasks in presence of competing factors (e.g. simultaneous of DRAM optimization for both read/program speed and retention robustness). Finally, the impact of discrete dopants and traps is evaluated using Synopsys GarandVE **[13]**, which has been previously presented for planar DRAM applications **[14]**.

III. Simulation Results

In the remainder of the manuscript, we will report quantitative results for the 3D-DRAM GAA transistor (**Fig.1b**), but similar conclusions hold true for the $4F^2$ structures (**Fig.1c**). The physical mechanisms involved in the FBE during a high-to-low BL switch are sketched in **Fig.2**. During a retention phase (high BL), holes (generated by BTB and TAT) accumulate in the transistor's body because of the absence of a body contact. These holes lower the source-to-drain electrostatic barrier and, once the BL is switched to low, create a significant hole current at the BL contact with a bipolar action that is deleterious on the storage node (SN) data retention. It is worth to highlight a point that is rarely considered: the strength of the floating-body bipolar action depends on the BL fall time in relation to the body Shockley–Read–Hall (SRH) recombination time. Indeed, if the BL transient from high to low is slower than the typical SRH recombination times (scale of microseconds), the holes will have a very high chance to recombine with the electron current, therefore limiting the strength of FBE. On the other hand, if the BL switches in the order of nanoseconds (as it is the case in real DRAM circuits) then the bipolar action can take place at its maximum strength, as the internal recombination phenomenon becomes negligible. This is confirmed in **Fig.3**, where we report the transient simulation results for the drain current during a BL switch for several BL fall-time amplitudes (from ns to ms), highlighting that a bipolar action is observed only at BL timescales of 10µs and below. These results are further analyzed in **Fig.4** and **Fig.5**, were we report snapshots of current, bands and carrier densities during the BL transient for

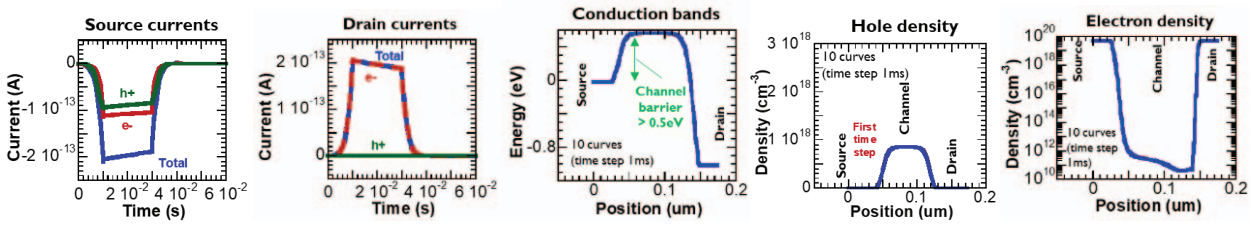

Figure 4: *For a slow (10ms) BL fall time, no FBE is observed: a) Source current during transient is a combination of electron and hole current, while b) drain current vs time is composed completely of electron current, consistent with a quasi-static model where off-current is a mix between BTB and conduction current. Ten curves for c) conduction bands, d) hole densities and e) electron densities with a time step of 0.1ms are plotted during transient while the bitline is a 0V. No time dependence is observed.*

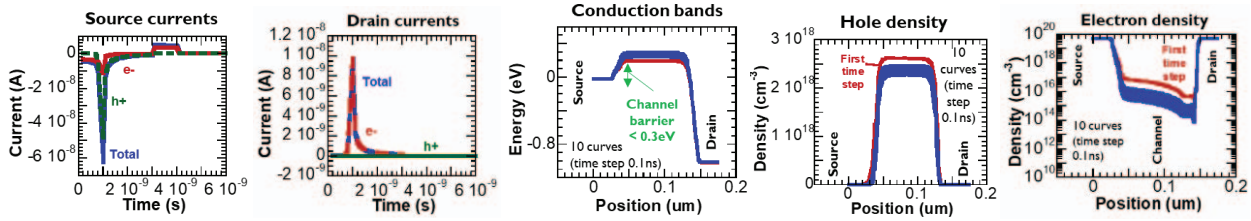

Figure 5: For a s*hort (1ns) BL fall time, significant FBE is observed: a) source currents vs time show a discharging hole peak, as well as b) significant drain current enhancement (compared to Fig. 4b) due to electron conduction current, caused by c) the lower channel barrier than in the quasi-static state (cf. Fig 4c). This barrier decrease is the consequence of d) hole accumulation in the channel which dominates the e) electron accumulation*

