IntechOpen

Complementary Metal Oxide Semiconductor

Edited by Kim Ho Yeap and Humaira Nisar

COMPLEMENTARY METAL OXIDE SEMICONDUCTOR

Edited by **Kim Ho Yeap** and **Humaira Nisar**

Complementary Metal Oxide Semiconductor

http://dx.doi.org/10.5772/intechopen.71097
Edited by Kim Ho Yeap and Humaira Nisar

Contributors

Kuan W. A. Chee, Tianhong Ye, Rawid Banchuin, Bartomeu Alorda, Gabriel Torrens, Sebastia Bota, Guilei Wang, Henry Radamson, Kim Ho Yeap, Mario Alberto Garcia-Ramirez, José Trinidad Guillen-Bonilla, Maria Esther Macias-Rodríguez, Miguel Bello-Jiménez, Barbara Cortese, Rosa Elvia Lopez-Estopier, Juan Carlos Gutierrez-Garcia, Everardo Vargas Rodriguez, Huaxiang Yin, Klaus Hempel, Dina Triyoso, Elke Erben

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2018 by IntechOpen
eBook (PDF) Published by IntechOpen, 2019
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, The Shard, 25th floor, 32 London Bridge Street
London, SE19SG – United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Complementary Metal Oxide Semiconductor
Edited by Kim Ho Yeap and Humaira Nisar

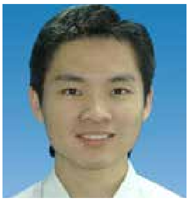
p. cm.

Print ISBN 978-1-78923-496-1

Online ISBN 978-1-78923-497-8

eBook (PDF) ISBN 978-1-83881-512-7

Meet the editors



Kim Ho Yeap is an Associate Professor at Universiti Tunku Abdul Rahman, Malaysia. He is an IEEE senior member, a Chartered Engineer registered with the UK engineering council, and a Professional Engineer registered with the Board of Engineers, Malaysia. He received his MSc in microelectronics from Universiti Kebangsaan Malaysia in 2005 and a PhD from Universiti Tunku Abdul Rahman in 2011. In 2008 and 2015, respectively, he underwent research attachment in University of Oxford (UK) and Nippon Institute of Technology (Japan). He is the external examiner of Wawasan Open University. He is also the Editor in Chief of the i-manager's *Journal on Digital Signal Processing*. He has also been a guest editor for the *Journal of Applied Environmental and Biological Sciences* and *Journal of Fundamental and Applied Sciences*. When working at Intel Corporation, he was involved in the design of the Pentium IV PSC and Celeron NWD-V microprocessors. This earned him 4 Kudos awards from Intel Microelectronics. He has also been given the university teaching excellence award and 16 research grants. He has published more than 100 scientific articles, which include refereed journal and conference papers, books and book chapters.

Humaira Nisar has a BS in Electrical Engineering from the University of Engineering and Technology, Lahore, Pakistan, an MS in Nuclear Engineering from Quaid-i-Azam University, Islamabad, Pakistan, another MS in Mechatronics, and a PhD in Information and Mechatronics from Gwangju Institute of Science and Technology, Gwangju, South Korea. She has more than 15 years of research experience. Currently, she is working as an Associate Professor in the Department of Electronic Engineering, Universiti Tunku Abdul Rahman, Malaysia. She is also the Head of Programme for the Master of Engineering Science Programme. She is a senior member of IEEE. Her research interests include signal and image processing, bio-medical imaging, brain signal and image analysis, and image analysis for wastewater treatment. She has published a number of international journal and conference papers. She has also served on technical committees of various conferences and journals.

Contents

Preface

In this book, *Complementary Metal Oxide Semiconductor* or CMOS devices are extensively discussed. The topics encompass the technology advancement in the fabrication process of metal oxide semiconductor field effect transistors or MOSFETs (which are the fundamental building blocks of CMOS devices) and the applications of transistors in the present and future eras.

Chapter 1 gives an overview of CMOS devices. A brief historical development of field effect transistors is first presented. This is then followed by a general illustration on the PMOS and NMOS transistors. The final part of the chapter briefly discusses the reasons that prompted the integration of both transistors, forming CMOS devices in integrated circuits.

Transistor performance encounters great technical challenges as the feature size shrinks below 32/28 nm. Chapter 2 gives a review of the various process technologies that have been introduced to overcome these challenges. These include the high-k/metal gate, strain engineering, and FinFET structure. A more detailed explanation of some of these methods is covered in the subsequent chapters.

As the size of a transistor continues to shrink, the SiO_2/polysilicon gate stack has been replaced by the high-k/metal gate to enable further scaling. Chapter 3 illustrates the two different high-k/metal gate integration approaches—the gate-first and gate-last approaches (the latter is also known as the replacement gate approach). In both integration schemes, getting the right work functions and threshold voltages for NMOS and PMOS transistors is critical. Studies have shown that the threshold voltage of the transistors is highly dependent not just on the deposited material properties but also on the subsequent device processing steps. This chapter includes a description of the different mechanisms of work function setting in gate-last and gate-first technologies, the sensitivities of the devices on different manufacturing conditions, as well as various special measurement techniques for gate stack analysis.

Chapter 4 presents an overview of the implementation, modeling and pattern dependency of selective epitaxy at the source and drain (S/D) regions in CMOS. Selective epitaxy is applied to these regions so as to induce strain in the channel region. The chapter also discusses wafer in- and ex-situ cleaning prior to epitaxy, the integrity of gates and the selectivity modes for transistors.

Chapter 5 introduces hybrid structures as an alternative method to overcome the scaling limitation of transistors. A detailed elaboration of the two popular hybrid structures, i.e., the nano-electromechanical systems and metal oxide technology, is given here.

Sub-threshold MOSFETs have been widely employed in low-power VHF circuits/systems. The performances of these transistors are mainly determined by three major high-frequency characteristics of intrinsic sub-threshold MOSFETs, i.e., gate capacitance, transition frequency and maximum frequency of oscillation. Due to the physical-level imperfections and variations in the manufacturing process, variations exist in the electrical characteristics of these transistors**.** These variations may affect the performance of the VHF circuits/systems. To minimize the variations, statistical/variability aware analysis and designing strategies have been implemented. Chapter 6 gives a comprehensive review of these analytical models. A novel improved model, which is based on the variation in maximum frequency oscillation, has also been proposed in the chapter.

Digital technology in the nanoelectronic era is based on intensive data processing and battery-based devices. As a consequence, the need for larger and more energy-efficient circuits with large embedded memories is growing rapidly in current system-on-chip (SoC) devices. In this context, where embedded SRAM yields dominate the overall SoC yield, memory sensitivity to process variation and aging effects has aggressively increased. In addition, long-term aging effects introduce extra variability reducing the failure-free period. Therefore, although stability metrics are used intensively in the circuit design phases, more accurate and non-invasive methodologies must be proposed to observe the stability metric for high-reliability systems. Chapter 7 reviews the most extended memory cell stability metrics and evaluates the feasibility of tracking SRAM cell reliability evolution by implementing a detailed bit-cell stability characterization measurement. The memory performance degradation observation is focused on estimating the threshold voltage drift caused by process variation and reliability mechanisms. A novel SRAM stability degradation measurement architecture is proposed to be included in modern memory designs with minimal hardware intrusion. The new architecture may extend the failure-free period by introducing adaptable circuits depending on the measured memory stability parameter.

Chapter 8 illustrates the development of a high-performance low-voltage rating power MOSFET. Power transistors possess low on-resistance and excellent avalanche current capability. Hence, they are very well suited for building automotive electric power steering systems (EPS). In this chapter, planar- and trench-technology power MOSFETs have been designed, modeled, simulated and compared using industry-standard technology computer-aided design (TCAD) tools. The specific on-resistance due to the different device structures is surveyed and analyzed, and various methods are highlighted and compared so that their benefits can be better understood and adopted. Additionally, device ruggedness has been investigated and its improvement was evaluated and established for the trench MOSFET due to gate corner smoothing.

Kim Ho Yeap, Associate Professor, and Humaira Nisar, Associate Professor
Department of Electronic Engineering
Faculty of Engineering and Green Technology
Universiti Tunku Abdul Rahman
Malaysia

Section 1

Introduction

Introductory Chapter: Complementary Metal Oxide Semiconductor (CMOS)

Kim Ho Yeap and Humaira Nisar

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73145

1. Introduction

In 1970s, the number of transistors in an integrated circuit (IC) chip was not more than 10,000 and the feature lengths of the transistor were larger than 1 μm. The Motorola 6800 microprocessor, for instance, had only a count of 4100 transistors in it with a feature length of 6.0 μm. In less than half a century time, however, the IC industries have undergone a dramatic revolution. Nowadays, the number of transistors in a chip can possibly hit 10 billion and the feature length may be as small as 10 nm. The significant increase in the number of transistors has enabled more functionalities to be installed in a chip. This is to say that, the chip found in an electronic device today is much smaller and, yet, more powerful [1]. Since the circuits in a typical chip are designed by incorporating two types of transistors that complement each other, the fundamental building block that powers up electronic circuits is known as a complementary metal oxide semiconductor field effect transistor or CMOS device. To provide readers with an overview of the CMOS device, this chapter gives a concise but complete illustration on the historical development and the operation of the device.

2. A brief history

When transistors were first introduced in early 1900s, they were actually made of vacuum tubes. The vacuum tube transistors were large and cumbersome to be used. In December 1947, John Bardeen, Walter Brattain and William Shockley from the Bell laboratory invented the point-contact germanium transistor. As can be seen in **Figure 1**, this transistor was much smaller in size. It also consumed significantly less power, operated at lower temperature and gave quicker response time. Because of this reason, the vacuum tube transistor was swiftly replaced by its solid-state counterpart. The solid-state transistor is obviously more convenient to be used.

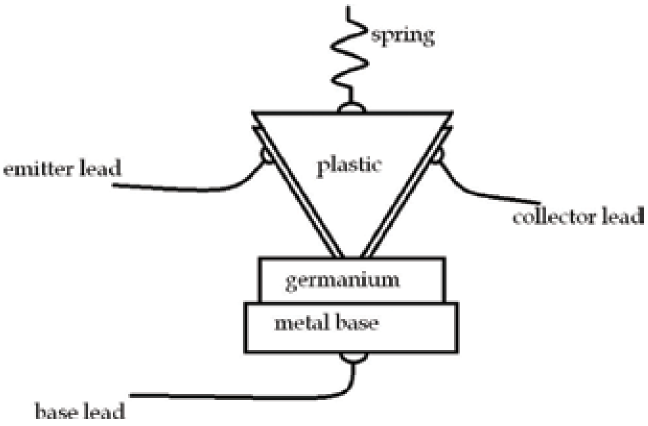


Figure 1. A point-contact transistor.

Very soon after its introduction, the electronic industries went through a dramatic revolution. Because of this significant contribution, the three scientists from the Bell laboratory shared the Nobel Prize in Physics in 1956.

The first commercially available silicon transistors were manufactured by Gordon Teal in 1954. Since silicon gives better performance than germanium, the substrate material for transistors was gradually changed to silicon. In 1955, the first diffused silicon transistor made its appearance. To reduce the resistivity of the collector, an epitaxy was deposited onto the transistor in 1960. In the same year, the planar transistor was proposed by Jean Hoerni [2, 3].

Without knowing each other and using their own methods, Jack Kilby from Texas Instruments and Robert Noyce from Fairchild invented independently the integrated circuits (ICs) in late 1950s. Kilby's IC was merely a simple 0.5 inch germanium bar, with a transistor, a capacitor and three resistors connected together using fine platinum wires; whereas, Noyce's was closer to the look of an IC today – the transistors were etched on a 4-inch silicon wafer. Both Kilby and Noyce shared the patent right for the invention of the integrated circuit. In 2000, Kilby was awarded the Nobel Prize in Physics "for his part in the invention of the integrated circuit".

3. MOSFET

The Metal Oxide Semiconductor Field Effect Transistor or MOSFET acts as an electronic switch or amplifier in circuitries. There are two types of MOSFETs, namely the enhancement-type MOSFET (E-MOSFET) and the depletion-type MOSFET (D-MOSFET). **Figure 2** depicts the basic structure for both types of MOSFETs. As can be observed from the figure, both devices are similar to each other. They comprise four terminals: the drain, source, gate and substrate terminals. The drain and source terminals of the E-MOSFET are separated apart from each other. Unlike the E-MOSFET, however, a channel connecting the two terminals is physically implanted in the D-MOSFET.

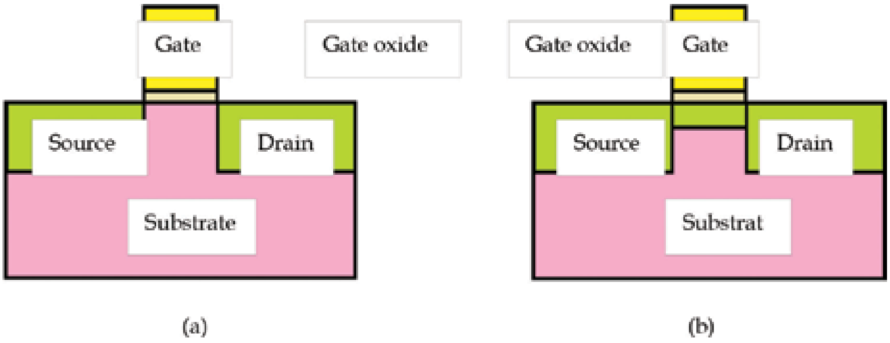


Figure 2. The cross sections of an (a) enhancement mode and a (b) depletion mode MOSFET.

When no voltage is applied to the gate terminal, the E- and D-MOSFETs act like an open and a closed switch, respectively. This is to say that voltage is to be applied to the D-MOSFET in order to have it switched off. Since the E-MOSFET does not require this additional voltage to be switched off, it consumes less power and is popularly used in the IC industries. Hence, the term MOSFET is generally used to refer to the E-MOSFET.

Basically, the device is composed of three layers: a polysilicon layer (i.e. the gate terminal), an oxide layer (i.e. the gate oxide) and a single crystal semiconductor layer (i.e. the substrate). In the early days, the gate terminal was made of aluminum. Indeed, the term MOSFET is coined from these three layers of materials and the fact that it relies on electric field to dictate its switching function. In mid 1970s, however, the gate material was replaced with polysilicon. The high temperature stability of the polysilicon gate is used as a mask to form the self-aligned source and drain terminals via ion implantation, rendering higher accuracy for the formation of these two terminals. Although the gate today is no longer made of aluminum, the term MOSFET has been so widely accepted that it stays until today [2].

3.1. NMOS and PMOS transistors

A MOSFET can be classified into two types, depending on the dopant at the drain, source and substrate regions. When the drain and source terminals are heavily doped with donator ions, such as phosphorous and arsenic, while the substrate is a p-type semiconductor material, the device is known as a negative channel MOSFET or NMOS transistor. On the other hand, when the two terminals are heavily doped with acceptor ions such as boron, and the substrate is an n-type, the device is known as a positive channel MOSFET or PMOS transistor. **Figures 3** and **4** show the symbols of the NMOS and PMOS transistors, respectively. Although the figures show that various symbols have been used to represent the transistors, the third from the left in both figures have been more popularly used in the IC industries.

When voltage V_{DS} is applied to the drain and source terminals, it requires a conducting channel between the two terminals to form a close circuit. Voltage V_{GS} connected between the gate and source terminals control the formation of this channel. It therefore acts like a switch of the transistor. When a positive V_{GS} greater or equivalent to the threshold voltage $V_{GS(th)}$ is applied to an NMOS transistor, the positive carriers (i.e. holes) accumulated at the gate terminal would

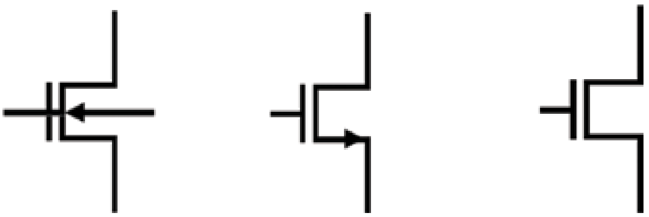


Figure 3. Different symbols for NMOS transistors.

be sufficiently strong to repel holes and attract electrons to form a channel at the substrate-oxide interface. The channel connects both source and drain terminals, forming a closed circuit for electrons to flow. Like the case of the NMOS transistor, a voltage applied at the gate to source terminal is required to form a channel in the PMOS transistor. However, unlike the NMOS transistor, voltage V_{GS} of the PMOS transistor is negative. This allows negative carriers (i.e. electrons) to be accumulated at the gate terminal. When the magnitude of V_{GS} exceeds its threshold, electrons at the oxide-substrate interface would be repelled and holes would be attracted to the interface. A conducting channel made of positive carriers is thus formed between the source and drain terminals.

3.2. CMOS devices

Although NMOS and PMOS transistors have been used independently in electronic circuits, they have their own limitations. A PMOS transistor is unable to produce an exact zero output voltage when a logic 0 is required, whereas an NMOS transistor fails to give a full V_{DD} voltage at the output when a logic 1 is required. Failure to give a full swing from 0 to V_{DD} has resulted in power loss in circuitries. In order to solve this problem, both NMOS and PMOS transistors are integrated together in IC designs. By connecting the source of the PMOS transistor to the V_{DD} input voltage and that of the NMOS transistor to the ground, output V_{DS} can be completely pulled up to V_{DD} and pulled down to ground when a logic 1 and 0 is to be generated, respectively. Because of this reason, the part of the circuit that is made from PMOS transistors is known as the pull-up network, whereas the part that comprises NMOS transistors is known as the pull-down network. Since these two transistors complement each other, a circuit which is designed from a combination of both is therefore known as a Complementary MOS circuit or CMOS circuit, in short. Each time a CMOS circuit operates, only either of the pull-up or

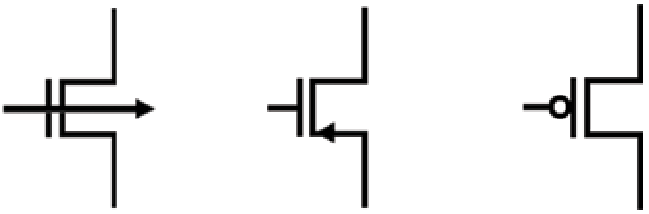


Figure 4. Different symbols for PMOS transistors.

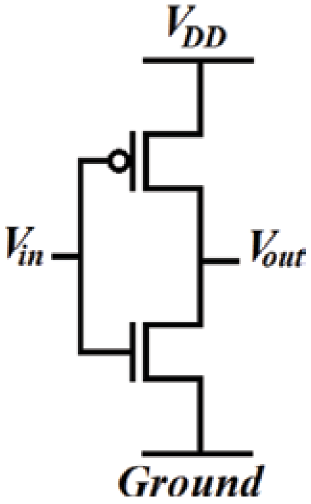


Figure 5. Schematic of a CMOS inverter.

pull-down network conducts. Take for instance the operation of a CMOS inverter, such as that shown in **Figure 5**. When a logic 1 is to be generated at the output, the PMOS transistor acts like closed switch and the NMOS transistor acts like an open switch. Similarly, when a logic 0 is to be generated, the PMOS and NMOS transistors act like an open and a closed switch, respectively. This allows almost zero power loss during steady states of the circuit.

Author details

Kim Ho Yeap* and Humaira Nisar

*Address all correspondence to: yeapkimho@gmail.com

Tunku Abdul Rahman University, Jalan Universiti, Kampar, Perak, Malaysia

References

[1] Ahmad I, Ho YK, Majlis BY. Fabrication and characterization of a 0.14 μm CMOS device using ATHENA and ATLAS simulators. International Scientific Journal of Semiconductor, Physics, Quantum Electronics, and Optoelectronics. 2006;**9**:40-44. DOI: https://doi.org/10.15407/spqeo

[2] Yeap KH, Nisar H. Very Large Scale Integration. InTech: Croatia; 2018

[3] Lukasiak L, Jakubowski A. History of semiconductors. Journal of Telecommunications and Information Technology. 2010;**1**:3-9. DOI: 10.1088/0031-9120/40/5/002

Section 2

Advancement in the Fabrication Process

Chapter 2

Advanced Transistor Process Technology from 22- to 14-nm Node

Huaxiang Yin and Jiaxin Yao

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.78655

Abstract

Transistor performance meets great technical challenges as the critical dimension (CD) shrinking beyond 32/28-nm nodes. A series of innovated process technologies such as high-k/metal gate, strain engineering, and 3D FinFET to overcome these challenges are reviewed in this chapter. The principle, developing route, and main prosperities of these technologies are systematically described with theoretical analysis and experimental results. Especially, the material choice, film stack design, and process flow integration approach with high-k/metal gate for sub-22-nm node is introduced; the film growth technique, process optimization, and flow integration method with advanced strain engineering are investigated; the architecture design, critical process definition, and integration scheme matching with traditional planar 2D transistor for 14-nm 3D FinFET are summarized.

Keywords: CMOS, high-k/metal gate, strain, FinFET, process

1. Introduction

The metal-oxide-semiconductor field effect transistors (MOSFETs) are core switch devices of current large-scale complementary-metal-oxide-semiconductor integrated circuits (CMOS ICs). The performance of the transistor has a critical effect on the performance of IC. As the continuous scaling of the transistor CD for a higher IC performance and integration density, the fabrication process technologies and methods of the transistor are fast changing and becoming relatively complicated. To suppress the short-channel effect as well as the performance degradation of devices, three main new technologies, including strain engineering, high-k/metal gate (HKMG), and FinFET of MOSFETs, are implemented into state-of-art IC manufacture technology. The three technologies are quite important and firstly applied in the

CMOS IC manufacturing process by Intel Corporation in the years 2003, 2007, and 2011 at 90-, 45-, and 22-nm node, respectively, which is developed into the industrial standards and widely adopted by other IC manufacturing corporations including TSMC, and Samsung. While the process node of CMOS IC scaling from 22- into 14-nm node, advanced technologies such as film growth, structure design, process optimization, and integration flow of them become more complicated, which often need elaborated process development with diversified knowledge and techniques from different fields.

2. Strain engineering

The effective carrier mobility in the channel of the transistor is crucial to the device's performance. With the gate length aggressively shrinking down, the electric field magnitude in the channel is strengthened with channel-doping concentration rising, resulting in obvious degradation in effective carrier mobility due to ionized impurity scattering. Mobility both for electron and for hole can be enhanced by changing the silicon atom arrangement of crystal lattice in the channel through the external stress. It is investigated that the tensile and compressive strain for silicon can improve the electron and hole mobility, respectively.

Mobility is closely dependent on the mean free time and the effective mass of the carrier. As we consider the simplified band structure of silicon, there are six equivalent minima at k = (x, 0, 0), (−x, 0, 0), (0, x, 0), (0, −x, 0), (0, 0, x), (0, 0, −x) with x = 5 nm^{-1} for the conduction band in Ref. [1]. There is one maximum containing two sub-bands at k = 0 for the valence band. These two sub-bands are referred to as the light and heavy hole bands with a light hole effective mass and a heavy hole effective mass. Therefore, the effective mass of these anisotropic minima is characterized by a longitudinal mass along the corresponding equivalent (1, 0, 0) direction and two transverse masses in the plane perpendicular to the longitudinal direction. For the electron in conduction band, the external stress can cause tensile or compressive strain in the silicon lattice. The longitudinal band valley will change. Thus, the corresponding longitudinal mass is changed leading to the mean free time increasing or decreasing for the carriers. For hole in the valence band, the strain effect on the light hole band and heavy hole band is familiar with electron.

For the device, several of strain types for the mobility enhancement are listed in **Table 1**. Different axial tensile and compressive strain can introduce different mobility enhancement or

Direction	NMOS	PMOS
Channel length	Tensile	Compressive
Channel width	Tensile	Tensile
Perpendicular to channel plane	Compressive	Tensile

Table 1. Strain type for carrier mobility enhancement.

degradation for electron and hole. Therefore, strain technique is very practical and important for device's performance promotion. In the point of IC process, this technique is called as strain engineering, which is divided into global strain stress and local strain stress. Global strain stress is less employed in the manufacturing process due to the application regime. Local strain stress can be targeted to enhance carrier's mobility in the specified region and widely used in the modern IC process flow. Local strain stress can be induced and achieved by IC processes, such as selective epitaxy growth (SEG) of silicon-germanium (SiGe) source/drain, dielectric etch-stop layer (ESL), metal gate, and contact. For PMOS, the most important strain engineering technology is to use selective source/drain epitaxy of SiGe with large lattice constant in order to provide channel with axial compressive stress for hole mobility enhancement as shown in **Figure 1**.

