# Modeling of Cross-Talk Effects in Floating-Gate Devices Using TCAD Simulations

Yv. Saad [1 2], M. Ciappa [2], P. Pfäffli [1], L. Bomholt [1] and W. Fichtner [1 2]
[1] Synopsys Switzerland LLC
Affolternstrasse 52, CH-8050 Zurich, Switzerland
[2] Swiss Federal Institute of Technology (ETH), Integrated Systems Laboratory
ETH-Zentrum, CH-8092 Zurich, Switzerland
E-mail: yves.saad@synopsys.com

***Abstract*—Technology CAD (TCAD) modeling is used to develop, analyze, and optimize flash memory devices under all operating conditions, taking into account three-dimensional effects such as cross-talk between the cells. A methodology for structure generation, meshing, device simulation, and characterization of flash memory devices is proposed. The results demonstrate the effectiveness of full 3D simulation models for flash memory cells, which capture the geometrical, physical, and electrostatic effects.**

**Keywords-component; TCAD simulations, flash memories, cross-talk.**

## I. INTRODUCTION

Nonvolatile memories are designed as single cells packed in a very dense array. This peculiarity has relevant impact on the characteristics of each cell due to the strength of the capacitive coupling between adjacent cells, which increases with increasing integration [1]. A major challenge in the design of flash memories is to minimize the mutual capacitive coupling among the cells, so that the potential of the floating gate of a given cell is not affected by the state of the adjacent cells [2]. A comprehensive 3D model is needed to capture all coupling effects, which takes into account both the microscopic features of the device and the long-range electrostatic interaction [3].
In this paper, a generalized methodology is demonstrated to analyze different configurations of the cells. The generation and the characterization of flash memory cells are carried out using Sentaurus TCAD tools [4].
Process simulation is introduced to generate the model, which is subsequently meshed by an appropriate strategy.
Finally, the simulation model is used to predict both the electrical characteristics and the strength of the electrostatic interference effects.

## II. MODEL

The model is generated by appropriately combining process simulation and process emulation. The shallow trench isolation (STI) profile is created by a process simulator where the full process steps can be reproduced such as oxidation, diffusion, annealing, and stress calculations. Subsequently, the structure is completed with geometrical operations in the process emulator. The floating gate and the control gate, which are separated by the ONO layer, are deposited and etched depending on the masks' descriptions. An oxide layer is deposited to separate the gates from the nitride cap layer. The nitride thickness is assumed as a parameter for the following simulations. The additional cell descriptions are listed in Table 1.

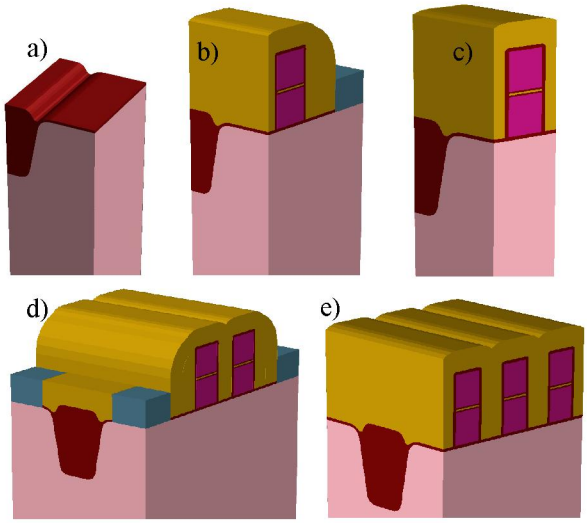


**Figure 1** The steps of the structure generation are represented: (a) the substrate and the STI extruded, (b) one half NOR cells, (c) one half NAND cell, (d) NOR block of four cells, and (e) NAND block of six cells.

Different strategies are used depending on the considered technology. For the 90 nm NAND technology, one quarter of the real cell is built and, after reflection, operations are performed that take advantage of the symmetrical positions and the identical structure of the transistors. For the 110 nm NOR technology, one half of a cell that consists of a drain extension and a source side is generated. Based on these models, blocks of six NAND flash cells and four NOR flash cells are generated.

## III. MESHING STRATEGY

The critical steps for 3D device simulations are the mesh generation and mesh optimization, which influence the quality of the results. Different mesh engines can be used depending on the complexity of the model and the effects to be captured.

