Diss. ETH No. 17435

TCAD-Based Three-Dimensional Modeling of Nonvolatile Memories

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Technical Sciences

presented by

YVES SAAD

Dipl. d'Etude Approfondie (DEA)
born 04 January 1979
citizen of Lebanon

accepted on the recommendation of

Prof. Dr. Wolfgang Fichtner, examiner
Dr. Hervé Jaouen, co-examiner

2007

Abstract

Nonvolatile memories are integrated into almost all electronics systems to store data. As their market is one of the most competitive domains in semiconductor industry, TCAD simulations are used in this field to improve the production ramping and to increase the yield. They help to reduce the costly cycle of the design, to model and optimize a technology, and to improve the efficiency of the research and development. In addition, TCAD simulations can assist to improve the reliability and give an inside view of the behavior of the device.

The actual size of flash cell transistors is one of the smallest devices in mass production. They are based on short channel transistors combined to narrow widths. With the continuing miniaturization of the cells, the 3D effects are becoming more prominent. In addition, the structure of a floating-gate device includes an isolated region that is coupled to other regions by capacitance that has a 3D shape. Due to this, the 3D simulations are mandatory to study the behavior of the cells. Moreover, the transistors are positioned in a highly dense organization, creating electrostatic interactions between the cells.

This work proposes a methodology to achieve realistic simulations of floating-gate devices, based on demonstrating the capabilities of TCAD to handle the investigation of similar types of device.

Different possibilities exist to generate the structure of a flash memory device. One technique is based on the use of a process emulator, which can generate the shape of the device using geometrical operations. Another approach consists of using a process simulator that can reproduce the process steps of the flow. In addition, the combination of both techniques is illustrated.

When performing TCAD simulations, the meshing of a 3D structure is one of the most critical steps. Three types of core algorithm and three meshing engines to mesh the structures are compared, such as the hybrid octree modified method using Mesh, the axis-aligned Delaunay method using Sentaurus Mesh and the conformal offsetting method using Noffset3D, which has been selected as the main meshing engine. A detailed explanation of the mesh generation and the criteria needed to build a stable mesh are described and applied to realistic devices.

One of the most interesting advantages of using TCAD is the capability to investigate and extract the capacitances and the coupling coefficients between the terminals in an accurate, simple, and easy technique under different bias conditions. This is performed on different types of structure and is applied to extract the coupling between the adjacent cells of a NOR and NAND block.

The operation of the flash cell is examined based on one isolated cell in 2D and 3D. An introduction to the principles of the physics involved in the read, program, and erase operations of the floating-gate device are presented. The read operation consists of extracting the DC characteristics of the cell. The program operation is simulated by hot carrier injection or Fowler–Nordheim tunneling, which is also used to simulate the erase operation of the cell.

Afterwards, the methodology is applied to a 3D isolated floating-gate memory where all the operation modes are investigated. Added to this, the influences of the narrow channel effect and the overetch effect on the transistor operations are considered. Furthermore, the methodology is extended to study the operation of the cell in a block such as in a NOR and NAND organization. In addition, a full row of 32 transistors that represent a NAND row is simulated in the read operation based on the same techniques. This method is suitable to investigate nonvolatile memories and can be generalized to any MOS technology.

Avant-propos

Les mémoires non volatiles sont intégrées dans la majorité des systèmes électroniques pour sauvegarder les données. Comme leur marché est l'un des plus compétitifs domaines de l'industrie de la microélectronique, la TCAD est utilisée dans ce secteur pour améliorer la production et pour augmenter le rendement. Elle aide à réduire le couteux cycle du design, à modéliser, à optimiser une technologie et à améliorer l'efficacité du département de la recherche et développement. En plus, les simulations TCAD peuvent assister pour améliorer la fiabilité et elles donnent une vue interne sur le comportement du dispositif.

Les dimensions actuelles du transistor de la cellule flash sont l'une des plus petites des dispositifs en production. Ils sont basés sur le transistor à canal court combiné à une largeur rétrécie. Avec la miniaturisation continue des cellules mémoires, les effets 3D deviennent plus proéminents. En plus, la structure du transistor à grille flottante contient une région isolée qui est couplée aux autres régions par des capacitances dont elles ont une forme 3D. A cause de ça, les simulations 3D sont essentielles pour étudier le comportement des cellules flash. De plus, les transistors sont positionnés dans une organisation très dense, qui crée une interaction électrostatique entre les cellules.

Ce travail propose une méthodologie pour effectuer des simulations réalistes sur des dispositifs à grilles flottantes, qui est basée sur les capacités de la TCAD à traiter des études sur des types similaires de dispositifs.

Différents possibilités existent pour générer la structure d'une cellule de mémoire flash. Une technique est basée sur l'utilisation d'un émulateur de process, qui peut reproduire la forme du dispositif en utilisant des opérations mathématiques. Une autre approche consiste à utiliser un simulateur de process qui peut reproduire les étapes de la recette de production. En plus, la technique de combiner l'émulateur de process et du simulateur de process est illustrée.

En réalisant des simulations TCAD, la génération du maillage de la structure 3D est l'une des plus importantes étapes. Trois types de algorithmes et trois outils de maillages sont comparés tel que l'"hybrid octree modified method" en utilisant Mesh, l'"axis-aligned Delaunay method" en utilisant Sentaurus Mesh et le "conformal offsetting method" en utilisant Noffset3D. Ce dernier a été choisi comme le principal outil de

maillage. Une explication détaillée sur la génération du maillage et la sélection des critères sollicités pour la construction du maillage est décrite tout en les appliquant sur des cellules flash.

L'un des plus importants avantages de l'utilisation de la TCAD est la capacité d'examiner et extraire les capacitances et les coefficients de couplages entre les terminaux avec une technique précise, simple et facile, sous différentes conditions. Ceci est accompli sur différents types de structures et appliquer pour extraire le couplage entre les cellules adjacents dans un block NOR et NAND.

L'opération de la cellule flash est examinée dans le cas isolé en 2D et 3D. Une introduction sur la physique impliquée dans la lecture, la programmation et l'effacement du dispositif à grille flottante est présentée. La lecture consiste à extraire les caractéristiques DC de la cellule. La programmation est simulée par l'injection des porteurs chaux ou par Fowler-Nordheim, qui est encore utilisé pour simuler l'opération de l'effacement de la cellule.

Après, la méthodologie est adoptée dans le cas 3D d'un transistor isolé a grille flottante ou tout les modes d'opérations sont examinés. Les influences des effets du canal étroit et de l'"overetch" sur l'opération du transistor sont considérées. De plus, la méthodologie est étendue pour étudier l'opération de la cellule dans un block comme dans le cas d'organisation NOR et NAND. Egalement, un rand de 32 transistors, qui représentent un rand NAND, est simulé dans le mode de la lecture tout en appliquant la même méthodologie de travail. Cette technique convient pour l'étude des mémoires non volatiles et encore elle peut être généralisé et utilisé pour n'importe quelle technologie de MOS.