The embedded SiGe in source/drain (S/D) region has been widely used to induce uniaxial strain in the channel, especially for the sigma SD epitaxy in Ref. [2]. The strain engineering in 22-nm planar transistor is becoming more complicated. The film growth technique often requires relatively low temperature and the decreasing of pattern dependency. In 22-nm node, the SiGe layers need to be grown at 650°C in a reduced-pressure chemical vapor deposition (RPCVD) reactor and with a series of complicated steps. First, in situ cleaning is performed by annealing in the range of 740–825°C for 3–7 min. Then, dichlorosilane (SiH_2Cl_2), 10% germane (GeH_4) in $H_{2,}$ and 1% diborane (B_2H_6) in H_2 are used as Si, Ge, and B precursors, respectively. Moreover, HCl is utilized as Si etchant to obtain selectivity during the epitaxy. The SiGe growth rate can be denoted by an empirical model (Eq. (1)) in Refs. [3, 4], which considers the contribution of a variety of molecule fluxes coming from different directions toward Si planes during epitaxy in Ref. [5]:

$$\begin{aligned} R_{total} = R_{Si}^{V} + R_{Si}^{LG} + R_{Si}^{SO} + R_{Si}^{CO} + R_{Si}^{IP} + R_{Ge}^{V} + R_{Ge}^{LG} + R_{Ge}^{SO} + R_{Ge}^{CO} \\ + R_{Ge}^{IP} - R_{HCl}^{V} - R_{HCl}^{LG} - R_{HCl}^{SO} - R_{HCl}^{CO} + R_{HCl}^{IP}, \end{aligned} \tag{1}$$

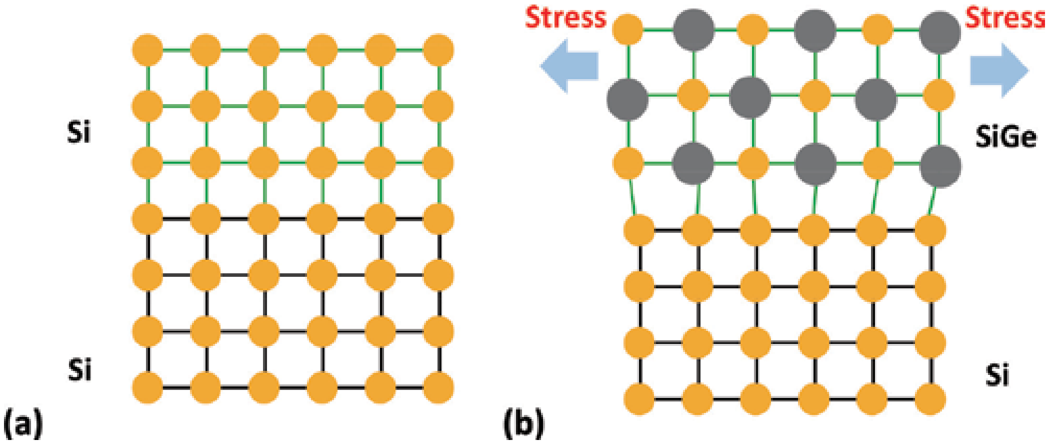


Figure 1. (a) Homoepitaxy and (b) heteroepitaxy: SiGe with high stress growth on the Si substrate.

where R^{V} and R^{LG} for Si, Ge, and HCl are the contribution of gas molecules in vertical and lateral directions; R^{SO} and R^{CO} are the mobile reactant molecules on the oxide surface surrounding or within a chip; R^{IP} is the contribution from atoms which diffuse from the edges toward the Si plane. After considering the reaction species and atom activation energy, gas partial pressure, and growth temperature, the final expression for the total growth rate is given by (Eq. (2))

$$\begin{aligned} R_{Total} = {} & \beta\frac{\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{\mathrm{N}_0}\times\frac{P_{SiH_2Cl_2}}{(2\pi m_{SiH_2Cl_2}k_bT)^{\frac{1}{2}}}\left(\frac{E_{SiH_2Cl_2\ on\ Si}}{k_bT}+1\right)\exp\left(-\frac{E_{SiH_2Cl_2\ on\ Si}}{k_bT}\right) \\ & +\chi\frac{(1+m_r)\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{N_0}\times\frac{P_{GeH_4}}{(2\pi m_{GeH_4}k_bT)^{\frac{1}{2}}}\left(\frac{E_{GeH_4\ on\ Si}}{k_bT}+1\right)\exp\left(-\frac{E_{GeH_4 on\ Si}}{k_bT}\right) \\ & +\chi\frac{(1+m_r)\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{N_0}\times\frac{\left(BP_{GeH_4}\ln\left(\frac{1}{c}\right)\right)}{(2\pi m_{GeH_4}k_bT)^{\frac{1}{2}}}\left(\frac{E_{GeH_4\ on\ Si}+0.1eV}{k_bT}+1\right) \\ & \times\exp\left(-\frac{E_{GeH_4 on\ Si}+0.1eV}{k_bT}\right)-\frac{\gamma}{N_0}\frac{P_{HCl}^{0.596}}{(2\pi m_{HCl}k_bT)^{\frac{1}{2}}}\left(\frac{E_{Etching}}{k_bT}+1\right)\exp\left(-\frac{E_{Etching}}{k_bT}\right) \end{aligned} \tag{2}$$

where θ_{Cl} and θ_H parameters stand for the occupied dangling bonds by hydrogen and chlorine atoms on Si; N_0 is the number of atoms per unit volume for Si; E and P are activation energy and partial pressure for different reactant molecules, respectively. The variable c is the exposed Si coverage of Si chip where B is a unit-less constant which is dependent on the architecture of the mask. The equation constants, β, χ, and γ, are tooling factors which depend on the temperature distribution and gas kinetic over the susceptor in the CVD reactor. Therefore, a series of process parameters are affecting the growth of SEG SiGe in 22-nm PMOS transistor, resulting in different growth rates, Ge content, film quality as well as the compressive stress to the channel. The stress distribution has also a strong relationship to the growth area, pattern intensity, and locations around the wafer.

For integrating SEG SiGe into 22-nm PMOSFET, a sacrificial epitaxy-block Si3N4 layer is deposited on whole wafer after the formation of dummy polysilicon gate and spacers. In the next step, the block layer is selectively opened and low-temperature epitaxy SiGe with high stress is performed at the source/drain region of PMOS.

The amount of strain induced by SiGe is dependent on the initial recess shape, interfacial quality of SiGe/Si, and defect density in the epilayers. Sigma-shaped recesses with (100) and {111} planes are very suitable shape for embedded SiGe in source/drain regions with the highest stress. Moreover, in such transistors, shorter distance between sigma-shaped recesses and channel region can generate a higher stress to the channel region. By applying dry etch together with wet etch in Si substrate, the sigma-shaped recesses turn more large and induce a stronger stress of embedded SiGe with a closer distance to the channel in Ref. [5].

For NMOS, PMD (pre-metal dielectric) layer is deposited as ESL, which can offer axial tensile stress to the channel for electron mobility enhancement [6]. In the modern IC manufacturing integrated process, more and more strain process is employed for the devices, which is of great significance to suppress the device's performance degradation. However, the film thickness of ESL is limited due to the scaling of gate pitch between transistors. New techniques, such as metal gate and contact electrode stress of NMOS, are necessary. The TiN metal gate and the W plug often bring effective tensile stress into the channel of NMOSFET, resulting in the enhancement of motilities for electrons.

3. High-k/metal gate

High-k/metal gate (HKMG) is a very important technique for modern CMOS IC manufacturing process. While the transistor CD scaling down, conventional oxide dielectric/polysilicon gate was formally replaced by high-k dielectric/metal gate, in order to suppress the unbearable leakage in the ultra-thin oxide dielectric film in Ref. [7]. HKMG technique has found a new effective path for equivalent oxide thickness (EOT) scaling tendency, which is of deep significance to continuous scaling of MOS transistors. However, HKMG brings about a series of challenges, including new high-k dielectric and metal gate materials, threshold voltage modulation, and process integration scheme. To some extent, the scaling of MOS transistor relies on the scaling of EOT of gate dielectric. When the conventional oxide film thickness shrinks to about 11–12 A, the transistor shrinking cannot be continued due to the extremely large leakage current from the gate to the substrate by the electron direct tunneling through the ultra-thin oxide film. EOT is defined as (Eq. (3)).

$$\mathrm{EOT} = T_{HK} \cdot \frac{\varepsilon_{OX}}{\varepsilon_{HK}} \tag{3}$$

where T_{HK} is the physical thickness of high-k dielectric, ε_{OX} is the SiO_2 dielectric constant, and ε_{HK} is the high-K dielectric constant. When it comes to high-k dielectric materials, the physical thickness of the gate dielectric is increased because of the high value of dielectric constant parameter. Hence, the gate leakage current induced by direct tunneling is reduced dramatically to continue the scaling of EOT.

Many high-k materials have been investigated for CMOS devices including metal oxide (HfO_2, ZrO_2, Al_2O_3, etc.) as shown in **Table 2**. Among these metal-oxide materials, HfO_2 has the advantages of the moderate relative permittivity value, the basically symmetrical energy band offset to silicon conduction band and valence band, and the uniformly amorphous structure. Therefore, HfO_2 material is applied in the IC manufacturing production.

In the early 1990, it is reported that the integration of polysilicon gate with HfO_2 dielectric results in serious Fermi Level Pinning (FLP) phenomenon, where the Fermi level of polysilicon gate is fixed at the poly/HfO_2 interfacial energy level. Although some theories including oxygen vacancy model, and dipole formation, are put forward to explain the pinning effect, the process cannot successfully release the effect of FLP, which causes huge difficulties on the device's threshold voltage modulation. Therefore, different metal gates with the high-k material are corresponded to different threshold voltage modulation regimes for PMOS and NMOS.

Material	Dielectric constant	Material	Dielectric constant
Al_2O_3	8–11.5	$NdAlO_3$	22.5
(Ba, Sr)TiO_3	200–300	$PrAlO_3$	25
$BeAl_2O_4$	8.3–9.43	Si_3N_4	7
CeO_2	16.6–26	$SmAlO_3$	19
HfO_2	26–30	$SrTiO_3$	150–250
Hf silicate	11	Ta_2O_5	25–45
La_2O_3	20.8	TiO_2	86–95
$LaAlO_3$	23.8–27	Y_2O_3	8–11.6
$LaScO_3$	30	ZrO_2	22.2–28

Table 2. High-k dielectric constant.

3.1. HKMG film stack

The introduction of high-k/metal gate provides great potential of transistor's scaling down under 45-nm node. Metal gate can reduce oxide thickness by eliminating polysilicon gate depletion effect. Metal gate has a low gate resistance and can suppress boron penetration to the substrate in Refs. [8–10].

In 22-nm node, the main challenge and research hotpot for HKMG stack lie in the effective work function (EWF) modulation of metal gate. The EWF can be defined as the value between Fermi level of metal gate and vacuum level in the metal-oxide-semiconductor system. As shown in **Figure 2**, EWF is defined as (Eq. (4))

$$\mathrm{EWF} = E_0 - E_{FM}, \tag{4}$$

where E_0 is the vacuum energy level, and E_{FM} is the Fermi level of metal gate. When E_{FM} is close to the conduction band edge E_C or valence band edge E_V of Si substrate, EWF will get the minima or the maximum value for the MOS device. Generally, the mid-gap EWF is around 4.6 eV, and the band edge EWF is less than 4.4 for conduct band in NMOS or is high than 4.8 eV for valence band in PMOS.

Fermi level of metal gate can be shifted both upwards and downwards. The Fermi level (E_F) of metal gate is set to be in the position of mid-gap in the substrate. When Fermi level shifts to the conduction band of Si substrate, the effective work function of metal gate decreases. On the other hand, when Fermi level shifts to the valence band of Si substrate, the effective work function of metal gate increases. The EWF movement behaviors can directly drive the threshold voltage (Vt) modulation for the MOS devices.

The most effective method to manipulate EWF is the selection of metal gate. For PMOS, a large EWF is preferred to achieve high Vt for low-power (LP) IC performance. Hence, Fermi level of PMOS metal gate is ideally near to the valence band maximum of silicon substrate, where the position in the valence band minima of silicon is the best choice for Fermi level of metal gate.

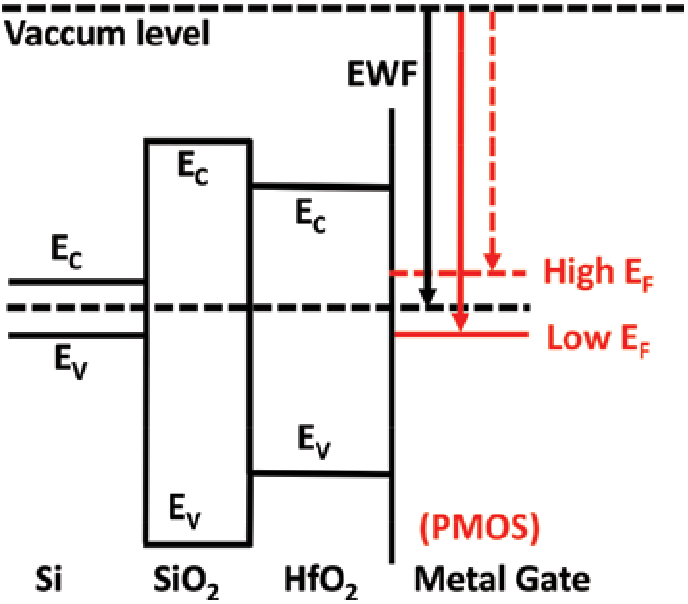


Figure 2. Illustration of band edge EWF of metal gate for PMOS.

For NMOS, a small EWF is preferred to achieve high Vt for LP IC performance, where the Fermi level of NMOS metal gate is ideally near to or at the conduction band minima of silicon substrate. Therefore, the demand of the large EWF for PMOS and small EWF for NMOS is selecting TiN with a high work function and TiAl with a low work function metals.

For the multi-Vt modulation of PMOS and NMOS, the most general method is to tune the gate-stack thickness control in Refs. [11, 12], in order to realize the regular, low, and high Vt levels. As shown in **Figure 3**, the metal gate stack can be divided into three layers: the first is the bottom-capping layer for the high-k dielectric, the second is exactly the work function layer, and the last is the top-capping layer for the contacted metal. Moreover, the etch-stop layer should be considered for the dual work function metal integration of PMOS and NMOS. Although the gate stack contains three parts, the effective work function of the entire gate electrode is dominated by the work function layer metal, where EWF sensitivity is strictly limited by the bottom-capping layer thickness, and the top-capping layer acts as the barrier layer for the contacted metal (tungsten). Therefore, the thickness control of metal gate-stack design is exactly of precision and significance.

3.2. Gate-first and gate-last integration scheme

The novel gate-stack structure of HKMG has been implemented for MOSFETs to promise conventional scaling of the high-performance CMOS process down to the 45/32-nm node. Two completely different integration schemes were proposed [13]. With the HKMG in the IC process flow, a big question arises that the module of HKMG structure formation is ahead of or after the module of source/drain process. Gate-first process integration scheme is familiar with

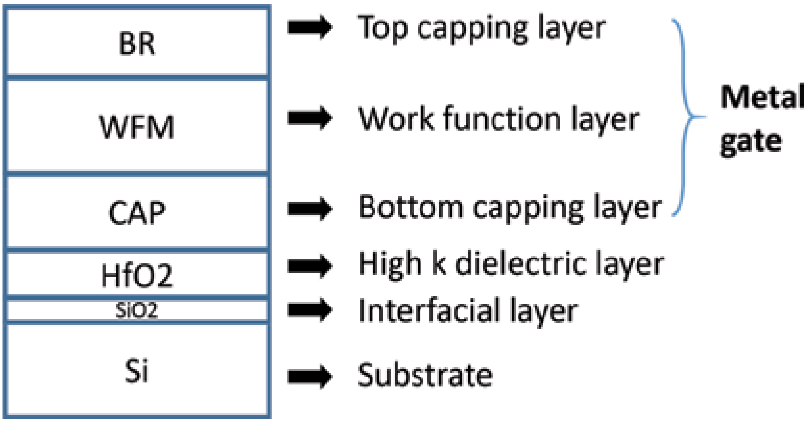


Figure 3. HKMG gate stack.

poly-Si/SiO2 process flow. HKMG module is firstly deposited after the active-region formation module, and then source/drain module formation module is following until the end. However, with the source/drain formation later than HKMG formation module, the high annealing temperature for the S/D doping profile has a serious impact on HKMG characteristics and its reliability.

To overcome shortcoming caused by gate-first integration scheme, gate-last integration scheme is put forward. In the gate-last process, conventional poly-Si/SiO2 is still formed on the wafer substrate firstly. After poly-Si/SiO2 formation module, it is followed by the S/D impurity doping and its activation with annealing process at high temperature ambient. Then, PMD layer is deposited on the poly-Si dummy gate, where PMD is also called an inter-layer-dielectric zero layer (ILD0). With poly-Si-open planarization (POP) chemical mechanical polishing (CMP), poly-Si gate is exposed for the following removal of poly-Si/SiO_2 process. Finally, HKMG is deposited in the position where poly-Si dummy gate previously existed, which is called gate-last process due to HKMG module later than the middle end of line (MEOL) process. The implement of gate-last integration scheme avoids the damage to devices by the high annealing temperature of S/D process. Therefore, gate-last integration scheme has obvious performance advantages for HKMG devices and becomes popular technique applied beyond 28 nodes.

In gate-last technique, it is divided into two integration schemes: high-k first/metal-gate last and high-k last/metal-gate last. In the first approach, the high-k layer is deposited together with the formation of dummy gate and before the annealing of source/drain, where only metal gate stack is formed with gate-last scheme. In the second approach, both high-k and metal gate are formed after the annealing of source/drain, which is also called all gate-last integration scheme. It has better film quality and process adjustment window than the former and is widely adopted for CMOS IC fabrication process in 22 nm and beyond node. In this integration scheme, multilayer HKMG stacks are IL/HfO_2/TiN/TaN/TiN/W and IL/HfO_2/TiN/TiAlC/TiN/W for PMOS and NMOS, respectively. IL layer is an interfacial layer between HK and substrate and is normally SiO_2 forming by chemical oxidation method. All HKMG depositions are finished by atomic layer deposition (ALD) approach with a high conformality and a precise thickness control ability.

4. FinFET technology

While process node scaling from 22 to 14 nm, the basic architecture of the transistor is changing from 2D planar device to 3D volume inversion device for a better control of SCE in channel with less leakage. The device design as well as the process techniques turns more complicated and needs a more elaborated technologies.

4.1. FinFET transistors

With feature size of CMOS IC shrinking to 20-nm node and beyond, the structure of the conventional planar MOSFET consisting of single-gate electrode to control channel potential distribution and the flow of current in the channel region is faced with the undesirable parasitic effects called short-channel effect (SCE) and drain-induced barrier lowering (DIBL) effect. Via voltage-doping transformation (VDT) model [14], the device's structure and material parameters can be translated into electrical parameters with electrostatic integrity (EI) (Eq. (5)). SCE and DIBL can be derived as (Eqs. (6) and (7))

$$\mathrm{EI} = \left[1 + \frac{{x_j}^2}{{L_{ch}}^2}\right] \frac{t_{ox}}{L_{ch}} \frac{t_{dep}}{L_{ch}} \tag{5}$$

$$\mathrm{SCE} = 0.64 \frac{\varepsilon_{Si}}{\varepsilon_{ox}} EI\ V_{bi}, \tag{6}$$

$$\mathrm{DIBL} = 0.8 \frac{\varepsilon_{Si}}{\varepsilon_{ox}} EI\ V_{ds} \tag{7}$$

where L_{ch} is the effective channel length, V_{bi} is the source or drain built-in potential, t_{ox} is the gate oxide thickness, x_j is the source/drain junction depth, and t_{dep} is the penetration depth of the gate electric field in the channel region. The parameter EI is denoted as electrostatic integrity factor.

The threshold voltage of MOSFET can be denoted as (Eq. (8))

$$V_T = V_{T\_long} - SCE - DIBL \tag{8}$$

According to the above expression, SCE can be minimized by reducing the junction depth, gate oxide thickness, and depletion depth via increasing the doping concentration in the channel region. However, the limits on the reducing junction depth and gate oxide thickness have become very toughly serious in the practical device. Hence, SCE and DIBL values of the planar MOSFET are not controlled well in the ultra-short-channel length.

The most efficient and direct way to suppress SCE is to strengthen the gate electric field control capability by double-gate (DG) or multi-gate (MG) structure. DG or MG structures on thin Si channel improve the electrostatic integrity of MOSFET (Eq. (9)) with the transistor working in a volume inversion mode due to a reduced device structure parameter, which decreases the SCE and DIBL effects on the device electric parameters, such as threshold voltage, subthreshold slope (SS), and DIBL voltage. In the equation, since the thickness of Si is much smaller than that of depletion region in planar transistor, EI is obviously improved. The whole new structures of MOSFET extend the shrinking boundary of the ultra-short gate length

$$\mathrm{EI} = \frac{1}{2}\left[1 + \frac{t_{Si}^2/4}{{L_{ch}}^2}\right]\frac{t_{ox}}{L_{ch}}\frac{t_{Si}/2}{L_{ch}} \tag{9}$$

FinFET is a typical double-gate or multi-gate device with a three-dimensional channel structure, as shown in **Figure 4**. The FinFET is made of a tall and narrow silicon island. The 3D channel is standing above the silicon substrate, where the ultra-thin silicon body is familiar with the fin of the fish. The fin channel under the gate can be fully depleted by electrostatic potential, providing a strong ability of controlling the carriers' behaviors in the channel. FinFET can really expand the limit of the shrinking size and is widely adopted for the 16/14-nm technology node and beyond. FinFET can effectively suppress the leakage of the sub-surface channel, which can obviously reduce the off-state current for the device's current-voltage transfer characteristic. In the meantime, the fully depleted channel can obtain benefit of carriers' mobility with less scattering. For the 3D fin structure, the transistor's width can be doubled compared to the planar one in the projected plane, which can improve the driving current at on-state in the saturation regime. With the same drive current, the supply voltage of FinFET can be significantly reduced regardless of the planar transistor's power limit, where the suppression of power consumption in modern integrated circuits emphasizes energy efficiency ratio.

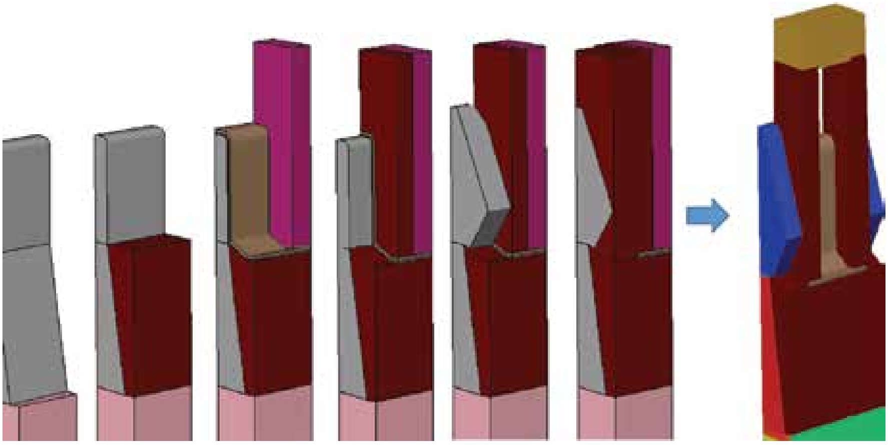


Figure 4. FinFET from fin to whole device.

4.2. FinFET integration process

Since 22-nm technology node, FinFET has been utilized for several process nodes [15–17]. It is firstly introduced by Intel in 22-nm node and widely adopted by different companies in 16- or 14-nm process node. The process integration scheme of FinFET is compatible with that of the planar transistor. In a general way, the critical fabrication steps of FinFET transistor include silicon fin formation on the substrate by the spacer-transfer lithography (STL), shallow trench isolation (STI) formation and recess, 3D dummy gate formation and planarization, 3D spacer formation, source/drain with 3D selective SEG, 3D HKMG formation, and back-end-of-line (BEOL) metallization and contact techniques. It added a little extra process steps than those of planar transistor fabrication. It is very meaningful to understand the integration process of FinFET. In future, the next-generation devices, such as gate-all-around nanowire transistor or nanosheet FET, are still dependent on current FinFET integration flow [18, 19].

4.2.1. Spacer-transfer lithography for bulk fin formation

Oxide by plasma-enhanced CVD (PECVD), poly-Si by low-pressure CVD (LPCVD), and SiNx by PECVD are sequentially deposited in the substrate for the formation of etch-hard-mask (EHM). After etching EHM with pattern, another SiNx is deposited as the spacer of the core layer of oxide/poly-Si/SiNx structure. After spacer and Si dry etch, the 3D Si fin is formed and the Si fin width depends on the SiN spacer thickness, as shown in **Figure 5**. The fin width may be beyond the lithography resolution limit and often smaller than 10 nm.

4.2.2. STI formation and recess

For adjacent fins isolation, high-aspect-ratio-process (HARP) oxide deposition is widely used with a good step coverage on 3D fins. The oxide for HARP STI is deposited by sub-atmospheric CVD (SACVD) with the reaction by tetraethoxysilane (TEOS) precursor and O3. After the isolation oxide annealing, chemical mechanical polishing is utilized for the planarization of deposited dielectric on 3D fins. In following steps, the oxide is precisely etched back and making the fin final formation with shallow trench isolation structures as shown in **Figure 6**.

4.2.3. 3D dummy gate formation

On 3D fins with STI, thin oxide is firstly formed on the surface. Then, amorphous-Si (α-Si) is deposited as dummy gate on the fin. However, the dummy gate etch is the most challenging, for which the top dummy gate needs to be protected during the etching and sidewall and the foot of the dummy gate needs strong etching capability to prevent the residue of Si and no process damage on the exposed fin tip (**Figure 7**).