 

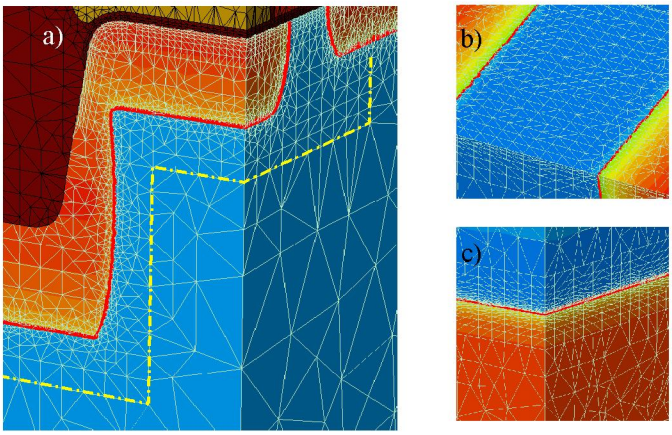


**Figure 2** Different types of mesh used where appropriate are represented: (a) anisotropic and isotropic mesh are used as layering from some interfaces, (b) the regular grid at the surface of the channel, and (c) the junction detection and layering algorithm result.

TABLE 1
CELL DESCRIPTION AND PARAMETER DEFINITIONS

| Description | Default values [nm] |
|---|---|
| ONO thickness | 5/5/5 |
| Nitride spacer thickness | 20 >> 100 |
| Oxide sidewall thickness | 10 |
| Tunnel oxide thickness | 10 |
| Control gate polysilicon height | 100 |
| Floating gate polysilicon height | 100 |
| NAND channel length | 90 |
| NOR channel width | 110 |
| STI height | 300 |
| Channel width | 150 |
| STI width | 100 |