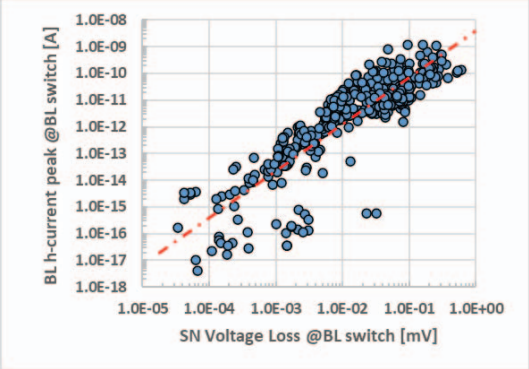

Figure 6: *Hole current peak at the BL contact vs SN voltage drop during a BL switch.*

the case of BL fall-time =10ms and 1ns, respectively. For very slow BL switching, there's barely any movement of carrier or bands profiles and no hole current peak is collected at the BL contact (**Fig.4**). For the case of nanosecond switch we can instead observe a significant hole peak at the BL contact as well as a significant band and carrier profile variation during the transient discharge (**Fig.5**). *These novel results highlight that a TCAD simulation of the FBE must be carried out with transients in the timescale of nanoseconds if we want to have a quantitative evaluation of the charge loss in a realistic DRAM circuit*. We have adopted this in the next step of our analysis, where a Multi-Objective Optimization (MOO) of the cell is carried out to find the best set of parameters (here we consider Gate Work-function (WF), Extension Doping (NExt), Channel Doping (NChan), Write Voltage (VGwrite), Channel thickness (Tnm) and Spacer Length (LSPnm)) to simultaneously optimizing conflicting targets such as the retention time of "0" (t_{ret0}) and "1" (t_{ret1}) bits (including FB effect) and the write time. The first task of this optimization analysis is to identify a figure of merit (FOM) able to quantify the FBE and apt to be assigned to the automated optimization engine. The hole current peak collected at the BL contact during the BL switch is a major fingerprint of the FBE and a good candidate FOM. **Fig.6** shows that this hole peak current correlates well with the SN voltage loss during a BL switch, across a large range of parameters variation. However, the correlation is not perfectly 1:1, therefore highlighting that by optimizing against the BL hole current (instead of the SN loss) we can isolate the specific FBE contribution, removing the voltage loss components due to electrostatic effects. Our optimization is therefore carried out with the following targets: t_{ret0}, t_{ret1}, t_{write} and $I_{BL\_h\_max}$, where the first three are computed as integrals over I_dV_d static characteristics, and the fourth target is calculated from transient simulations. **Fig.7** shows the optimization results using a parallel coordinate plot. It is worth noting that a MOO algorithm explores the parameters space in search of a set of optimal points, where the optimality is reached for a full Pareto front of *non-dominated* solutions **[15]**. The designer can then explore different pathways among the multiple solutions of a Pareto front and select a subset by assigning weights to the conflicting targets. In **Fig.7** we are highlighting the optimization solutions that represent good compromises in maximizing the bit "0" and bit "1" retention whilst minimizing the writing time and FBE. We can see these are achieved by design paths featuring WF>4.6eV (mainly constrained by bit "0" retention), NExt around 1e18cm-3 (compromise between access resistance for writing and leakage for retention), NChan around 5e16cm-3 (compromise between electrostatic control and leakage), VGwrite> 2.5V (write time constraints), Tnm<10nm (electrostatic control), LSPnm>17nm (leakage constraints). Moreover, in **Fig.8** we report the data across the optimization space for the data "1" retention and the FBE, highlighting a very strong anti-correlation*: hence, FBE can be minimized by maximizing the static data "1" retention, i.e. minimizing the drain-to-channel junction BTB and TAT leakage*. Finally, we complete our analysis studying the impact of statistical variability induced by doping and defects discretization. **Fig.9** show an example of a statistical instance