4.2.4. Source/drain 3D SEG

On 3D fin, it often needs SEG on source/drain regions for less contact resistance. Source/drain selective epitaxy growth normally employs SiH2Cl2, GeH4, and HCl gases. Especially, for PMOS source/drain, B2H6 is mixed into the carrier gas of the reaction. The selectivity of SiGe

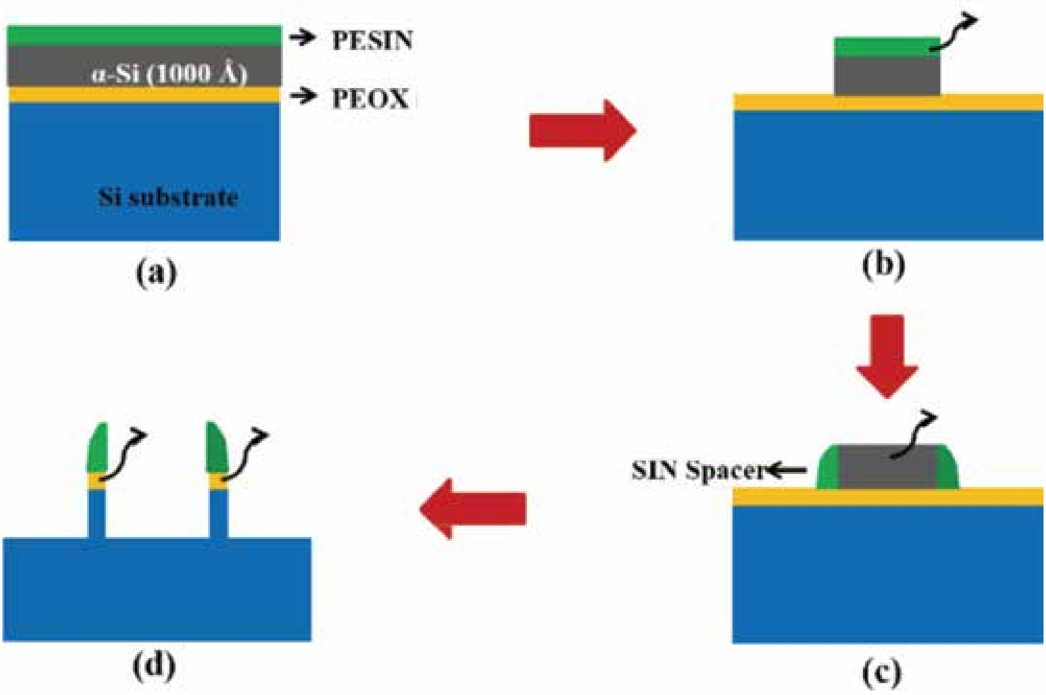


Figure 5. STL for bulk fin formation (a) Hard mask deposition; (b) Hard mark etch; (c) SiN spacer deposition and etch; (d) Fin structure etch.

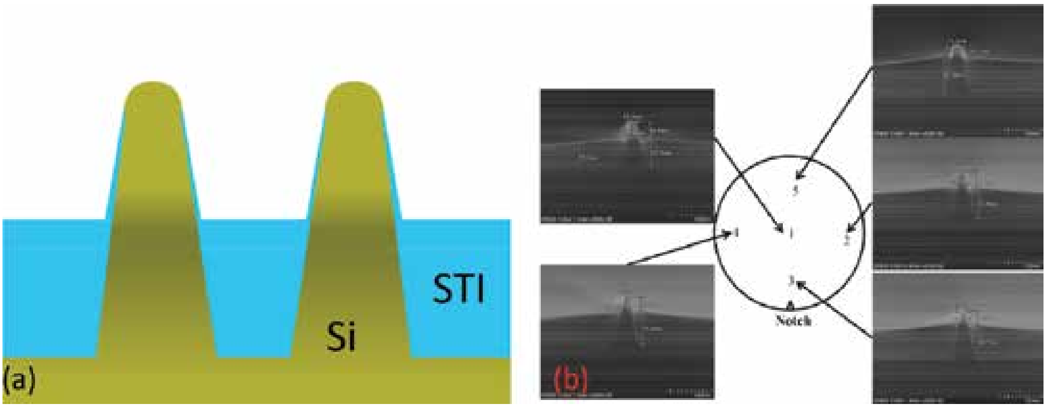


Figure 6. STI formation and recess on 3D fins (a) Fin and STI structure after recessng; (b) SEM images for Fin and STI after recessing.

epitaxy is mainly due to the function of HCl gas, where the etch rate of polycrystalline SiGe is higher than that of single crystalline of SiGe by HCl. In the whole process, the dilution protective gas contains N_2 or H_2 all the time. Due to the slowest growth rate on Si (111) lattice plane, as shown in **Figure 8**, the final formed SiGe shape on 3D fin is more like a diamond. The film stress not only depends on the process conditions but also is strongly affected by the surface quality of fins.

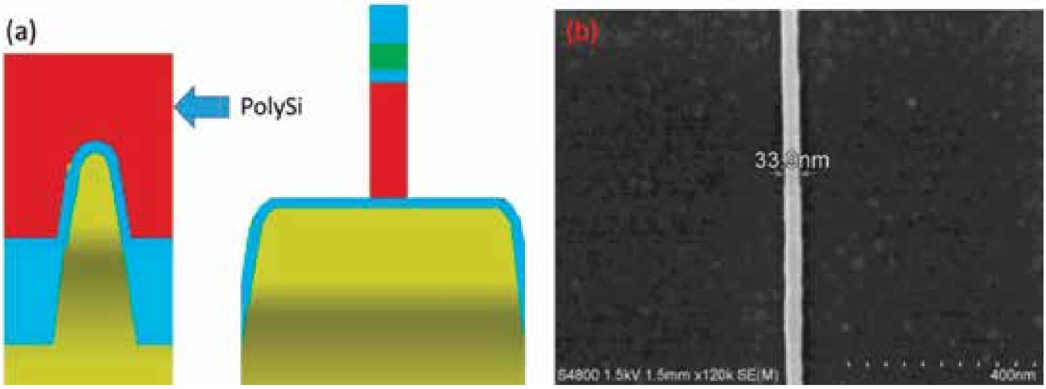


Figure 7. 3D dummy gate formation (a) PolySi deposition, planarization and dummy gate etch; (b) SEM image after dummy gate formation.

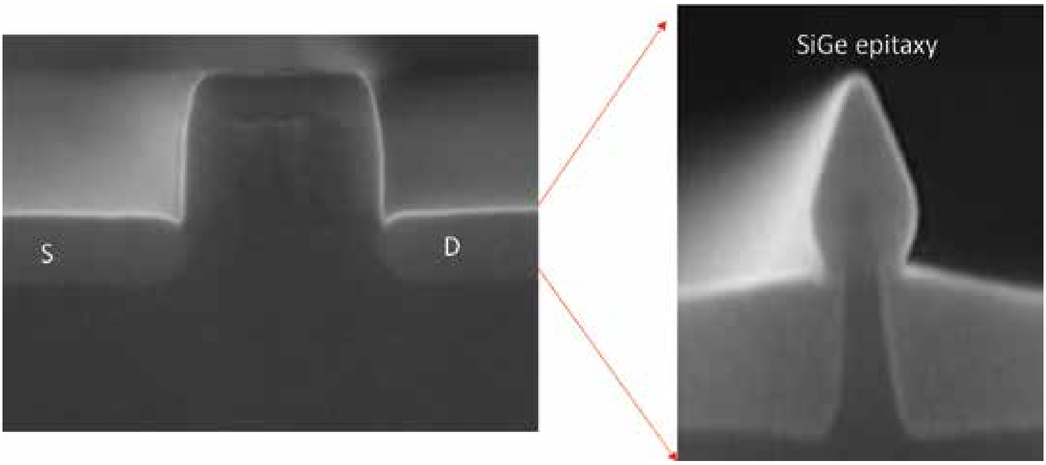


Figure 8. SEG SiGe on 3D fin.

5. Conclusions

Advanced transistor technologies were extensively implemented into the CMOS IC manufacture with the process node scaling from 22 to 14 nm. They require new materials and novel structures as well as complicated process techniques and different device integration flow. This chapter presented a summary on the three important techniques, strain engineering, high-k/ metal gate, and FinFET. Both the process theory related to the suppress on SCE for device's shrinking and the detailed illustration on material choice, film growth method, architecture design, critical process definition, and integration are presented in a comprehensive and systematic manner. The process condition optimizations for suppressing stress release are key technologies of strain engineering. The high-k/metal gate needs multilayer structure for modulating Vt in a different manner for PMOS and NMOS, respectively. The integration scheme is also changed from gate-first to all-last integration. FinFET requires a sophisticated device integration structure and a flow design with less extra process cost. It also has some new fabrication techniques, such as ultra-thin fin formation with STL and improved process methods, including HKMG and SiGe SEG in 3D approach.

Acknowledgements

The reported work was supported by "16/14nm Basic Technology Research" of national 02 IC R&D projects in China (No. 2013ZX02303). The authors would like to thank all the colleagues in Integrated Circuit Advanced Process Center, IMECAS, for their kind and great support.

Conflict of interest

The authors declare that they have no competing interests.

Author details

Huaxiang Yin* and Jiaxin Yao

*Address all correspondence to: yinhuaxiang@ime.ac.cn

Key Laboratory of Microelectronics Devices and Integrated Technology, Institute of
Microelectronics of Chinese Academy of Science, Beijing, China

References

[1] Vasileska D, Goodnick SM. Computational electronics. Synthesis Lectures on Computational Electromagnetics. 1st ed. San Rafael, CA, USA: Morgan & Claypool Publishers; 2006. 20 p. 1-216. DOI: 10.2200/s00026ed1v01y200605cem006

[2] Qin C, Yin H, Wang G. Study of sigma-shaped source/drain recesses for embedded-SiGe pMOSFETs. Microelectronic Engineering. 2017;**181**:22-28. DOI: 10.1016/j.mee.2017.07.001

[3] Wang G, Abedin A, Moeen M. Integration of highly-strained SiGe materials in 14nm and beyond nodes FinFET technology. Solid-State Electronics. 2015;**103**:222-228. DOI: 10.1016/j.sse.2014.07.008

[4] Radamson HH, Kolahdouz M. Selective epitaxy growth of Si1−xGex layers for MOSFETs and FinFETs. Journal of Materials Science: Materials in Electronics. 2015;**26**(7):4584-4603. DOI: 10.1007/s10854-015-3123-z

[5] Qin C, Wang G, Kolahdouz M. Impact of pattern dependency of SiGe layers grown selectively in source/drain on the performance of 14nm node FinFETs. Solid-State Electronics. 2016;**124**:10-15. DOI: 10.1016/j.sse.2016.07.024

[6] Yin H, Meng L, Yang T. CMP-less planarization technology with SOG/LTO etchback for low-cost high-k/metal gate-last integration. ECS Journal of Solid State Science and Technology. 2013;**2**(6):268-270. DOI: 10.1149/2.011306jss

[7] Auth C, Cappellani A, Chun J-S. 45nm high-k + metal gate strain-enhanced transistors. In: Symposium on VLSI Technology (VLSI '08); 17-19 June 2008; Honolulu. New York: IEEE; 2008. pp. 128-129

[8] Robertson J. High dielectric constant gate oxides for metal oxide Si transistors. Reports on Progress in Physics. 2006;**69**(2):327-396. DOI: 10.1088/0034-4885/69/2/r02

[9] Wilk GD, Wallace RM, Anthony JM. High-κ gate dielectrics: Current status and materials properties considerations. Journal of Applied Physics. 2001;**89**(10):5243-5275. DOI: 10.1063/1.1361065

[10] Choi J, Mao Y, Chang J. Development of hafnium based high-k materials—A review. Materials Science and Engineering: R: Reports. 2011;**72**(6):97-136. DOI: 10.1016/j.mser.2010.12.001

[11] Ma X, Yang H, Wang W. An effective work-function tuning method of nMOSCAP with high-k/metal gate by TiN/TaN double-layer stack thickness. Journal of Semiconductors. 2014;**35**(9):096001-096004

[12] Xu J, Wang A, He J. 14nm metal gate film stack development and challenges. In: China Semiconductor Technology International Conference (CSTIC '2017); 12-13 March 2017; Shanghai. New York: IEEE; 2017. p. 1-3

[13] Veloso A, Ragnarsson L-A, Cho M-J. Gate-last vs. gate-first technology for aggressively scaled EOT logic/RF CMOS. In: Symposium on VLSI Technology (VLSI '11); 14-16 June 2011; Honolulu. New York: IEEE; 2011. pp. 34-35

[14] Skotnicki T, Merckel G, Pedron T. The voltage-doping transformation: A new approach to the modeling of MOSFET short-channel effects. IEEE Electron Device Letters. 1988;**9**(3): 109-112. DOI: 10.1109/55.2058

[15] Auth C, Allen C, Blatter A. A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors. In: Symposium on VLSI Technology (VLSI '12); 12–14 June 2012; Honolulu. New York: IEEE; 2012. pp. 131-132

[16] Natarajan S, Agostinelli M, Bost M. A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588um2 SRAM cell size. In: IEEE International Electron Devices Meeting (IEDM '14); 15-17 Dec. 2014; San Francisco. New York: IEEE; 2014. pp. 3.7.1-3.7.3

[17] Auth C, Aliyarukunju A, Asoro M. A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects. In: IEEE International Electron Devices Meeting (IEDM '17); 2-6 Dec. 2017; San Francisco. New York: IEEE; 2017. pp. 29.1.1-29.1.4

[18] Loubet N, Hook T, Montanini P. Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET. In: Symposium on VLSI Technology (VLSI '17); 5-8 June 2017; Kyoto, Japan. New York: IEEE; 2017. pp. T230-T231

[19] Zhang Q, Yin H, Meng L. Novel GAA Si nanowire p-MOSFETs with excellent short-channel effect immunity via an advanced forming process. IEEE Electron Device Letters. 2018;**39**(4):464-467. DOI: 10.1109/LED.2018.2807389

Chapter 3

Work Function Setting in High-k Metal Gate Devices

Elke Erben, Klaus Hempel and Dina Triyoso

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.78335

Abstract

As transistor size continues to shrink, SiO_2/polysilicon gate stack has been replaced by high-k/metal gate to enable further scaling. Two different integration approaches have been implemented in high-volume production: gate first and gate last; the latter is also known as replacement gate approach. In both integration schemes, getting the right work functions and threshold voltages for N-type metal-oxide-semiconductor (NMOS) and P-type metal-oxide-semiconductor (PMOS) devices is critical. A number of recent studies have shown that the threshold voltage of devices is highly dependent on not just the deposited material properties but also on subsequent device processing steps. This chapter contains a description on the different mechanisms of work function setting in gate last and gate first technologies, the sensitivities to different process conditions and special measurement techniques for gate stack analysis is shown.

Keywords: complementary metal-oxide-semiconductor, NMOS transistor, PMOS transistor, high-k metal gate, work function, gate first, gate last, replacement metal gate, low energy ion scattering, electron energy loss spectroscopy

1. Introduction

The basic principle of metal-oxide-semiconductor field-effect transistor (MOSFET) function has not been changed since the introduction of this transistor type ~40 years ago. The control of charges close to the silicon surface by an applied voltage to the gate electrode turns the transistor channel on and off. The required gate voltage to turn the transistor on (to form the inversion channel)—the threshold voltage Vt—is defined by the work functions of the transistor channel semiconductor and the gate electrode and by additional charges at the transistor channel-dielectric interface and distributed charges through the dielectric. The work function difference between channel material and gate electrode should be small to ensure a low threshold voltage (**Figure 1**). The Si/SiO_2-dielectric/polysilicon-electrode gate stack is optimized to

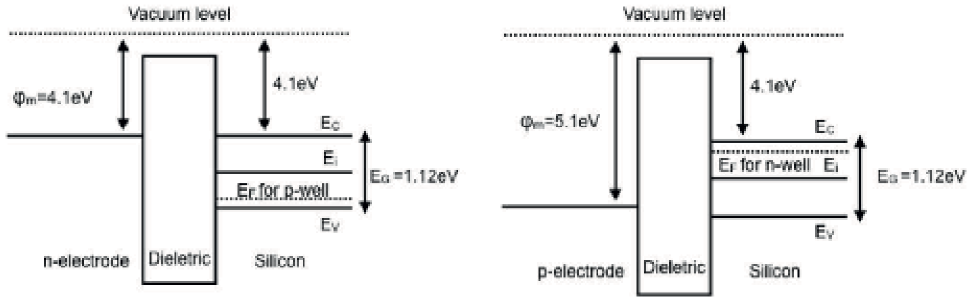


Figure 1. Energy diagrams for NMOS (left) and PMOS (right). φ_m is the work function of the gate electrode, E_C the conduction band, E_V the valence band, and E_F the Fermi level energy.

fulfill these requirements. Doping of the polysilicon can tune the work function for N-type metal-oxide-semiconductor (NMOS) and P-type metal-oxide-semiconductor (PMOS) transistors accordingly. In the early years of MOSFET technology, typical gate length was around several micrometers and the thickness of the dielectric between the silicon and the gate electrode was above 10 nm. Today's leading edge technologies have a gate length of below 20 nm, a shrink by a factor of 100 and almost in the range of gate oxide thickness from the early technologies. This scaling of the gate length also requires a significantly thinner gate dielectric for gate control and to mitigate short channel effects. The desired electrical thickness of the gate dielectric became less than 2 nm. The well-established SiO_2 dielectric became too leaky for this thickness range, since tunnel leakage became the dominating leakage path. Therefore, the SiO_2 with dielectric constant $k = 3.9$ had to be replaced by a dielectric material with higher dielectric constant, a so-called high-k material such as HfO_2 ($k = 20$). The introduction of this new material requires also a change in the material for the gate electrode. Direct contact of HfO_2 with polysilicon leads to oxygen and electron transfer through this interface. As a result, the Fermi level of p^+ polysilicon increases significantly and the Fermi level of n^+ polysilicon decreases, causing high threshold voltages. This effect is known as Fermi pinning [1]. The work function difference between n and p-gate electrodes becomes very small. To avoid this effect, the gate electrode on top of the high-k dielectric must be a metal electrode. Two different integration approaches for high-k metal gate have been developed and implemented in high-volume production: gate first and gate last; the latter is also known as replacement gate approach. In both integration schemes, getting the right threshold voltage for NMOS and PMOS devices is a challenge. In gate first technology, the complete gate stack is formed before gate patterning and has therefore to withstand the high thermal budget of all subsequent processes which are required for transistor formation, including dopant activation. This exposure to high temperature limits the material choice for the gate stack [2]. The work function of metals used in the gate stack is shifted toward mid-gap for temperatures above ~500°C. The required work functions for NMOS and PMOS have to be set by careful optimization of thermal treatments or anneals. The details of this approach are given in Section 2.1. For gate last approach, a polysilicon dummy gate is formed as in the classical SiO_2/polysilicon technology, and all process steps with high thermal budget will be performed with this dummy gate in place. The dummy gate will be removed after completion of all implant and high thermal budget processes and replaced by a metal gate electrode. The work function of this gate stack will be defined by the used metals, their thickness values and deposition conditions. Details will be described in Section 2.2.

2. Metal gate technologies

2.1. Gate first technology

The metal gate for NMOS transistors requires a work function close to the conduction band of Si (~4.1 eV) and the PMOS transistor needs a metal gate with a work function close to Si valence band (~5 eV). There are known metals with the right work functions, TiN for PMOS and Al for NMOS [2]. But the high temperature processes required for several steps post gate patterning will shift the work functions of these metals toward mid-gap, causing unacceptably high threshold voltages. The reason for this work function shift is the diffusion of oxygen from the metal layer to the interface of the high-k dielectric and the metal electrode [3]. Therefore, another way has to be found to set the work function of a Si-high-k dielectric/metal gate stack. The interface between the transistor channel and the high-k dielectric plays a critical role for this purpose. The direct contact of Si to the high-k material decreases carrier mobility and creates defects impacting the electrical characteristics of the transistors, including reliability. A roughly 1 nm thick interfacial SiO_2 layer, grown by wet chemical oxidation, prevents these effects. The properties of this interfacial layer are not stable against several subsequent processes. A stabilization of the oxide interfacial layer by nitridation is required, leading finally to a SiON interfacial layer.

The effective work function of such a gate stack is defined by a dipole layer consisting of metal and oxygen atoms at the interface between the SiON interfacial layer and high-k dielectric [4]. To form this dipole layer, a thin capping metal layer has to be deposited on top of the high-k dielectric; TiN is used for this purpose. The atoms for NMOS- and PMOS-specific dipole formation are deposited by plasma enhanced vapor deposition (PVD) as very thin layers on top of this capping TiN and have to be driven into the high-k dielectric by high temperature anneal. Lanthanum (La) and Aluminum (Al) have been found to be suitable materials for dipole formation in NMOS and PMOS transistors, respectively [4]. The achievable threshold voltage depends on the number of metal atoms available for dipole formation at the interface between the SiON interfacial layer and high-k dielectric. La and Al have different saturation behavior of surface coverage through deposition. Therefore, La allows a wide range of Vt tuning. Al content, in contradiction, saturates quickly and therefore PMOS has a significantly smaller range of Vt tuning by the Al-based dipole formation. **Figure 2** shows the gate stack for NMOS and PMOS devices before the drive-in anneal. The dipole formation alone is not sufficient to achieve the PMOS band edge.

Replacing the Si in the transistor channel by SiGe lowers the achievable Vt significantly, due to the different energy band structure of SiGe versus Si, see **Figure 3**.

The valence band energy of SiGe is significantly higher compared to Si, the band gap is slightly smaller. The p-type field effect transistor (PFET) work function can then be tuned in addition

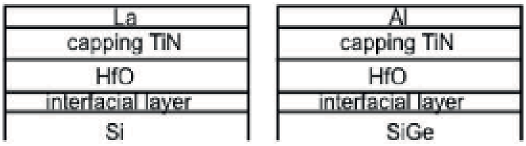


Figure 2. Gate stack for NMOS (left) and PMOS (right) devices before the drive-in anneal.

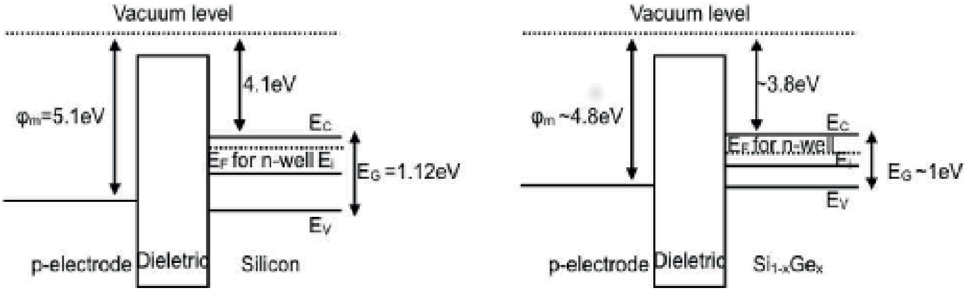


Figure 3. Energy diagrams for PMOS on Si (left) and PMOS on SiGe with 25% Ge content (right). φ_m is the work function of the gate electrode, E_C the conduction band, E_V the valence band, and E_F the Fermi level energy.

by the Ge content of the SiGe channel [5]. Then, higher the Ge concentration, lower the Vt (~10 mV Vt shift by 1% change of Ge concentration).

The temperature range for the drive-in anneal must be chosen carefully, since interfacial regrowth may occur causing an increase of electrical interface thickness (CET). At the same time, oxygen scavenging will be observed, since the high temperature drive-in anneal is performed with a TiN layer on top of the stack [6]. Indirect scavenging will cause a thinning of SiON interface by N—O exchange in the interfacial layer and partially oxidation of TiN. This effect has a minor impact on CET reduction in comparison with other effects within complementary metal-oxide-semiconductor (CMOS) process but is also detrimental in terms of defect formation (oxygen vacancies) at the interface to the channel, leading to threshold variation and in worst case to reliability problems. Interfacial layer scavenging is observed for a couple of materials in gate first approach and happens either in a direct way by diffusing scavenging elements toward the HfO_2/SiON interface (like La and Al) or remote by isolating the scavenging elements from the interface (like TiN) [6]. Besides the temperature range and the scavenging element, the composition and deposition method of the interfacial layer contributes to the overall scavenging amount that could be achieved. The scavenging effect is basically a reduction of interface thickness by oxidation of metal dopants at the interface between high-k and interfacial layer and/or oxidation effects at the interface between HfO_2 and TiN top layer, see **Figure 4**.

After the drive-in anneal, the NMOS- and PMOS-specific capping layers have to be removed from the high-k dielectric and replaced by a common final metal electrode for both transistor flavors to avoid the Fermi pinning. A polysilicon layer is deposited on top of the metal gate electrode for a proper contact formation by silicidation. **Figure 5** shows the final gate stack after drive-in anneal and capping layer and polysilicon deposition.

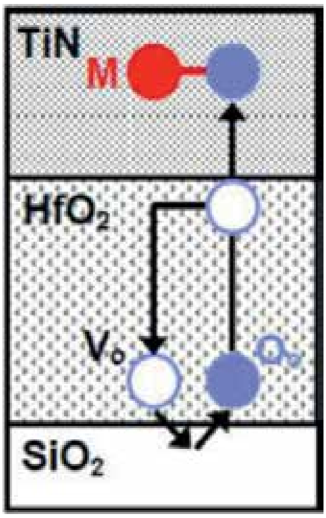


Figure 4. Mechanism of remote interfacial layer scavenging by high temperature anneal (T ≥ 850°C) with TiN on top of high-k/SiO stack. M, V0, and O0 represent the scavenging element, the oxygen vacancy in HfO_2, and the oxygen atom in the lattice position of HfO_2, respectively [6].

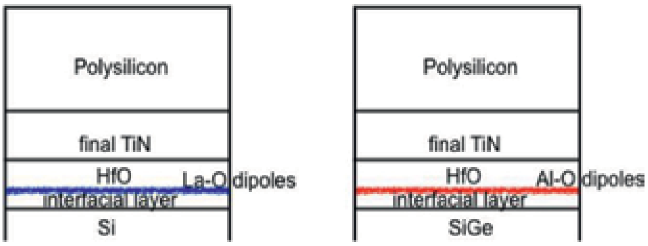


Figure 5. Final gate stack of NMOS transistor (left) and a PMOS transistor (right) with the corresponding dipole layers in place.

After formation of the gate stack, all following process steps are comparable to those in conventional SiO_2-polysilicon technology. Special care hast to be taken to avoid any oxygen ingress into the gate stack, since this will cause uncontrolled Vt shifts of the devices.

2.2. Gate last technology

In planar gate last technology, the high k metal gate stack is built after completion of all processes up to silicidation in the front end of line (FEOL) of the whole CMOS flow, including high-temperature processes. There have been two options developed, either the high-k gate stack is deposited prior to gate patterning and the metal gate stack is deposited after removal of the polysilicon gate or the complete high-k metal gate stack is deposited after removal of the polysilicon gate [7]. The mechanism of work function setting does not differ between these two options.