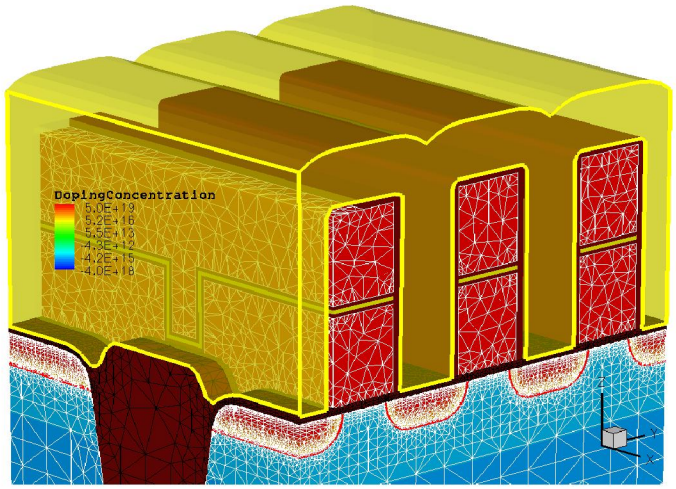


**Figure 3** The final meshed structure of six cells in a NAND blocks is represented.

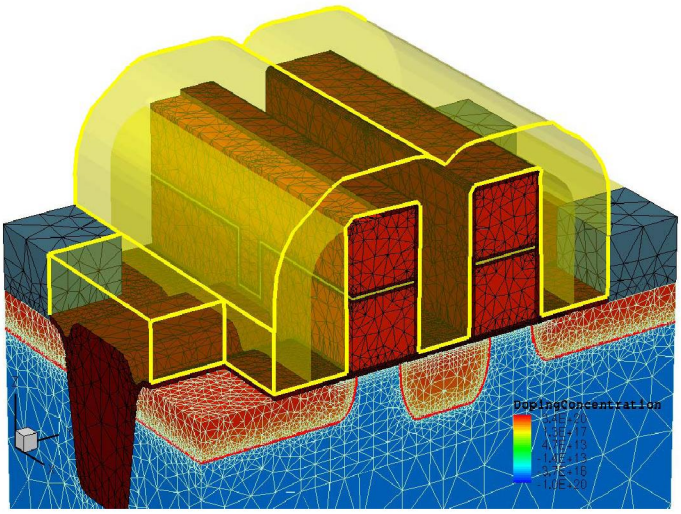


**Figure 4** The final meshed structure of four cells in a NOR block is represented.

In our case, a boundary-offsetting mesh approach is used that generates conformal layers from the interfaces into the bulk. This type of mesh is the most suitable to refine nonplanar surfaces, rounded corners, overetch regions, STI/silicon interface, and the tunnel oxide layer. In addition, it is possible to combine isotropic and anisotropic refinement. For the bulk, an anisotropic regular grid is applied, while for the doping refinements of the drain/source, an isotropic mesh is applied. For the n-well/p-well junction, an algorithm is used that detects the junction and generates layers parallel to it. On the top substrate surface, a regular surface mesh is enforced to form a better bulk layering. Layering is performed from the oxide/nitride interfaces to ensure the capture of the electrostatic coupling between the cells. Examples for the different types of mesh are shown in Figure 2. For the NAND six-cell block, the entire structure is meshed with the doping defined by analytical functions. For the NOR four-cell block, a single cell is meshed, using analytical functions for the doping distribution. The meshed model is subsequently mirrored to create the full block.

The final meshed structures of six NAND and four NOR block cells are shown in Figure 3 and Figure 4 respectively.
Additional criteria influence convergence of the device simulations. The layer distance from the interface must be about ten times the surface element size. In addition, element connectivity must to be minimized. It is important to check not only the mesh quality on the surface but also the internal structure nodes, especially in the layered zones.

## IV. DEVICE SIMULATIONS

For typical 3D simulations, the number of mesh nodes exceeds 100 000. With this mesh, it is important to control all the criteria needed to obtain good convergence, to speed up the run-time, and to limit the memory size.
The physical phenomena in a semiconductor device are governed by partial differential equations (PDEs), which solve the transport of the carriers. The PDEs are difficult to solve due to the high degree of the nonlinearity involved. The success in resolving the problem strongly depends on the quality of the mesh and the solver conditions [5].

Important information on the mesh quality can be obtained from the device simulator output when applying the average box method algorithm for the discretization that compares and report the box method volume to the physical volume of the structure.

The solution of large sparse linear systems that arise from the PDE is carried out by direct or iterative methods. The advantage of the direct methods is in their high reliability and accuracy, while their disadvantages are the large amount of memory and run-time that could make them impractical. An alternative is to use the iterative solvers that require reasonable memory space and run-time. Special care must be taken to set up their parameters. The iterative linear solver (ILS) is used in this work in combination with GMRES, which solves the unsymmetrical sparse linear systems [6]. This combination was found to be superior to other methods.

The generated models are not only suitable for simulating electrical characteristics. They are also applicable for operation characterization, capacitance extraction, parasitical quantification, mechanical stress simulations, and estimation of the impact of process variations.

### *A. Electrostatic simulations*

We perform electrical simulations to extract the capacitance between different regions. This type of simulation is helpful to extract in a fast, accurate, and inexpensive procedure some important input parameters used in SPICE models.

The ONO equivalent capacitance is determined and the parasitic capacitances between the different floating gates are evaluated as a function of the spacer width.

The capacitances between all contacts are extracted by small-signal AC device simulation at a frequency of 100 MHz [7]. The accuracy of this method depends on the mesh quality and the model authenticity. The run-time of an electrostatic simulation is about 10 minutes for the largest structure.

### *B. Electrical simulations*

Simulations of single flash memory cells (*isolated*) and blocks of flash memory cells (*coupled*) have been performed.

The NOR block consists of four transistors placed in two rows that are connected by a doping underline contact from the source side. Each transistor has a drain extension contact where their gates are separated by nitride layers. The NAND block consists of six flash cells grouped as two rows. The gates are separated by nitride layers. The thickness of the nitride layer is considered to be a parameter in investigating the cross-talk effects in the cells blocks. Only half of the transistor is simulated due to its symmetry.

#### *1) Read operations*

During the reading operation, the potential of the floating gate is sensed by applying a voltage to the control gate and by recording the drain current of the flash cell transistor. The transistor characteristic depends on the capacitive coupling with the neighboring regions as follows:

$$Q = C_{fg}.(V_{fg} - V_{cg}) + C_s.(V_{fg} - V_s) + C_d(V_{fg} - V_d) + C_b.(V_{fg} - V_b) + \sum C_{neighbor}.(V_{fg} - V_{fgneighbor}), \quad (1)$$