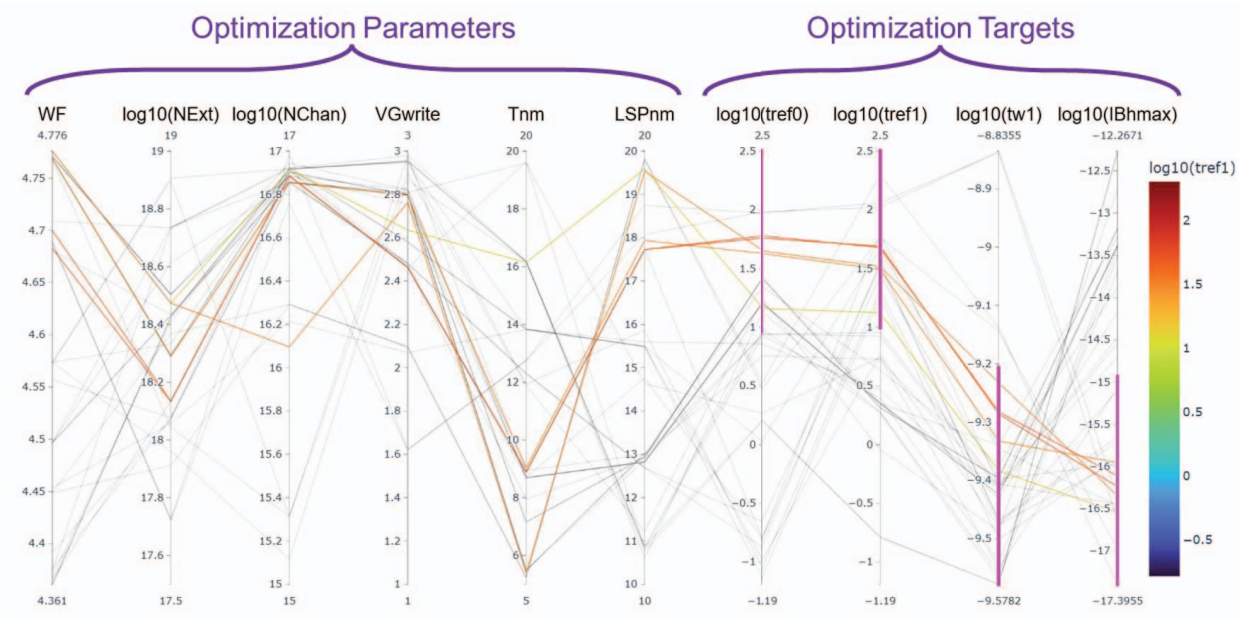

Figure 7: *Parallel coordinates plot showing the Pareto solutions from the MOO. The pink bars are filtering the solutions that maximize retention and minimize FBE and write time.*

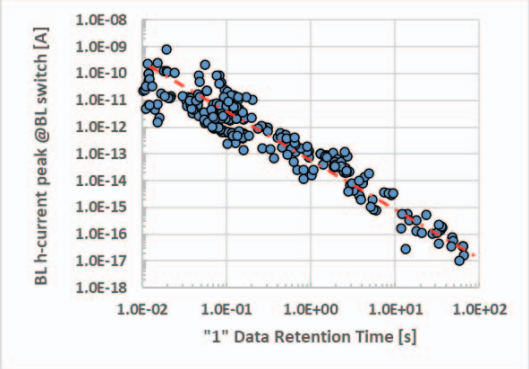

Figure 8: *Hole current peak at the BL contact vs "1" data retention.*

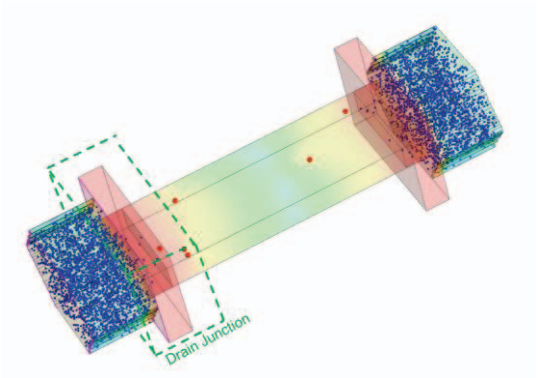

Figure 9: *3D-DRAM device with random discrete dopants and discrete traps.*

of the DRAM transistor as simulated by GarandVE. The discrete dopants close the SN junction induce stochastic peaks in the electric field, which can enhance the TAT leakage in nearby traps. This explains the results in **Fig.10**, where we show that increasing the junction spacer is a less effective solution to reduce junction leakage in the case of statistical simulation with random dopants with respect to the conventional simulation of continuously doped transistors. *Thus, stochastic simulation becomes a necessary tool to obtain the correct directionality in the optimization of ultra-large-scale-integrated 3D-DRAMs, where the low probability tails of the leakage distributions limit both performance and yield*.