In addition to delivering the required work functions, the gate materials have to be compatible to the CMOS process flow, must not cause danger of uncontrolled metal contamination of wafers and tools or cause reliability problems. Aluminum with a work function of 4.1 eV is a suitable material for NMOS transistors. One possible material for PMOS transistors is TiN. The work function of TiN can be tuned close to 5 eV depending on the detailed composition of TiN, like the Ti to N ratio, the TiN thickness, and the deposition techniques.

The direct contact of the gate metals to the high-k dielectric or to a too thin capping layer on top of the dielectric may damage the dielectric and create leakage paths. One example of Al spiking through an insufficient protected gate dielectric is shown in **Figure 6**.

To avoid this, the high-k dielectric has to be protected against the gate metal, especially the Al for the NMOS, by a protection layer of certain thickness. A roughly 2 nm thick TiN layer on top of the high-k dielectric protects the high-k dielectric against damage due to metal gate deposition. CMOS technology sets an additional boundary condition to the processes of metal gate formation. Since there must be different gate materials for NMOS and PMOS transistors close to each other, a process sequence had to be developed allowing the deposition of the different gate metals without disturbing the gate stack of the complementary transistor. This could be realized by first depositing the metal gate for PMOS transistors on all devices, including NMOS, and then removing the PMOS metal gate (TiN) from the NMOS. To achieve this, the removal of the TiN has to be well controlled and selective to the TiN protection layer on top of the high-k dielectric. This can be realized by introducing a thin stopping layer on top of the protection TiN layer. TaN was found as a suitable material for this purpose. The removal of the TiN metal gate on the NMOS transistors stops on the TaN layer, and then the metal gate for the NMOS transistors is deposited. Once this has been completed, the gate electrodes of both NMOS and PMOS devices will be finalized with deposition of aluminum. As a result, the complete gate stack becomes quite complex, and it becomes difficult to ensure reproducible and reliable work functions. **Figure 7** shows the resulting gate stack for NMOS and PMOS transistors, respectively. The effective work function of both device flavors is defined rather by a multilayer gate stack then by a single metal with a clearly defined work function.

The thickness of the single metal layers is only a few nanometer each, so the gate stack is more a sequence of several interfaces than a stack of different bulk metal layers. A metal to metal interdiffusion of atoms from the single layers into neighboring layers takes place, resulting in an effective work function for the whole stack [8].

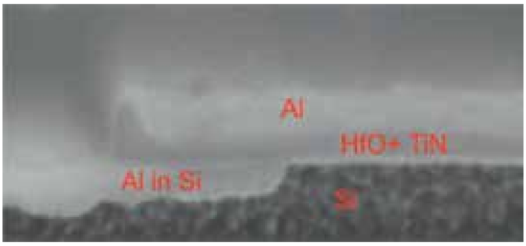


Figure 6. SEM image of a transistor heavily affected by Al spiking.

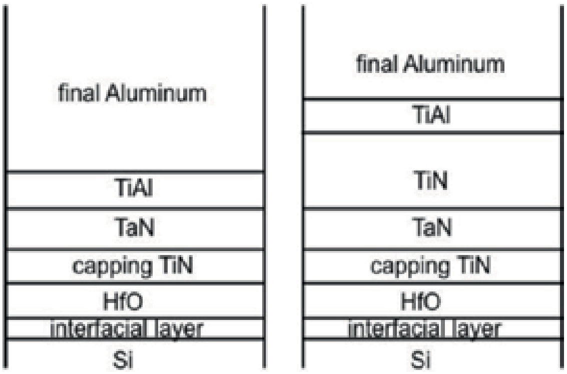


Figure 7. Gate stack of NMOS transistor (left) and a PMOS transistor (right).

3. Analytical characterization of the gate stack

The effective work function of high-k metal gate transistors is defined by complex gate stacks in both gate first and gate last technologies. The analytical characterization of these gate stacks is challenging, but required for process development and optimization. Gate first technology requires a detailed quantitative mapping of the dipole forming metals close to the interfacial SiON-layer—high k dielectric interface. Since the deposition of these metals is realized on top of the TiN capping layer, a surface sensitive analytical technique must be applied for this task. The required spatial resolution is not high, since the deposition of the work function metals is done before gate patterning.

The characterization of the gate stack in gate last technology has to fulfill different requirements. Since the gate stack formation is done after gate patterning, it has to be applicable to real device structures, meaning it requires a very high spatial resolution combined with depth profiling for different elements through several metal layers down to the high-k dielectric.

3.1. LEIS for gate first technology

Low energy ion scattering (LEIS) is a unique tool in surface analysis, since it provides the atomic composition of the outer surface as well as a nondestructive ("static") in-depth profile (0–10 nm) for the heavier elements [9]. In LEIS, the surface of a solid is bombarded with noble gas ions such as 4He+ and 20Ne+ with energies between 1 and 10 keV. For a fixed scattering angle, the energy distribution of the backscattered ions is measured. The interaction of the ion with the surface can be considered as an elastic collision with a single surface atom at rest. For a given primary ion and energy, the energy Ef of the backscattered ion is determined by the mass of the (unknown) surface atom and the scattering angle. Conservation of energy and momentum results in a higher Ef for scattering by a heavier target atom. Using noble gas ions makes LEIS extremely surface sensitive because most of the ions that penetrate the outer monolayer are neutralized and therefore do not contribute to the scattered ion spectrum. In general, there are no matrix effects in LEIS; the ion yield for scattering by one atomic species does not depend on the other atomic species in the surface. A LEIS analysis thus gives the atomic composition of the outer atomic layer of a surface. For atomic layer deposition (ALD) growth, the extreme surface sensitivity of LEIS allows to characterize the progress in the closure of the layer for every deposition cycle, precisely during the transient regime that is responsible for the thickness inhomogeneity of the final layer. **Figure 8** shows the different surface saturation behavior for sputtered La versus Al. The different substrates had been chosen since La signal was better detectable on Hf covered surface than on Si substrate. **Figure 9** illustrates the effect of surface treatment prior to ALD process. Surface coverage saturates already between 10 and 15 cycles in case of pretreated surface in comparison with >20 cycles for surface without pretreatment.

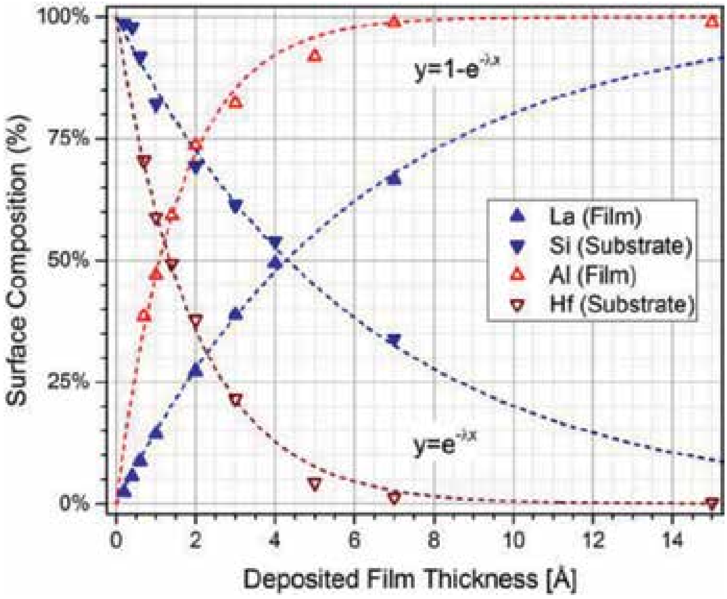


Figure 8. Surface coverage La versus Al deposited by PVD on pure Si substrate and on Si substrate covered by thin Hf layer for better solution of La signal [10].

3.2. EDX and EELS for gate last technology

The resulting work function and therefore threshold voltage are very sensitive to the details of the gate stack composition and deposition conditions. An analytical technique is required to investigate the impact of process details on the resulting threshold voltages. In order to accurately account for all of these impacts, there is a need to apply analytical methods which accurately measure material composition on real device structures, rather than on unpatterned wafers.

Energy-dispersive X-ray spectroscopy (EDX) is a well-established analytical method in conjunction with transmission electron microscopy (TEM) for the detection of metals used in the gate stack. TEM has the required spatial resolution to investigate the gate stack on real transistors. Electron energy loss spectroscopy (EELS) is a technique preferably used for the

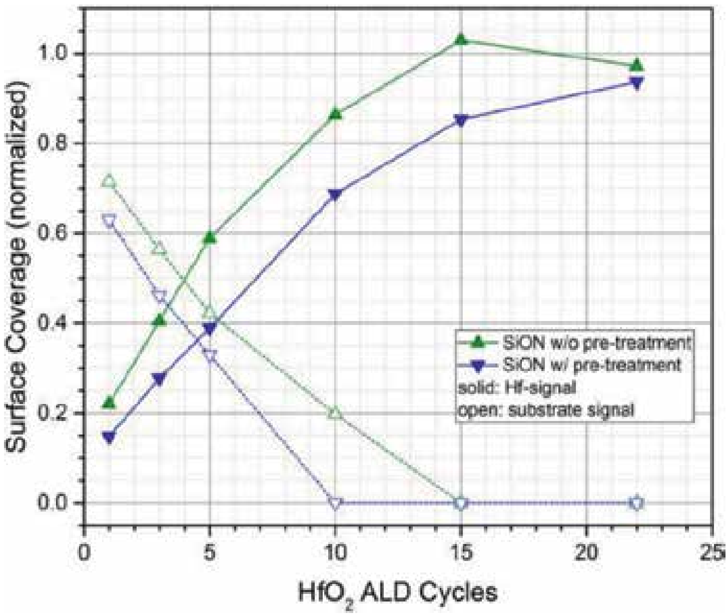


Figure 9. Surface coverage by HfO_2-ALD process in dependence on ALD cycle number and surface pretreatment [9].

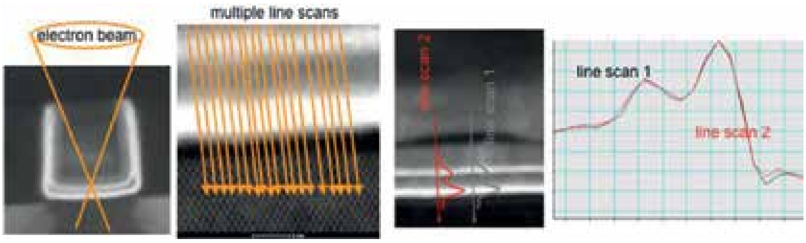


Figure 10. High-resolution EELS method with multiple line scans to improve SNR.

detection of lighter elements in the gate stack, like oxygen and nitrogen, which play also an important role in work function setting. An advanced high-resolution EELS method capable of accurate measurement of material composition on device structures can be applied for this task [11]. The standard EELS measurement has too low probe intensity for high enough signal-to-noise ratio (SNR). In order to improve the SNR, multiple line scans have to be done across the layers of the gate stack and then aligned to each other and integrated, see **Figure 10**.

The standard and high-resolution EELS profiles of a sample are shown in **Figure 11**. The gate stack is shown from right to left, bulk silicon, interfacial SiON, high-k dielectric, capping TiN layer, TaN stopping layer, TiN layer for PMOS. No details of the oxygen and nitrogen profiles in the lower part of the gate stack can be resolved. The high-resolution EELS spectrum of the same sample shows the profiles of both oxygen and nitrogen through the gate stack. A dip in the oxygen profile at the high-k dielectric-capping layer interface can now be resolved.

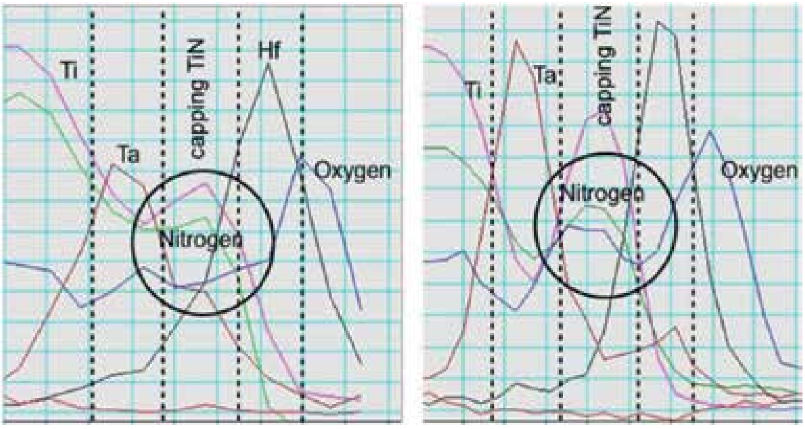


Figure 11. Standard resolution EELS spectrum (left), no details of the oxygen and nitrogen profiles in the lower part of the gate stack can be resolved. A dip in the oxygen profile at the interface capping TiN-HfO_2 can be resolved for a high-resolution EELS profile (right).

This technique can be used to understand the observed differences in threshold voltage between devices processed with slightly different formation of the gate stack.

3.2.1. N-type metal-oxide-semiconductor

The gate stack of NMOS transistors consists of interfacial oxide—high-k dielectric (HfO_2)—TiN—TaN—TiAl—final Al. The desired 4.1 eV work function from the Al has to be achieved for the whole stack of the gate electrode, even if the Al is not in direct contact with the HfO_2. Therefore, the TiN protection layer and the TaN stopping layer have to be thin enough not to screen the work function of the Al from the HfO_2. An interdiffusion of atoms occurs between the different metal layers resulting in the formation of the effective work function. This interdiffusion can be enhanced by thermal processes, like the reflow of the final aluminum with 400–480°C. If this interdiffusion is too strong or the TiN protection layer is not stable enough, aluminum atoms may penetrate the HfO_2 and cause leakage paths in the device, as shown in **Figure 6**. **Figure 12** shows the EDX and EELS profiles of two samples with different reflow temperatures for the final aluminum. A 50°C higher reflow temperature causes several orders of magnitude higher leakage current. One way to avoid this detrimental Al penetration is a thicker TaN layer on top of the capping TiN. However, if the thickness of the protection and stopping layers is too large, the effective work function of the gate electrode becomes too high, resulting in higher threshold voltage. This is also reflected in the combined EDX-EELS spectra shown in **Figure 13**. The optimum conditions require a tight balance between the thicknesses of the different metal layers, the deposition details and reflow temperatures.

3.2.2. P-type metal-oxide-semiconductor

The desired work function for PFETs is in the range of 5 eV. This requires a different gate stack composition as for the NFET, especially the impact of the final Al to the effective work function

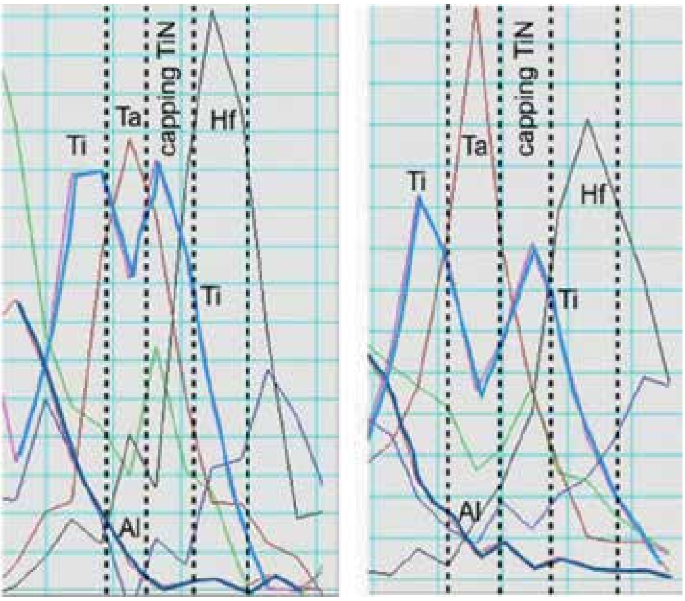


Figure 12. Combined EDX and EELS profiles of the different materials of the NMOS gate stack for a sample with lower reflow temperature (left) and higher reflow temperature (right). The Al tail in the region of capping TiN and HfO_2 layers is more pronounced for the sample with higher reflow temperature for the final Al.

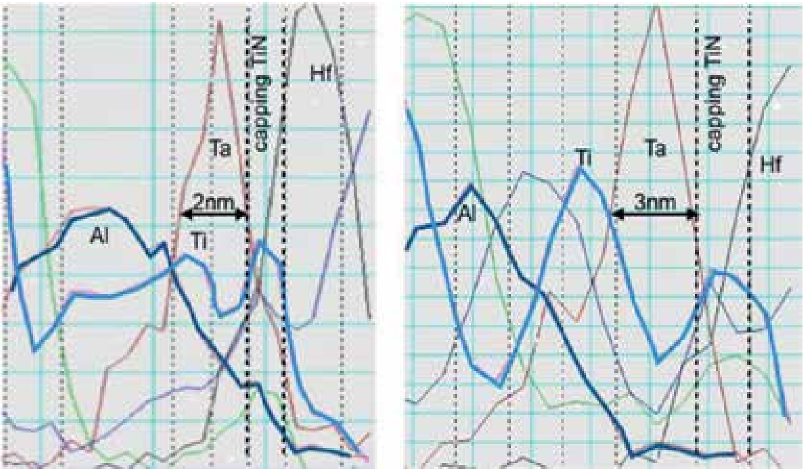


Figure 13. Combined EDX and EELS profiles of the different materials of the NMOS gate stack for a sample with low Vt (left) and 50 mV higher Vt (right). The thicker TaN layer and the lower Al content close to the HfO_2 can be seen.

must be screened from the lower part of the gate stack. TiN has been found as a suitable material for this purpose. The thickness of this second TiN in the gate stack must be larger than that of the capping TiN in order to keep the Al away from the high k dielectric. In addition to the composition of the metal electrode the resulting Vt for PMOS also depends

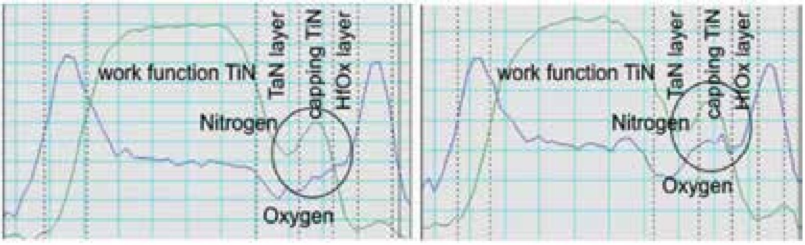


Figure 14. High-resolution EELS spectrum of a sample with high Vt (left), a low oxygen concentration in the capping layer with a strong gradient toward the high-k dielectric is detected. High-resolution EELS spectrum of a sample with low Vt (right). The dip of the oxygen concentration at the interface capping layer—high-k dielectric is well resolved.

strongly on the concentration of nitrogen and oxygen at the interface of high k dielectric and metal layer [12].

The responsible mechanism for the work function setting corresponds therefore more to the dipole engineering at the interfacial-high-k dielectric interface. The required PMOS Vt can only be achieved by having the correct concentration of nitrogen and oxygen at the high-k dielectric-capping layer interface. Again, as for the NFET, the dependency of the Vt from the gate stack composition can be checked by combined EDX/EELS spectra of the gate stack. High-resolution EELS spectra from two samples with different Vts are shown in **Figure 14**. There is reasonably a high oxygen level in the capping layer with a dip of the oxygen concentration at the interface cap layer—high-k dielectric for the transistor with reasonable low Vt. The EELS spectrum of a sample with high Vt shows a low oxygen concentration in the capping layer with a strong gradient toward the high-k dielectric.

4. Conclusion

As a consequence of the aggressive scaling of transistor dimensions, the engineers have developed two quite different approaches to address the integration of high-k gate dielectric into the very complex CMOS process flows. Gate first and gate last technologies use different mechanisms to set the work functions required to achieve the desired threshold voltage. Gate first technology is based on dipole formation at the interfacial layer-high-k dielectric interface, whereas gate last technology uses metal-metal interdiffusion within the metal gate electrode to tune the work function. Both technologies are capable to deliver high performance devices in high-volume production. To date, gate first technology is targeted more for low leakage, low-power applications and is applied, for example, in fully depleted SOI technology [13], whereas gate last is used in FINFET technology for high performance products [14]. The final gate stacks are quite complex for both approaches and involve a large number of process steps, which had to be optimized carefully to achieve the desired result. Advanced analytical techniques had to be adapted to meet the specific requirements for process characterization of gate first and gate last technologies.

Acknowledgements

The authors would like to thank Pavel Potapov and Kornelia Dittmar for their support for the analytical characterization of the gate stacks and Joachim Metzger and Robert Binder for the development of required PVD processes.

Author details

Elke Erben[1], Klaus Hempel[1]* and Dina Triyoso[2]

*Address all correspondence to: klaus.hempel@globalfoundries.com

1 GLOBALFOUNDRIES Fab1 LLC & Co. KG, Dresden, Germany

2 GLOBALFOUNDRIES, Malta, NY, USA

References

[1] Shiraishi K, Akasaka Y, Umezawa N, et al. Theory of fermi level pinning of high-k dielectrics. In: Proceedings of IEEE Simulation of Semiconductor Processes and Devices, 2006 International Conference. DOI: 10.1109/SISPAD.2006.282897

[2] Tseng H-H. The progress and challenges of applying high-k/metal-gated devices to advanced CMOS technologies. In: Swart JW, editor. Solid State Circuits Technologies. Rijeka: InTech; 2010. ISBN: 978-953-307-045-2. Available from: http://www.intechopen.com/books/solid-state-circuits-technologies/theprogress-and-challenges-of-applying-high-k-metal-gated-devices-to-advanced-cmos-technologies

[3] Schaeffer JK, Capasso C, Fonseca LRC, et al. Challenges for the integration of metal gate electrodes. In: Proceedings of IEEE International Electron Devices Meeting. 2004. pp. 287-290. DOI: 10.1109/IEDM.2004.1419135

[4] Gilmer DC, Schaeffer JK, Taylor WJ, Spencer G, Triyoso DH, Raymond M, Roan D, Smith J, Capasso C, Hegde RI, Samavedam SB, et al. In: Proceedings of 2006 European Solid-State Device Research Conference. DOI: 10.1109/ESSDER.2006.307710

[5] Gilmer DC, Schaeffer JK, Taylor WJ, et al. Strained SiGe Channels for Band-Edge PMOS Threshold Voltages With Metal Gates and High-k Dielectrics in IEEE Transactions on Electron Devices. 2010;**57**(4):898-904. DOI: 10.1109/TED.2010.2041866

[6] Ando T. Ultimate scaling of high-κ gate dielectrics: Higher-κ or interfacial layer scavenging. Materials. 2012;**5**:478-500

[7] Packan P, Akbar S, Armstrong M, et al. High performance 32 nm logic technology featuring 2nd generation high-k + metal gate transistors. In: Proceedings of IEEE International Electron Devices Meeting. 2009. DOI: 10.1109/IEDM.2009.5424253

[8] Lu et al. Characteristics and mechanism of tunable work function gate electrodes using a bilayer metal structure on SiO_2 and HfO_2. IEEE Electron Device Letters. 2005;**26**(7):445-447

[9] Dittmar K, Triyoso DH, Erben E, et al. The application of low energy ion scattering spectroscopy (LEIS) in sub 28-nm CMOS technology. Surface and Interface Analysis. 2017;**49**:1175-1186. DOI: 10.1002/sia.6312LEIS

[10] Drescher M. PhD (to be published end of 2018)

[11] Hempel K, Erben E, Binder R, Triyoso D, et al. Impact of both metal composition and oxygen/nitrogen profiles on p-channel metal-oxide semiconductor transistor threshold voltage for gate last high-k metal gate. Journal of Vacuum Science & Technology B. 2013; **31**(2):2202

[12] Hinkle et al. Interfacial oxygen and nitrogen induced dipole formation and vacancy passivation for increased effective work functions in TiN/HfO_2 gate stacks. Applied Physics Letters. 2010;**96**:103502

[13] Carter R, Mazurier J, Pirro L, et al. 22 nm FDSOI technology for emerging mobile, internet-of-things, and RF applications. In: The Proceeding of 2016 IEEE International Electron Devices Meeting (IEDM). DOI: 10.1109/IEDM.2016.7838029

[14] Narasimha S, Jagannathan B, Ogino A, et al. A 7 nm CMOS technology platform for mobile and high performance compute application. In: The Proceedings of 2017 IEEE International Electron Devices Meeting (IEDM). DOI: 10.1109/IEDM.2017.8268476

Chapter 4

Selective Epitaxy of Group IV Materials for CMOS Application

Guilei Wang, Henry H. Radamson and
Mohammadreza Kolahdouz

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76244

Abstract

As the International Technology Roadmap for Semiconductors (ITRS) demands an increase of transistor density in the chip, the size of transistors has been continuously shrunk. In this evolution of transistor structure, different strain engineering methods were introduced to induce strain in the channel region. One of the most effective methods is applying embedded SiGe as stressor material in source and drain (S/D) regions by using selective epitaxy. This chapter presents an overview of implementation, modeling, and pattern dependency of selective epitaxy for S/D application in CMOS. The focus is also on the wafer in and ex situ cleaning prior to epitaxy, integrity of gate, and selectivity mode.

Keywords: selective epitaxy, SiGe, RPCVD, CMOS

1. Introduction

Selective epitaxy of SiGe material is considered as one of the most important steps in the CMOS processing. This type of growth method was already discovered and highlighted in the 1990s with a focus on optimizing the growth parameters to improve the layer quality of Si/SiGe multi-layers [1–4]. The outcome of the initial works showed that {113} and {110} facet planes were mostly dominant, whereas other facet planes, e.g., {119} and {018}, could also appear at certain growth conditions. The activation energy for the growth rate on facet planes (Rhkl) was also calculated for growth temperature of 700–850°C. No significant change in the activation energy of deposition on {hkl} surfaces compared to (100) one was observed showing a similar kinetic growth ruled over these facet planes [1].