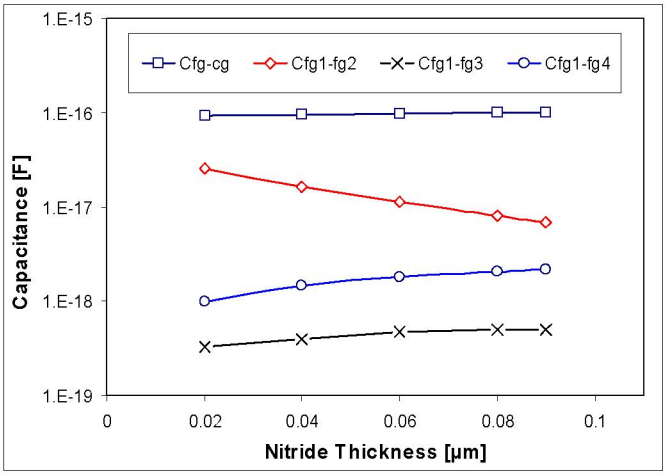


**Figure 5** The capacitance between the adjacent cells for a NAND block. Cfg-cg is the capacitance between the floating gate and the control gate, Cfg1-fg2 is the capacitance between the cells of the other row, Cfg1-fg3 is the capacitance between the adjacent cells, and Cfg1-fg4 is the capacitance between the cells with the same control gate.

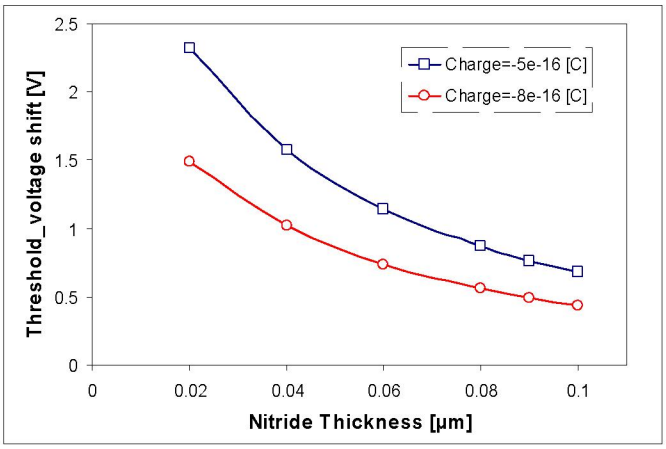


**Figure 6** Threshold voltage shift is represented as a function of the nitride thickness and charge state of the neighboring cells in NAND block.

where $Q$ is the floating-gate charge, $C_{fg}$ is the floating gate/control gate capacitance, $C_s$ is the floating gate/source capacitance, $C_d$ is the floating gate/drain capacitance, $C_b$ is the floating gate/substrate capacitance, and $C_{neighbor}$ are the floating gate/floating gate of the neighbors' cells.

The term $\sum C_{neighbor}.(V_{fg}\text{-}V_{fgneighbor})$ represents the cross-talk effect between the gates of the block.

The neighboring capacitances vary depending on their distance from the cell under investigation such that, in both configurations, there are generally three different capacitances due to the symmetry of the model.

The simulation enables to quantify the dependency of the threshold voltage of the flash cell on the charge and on the potential of the neighboring cells. The cross-talk between the gates increases with increasing charge on the floating gates and decreases with increasing distance between the cells in the block.

The simulations of the DC characteristics have been performed on both the *isolated* and the *coupled* cells.

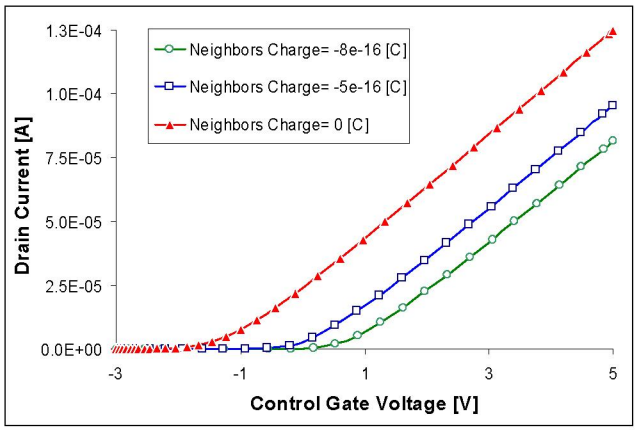


**Figure 7** The characteristics of a NAND flash cell depending on the charge state of its neighbors is represented. The nitride isolation thickness is 40 nm and the drain bias is 1 V.