CONCLUSIONS

In this paper we have presented a simulation study of the floating body effect (FBE) impacting the dynamic retention of novel 4F2 and 3D DRAMs. Our work highlights three main conclusions: (i) a TCAD simulation of the FBE must be carried out with transients in the timescale of nanoseconds for a quantitative evaluation of the charge loss in a realistic DRAM circuit; (ii) FBE can be minimized by maximizing the static data "1" retention, i.e. a minimization of the drain-to-channel junction BTB and TAT leakage optimizes both static and dynamic retention; (iii) statistical variability plays a major role in driving the optimization of ultra-large-scale-integrated 3D-DRAMs, where the low probability tails of the leakage distributions limit both performance and yield, therefore calling for the adoption of statistical simulation methods and tools.

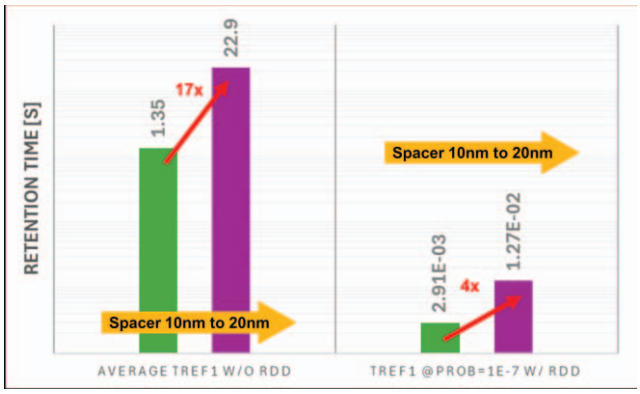

Figure 10: Impact of spacer length optimization from simulation without (left) and with (right) random discrete dopants and traps.

REFERENCES

[1] J. W. Han, et al., "Ongoing Evolution of DRAM Scaling via Third Dimension -Vertically Stacked DRAM", IEEE VLSI Tech Dig., 2023, DOI:10.23919/VLSITechnologyandCir57934.2023.10185290

[2] J-Y Kim et al., "The breakthrough in data retention time of DRAM using Recess-Channel-Array Transistor(RCAT) for 88 nm feature size and beyond" VLSI Tech Dig., 2003, DOI:10.1109/VLSIT.2003.1221061

[3] T. Schloesser et al., "6F2 buried wordline DRAM cell for 40nm and beyond", IEDM Tech. Dig., 2008, DOI:10.1109/IEDM.2008.4796820

[4] H. Chung et al., ''Novel 4F2 DRAM cell with vertical pillar transistor(VPT),'' in Proc. Eur. Solid-State Device Res. Conf. (ESSDERC), 2011, DOI:10.1109/ESSDERC.2011.6044197.

[5] K.S. Choi et al., "A Three Dimensional DRAM (3D DRAM) Technology for the Next Decades" IEEE VLSI Tech Dig. 2024, DOI:10.1109/VLSITechnologyandCir46783.2024.10631471

[6] X. Wu et al., "Signal Margin, Density, and Scalability of 3-D DRAM: A Comparative Study of Two Bitline Architectures" IEEE-TED, 671-677, 2025, DOI:10.1109/TED.2024.3520074

[7] Y. Cho et al., "Suppression of the Floating-Body Effect of Vertical-Cell DRAM With the Buried Body Engineering Method" IEEE-TED, 3237, 2018, DOI:10.1109/TED.2018.2849106

[8] Sentaurus Process User Guide W-2024.09, Synopsys;

[9] Sentaurus Device User Guide W-2024.09, Synopsys;

[10] Sentaurus Process Explorer User Guide W-2024.09, Synopsys;

[11] Sentaurus Calibration Workbench User Guide W-2024.09, Synopsys;

[12] K. Deb et al., "A fast and elitist multi-objective genetic algorithm: NSGA-II", IEEE Transactions on Evolutionary Computation, 6(2), 182-197. DOI:10.1109/4235.996017

[13] Garand-VE User Guide W-2024.09, Synopsys;

[14] S.M. Amoroso et al., "High-sigma analysis of DRAM write and retention performance: a TCAD-to-SPICE approach" IEEE SISPAD 2020, DOI:10.23919/SISPAD49475.2020.9241690

[15] I. Giagkiozis, P.J Fleming, "Methods for multi-objective optimization: An analysis" Inf. Sci., 338, 2015, DOI:10.1016/j.ins.2014.08.071