The facet formation results in a pileup shape at the edge of the epi-layer close to the oxide (see **Figure 1a–c**) [5–7]. The reason behind forming this pile shape is the diffusion of incoming Si or Ge atoms on the faceted surface is higher than those on the (001) surface at the central part of the oxide opening. This means that the molecules may move toward the central part, and if their diffusion length at a certain growth temperature is shorter than the opening size, a pileup shape is formed.

More detailed studies about facet formation showed the existence of chlorine during the epitaxy is the main factor dominating the growth rate for both Si and SiGe; however, the formation of the facets is correlated with the growth conditions and pattern geometry.

The application of SiGe as embedded layer source and drain (S/D) of pMOS was presented in year 2000 when (S/D) junctions were formed by a dry etch followed by the growth of highly B-doped SiGe [8]. This idea arose attention since no dopant implantation and post annealing for activation were necessary when the in situ doping of SiGe layer could provide high-quality epi-layers.

Later, the embedded SiGe layers were used as stressor material in the S/D regions to create uniaxial strain in the channel region. As the result of induced strain, carrier mobility in the channel is significantly improved. For the first time, $Si_{0.83}Ge_{0.17}$ layers were integrated by Intel in 90 nm technology node in 2003 [9], and since that year, the Ge content (or strain amount) was continuously increased in each new coming technology node up to 45% in 22 nm technology node [10, 11]. During this evolution of CMOS, a revolutionary design was introduced when the planar type of transistors was abandoned and three-dimensional (3D) transistors were implemented [12]. Such nanoscale 3D transistors were initially called Tri-Gate, but later

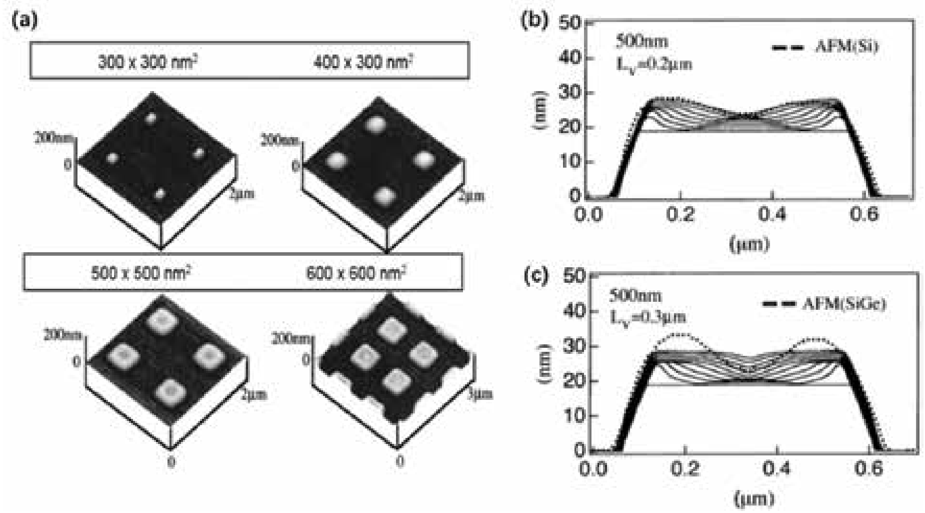


Figure 1. (a) AFM pictures of 6 nm selectively grown SiGe on 18 nm Si buffer layers grown at 740°C. The oxide layer has been removed and SiGe layers are in mesa shape. (b) and (c) show the cross section of simulated and the experimental cross-sectional profile (dotted line) for Si and SiGe layers, respectively [5].

this notation was changed to FinFETs to merge with the other groups' suggestion. In these transistors, selective epitaxy of SiGe layer was used to raise the S/D regions.

A drawback with selective epitaxy growth is that the SiGe layer profile is dependent on shape, size, and density of the oxide openings of S/D in a chip. This problem affects also the B concentration in SiGe since the incorporation of B is dependent on both the growth rate and the Ge content [13–21].

SiGe as channel material has been also proposed in the SiGe/Si vertical nanowire transistors when the lateral downscaling will finally reach the end of technology roadmap. In order to have a full control on the carrier transport in the channel region, gate-all-around (GAA) design has been proposed [22].

2. Chemical vapor deposition (CVD) technique

In the semiconductor industry, there are many different materials need to be grown during the device processing. CVD is also a process in which gaseous chemical precursors have chemical reactions on the wafer surface and deposit a layer of solid thin film. The rest of the by-products that are in gas phase can be easily pumped and leave the reaction surface. Among these CVD deposited thin films, SiGe is a key material, which offers the applications in a wide range of CMOS devices.

CVD is practiced in a variety of types, which are classified by the operating pressure:

1. Atmospheric pressure CVD (APCVD)—CVD at atmospheric pressure
2. Reduced-pressure CVD (RPCVD)—CVD at torr pressure
3. Low-pressure CVD (LPCVD)—CVD at mTorr pressure
4. Ultrahigh vacuum CVD (UHVCVD)—CVD at 10^{-8} torr pressure

Among these CVD techniques, RPCVD has shown the highest output, and it is accepted by semiconductor industry for epitaxy of Si and SiGe films for IC mass production.

For RPCVD, several precursors are available in the market for the Si growth such as silane (SiH_4), dichlorosilane (SiH_2Cl_2), trichlorosilane ($SiHCl_3$), tetrachlorosilane ($SiCl_4$), disilane (Si_2H_6), and trisilane (Si_3H_8). Germane (GeH_4) and digermane (Ge_2H_6) are the common sources for Ge to grow $Si_{1-x}Ge_x$. The most commonly used precursors for p- and n-type doping are diborane (B_2H_6), phosphine (PH_3), and arsine (AsH_3), respectively. $As(GeH_3)_3$ is a gas source which is also used for n-type doping of SiGe layers at low growth temperature. These sources are usually diluted in H_2. For Sn growth, SnD_4 and $SnCl_4$ are the most practical and common sources [23]. Methylsilane (SiH_3CH_3) is widely used for carbon doping in Si and SiGe layers. Meanwhile, HCI and Cl_2 are the precursors used as the etch reactant to obtain the selectivity during the selective process.

3. Ex- and in-situ cleaning

The wafer cleaning process before the epitaxy has an important role for the epitaxy quality. The purpose of epitaxy is to duplicate the substrate atomic columns in deposition of a layer. Therefore, the presence of SiO_2, carbon and polymer residuals has to be removed from the Si surface [24].

As an example, **Figure 2** shows the micrographs of the samples prior and after the SiGe selective epitaxy. The presence of carbon residuals on Si surface is commonly formed during the plasma dry etch of oxide openings. The SiGe epitaxial layer can be only grown on the clean surfaces of Si, and the growth occurs through nucleation as shown in **Figure 2b** and **d**. The TEM EDS mapping analysis from the cross section of S/D areas in **Figure 2e** confirms the presence of carbon and oxygen residuals on the initial Si surface. Meanwhile, a standard chemical cleaning will remove all undesired impurities, and a two-dimensional SiGe layer could be grown successfully as shown in **Figure 2c** [25]. There are many different cleaning methods for Si wafers as following:

SPM: $H_2SO_4 + H_2O_2$ (4:1)

DHF: HF + DIW (1–2%)

HPM: HCl + H_2O_2+ DIW (1:1:6)

APM: $NH_4OH + H_2O_2$ + DI-H_2O (1:1:5)

where DIW stands for deionized water. A diluted hydrofluoric acid in DIW (1% DHF) removes the native oxide, and the wafers are ready to be placed in the load locks of epi-reactor [26].

Later, an in situ cleaning in RPCVD reactor is required to remove the native oxide on the exposed Si surface. This process usually occurs at high temperature such as 1000–1100°C for monitoring Si wafers; meanwhile, for the patterned substrates, this is remarkably at lower temperatures (850–950°C).

SiGe material is grown in a recess in S/D regions in nanoscale transistors. The recess is formed by wet etch, and its shape can be rounded, sigma, or trapeze, and then the in situ is needed not only in removing the contaminants from the Si surface but also in preserving the recess shape. High annealing treatment results in Si loss and affects the recess shape. This is due to the thermal mismatch between Si and SiO_2 where Si reflows in the recess region. Si loss is a critical problem for the short-channel length transistors.

In conclusion, an appropriate low annealing temperature is sought to have a trade-off between all these requirements. **Figure 3** shows the results of in situ experiments on the quality of epi-layers and the recess shape at different annealing temperatures. In these experiments, 800°C for 7 min is recognized as the minimum temperature (and enough annealing time) to preserve the shape and improve the high quality of epi-layer [27].

The in situ annealing for three-dimensional (3D) transistors, e.g., FinFET, becomes more critical where the Si fin is small and any damage has a significant effect on the transistor performance.

High-quality SiGe is required to be selectively grown on the S/D to induce the strain into the channel. Before the growth, the surface of Si fin has to be free of native oxide or any residual of impurities [28].

Figure 4a shows the HRTEM cross-sectional image of as-processed Si fin. Then the prepared samples were baked at different temperatures ranging from 740 to 825°C prior to epitaxy [29].

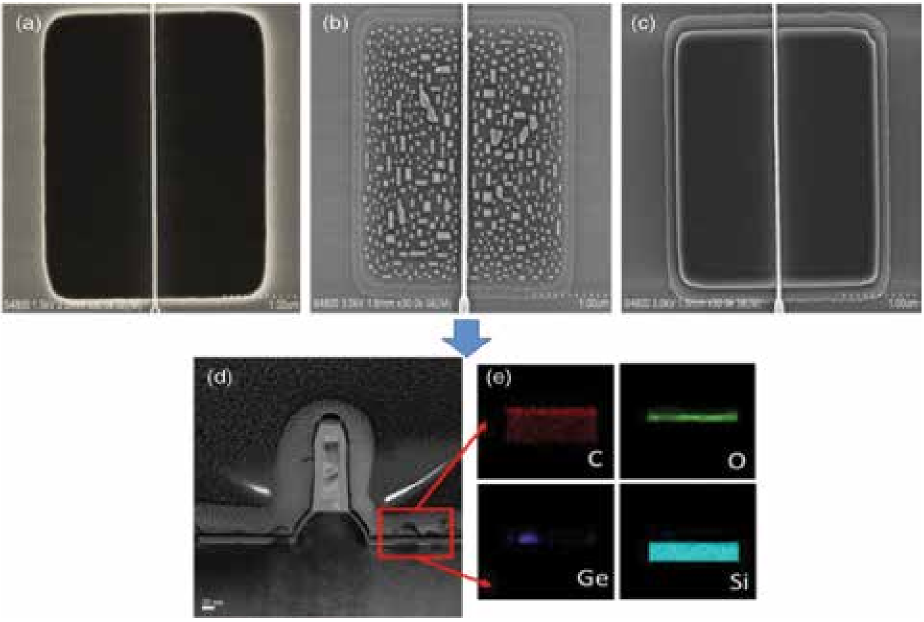


Figure 2. HRSEM images showing cross section of a planar transistor with 22 nm gate length (a) prior to SiGe growth and after SiGe epitaxy (b) with poorly cleaned Si surface and (c) with cleaned Si surface and (d) TEM cross section of sample in (b) and (e) TEM EDX mapping of sample in (b) [25].

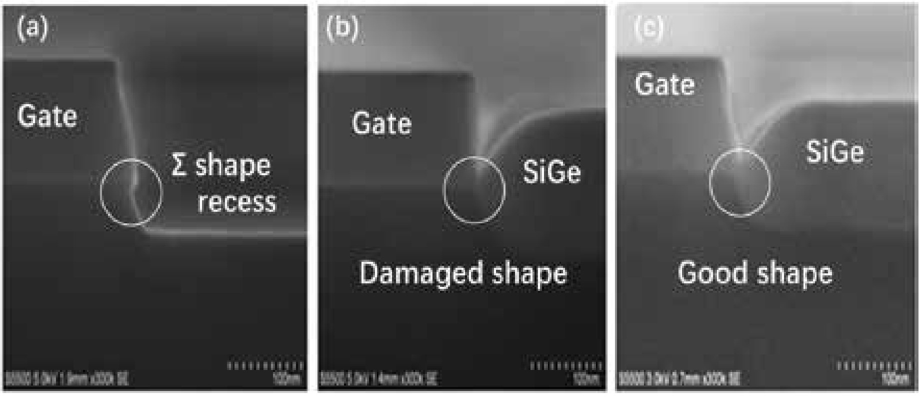


Figure 3. SEM images showing cross section of different prebaking condition samples (a) prior to prebaking, (b) prebaking temperature at 825°C, and (c) prebaking temperature at 800°C.

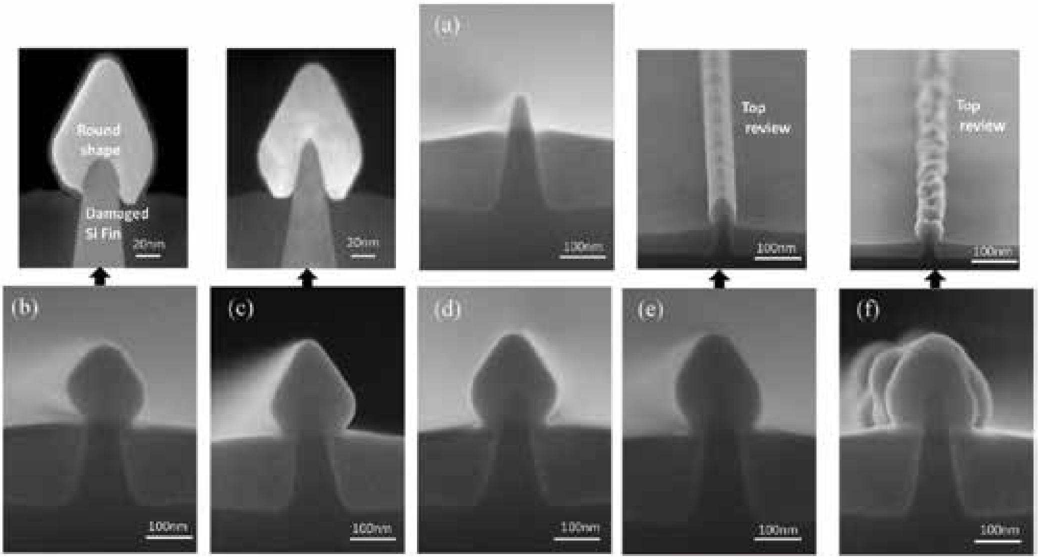


Figure 4. HRTEM and SEM cross-sectional images of the Si fins with different prebaking temperatures as follows: (a) the processed Si fin and annealed at (b) 825°C, (c) 800°C, (d) 780°C, (e) 760°C, and (f) 740°C [29].

Figure 4b reveals a serious damage on Si fin shape where the height of the fin has been shrunk and the top of the fin became rounded. Although the shape of Si fins was changed, SiGe layer could still be grown with reasonable quality. The irregularity of Si fins' shape at high temperature annealing was originated from Si migration and thermal mismatch between Si and SiO_2. Therefore, the lower baking temperature is chosen. The samples with 800 and 780°C prebaking in **Figure 4c** and **d**, respectively, had also high-quality Si fins and SiGe layers.

In general, an appearance for a successful SiGe growth is that the shape for layer coverage over the Si fin should be symmetric. The symmetry is important since it determines the uniformity of strain over the Si fins. Among the micrographs, **Figure 4e** has symmetric feature of SiGe layer, but this degrades by lowering the prebaking temperature to 760°C. **Figure 4f** shows that the surface roughness is the worst when the baking temperature is at 740°C. It is believed that asymmetric shape is a result of the residual native oxide remained on the surface of Si fins.

4. Gate integrity: HCl selectivity

The other importance of SiGe S/D epitaxy is selectivity of the growth. In the patterned structure, the surface of the gate sidewall consists of both oxide and nitride, which makes it difficult to obtain a selective growth, especially on the nitride spacer surface. However, in the worst case, at the top corners of the gate (poly-Si) would appear a "mushroom"-shaped deposition, which is difficult to be removed. To solve this issue and obtain a completely selective growth furthermore, HCl amount during the selective growth is needed.

Figure 5 shows the cross-sectional images from SiGe S/D with long channel gates. When the HCl partial pressure is 50 mTorr, small SiGe nuclides are formed, and ploy-SiGe with a mushroom shape appeared on the top of gate sidewall (**Figure 5a**). **Figure 5b** shows how the selectivity is improved by optimizing HCl partial pressure to 65 mTorr. Based on the previous reports, the high amount of HCl results in higher Ge content and lower growth rate in SiGe epitaxy. It has also demonstrated that an increase of HCl partial pressure reduces the pattern dependency behavior of the growth. However, as **Figure 5c** shows, a ditch forms in vicinity to the gate sidewall when HCl amount is further increased to 80 mTorr. One reason of the non-planar filling of SiGe in S/D regions is due to the (311) facets close to oxide where the growth rate of SiGe on crystal direction (100) is higher than (311).

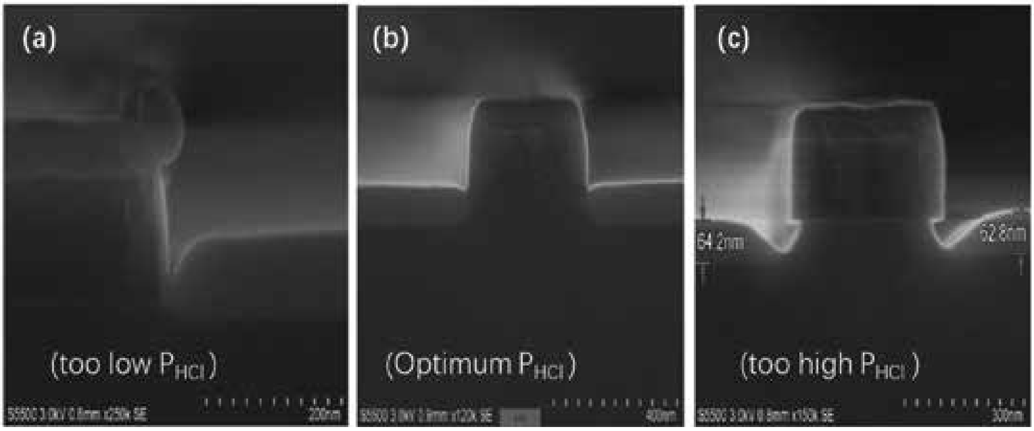


Figure 5. Cross-sectional SEM images of the gate and S/D regions after the SiGe growth when HCl partial pressures were (a) 50 mTorr, (b) 65 mTorr, and (c) 80 mTorr [25].

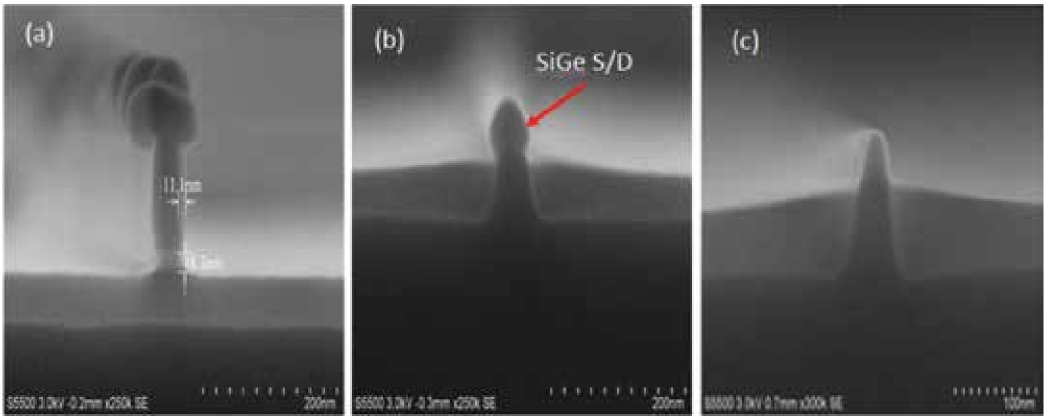

(a)
11.1nm
S5500 3.0kV -0.2mm x250k SE
200nm
(b)
SiGe S/D
S5500 3.0kV -0.3mm x250k SE
200nm
(c)
S5500 3.0kV 0.7mm x300k SE
100nm

Figure 6. Cross-sectional SEM images of Si fin/SiGe samples when HCl partial pressures were the following: (a) 50 mTorr, (b) 60 mTorr, and (c) 70 mTorr [30].

Optimizing HCl partial pressure for 3D FinFETs is a more sensitive task. **Figure 6a–c** shows the SEM images from SiGe/Si fins grown with different HCl partial pressures. As the HCl amount is not enough (50 mTorr) and "mushroom" defects occur on the top of gate sidewalls (**Figure 6a**). However, when the amount of HCI was enhanced to 60 mTorr, a good selectivity for SiGe epitaxy could be achieved (**Figure 6b**). As shown in **Figure 6c**, further increase of HCl partial pressure to 70 mTorr leads to the case when the etch rate is higher than the growth rate resulting in no depositions on the Si fins (**Figure 6c**).

5. Pattern dependency of selective epitaxy

In order to provide a broad knowledge about the local and global pattern dependency, an example is pointed out here when the SiGe layers were grown selectively on S/D regions in transistors with 22 nm gate length [31]. **Figure 7a** shows a schematic of an 8-inch Si wafer containing 112 processed chips. The chips contain arrays of different transistor sizes (or coverage of exposed Si area varies) as shown in **Figure 7b**. This layout has been repeated for all the chips over the wafer.

In this figure, the blue cross marks illustrate the position of a transistor in the chip where the electrical measurements were done in all 112 chips. According to performance of transistors, at least three groups of chips (A, B, and C) with similar electrical characteristics (poor, good, and excellent) over the wafer were distinguished.

The Ge content in SiGe layers in transistors in A, B, and C groups was estimated to be 38, 40, and 35%, respectively by using energy-dispersive spectroscopy (EDS) technique. The growth

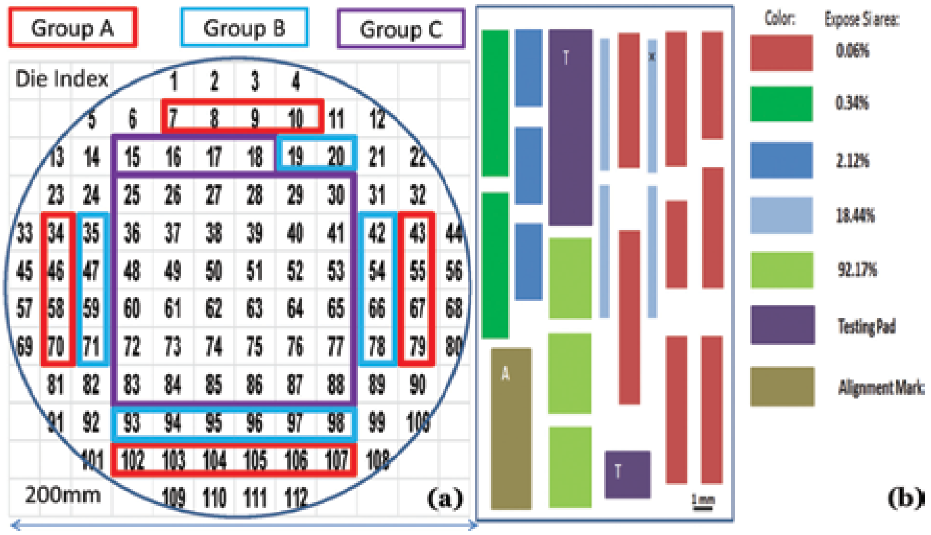


Figure 7. (a) The schematic picture of an 8-inch test wafer with 112 chips where three groups were marked after the performance of transistors. **Figure 7** (b) shows the layout of one chip where the exposed Si areas are nonuniform, and they are illustrated in different colors [31].

rate of SiGe was 0.58, 0.62, and 0.51 nm/s for transistors A, B, and C, respectively. The variation of Ge content and the growth rate affected also the threading dislocation density (TDD) in A, B, and C group transistors where TDD was estimated to be 3 × 10^9, 1 × 10^9, and 1 × 10^8 /cm^2, respectively. The amount of TDD is also related to strain relaxation in the epi-layers.

A more detailed information about the pattern dependency versus layout variation is demonstrated in **Figure 8**. The curves indicate that when the density of transistor arrays decreases, then both Ge content and the growth rate increase. The change of layer profile will directly affect the electrical characteristics of the transistor in the chip.

The transistor characteristics provide the data of V_{sat}, I_{on}, I_{off}, drain-induced barrier lowering (DIBL) and carrier mobility as shown in **Table 1** [31].

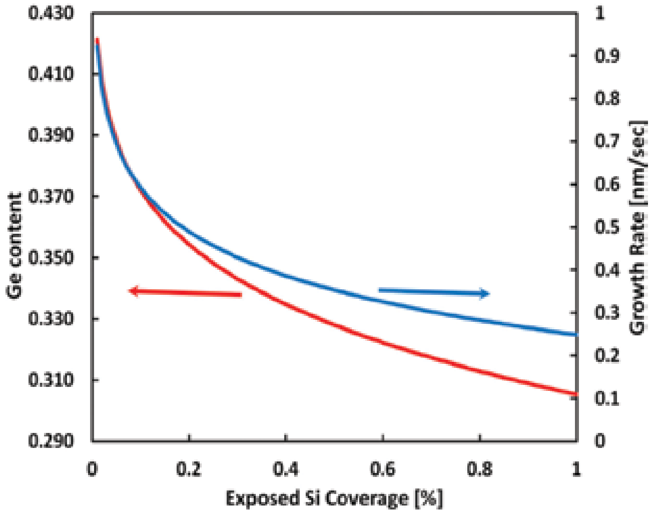


Figure 8. The calculated SiGe profiles for different exposed Si coverages [31].