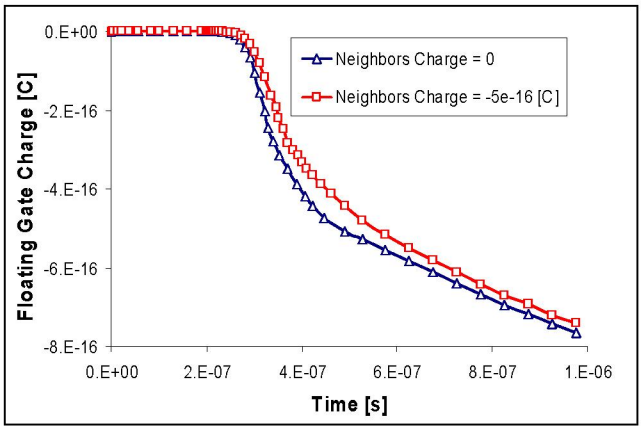


**Figure 8** The charge of the floating gate as a function of the programming time depending on the neighbors' charges in a NOR block of four cells is represented. The drain is ramped to 4 V and the control gate to 7 V.

The simulations of the *isolated* cell can be reproduced by the *coupled* cells when applying the same bias on the contacts and erasing the neighbors' cells.

For both *coupled* cells (NAND and NOR blocks), the threshold voltage strongly depends on the charge stored on the other floating gates. Additionally, as expected, the cross-talk becomes more prominent as the thickness of the nitride isolation decreases. Therefore, the control of the threshold voltage becomes prominent when the nitride thickness decreases below 60 nm. To overcome the cross-talk between the cells, it is necessary to reduce the floating gate height and adopt a low-*k* material [1].

*2) Erase operation*

During the erase operation electrons are removed from the floating gates in a block by Fowler-Nordheim (FN) tunneling across the tunnel oxide. This requires an oxide electric field of about 10 MV/cm. In NOR-type flash memories, such an electric field is usually obtained by applying a moderate positive bias at the source contact and a large negative bias on the control gate.

In NAND-type flash memories, tunneling is induced by applying a high bias on the substrate and by applying a zero bias to the gate.

The erase time of the block has been simulated under consideration of the capacitive interaction with the neighboring cells. The erase time has been observed to increase as a consequence of the cross-talk effect.

TABLE 2
RUN-TIME SIMULATIONS ON 2.4 GHZ AMD OPTERON™ SINGLE PROCESSOR

| | one cell Meshing | Block meshing | onecell Simulations | Block Simulations |
|---|---|---|---|---|
| NOR | 2 min 33000 nodes | 3 min 130000 nodes | 1 h to 2 h | 4 h to 7 h |
| NAND | 1.5 min 30000 nodes | 9 min 160000 nodes | 1 h to 2 h | 8 h to 10 h |

*3) Program operation*

During the program operation, electrons are injected into the floating gate from the substrate either by FN tunneling across the tunnel oxide or by channel hot-electron (CHE) injection.

The cross-talk effect influences the programming time for the NOR technology that is based on CHE and the NAND technology which is based on FN.

## V. CONCLUSION

The capabilities of TCAD simulations have been demonstrated on block flash memories by using a methodology to create a suitable model. The simulations have demonstrated the need for full blocks of flash memory to simulate correctly the characteristics of the cells taking into account the cross-talk effects and the geometrical effects that are combined to the electrical simulations. This model is not limited to perform electrical simulations of the operation of the cells but also degradation, radiation and stress. The simulation run-time is in the range of few hours, which is of great interest for industrial applications and optimization.

## ACKNOWLEDGMENT

The authors would like to thank Synopsys Switzerland support team for their contributions and D. Polimeni for her help.

## REFERENCES

[1] J.D. Lee et al., "Effects of Floating-Gate Interference on NAND Flash Memory Cell Operation," IEEE Electron Device Letters, vol. 23, no. 5, pp. 264-266, 2002.

[2] G. Atwood, "Future Directions and Challenges for ETox Flash Memory Scaling," IEEE Transactions on Device and Materials Reliability, vol. 4, no. 3, pp. 301-305, 2004.

[3] A. Ghetti et al., "3D Simulation Study of Gate and Noise Coupling in Advanced Floating Gate Non Volatile Memories," ICMTD, pp. 157-160, 2005.

[4] www.synopsys.com

[5] G. Garretón, "A Hybrid Approach to 2D and 3D Mesh Generation for Semiconductor Device Simulations," Hartung-Gorre, 1999.

[6] Y. Saad, "Iterative methods for sparse linear systems," Boston: PWS, 1996.

[7] Y. Saad et al., "TCAD Tools for Efficient 3D Simulations of Geometry Effects in Floating-Gate Structures," NVMTS, pp. 77-82, 2004.