Transistor group	Chip	Measured Ge content	VTsat (V) V_{DD} = 1 V	I_{on} (μA/μm)	I_{off} (nA/μm)	DIBL (mV)	Mobility (cm^2/Vs)
A	8		−0.46	263	0.34	75.4	24
	9	0.38	−0.54	111	0.24	91.3	13
	10		−0.56	86	0.47	112	9
B	71		−0.39	407	0.82	101	36
	82	**0.40**	−0.39	420	0.86	90	37
	92		−0.39	405	0.65	96	35
C	27		−0.32	598	4.8	115	65
	38	0.35	−0.30	618	9.8	123	71
	50		−0.28	619	10.2	119	75

Table 1. A summary of electrical data for three transistor groups A, B, and C.

It is well known that the presence of misfit dislocations is a sign of strain relaxation in the epilayer. In **Table 1**, transistors in group C have the best electrical performance compared to the others since the Ge content is lowest as well as the defect density. For example, the mobility values are remarkably high for this group.

It is important to emphasize here that the transistor profile is not optimized in all these groups; therefore, the electrical values are not impressive, but only the pattern dependency of epitaxy was in the interest of the discussions.

6. Kinetics of SiGe selective growth

The serious problem with integration of selective epitaxy growth is pattern dependency. The origin of pattern dependency behavior is the nonuniform consumption of reactant gas molecules over the patterned wafer [13–19, 21, 31]. The kinetics of growth is explained by gas molecules move in a laminar flow over the wafer forming boundaries [32].

In CVD, the gas steam through the reactor quartz is under a friction force with the stationary susceptor/substrate, creating stagnant gas boundaries during the gas flow. Gas molecules diffuse/are being attracted downward through the gas boundaries toward the susceptor and finally are consumed on the Si wafer. The length of the attraction force on the gas molecules over the wafer was estimated in the range of 10–15 mm for a total pressure of 20–40 torr in the CVD reactor [33].

Figure 9 illustrates a schematic view of the gas kinetics. In this figure, the black arrows illustrate the direction of molecule movement toward the exposed Si areas (transistor arrays in a chip) of a patterned substrate.

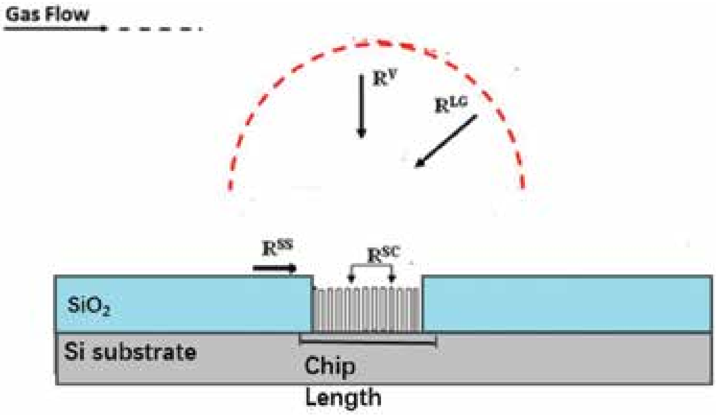


Figure 9. Schematic view of gas flow in different directions over a chip containing planar transistors.

The gas molecules are moved in four possible directions toward the exposed Si areas in a patterned substrate and contribute to the selective epitaxy growth as follows:

a. Vertically in gas phase (R^V)

b. Laterally in gas phase (R^{LG})

c. Surrounding oxide (or nitride) surface between the chips (R^{SS})

d. Oxide surface between the openings within a chip (R^{RC})

The total growth rate (R_{Tot}) can be written as a sum of contributions from incoming molecules from all four directions as follows:

$$R_{Tot} = R_{Si}^{V} + R_{Si}^{LG} + R_{Si}^{SS} + R_{Si}^{SC} + R_{Ge}^{V} + R_{Ge}^{LG} + R_{Ge}^{SS} + R_{Ge}^{SC} - R_{E}^{V} - R_{E}^{LG} - R_{E}^{SS} - R_{E}^{SC} \qquad (1)$$

In Eq. (1), R^{SS} and R^{SB} are changed due to layout variation of the chip, and they are the main components behind the pattern dependency behavior.

Many studies have tried to decrease the pattern dependency by optimizing the growth parameters e.g., applying high HCl partial pressure or low total pressure during epitaxy [34]. The effect of high HCl is to terminate the lateral components in Eq. (1). The success to reduce the pattern dependency by increasing HCl partial pressure is good; however, it is not effective to eliminate the pattern dependency problem. This method demands high HCl partial pressure which leads to very low growth rate as shown in **Figure 10**.

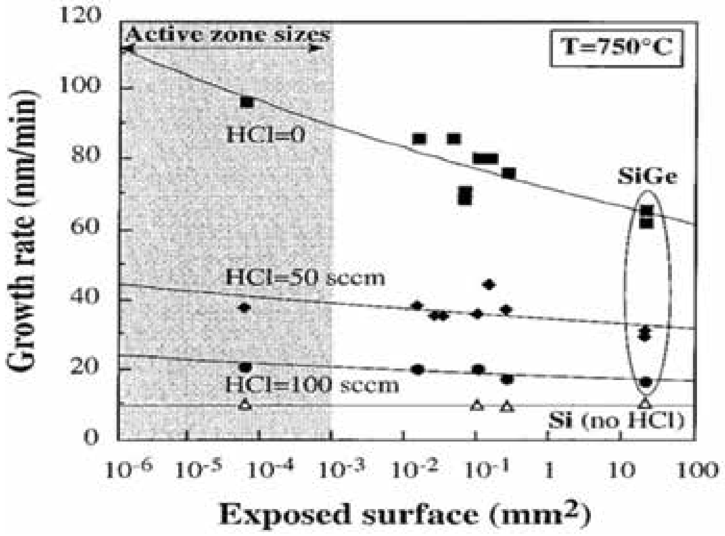


Figure 10. Growth rate of SiGe vs. exposed areas in the chip for the growth at 750°C. The precursor parameters for SiGe growth were the following: (Ge/Si) g = 0.0125, DCS = 200 sccm, and HCl = 0, 50, and 100 sccm [34].

Another method to decrease pattern dependency is proposed to increase the hydrogen carrier gas as well as apply low growth pressure. In this way the number of atoms coming laterally is decreased, but high hydrogen gas demands a better safety of the epi-tool [21].

There is another approach to deal with pattern dependency where an empirical model calculates the layer profile of SiGe and later the layout can be modified for a uniform deposition over the chip or wafer. Such a model has to calculate the components in Eq.(1) where diffusion, adsorption, and desorption of atoms during epitaxy have to considered.

7. Modeling of SiGe selective growth

One of the early works for modeling of Si epitaxy was presented by Meng Tao et al. [35]. The scaffold of the model is based on Maxwell distribution function for the molecules impinging to a surface during epitaxy. In this case, in epitaxy the number of the reactant molecules/unit time (*dn*) with an activation energy in an interval of ($\mathrm{E_A}$, $\mathrm{E_A} + \mathrm{dE_A}$) incoming to a unit area according to Eq. (2):

$$\mathrm{dn} = 8\pi \mathrm{N_m}\left(\frac{1}{2\pi \mathrm{m_m k_b T}}\right)^{\frac{3}{2}} \mathrm{m_m E_A} \exp\left(-\frac{\mathrm{E_A}}{\mathrm{k_b T}}\right)\mathrm{dE_K} \tag{2}$$

where $\mathrm{N_m}$ and $\mathrm{m_m}$ are the number of molecules/unit volume and are the mass of the reactant molecules in the gas. Si growth is the simplest epitaxy where dichlorosilane ($SiCl_2$) is used as precursor. For SiH_2Cl_2 epitaxy, the chemical reactions occur through Cl dissociation and are written as follows:

$$SiH_2Cl_2(g) \rightarrow SiCl_2 + H_2(g) \tag{R1}$$

$$SiCl_2 \rightarrow SiCl + Cl \tag{R2}$$

The growth rate can be calculated by integrating Eq. (2), and it is obtained by:

$$R = \frac{n}{N_0} = \beta \frac{\left(1 - \theta_{H(Si)} - \theta_{Cl(Si)}\right)}{N_0} \frac{P_{SiH_2Cl_2}}{\left(2\pi m_{SiH_2Cl_2} k_b T\right)^{\frac{1}{2}}} \left(\frac{E_{SiH_2Cl_2}}{k_b T} + 1\right) \exp\left(-\frac{E_{SiH_2Cl_2}}{k_b T}\right) \tag{3}$$

where β is a unitless constant and θ, $\mathrm{P_{SiH_2Cl_2}}$, m, $\mathrm{N_0}$, and $\mathrm{E_{SiH_2Cl_2}}$ are the coverage of hydrogen or chlorine on Si surface, partial pressure of DCS, molecular mass of DCS, number of Si atoms in a unit volume of crystal, and activation energy for the growth, respectively.

For selective epitaxy in the presence of HCl, the growth rate is decreased, and it can be:

$$R_T = R_{Si}^V - R_E^V \tag{4}$$

Experimental results show that the etch rate is not linearly dependent on HCl partial pressure ($\mathrm{P_{HCl}}$) parameter and it has a sublinear relationship ($\mathrm{P_{HCl}}^{0.596}$). This behavior could be referred to the fact that all chlorine atoms may not participate during the etch process. Therefore, Eq. (3) is modified as follows:

$$\begin{aligned} R_T = \; & \beta \frac{\left(1 - \theta_{H(Si)} - \theta_{Cl(Si)}\right)}{N_0} \frac{P_{SiH_2Cl_2}}{\left(2\pi m_{SiH_2Cl_2} k_b T\right)^{\frac{1}{2}}} \left(\frac{E_{SiH_2Cl_2}}{k_b T} + 1\right) \exp\left(-\frac{E_{SiH_2Cl_2}}{k_b T}\right) \\ & \times - \frac{\gamma}{N_0} \frac{P_{HCl}^{0.596}}{\left(2\pi m_{HCl} k_b T\right)^{\frac{1}{2}}} \left(\frac{E_{Etching}}{k_b T} + 1\right) \exp\left(-\frac{E_{Etching}}{k_b T}\right) \end{aligned} \tag{5}$$

where γ is a unitless S constant which relates to the HCl molecule distribution in the CVD chamber. The activation energy of etching part is estimated to be 37.5 Kcal/mol. This value lies between 22 and 44 Kcal/mol which is the needed energy to break one or two Si-Si bonds.

For the growth of SiGe layers in the presence of HCl, GeH_4 precursor has been introduced into the CVD chamber. In this case, Eq. (4) should be written as follows:

$$R_T = R_{Si}^V + R_{Ge}^V - R_E^V \tag{6}$$

The SiGe epitaxy is significantly different than Si epitaxy, since the presence of Ge atoms increases the growth rate.

The growth rate is increased due to two reasons; at first, the activation energy for SiGe deposition is lowered when Ge is added during epitaxy. The activation energy for Ge is 0.61 eV compared to 2.08 eV for Si, and for SiGe, this value should lie between Si and Ge ones. At second, the presence of Ge increases the available sites (or dangling bond sites) on the Si surface owing to increase of desorption energy of hydrogen and chlorine from Si surface. In this case, Si atoms can easily find the available sites and bind to the lattice, and therefore the growth rate is increased. The above reasons make the SiGe growth more complicated than Si, and therefore Eq. (6) cannot simply be used for SiGe epitaxy. In this case, a coefficient "m" is implemented, and the revised equation for SiGe growth in the presence of HCl is given by:

$$R_T = R_{Si/Si}^V + R_{Ge/Si}^V + mR_{Ge/Si}^V - R_E^V \tag{7}$$

The full form of equation for the SiGe growth rate can be obtained from:

$$\begin{aligned} R_T = \; & \beta \frac{\left(1 - \theta_{H(Si)} - \theta_{Cl(Si)}\right)}{N_0} \frac{P_{GeH_4}}{\left(2\pi m_{SiH_2Cl_2} k_b T\right)^{\frac{1}{2}}} \left(\frac{E_{SiH_2Cl_2 on\ Si}}{k_b T} + 1\right) \exp\left(-\frac{E_{SiH_2Cl_2\ on\ Si}}{k_b T}\right) \\ & + \chi \frac{(1+m)\left(1 - \theta_{H(Si)} - \theta_{Cl(Si)}\right)}{N_0} \frac{P_{GeH_4}}{\left(2\pi m_{GeH_4} k_b T\right)^{\frac{1}{2}}} \left(\frac{E_{GeH_4 on\ Si}}{k_b T} + 1\right) \exp\left(-\frac{E_{GeH_4 on\ Si}}{k_b T}\right) \\ & \times - \frac{\gamma}{N_0} \frac{P_{HCl}^{0.596}}{\left(2\pi m_{HCl} k_b T\right)^{\frac{1}{2}}} \left(\frac{E_{Etching}}{k_b T} + 1\right) \exp\left(-\frac{E_{Etching}}{k_b T}\right) \end{aligned} \tag{8}$$

where χ is a constant which depends on the gas property. In above equation, the m coefficient is estimated to be 2 for growth temperatures 600–725°C [36].

A series of input parameters, e.g., Ge, Si, and HCl partial pressures, can be inserted in Eq. (8), and the etch rates during SiGe epitaxy can be extracted [37]. The experimental data show Arrhenius curves, and the activation energy can be obtained from the slope of these curves. The results show that the activation energy is decreased with increasing Ge partial pressures as shown in **Figure 11**. This outcome could be predicted since the strength of atomic bond in Si matrix becomes weaker with increasing Ge content (or strain).

The dependence of activation energy to Ge partial pressure is expressed as:

$$E_{\mathrm{Etching}} = E_{\mathrm{a,\,Etching(Si)}} e^{-12.535 P_{\mathrm{GeH_4}}} \tag{9}$$

where $E_{\mathrm{a,Etching(Si)}}$ is the activation energy for etch of Si. It is worth mentioning here that the activation energy in **Figure 11** differs from the previously reported values to etch SiGe bulk materials [38]. This difference in activation energies can be explained by the fact that the energies to etch SiGe in bulk form and during SiGe epitaxy are entirely different processes.

The Ge content, x in $Si_{1\text{-}x}Ge_x$ layers, can be obtained from Ge and Si partial pressures using the following [16]:

$$\frac{x^2}{1-x} = \sigma\left(\frac{P_{GeH_4} - (1-\eta)P_{HCl}}{P_{SiH_2Cl_2} - \eta P_{HCl}}\right) \tag{10}$$

where σ is a constant which links to chemical reactions in CVD reactor and η is a reaction rate coefficient which lies in a range between 0.9 and 1 depending on HCl partial pressure. The experimental data demonstrate that η is 1 when HCl partial pressure is lower than DCS partial pressure; otherwise it is ~0.9 for higher HCl pressures. The constant σ is related to adsorption and desorption of the main species during CVD, and it is written in the following equation below:

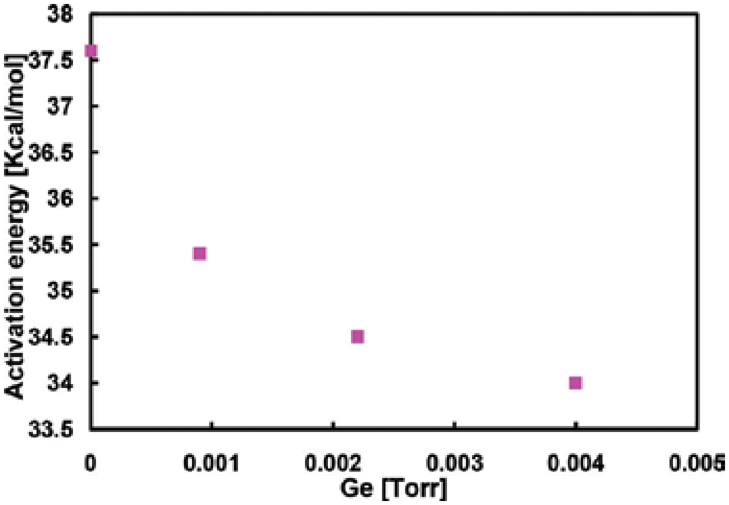


Figure 11. Activation energy vs. Ge partial pressures for etch part during SiGe epitaxy.

$$\sigma = \frac{k_{a,\,GeH_2} \times k_{d,\,H}}{k_{a,\,SiCl_2} \times k_{d,\,Cl}} = Aexp\left(\frac{E}{kT}\right) \tag{11}$$

The adsorption energy difference ($E_{a,SiCl2}-E_{a,GeH2}$) is ~0.1 eV [39], and the desorption energy difference is ~0.48 eV [40]. Then, the total activation energy is valued to 0.58 eV which is close to the derived energy value (0.697 eV) [37].

Until now, the vertical components in Eq. (1) have been discussed and calculated. The lateral components can be derived in the same way when these parameters are used for the patterned substrates [37]. A few assumptions have to be considered in order to make the calculations easier. At first, it is defined that a wafer has a global pattern when the chip layout is uniform and it is repeated over the entire wafer.

At second, the HCl partial pressure has to be enough to ensure the selectivity of the growth.

The results from Si deposition on patterned substrate have demonstrated that the growth rate is decreased when the coverage of the exposed Si areas becomes smaller. **Figure 12** demonstrates the growth rate from five globally patterned wafers with different exposed areas. The figure confirms the behavior of Si epitaxy is different than SiGe epitaxy [41, 42].

This discrepancy is due to the presence of R_{HCl}^{SC} in Eq. (1), which is formed through the lateral diffusion of Cl atoms on the oxide surface in a fully selective mode. The contribution of this component was found to be inversely related to the exposed Si areas:

$$P_{HCl}^{SC} = AP_{HCl}\ln\left(\frac{1}{c}\right) \tag{12}$$

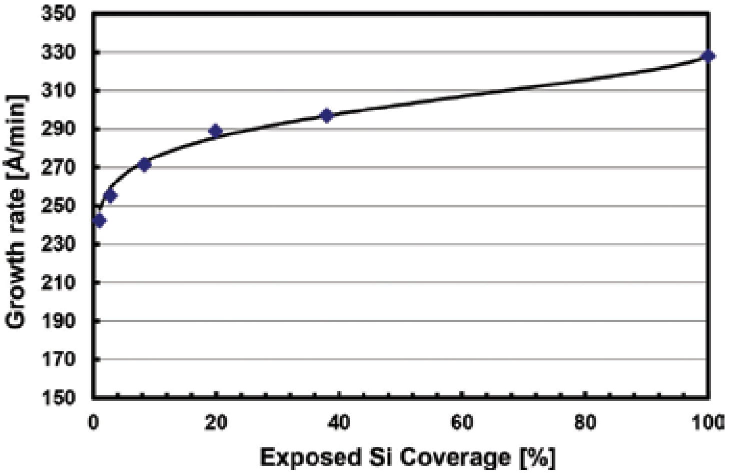


Figure 12. Growth rate vs. coverage of exposed Si for five different globally patterned wafers in total pressure of 20 torr. The P_{DCS} and P_{HCl} were 120 and 20 mTorr, respectively [37].

where c stands for the coverage of exposed Si areas of the chip and A is a parameter that depends on which type of mask is used for isolation material. Thus, Eq. (15) can be reformulated as follows:

$$\begin{aligned} R_T = \; & \beta\frac{\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{N_0}\frac{P_{SiH_2Cl_2}}{(2\pi m_{SiH_2Cl_2}k_bT)^{\frac{1}{2}}}\left(\frac{E_{SiH_2Cl_2 on\ Si}}{k_bT}+1\right)\exp\left(-\frac{E_{SiH_2Cl_2 on\ Si}}{k_bT}\right) \\ & \times-\frac{\gamma}{N_0}\frac{P_{HCl}^{0.596}}{(2\pi m_{HCl}k_bT)^{\frac{1}{2}}}\left(\frac{E_{Etching}}{k_bT}+1\right)\exp\left(-\frac{E_{Etching}}{k_bT}\right)-\frac{\gamma}{N_0}\frac{\left(AP_{HCl}\ln\left({}^{1}\!/_{c}\right)\right)_{HCl}^{0.596}}{(2\pi m_{HCl}k_bT)^{\frac{1}{2}}}\left(\frac{E_{Etching}}{k_bT}+1\right) \\ & \times\exp\left(-\frac{E_{Etching}}{k_bT}\right) \end{aligned} \tag{13}$$

For SiGe epitaxy, GeH$_4$ precursor is introduced to the reactant gases which increases the Cl desorption, and therefore, the lateral diffusion of Cl becomes minor [43].

In a similar way, the later component for Ge atoms on the oxide surface $\mathrm{P}_{\mathrm{Ge}}^{\mathrm{SC}}$ can be written as:

$$P_{Ge}^{SC}=BP_{GeH_4}\ln\left(\frac{1}{c}\right) \tag{14}$$

where B is a parameter similar to Eq. (15) which is dependent on the mask material and c is the exposed Si coverage of the chip. Due to the lateral diffusion of Ge atoms on the oxide surface, an activation energy of 0.1 eV is added to the activation energy of the growth. In this case, the total growth rate for SiGe epitaxy is written as:

$$\begin{aligned} R_T = & \beta\frac{\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{N_0}\frac{P_{GeH_4}}{(2\pi m_{SiH_2Cl_2}k_bT)^{\frac{1}{2}}}\left(\frac{E_{SiH_2Cl_2 on\ Si}}{k_bT}+1\right)\exp\left(-\frac{E_{SiH_2Cl_2\ on\ Si}}{k_bT}\right) \\ & +\chi\frac{(1+m)\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{N_0}\frac{P_{GeH_4}}{(2\pi m_{GeH_4}k_bT)^{\frac{1}{2}}}\left(\frac{E_{GeH_4 on\ Si}}{k_bT}+1\right)\exp\left(-\frac{E_{GeH_4 on\ Si}}{k_bT}\right) \\ & +\chi\frac{(1+m)\left(1-\theta_{H(Si)}-\theta_{Cl(Si)}\right)}{N_0}\frac{\left(BP_{GeH_4}\ln\left({}^{1}\!/_{c}\right)\right)}{(2\pi m_{GeH_4}k_bT)^{\frac{1}{2}}}\left(\frac{E_{GeH_4 on\ Si}+0.1eV}{k_bT}+1\right)\exp\left(-\frac{E_{GeH_4 on\ Si}+0.1eV}{k_bT}\right) \\ & -\frac{\gamma}{N_0}\frac{P_{HCl}^{0.596}}{(2\pi m_{HCl}k_bT)^{\frac{1}{2}}}\left(\frac{E_{Etching}}{k_bT}+1\right)\exp\left(-\frac{E_{Etching}}{k_bT}\right) \end{aligned} \tag{15}$$

The lateral contribution of Ge on the oxide surface has to be considered in the composition in Eq. (12) as well, and therefore the equation is modified as:

$$\frac{x^2}{1-x}=\sigma exp\left(\frac{0.7eV}{k_bT}\right)\left(\frac{P_{GeH_4}+\left(BP_{GeH_4}\ln\left({}^{1}\!/_{c}\right)\right)-(1-\eta)P_{HCl}}{P_{SiH_2Cl_2}-\eta P_{HCl}}\right) \tag{16}$$

Figure 13a and **b** show the experimental and calculated outcomes for the SiGe selective epitaxy. A good agreement between the experimental data and the calculated one is observed for patterned wafers.

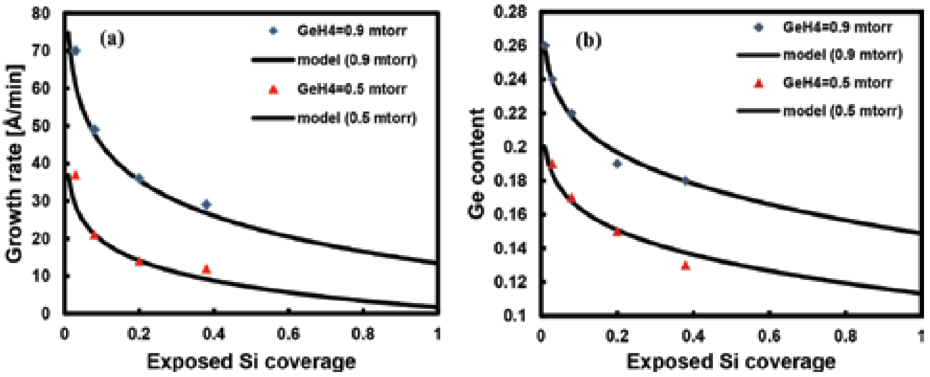


Figure 13. (a) Growth rate vs. chip exposed Si coverage and (b) Ge content vs. chip exposed Si coverage for SiGe layers grown at 20 torr on wafers with different global patterns. The applied $P_{SiH2Cl2}$ and P_{HCl} were 60 and 20 mTorr, respectively [37].

Until now, all calculations were for chips in wafers with global layout which in fact is an ideal case; however, for many cases, the layout of chips is nonuniform. When the layout varies the gas consumption over, the chip (wafer) becomes nonuniform. The part of chip with largest exposed area would attract stronger the surrounding atoms toward itself compared to the other part of the chip. This means that there is an interaction between different areas in the chip where the growth rate at part of the chip with highly exposed Si area (R_{Trap}) has an influence in neighboring parts (R_{surr}). The interaction range between two individual parts is denoted "τ(c)" where parameter "c" stands for exposed Si coverage in the chip. Then the growth rate of any part of chip which is located at a distance d (R(d)) can be written according to this interaction theory as follows [19]:

$$R(d) = R_{Trap} + \left(R_{Surr} - R_{Trap}\right)\left(1 - \exp\left(\frac{-d}{\tau(c)}\right)\right) \tag{17}$$

Parameter "τ" depends on the exposed coverage of the chip since the dangling bonds are creating an attraction force which drags the gas molecules.

Therefore, the gas consumption has to be uniformly over the chip in order for growth rate to be uniform. In this case, the trap parts have to be distributed over the chip and not isolated. As an example, **Figure 14** shows a nonuniform chip with six regions with exposed Si coverage areas of 0, 1, 3, 8, and 10%. At first, the trap regions are identified as 8 and %10 in this chip since the coverage is highest. Later, the interaction length between the six areas has to be calculated mutually. The condition for uniform gas consumption over this chip can be achieved either by inserting dummy features or by subdividing the trap areas over the chip or both methods.

The modeling of SiGe selective growth for advanced chip layout inaugurates a new window for chipmakers to deposit epi-layers with high quality and high uniformity over the wafer.

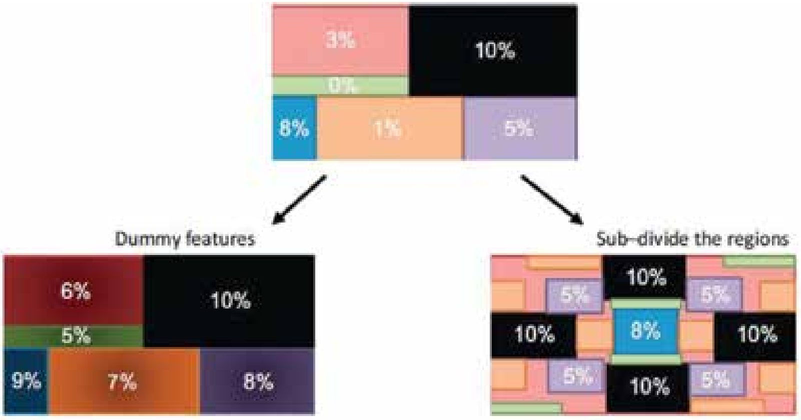


Figure 14. Design of chip layout to obtain uniform SiGe deposition.

8. Strain mechanism in group IV materials

Strain is a mechanical deformation which is resulted when a crystal with a lattice mismatch is epitaxially deposited on a substrate. The crystal of the deposited material has to align to the substrate, and as a result, the crystal is deformed. Strain is categorized in two types depending on whether the direction of the applied force is inward (compressive strain) or outward (tensile strain). Therefore, the strain is a hidden energy in the crystal which affects the electrical, mechanical, and optical properties of the semiconductor. Compressive strain, which is generated by SiGe layers by using selective epitaxy, has been the core discussion in this book chapter. Compressive strain is applied in pMOS to increase the hole mobility in the channel.

In general, the mobility is defined as:

$$\mu = \frac{q < \tau >}{m^*} \tag{18}$$

where m* stands for the effective mass and τ expresses the scattering time for the carriers [11].

The compressive strain splits the heavy and light hole (HH and LH) bands and changes the curvature of these bands. The latter effect is directly related to the decrease of effective mass for holes, whereas the first effect decreases the holes' scattering between the HH and LH bands. Both these effects have direct impact on $< \tau >$ and m* [11].

Other ways to describe the transport properties in the channel of transistor in the presence of compressive strain piezoresistance coefficients are commonly calculated. These coefficients are expressed in respect to mobility's fractional variations as:

$$\Delta\mu/\mu \approx |\pi//\sigma// + \pi_\perp\sigma_\perp| \tag{19}$$

where $\sigma_\perp$ and $\sigma_{//}$ are the transverse and longitudinal stresses and $\pi_{//}$ and $\pi_\perp$ denote for the piezoresistance coefficients in longitudinal and transverse directions. The piezoresistance coefficients can also be written in form of the three fundamental coefficients $\pi 11$, $\pi 12$, and $\pi 44$ [44].

The recent results have demonstrated that compressive strain along <110> has highest piezoresistance coefficients for both (001) and (110) wafers. As a result, <110> channel direction has been mainly applied for industrial applications [45, 46].

Author details

Guilei Wang[1,2]*, Henry H. Radamson[1,2,3] and Mohammadreza Kolahdouz[4]

*Address all correspondence to: wangguilei@ime.ac.cn

1 Key Laboratory of Microelectronics Devices and Integrated Technology, Institute of Microelectronics, Chinese Academy of Sciences, Beijing, People's Republic of China

2 University of Chinese Academy of Sciences, Beijing, People's Republic of China

3 KTH Royal Institute of Technology, Stockholm, Sweden

4 Thin Film Laboratory, Electrical and Computer Engineering Department, University of Tehran, Tehran, Iran

References

[1] Vescan L, Grimm K, Dieker C. Facet investigation in selective epitaxial growth of Si and SiGe on (001) Si for optoelectronic devices. Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena. 1998;**16**(3):1549-1554

[2] Dutartre D, Talbot A, Loubet N. Facet propagation in Si and SiGe epitaxy or etching. ECS Transactions. 2006;**3**(7):473

[3] Drowley CI, Reid GA, Hull R. Model for facet and sidewall defect formation during selective epitaxial growth of (001) silicon. Applied Physics Letters. 1988;**52**:546

[4] Aoyama T, Ikarashi T, Miyanaga K, Tatsumi T. Facet formation mechanism of silicon selective epitaxial layer by Si ultrahigh vacuum chemical vapor deposition Sf02. Journal of Crystal Growth. 1994;**136**:349-354

[5] Kawaguchi K, Usami N, Shiraki Y. Formation of relaxed SiGe® lms on Si by selective epitaxial growth. Thin Solid Films. 2000;**369**:126-129

[6] Ohtsuka M, Suzuki A. Modeling of molecular-beam epitaxy and metalorganic vapor-phase epitaxy on nonplanar surfaces. Journal of Applied Physics. 1993;**73**:7358

[7] Xiang Q, Li S, Wang D, Wang KL, Couillard JG. Interfacet mass transport and facet evolution in selective epitaxial growth of Si by gas source molecular beam epitaxy. Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures. 1996;**14**:2381

[8] Gannavaram S, Pesovic N, Ozturk MC. Low Temperature (<8OOC) Recessed Junction Selective Silicon-Germanium SourceDrain Technology for sub-70 nm CMOS. Electron Devices Meeting. IEDM Technical Digest. IEEE International 2000. p. 437; 2000

[9] Ghani T, Armstrong M, Auth C, Bost M, Charvat P, Glass G, Hoffiann T, Johnson K, Kenyon C, Klaus J, Mclntyre B, Mistry K, Murthy A, Silberstein M, Sivakumar S, Smith P, Zawadzki K, Thompson S, Bohr M. A 90 nm High Volume Manufacturing Logic Technology Featuring Novel 45 nm Gate Length Strained Silicon CMOS Transistors. Electron Devices Meeting, 2003. IEDM Technical Digest. IEEE International; 2003. pp. 978-980

[10] Packan P, Al E. High Performance 32 nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistor. Electron Devices Meeting. (IEDM). 2009, p. 1

[11] Radamson HH, Thylen L. Monolithic Nanoscale Photonics-Electronics Integration in Silicon and Other Group 1V Elements; September 1, 2014. Elsevier Science & Technology. San Diego, CA, USA: Elsevier

[12] Radamson HH, Zhang Y, He X, Cui H, Li J, Xiang J, Liu J, Gu S, Wang G. The challenges of advanced CMOS process from 2D to 3D. Applied Sciences. 2017;**7**(10):1047

[13] Hallstedt J, Kolahdouz M, Ghandi R, Radamson H, Wise R. Pattern dependency in selective epitaxy of B-doped SiGe layers for advanced metal oxide semiconductor field effect transistors. Journal of Applied Physics. 2008;**103**:0549071-0549077

[14] Fellous C, Romagna F, Dutartre D. Thermal and chemical loading effects in non selective Si/SiGe epitaxy. Materials Science and Engineering B. 2002;**89**:323-327

[15] Kolahdouz M, Hallstedt J, Khatibi A, Ostling M, Wise R, Riley DJ, Radamson H. Comprehensive evaluation and study of pattern dependency behavior in selective epitaxial growth of B-doped SiGe layers. IEEE Transactions on Nanotechnology. 2009;**8**:291-297

[16] Kolahdouz M, Maresca L, Ostling M, Riley D, Wise R, Radamson H. New method to calibrate the pattern dependency of selective epitaxy of SiGe layers. Solid-State Electronics. 2009;**53**:858-861

[17] Hartmann JM, Abbadie A, Vinet M, Clavelier L, Holliger P, Lafond D, Semeria MN, Gentile P. Growth kinetics of Si on fullsheet, patterned and silicon-on-insulator substrates. Journal of Crystal Growth. 2003;**257**:19-30

[18] Loo R, Wang G, Souriau L, Lin JC, Takeuchi S, Brammertz G, Caymax M. Epitaxial Ge on standard STI patterned Si wafers: High quality virtual substrates for Ge pMOS and III/V nMOS. ECS Transactions. 2009;**25**(7):335-350

[19] Kolahdouz M, Ghandi R, Hallstedt J, Osling M, Wise R, Wejtmans H, Radamson H. The influence of Si coverage in a chip on layer profile of selectively grown Si1−xGex layers using RPCVD technique. Thin Solid Films. 2008;**517**:257-258

[20] Ghandi R, Kolahdouz M, Hallstedt J, Wise R, Wejtmans H, Radamson H. Effect of strain, substrate surface and growth rate on B-doping in selectively grown SiGe layers. Thin Solid Films. 2008;**517**:334-336

[21] Loo R, Caymax M. Avoiding loading effects and facet growth key parameters for a successful implementation of selective epitaxial SiGe deposition for HBT-BiCMOS and high-mobility hetero-channel pMOS devices. Applied Surface Science. 2004;**224**:24-30

[22] Yakimets D, Eneman G, Schuddinck P, Bao TH, Bardon MG, Raghavan P, Veloso A, Collaert N, Mercha A, Verkest D, et al. Vertical GAAFETs for the ultimate CMOS scaling. IEEE Transactions on Electron Devices. 2015;**62**:1433-1439

[23] Radamson HH, Kolahdouz M. Selective epitaxy growth of Si 1−x Ge x, layers for MOSFETs and FinFETs[J]. Journal of Materials Science Materials in Electronics. 2015; **26**(7):4584-4603

[24] Bühler J, Steiner F-P, Baltes H. Silicon dioxide sacrificial layer etching in surface micromachining. Journal of Micromechanics and Micro Engineering. 1997;**7**(1):R1-R13

[25] Wang G, Luo J, Qin C, Liang R, Xu Y, Liu J, et al. Integration of highly strained SiGe in source and drain with HK and MG for 22 nm bulk PMOS transistors. Nanoscale Research Letters. 2017;**12**(1):123

[26] Abbadie A, Hartmann J-M, Holliger P, Semeria MN, Besson P, Gentile P. Low thermal budget surface preparation of Si and SiGe. Applied Surface Science. 2004;**225**(1–4):256-266

[27] Wang G, Moeen M, Abedin A, Kolahdouz M, Luo J, Qin CL, Zhu HL, Yan J, Yin H, Li J, et al. Optimization of SiGe selective epitaxy for source/drain engineering in 22 nm node complementary metal-oxide semiconductor (CMOS). Journal of Applied Physics. 2013; **114**:123511

[28] Wang G, Abedin A, Moeen M, Kolahdouz M, Luo J, Guo Y, et al. Integration of highly-strained sige materials in 14 nm and beyond nodes finfet technology. Solid State Electronics. 2015;**103**:222-228

[29] Wang G, Qin C, Yin H, Luo J, Duan N, Yang P, et al. Study of SiGe selective epitaxial process integration with high-k and metal gate for 16/14 nm nodes FinFET technology. Microelectronic Engineering. 2016;**163**:49-54

[30] Radamson HH, Luo J, Qin CL, Yin H, Zhu HL, Zhao C, et al. Optimization of selective growth of sige for source/drain in 14nm and beyond nodes FinFETs. International Journal of High Speed Electronics & Systems. 2017;**26**(01n02):1740003

[31] Wang G, Moeen M, Abedin A, et al. Impact of pattern dependency of SiGe layers grown selectively in source/drain on the performance of 22 nm node pMOSFETs. Solid-State Electronics. 2015;**114**:43-48

[32] Knutson KL, Carr RW, Liu WH, Campbell SA. A kinetics and transport model of dichiorosilane chemical vapor deposition. Journal of Crystal Growth. 1994;**140**:191-204

[33] Newman AJ, Krishnaprasad PS, Ponczak S, Brabant P. Modeling and Model Reduction for Control and Optimization of Epitaxial Growth in a Commercial Rapid Thermal Chemical Vapor Deposition Reactor. Technical Report 98–45. Institute for Systems Research; 1998

[34] Bodnar S, de Berranger E, Bouillon P, Mouis M, Skotnicki T, Regolini JL. Selective Si and SiGe epitaxial heterostructures grown using an industrial low-pressure chemical vapor deposition module. Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures. 1997;**15**:712

[35] Tao M. Growth kinetics and reaction mechanism of silicon chemical vapour deposition from silane. Thin Solid Films. 1993;**223**:201-211

[36] Kühne H. On a substituting, sticking and trapping model of CVD $Si_{1-x}Ge_x$ layer growth. Journal of Crystal Growth. 1992;**125**:291-300

[37] Kolahdouz M, Maresca L, Ghandi R, Khatibi A, Radamson HH. Kinetic model of SiGe selective epitaxial growth using RPCVD technique. Journal of the Electrochemical Society. 2011;**158**:H457

[38] Bogumilowicz Y, Hartmann JM, Truche R, Campidelli Y, Rolland G, Billon T. Chemical vapour etching of Si, SiGe and Ge with HCl; applications to the formation of thin relaxed SiGe buffers and to the revelation of threading dislocations. Semiconductor Science and Technology. 2005;**20**:127-134

[39] Ito S, Nakamura T, Nishikawa S. Reduced-pressure chemical vapor deposition. Applied Physics Letters. 1996;**69**:1098-1100

[40] Hierlemann M, Werner C, Spitzer A. Equipment simulation of SiGe heteroepitaxy: Model validation by ab initio calculations of surface diffusion processes. Journal of Vacuum Science B. 1997;**15**:935

[41] Radamson H, Kolahdouz M, Ghandi R, Hallstedt J. Selective epitaxial growth of B-doped SiGe and HCl etch of Si for the formation of SiGe:B recessed source and drain (pMOS transistors). Thin Solid Films. 2008;**517**:84-86

[42] Kolahdouz M, Adibi PTZ, Farniya AA, Shayestehaminzadeh S, Trybom E, Di Benedetto L, Radamson H. Selective growth of B- and C-doped SiGe layers in unprocessed and recessed Si openings for p-type metal-oxide-semiconductor field-effect transistors application. Journal of The Electrochemical Society. 2010;**157**:H633

[43] Suh KY, Lee HH. Ge composition in $Si_{1-x}Ge_x$ films grown from SiH_2Cl_2/GeH_4 precursors. Journal of Applied Physics. 2000;**88**:4044-4047

[44] Smith CS. Piezoresistance effect in germanium and silicon. Physical Review. 1954;**94**(1):42

[45] Giles MD, Armstrong M, Auth C, Cea SM, Ghani T, Hoffman T, et al. Understanding Stress Enhanced Performance in Intel 90 nm CMOS Technology[C]//VLSI Technology, 2004. Digest of Technical Papers; 2004. pp. 118-119

[46] Ghani T, Thompson SE, Bohr M, et al. A 90 nm high volume manufacturing logic technology featuring novel 45 nm gate length strained silicon CMOS transistors. In: IEDM Technical Digest; San Francisco, 2003. pp. 11.6.1-11.6.3

Chapter 5

MOS Meets NEMS: The Born of Hybrid Devices

Mario Alberto García-Ramírez,
Miguel Angel Bello-Jiménez,
María Esther Macías-Rodríguez, Barbara Cortese,
José Trinidad Guillen-Bonilla,
Rosa Elvia López-Estopier,
Juan Carlos Gutiérrez-García and
Everardo Vargas-Rodríguez

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.78758

Abstract

Nowadays, the semiconductor industry is reaching an impasse due to the scaling-down process according to Moore's Law, initiated back in 1960s, for the Metal-Oxide-Technology in use. To overcome such issue, the semiconductor industry started to foresee novel materials that allow the development of nanodevices with a broad variety of characteristics such as high switching speed, low power consumption, robust, among others; that can overcome the inherent issues for Silicon. A few "exotic materials" appear such as Graphene, MoS_2, BN-h, among others. However, the time for the novel technology to be mature is a few decades in the future. To allow the "exotic materials" to mature, the semiconductor industry requires of novel nano-structures that can overcome a few of the issues that Silicon-based technology is facing today. A key alternative is based on hybrid structures. Hybrid structures encompass two dissimilar technologies nano-electromechanical systems with the well known Metal-Oxide-Technology. The hybrid nano-structure provides a broad variety of options to be used in such as transistors, memories and sensors. These hybrid devices can give enough time for the technology based on "exotic materials" to be reliable as Silicon based is.

Keywords: hybrid devices, MEMS/NEMS, MOS technology, nano-electronics, exotic materials, bio-applications, aerospace, military applications

1. Introduction

The semiconductor industry has been paved the way for the development of science and technology for more than half-century. Within this time, the development of different sciences such as medicine, biology, archaeology, law, among others has been benefited by the semiconductor industry through the materials as well as the electronics devices/systems developed. By continuing allowing the development of such devices/systems, the semiconductor industry based whole scientific and technologic development in Moore's Law, established back in the 1960s [1]. By following it, nowadays it is possible to get high power processing computing at low cost, high definition graphics for portable video games that consider low power consumption as well as lightweight.

This romantic trend continued for a long time until the scaling-down process due to the Moore's Law became an impasse. According to the International Technology Roadmap for Semiconductors (ITRS) [2], within the "in-use" Metal-Oxide-Semiconductor technology (MOS), by continuing with the scaling-down trend, the tunnel oxide layer cannot be thinner than 7 nm in order to avoid a leakage issue. The issue should be appropriately addressed by migrating the silicon-based technology to other materials that can cope with the requirements that the semiconductor industry needs. It was found that a few materials match the requirements such as chalcogenides, Graphene, Diamond, CNTs, MoS_2, BN-h, among others [3–8]. However, this novel technology takes time to be developed as well as to be mature enough to be reliable. In the meantime, it is needed to fill the gap in time and technology by considering some novel structures that can be used to overcome the inherent MOS issues until the novel technology is available. An option that came across is the use of hybrid devices that encompasses the well-known MOS technology with a nanoelectromechanical systems (NEMS).

To understand the use of such technologies, this chapter is divided into a brief introduction, in Section 2 a review of MOS technology is to understand how does it work. Section 3 shows the characteristics of the nanoelectromechanical systems. Section 4 the hybrid structures are introduced as a merge between the two technologies giving several examples of the reliability of the hybrid structures. In Section 6, several examples that can be implemented for key industrial applications are exposed and finally, a resume of the hybrid devices and the importance of them for the semiconductor industry.

2. Metal oxide semiconductor technology

Metal oxide semiconductor (MOS) technology has been used to develop a wide variety of devices ranging from memories, sensors, clocks up to quite complicated systems such as mobile phones, personal computers, satellites or even fridges. The main aim to develop all those systems as well as the science and technology that made them possible are based on a statement made a few years back in the 1960s by Gordon Moore that is known as Moore's Law [1]. According to Moore's law, the number of devices fabricated should be twice the previous number every 24 months over the same area. By following such law, the semiconductor

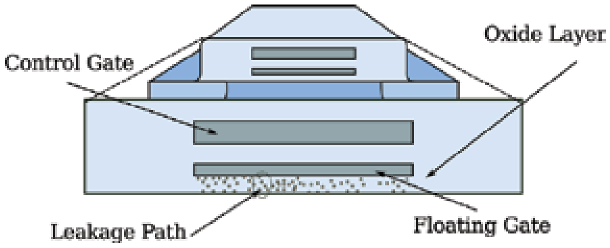


Figure 1. Schematic diagram of a metal-oxide-semiconductor transistor featuring the leakage path due to the scaling-down process.

industry has been capable of delivering, with small variations, this trend for a half century. However, by continuing this trend, MOS technology is reaching an impasse produced by the scaling-down process due to the tunnel oxide layer within the floating gate structure as depicted in **Figure 1**.

According to the International Technology Roadmap for Semiconductors (ITRS) [2], the tunnel oxide layer cannot be thinner than 7 nm due to leaking issues towards the substrate or to the control gate. To overcome such issue, it is required to foreseen for novel materials beyond Si or Ge. This is why, from 2007 the semiconductor industry started to search for materials that can fulfil critical requirements to develop novel devices with improved capabilities such as low-power consumption, high switching speed, scalable capabilities, robust, multi-functional, co-integration capabilities similar as for Si-based devices, among others.

As a result, several materials that present those characteristics such as Graphene, MoS_2, Diamond, BN-h were found [9–13]. Above mentioned materials are capable to deliver the requirements that the semiconductor industry desperately needs. However, there is a time that the "exotic materials" need to mature to be robust enough to feed the market with novel devices [8, 12, 14]. To continue feeding the market, the semiconductor industry requires using the whole set of tools and technology developed over a half-century to give enough time to improve the devices/technology for the emerging technologies.

A key technology that can allow to the semiconductor industry to give enough time to mature the emerging technologies is based on Micro/nanoelectromechanical systems (MEMS/NEMS) that can be co-integrated with the well-known MOS technology. The co-integration between those unique technologies can allow a broad variety of novel devices with the capabilities of robustness as well as maturity and improved characteristics as similar Si-based devices.

3. MEMS/NEMS

The micro-electromechanical systems or micro systems technology is a technology developed in the early 1980s. This technology appears as a result of a lecture given in Caltech 1959 by Prof. Richard Feynman "There's Plenty of Room at the Bottom" [15]. The lecture re-shape the Si-based semiconductor industry to foreseen novel applications for micromachines. MEMS technology encompasses a series of materials that interact with the media allowing to move some parts within it to detect or to have a response according to the media while biased. Typical characteristics for this technology feature components between 1 to 100 μm, a complete system can range from a few tens of microns up to 1 mm. The first MEM fabricated was a large mirror array that was capable to move while bias each axes. The fabrication process developed for such technology was in early stages. From this point, novel processes were proposed and mastered to remove key layers known as a sacrificial layer to free layers within the device or to create holes for particular purposes as well as to deliver a smooth material deposition to accurately shape the features required for the proper operation of the NEM under fabrication. A process needed to be standardised is based on the etching processes that encompasses both wet etching (KOH, TMAH, FNA, …) and dry etching (RIE, DRIE, … [16–18]).

At this point, MEMS technology needed to scale-down by following the Moore's law in order to become relevant for the semiconductor market. By improving the Si-based fabrication processes as well as the etching and lithography processes, the micro-electromechanical systems became the nanoelectromechanical systems (NEMS). NEMS feature a working range from few up to hundreds of nanometres, ultra-low power consumption, reliable, scalable, robust, among others. By considering that NEMS has been scaled-down by following Moore's law, as semiconductor industry states, it is possible to co-integrate them by the well-known MOS technology due to both consider the same substrate and can be merged within the same die (Si-based).

By developing the proper fabrication processes and due to the feature size for MOS as well as for NEMS devices, the co-integration for both dissimilar technologies it is now possible. As the development for such hybrid structures are in its early stages and there is no specialised software to analyse the nanostructure but software based on physic properties that encompassed a wide variety of scientific branches. It is a drawback that is being overcome by performing an algebraic analysis of the NEM structure coupled to the MOS device. This method has been widely used delivering accurate results as measurements can confirm [19, 20].

To deliver the set of hybrid devices that semiconductor industry requires to flow the market, we require a set of novel devices that can be co-integrated within the same die to reduce fabrication cost, improve reliability as well as to overcome previous drawbacks inherent to MOS technology. A few of the nanodevices already fabricated are based on simple structures such as single/double clamp beams, membranes and pillars [20]. Furthermore, one of the main drawbacks from MOS technology, scaling-down feature, has been successfully overcome as exposed elsewhere [21, 22].

4. NEMS-MOS: hybrid devices

As state-of-the-art devices, in this section we are going to analyse the most common nanostructures that encompass NEMS with a MOS technology such as transistors, sensors, nonvolatile memories and high-Q resonators, to name a few.

4.1. Hybrid transistor

In general, the MOS transistor works by biasing the source and drain, in order to generate the full channel, it is required a signal from the gate to close it up and connect both terminals as shown in **Figure 2**.

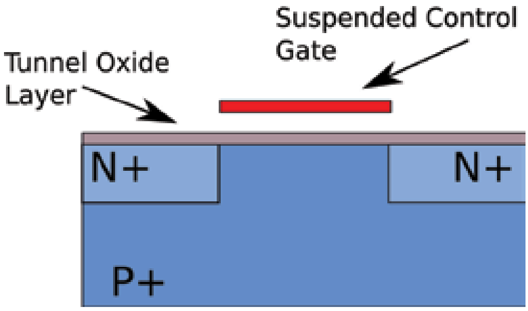


Figure 2. Pull-in curves featuring a scaling-down process. It is possible to observe that by reducing the key characteristics of the double-clamped beam, the pull-in as well as the applied voltage is also reduced.

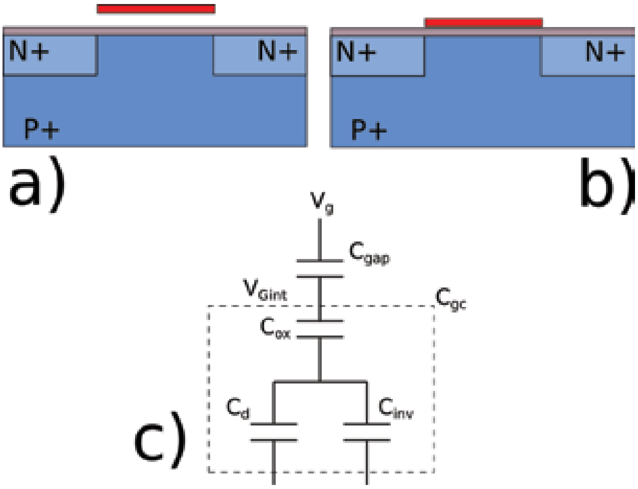


Figure 3. Set of images that depict the full operation of the suspended control gate transistor. a) Shows the suspended gate transistor unbiased. b) Depicts the SG-transistor biased and c) shows the equivalent model for the SG-MOS.

By performing the scaling-down process as Moore's law state, the MOS transistor will fence an impasse due to the tunnel oxide layer cannot be reduced further [2]. Therefore, the suspended gate MOS transistor (SG-MOS) reached the stage. **Figure 3** shows the schematic diagram of the full behaviour of the SG-MOS. A model that describe the full comportment of such device considers a set of capacitors.

The equation that describes the gate voltage according to the model is:

$$V_{Gin} = \frac{V_G}{1 + \frac{C_{gc}}{C_{gap}}} \tag{1}$$

4.1.1. Electro-mechanical modelling

To model the suspended gate with the bulk MOSFET, it is required to analyse it by considering the total energy between the conductive plates that store the energy defined as

$$E_{tot} = E_{elect} - E_{mech} = \frac{1}{2} C_{gap} V^2 - \frac{ky^2}{2} \tag{2}$$

At mechanical equilibrium, the displacement is zero, the gap capacitance is

$$C_{gap} = \frac{A \epsilon_r \epsilon_0}{t_{gap}} \tag{3}$$

where A is the plate area, t_{gap} is defined as the air-gap and ϵ_r and ϵ_0 are the material and space permittivity, respectively. While biased, the voltage between gate and substrate is coupled by the capacitance gate to channel as $V = V_G - V_{int}$ and the electrostatic force is defined as

$$F_{elec} = \frac{\epsilon_0 A}{2 y^2} (V_G - V_{int})^2 = ky \tag{4}$$

where y is the vertical gate displacement, A is the overlap area and k is the gate stiffness.

4.1.2. Pull-in and pull-out effect

When the transistor is biased, the gate is bent downwards due to the electrostatic force. By increasing it, the electrostatic force overcomes the material stiffness (k) that is function of shape defined as $F_k = ky$. By balancing both forces, the stiffness can be modelled as

$$k(t_{gap} - y) = \frac{\epsilon_0 A}{(t_{gap} - y)^2} (V_A - V_{int})^2 \tag{5}$$

where t_{gap} is the initial air-gap, material stiffness is geometry dependent that is defined as

$$k_0 = \frac{192 EI}{L_{beam}^3} \tag{6}$$

where E is the Young's modulus, I is the bending inertia moment for a rectangular beam shape. Restoring force is a linear displacement function that couples to the electrostatic force as an inversely quadratic function. Thus, there is an unstable point while the equilibrium point surpasses by biasing the structure defined as

$$y \geq \frac{2}{3} t_{gap} \tag{7}$$

The suspended gate deflection process is performed by increasing the voltage linearly until a point known as the pull-in voltage due to electrostatic force is reached. Beyond this point, the beam will collapse on the substrate due to electrostatic instability produced for the overcome of the material stiffness by the electrostatic force [23]. For a double-clamped beam, the pull-in voltage is defined as

$$V_{pull\text{-}in} = \sqrt{\frac{8\, k t_{gap}^3}{27 \epsilon_0 WL}} \tag{8}$$

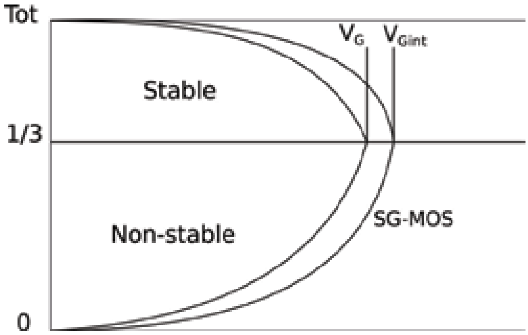


Figure 4. Schematic set of curves that define the stable and non-stable region for an initial air-gap. Approximately at one-third of the total air-gap the electrostatic force overcomes the material stiffness and the beam collapses producing the pull-in effect.

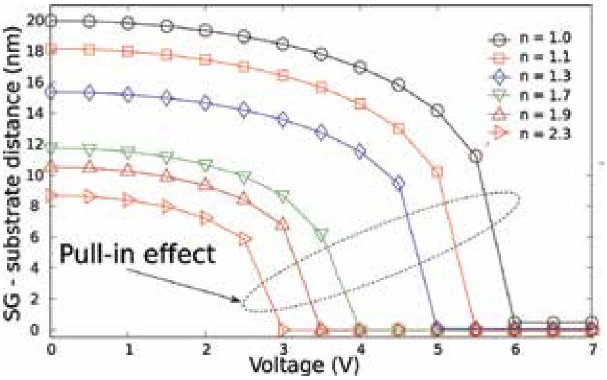


Figure 5. Pull-in curves featuring a scaling-down process. It is possible to observe that by reducing the key characteristics of the double-clamped beam, the pull-in as well as the applied voltage is also reduced.

where, W is the width, L is the channel length, $V_{pull-in}$ is the pull-in voltage. Stable and non-stable regions that define the pull-in voltage is depicted in **Figure 4**.

Pull-in voltage is strongly related to several key parameters such as beam thickness, permittivity and channel dimensions. **Figure 5** shows the pull-in effect for a set of parameters and how those are affected by scaling them down by a factor.

While the gate is collapsed on the substrate and by increasing the applied voltage, the contact area increases. By reducing it, the beam will remain attached to the substrate until the material stiffness overcomes the electrostatic force. When the beam is released due to the unbalance between those forces, the channel is interrupted. This point is it known as the pull-out effect as shown in **Figure 6**.

4.1.3. Low-range forces: Casimir and van der Waals forces

The electrostatic force is responsible for the suspended gate collapse on the substrate and as a consequence generate the channel when biased. To reduce the applied voltage, dimensions

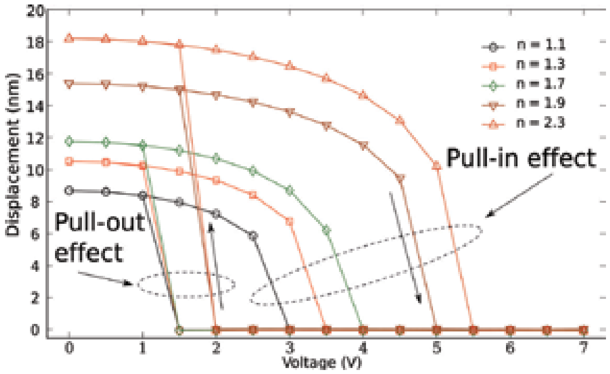


Figure 6. Pull-in curves featuring a scaling-down process. It is possible to observe that by reducing the key characteristics of the double-clamped beam, the pull-in as well as the applied voltage is also reduced.

are shrunk and as consequence other forces that only were considered to appear in systems with low dimensionality are now key for the optimal behaviour of the nanodevices: Casimir and van de Waals Forces.

In general terms, Casimir effect is strongly related to the field radiation pressure that can be generated by an electromagnetic field on every plate surface. While in contact, the Casimir force is stronger than the electrostatic force and the restitution force produced by the material stiffness.

$$F_{Casimir} = -\frac{\pi^2 hcA_{plates}}{480\, d_o^4} \tag{9}$$

where, A_{plates} is the contact area between plates, h is the Plank constant and c is the speed of light. Moreover, due to the low proximity between surfaces, the van der Waals force also appears. Van der Waals force occurs at low proximity, usually between 1 and 2 nm of separation. This force is shape dependent and it is strongly related to the Hamaker constant that encompasses the material behaviour as permittivity (ϵ) and refractive index (n) [24].Once the whole set of forces that intervene in the SG-MOS operation are put, it is possible to numerically analyse the hybrid device to later on, continue with the fabrication process and characterisation of the device.

4.2. Non-volatile memory

Another key device that has allowed the semiconductor industry to overcome a few issues such as programming/erasing speed as well as scaling-down process and low power consumption, the suspended gate silicon nanodot memory (SGSNM). The SGSNM is a hybrid device that encompasses dissimilar technologies to overcome the issues inherent to MOS technology. The non-volatile memory features a MOS transistor as readout element, a memory node fabricated with a silicon nanodots monolayer and a control gate that is double-isolated by a thin tunnel oxide layer and an air-gap as shown in **Figure 7**.

Similarly, as for the SG-MOS transistor, the SGSNM is driven by the suspended control gate to either inject or retract electrons from the memory node through the tunnel oxide layer as shown by the schematic diagram in **Figure 8**.

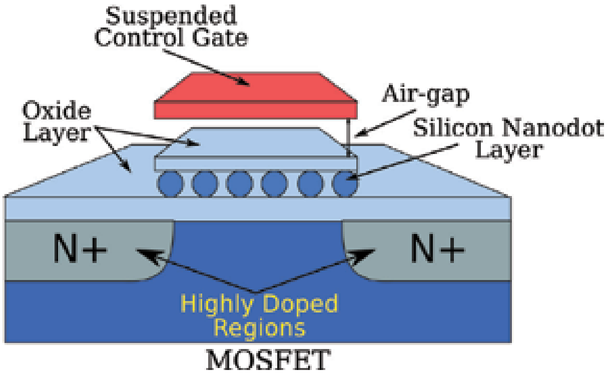


Figure 7. Schematic diagram of a hybrid nanostructure that features a non-volatile memory device. The memory features a MOSFET as readout element, a memory node fabricated with a monolayer of silicon nanodots embedded within a SiO_2 layer. The control gate is doubly-isolated by an air-gap and by a thin tunnel oxide layer.

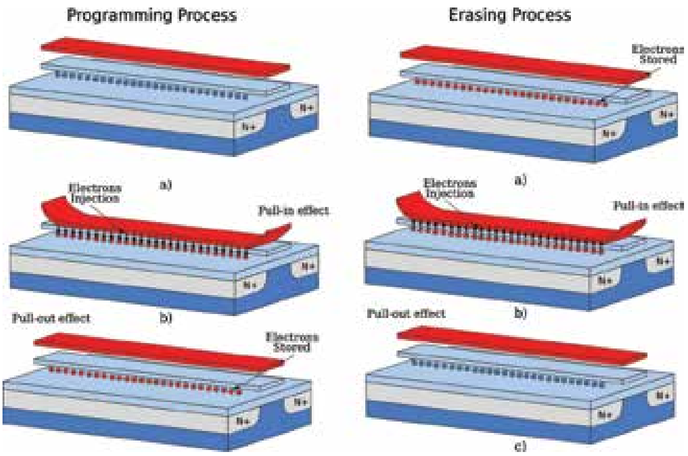


Figure 8. Schematic programming and erasing diagram for the hybrid nanodevice structure. In here, the programming and erasing feature of the nanodevice is defined. While applying a negative voltage, the suspended control gate will collapse on the tunnel oxide layer due to the electrostatic force (pull-in effect). Once in contact, the electrons will be injected into the memory node and by reducing the applied voltage, the pull-out voltage will allow to the control gate to return to its initial isolated position. It is possible to see that in the memory node the electrons are stored. On the other hand, by applying a positive voltage, the control gate will collapse and the electrons will be removed from the memory node until the pull-out voltage is reached. As a result, the memory node is empty.

The non-volatile hybrid device requires to improve a few of the inherent issues that MOS technology has. Therefore, a robust numerical analysis is needed. As point out elsewhere, there are not specific software available for such analysis. Hence, a combination of the commercial software available the set of numerical analyses is performed. To get the entire behaviour, it is needed to analyse the suspended control gate under different bias. The injection of electrons from the control gate towards the memory node and inversely is also considered. The above mentioned behaviour is required to be implemented as a library within a robust commercial software standard for the circuit simulation such as Spice [25].

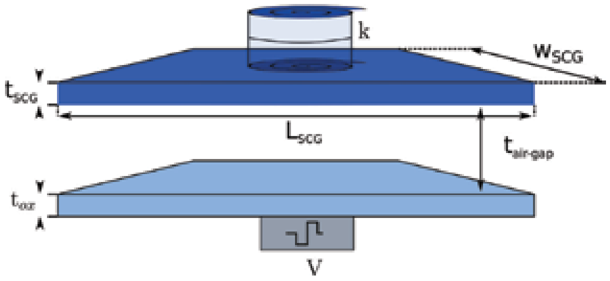


Figure 9. Schematic diagram of a two-plate capacitor that features the critical parameters to analyse the pull-in voltage.

4.2.1. Suspended control gate

The suspended control gate for the SGSNM is a double-clamped beam as featured in **Figure 9**. The beam can be modelled considering a few essential characteristics such as beam permittivity, thickness, air-gap space and substrate permittivity. As shown in **Figure 9**, a two-plate capacitor model is considered to obtain algebraically the pull-in as well as the pull-out voltages.

The key parameter is the spring constant (k) defined as

$$k = \frac{16\, E W_{SCG}\, t_{SCG}^3}{L_{SCG}^3} \tag{10}$$

where, E is defined as the Young's modulus, W_{SCG}, t_{SCG} and L_{SCG} are the width, thickness and length of the suspended control gate, respectively. As the pull-in equation has been obtained elsewhere (Eq. (8)), it will be modified according to the double-plate capacitor model. The pull-in equation is defined as

$$V_{pull-in} = 8\sqrt{\frac{2\, E t_{SCG}^3\, t_{air-gap}^3}{27\, \epsilon_0\, L_{SCG}^4}} \tag{11}$$

where ϵ_0 is the space permittivity, $t_{air-gap}$ is the air-gap separation. In the other hand, the pull-out effect is to be calculated. The pull-out effect considers that both plates are initially in contact. Both the electrostatic and electromechanical forces are driven by the applied voltage. By reducing the applied voltage, the double-clamped beam stiffness increases its presence. A further reduction allows to overcome the electrostatic force and is in here that the top plate detaches from the bottom returning to the initial isolated position. **Figure 10** shows the schematic diagram analysed to calculate the pull-out voltage.

The characteristics needed to obtain the pull-out effect considers, as the initial condition, that both plates are in contact as the initial spring constant. Due to the force that act is a combination of electrostatic, Casimir and van der Waals forces, the pull-out voltage is mathematically defined as

$$V_{pull-out} = \sqrt{\frac{2\, k t_{ox}^2}{\epsilon_0\, \kappa_{ox} A}\left(t_{SCG} - t_{ox}\right) - \frac{A_h}{3\pi k_{ox}\, t_{ox}}} \tag{12}$$

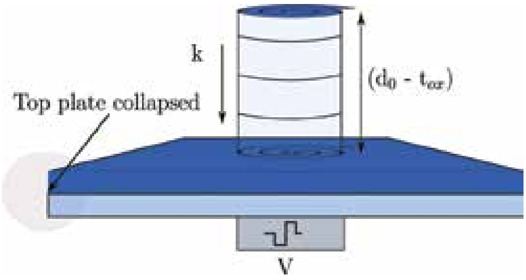


Figure 10. Schematic diagram of a two-plate capacitor that features the key parameters to analyse the pull-out voltage.

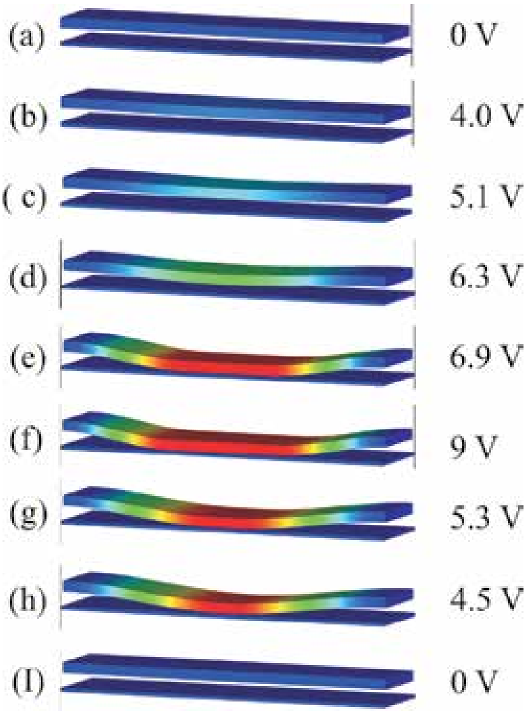


Figure 11. Set of images that describe the beam while bias (a) to (c) until it is being trapped by the pull-in voltage (d) and collapsed on the substrate (e). By increasing the applied voltage, contact area increases as well as the current density (f). By reducing the applied voltage, the contact area is reduced (g) & (h) until it reached the pull-out point and the beam return to its initial isolated position (i).

where, κ_{ox} and t_{ox} are defined as the dielectric constant, thickness of the dielectric material, respectively and A_h is defined as the Hamaker constant. By considering the above mentioned equations for pull-in and pull-out voltages (Eqs. (11) and (12)), the voltage obtained can be used as a guide to find those voltages. Further analysis is strongly suggested by using commercial software such as Comsol or CoventorWare to corroborate the pull-in and pull-out voltages as shown in **Figure 11** [26, 27].

4.2.2. Programming and erasing processes

The programming and erasing processes occur due to a combination of the applied voltage to the control gate (pull-in effect) and the injection of electrons through the tunnel oxide layer. The injection of electrons to program and to erase the memory node are through the movement of the suspended control gate (top injection). In contrast, the typical memory devices that require the channel formation and the injection from the bottom, i.e., flash memory [28]. The tunnelling process starts once both layers are in contact (after the pull-in process occurs). **Figure 12** depicts a set of schematic diagrams for the programming and easing processes by using energy band plots.

Above figure (**Figure 12**) describes the tunnelling process that mathematically co-integrates in the transfer Matrix method, the Tsu-Esaki equations in a finite element method based in homemade algorithm. In this algorithm, the Poisson's equation and the Schrödinger equation are co-solved simultaneously to obtain the current density curve.

The model that is being considered to implement assumes that the energy and momentum are kept due to there is no energy dissipation process considered. Hence, the total energy can be divided into lateral and vertical components

$$E(\vec{k}) = \frac{\hbar^2\left(k_x^2 + k_y^2\right)}{2\,m^*} + E_z \tag{13}$$

where, m^* is defined as the effective mass of the electron, $\hbar$ is defined as the Planck constant and $\vec{k}$ represents the lateral wave vector. The Tsu-Esaki equation at finite temperature is defined as

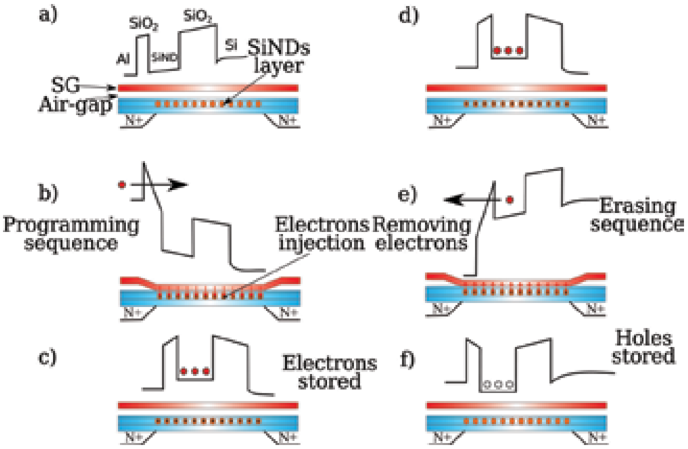


Figure 12. A schematic diagram of the quantum-mechanical tunnelling process for the programming and erasing processes according to the band energy diagram. (a) Shows the energy band diagram for the SGSNM device. While applying a negative voltage the beam collapsed on the tunnel oxide layer and the electrons are injected due to the band diagram became triangular (b). By removing the voltage, the beam returns to its initial flat position and the electrons are trapped within the memory node (c). By applying a positive voltage the energy bands are bent in the opposite direction and the electrons are removed from the memory node and the removing the applied voltage, the memory node is empty as shown in (f).

$$J = J_{\rightarrow} - J_{\leftarrow} \tag{14}$$

$$J_{\rightarrow} = 2\sum_{k_x,k_x,k_x>0} ev_z\, T(E_z)\left[f_L(\vec{k})\left[1 - f_R(\vec{k})\right]\right] \tag{15}$$

$$J_{\leftarrow} = 2\sum_{k_x,k_x,k_x<0} ev_z\, T(E_z) f_R(\vec{k})\left[1 - f_L(\vec{k})\right] \tag{16}$$

where, $T(E_z)$ is the transmission probability function, f_L and f_R are the Fermi distribution functions at barrier sides called emitter and collector regions.

$$f_{L,R}(\vec{k}) = \frac{1}{1 + \exp\left(\frac{E(\vec{k}) - E_F^{L,R}}{K_B T}\right)} \tag{17}$$

where.

$E_F^L = E_F^R + V$, V is defined as an external voltage applied to the barrier Integrating over the regions perpendicular to z

$$J = \int_0^\infty dE_z\, T(E_z) S(E_z) \tag{18}$$

Tsu-Esaki equation and the transfer matrix method consider the Schrödinger equation as time independent and in one dimension (1D).

$$-\frac{\hbar^2}{2}\nabla\left(\frac{1}{m^*(z)}\nabla\right)\Psi(z) + V(Z)\Psi(z) = E_z\,\Psi(z) \tag{19}$$

where $m^*(z)$ is defined as the z-dependent conduction-band effective mass, $V(z)$ as the potential energy used and $\Psi(z)$ is defined as the wave function. The solution for the wave function has the form of

$$\Psi^{(i)}_{k_z^{(i)}}(z) = A^{(i)}_{k_z^{(i)}}\exp\left(ik_z^{(i)} z\right) + B^{(i)}_{k_z^{(i)}}\exp\left(-ik_z^{(i)} z\right) \tag{20}$$

As a result, we obtained a voltage-current density curve (V-J) for a particular substrate, in this case SiO_2. Figure shows a set of curves for a set of SiO_2 thicknesses.

4.2.3. Circuit simulation

Once the pull-in and pull-out voltages as well as the tunnelling process through the current density curve have been obtained, those can be implemented within a commercial simulation software such as Spice as external libraries. **Figure 13** shows the algorithm that is considered to be implemented for the correct behaviour of the non-volatile memory in particular for the memory node.

The set of libraries added to Spice are based on the models depicted in **Figure 14**. Curves are coded by using a state-of-the-art language such as Verilog-AMS [29].

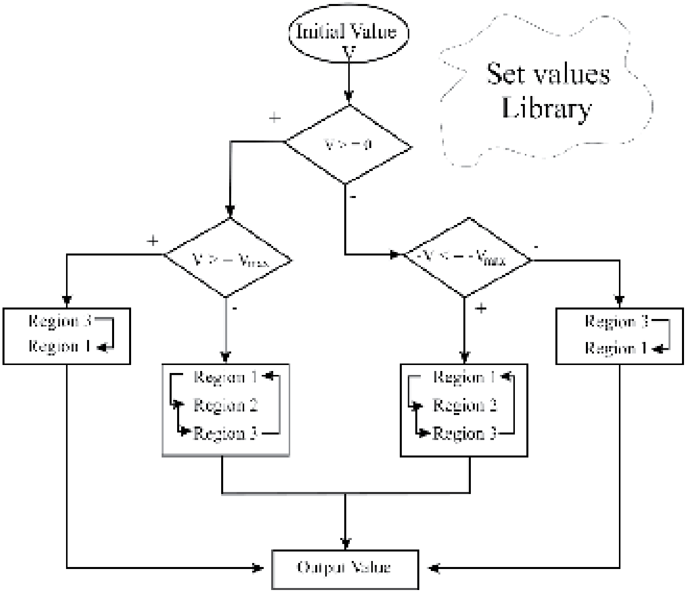


Figure 13. Schematic diagram of the libraries considered to be implemented within the circuit simulation.

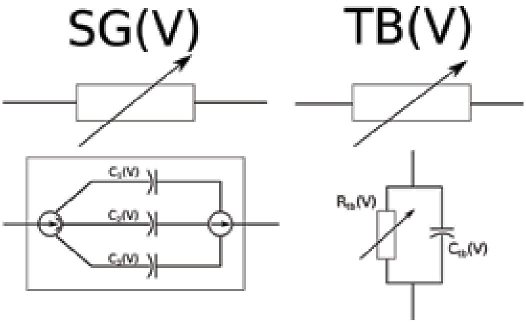


Figure 14. Schematic diagram of the libraries considered to be implemented within the circuit simulation.

Once both set of curves are implemented, the equivalent circuit considered is shown in **Figure 15**. Before analysing the cell, it is required to define the bias source for the suspended control gate as a piece-wise linear source (PWL) and for the MOS transistor a normal bias source.

By simulating the SG-MOS transistor cell, the set of curves obtained are displayed in **Figure 16**. The chart is divided in four arrows in which the PWL source, bias source, memory node and the readout element to identify the node state. The PWL source follows a cycle of four sections. The cycle starts with a negative voltage applied to the control gate, in the memory node it is possible to observe how the electrons are being injected from the control gate towards the thin tunnel oxide layer into the memory node. While the PWL source shows zero volts, the memory node displays a negative charge level. By biasing the MOS transistor, it shows a current level in the order of 10^{-12} A indicating that the memory has been programmed. On the

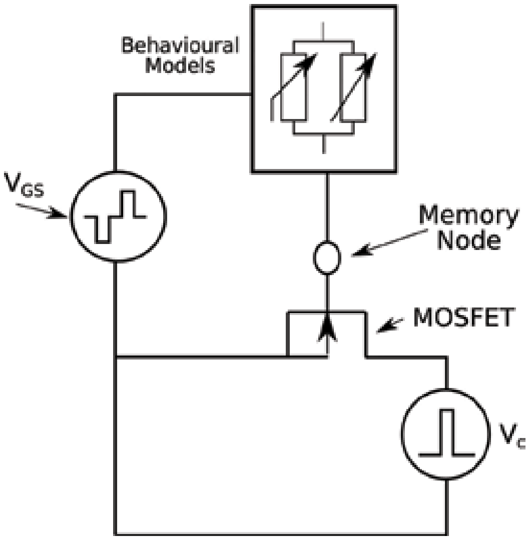


Figure 15. Equivalent model for the memory cell to be implemented within the commercial circuit simulator.

other hand, by applying a positive voltage on the suspended control gate, the memory node shows how the electrons are being retrieved by the gate. When the PWL source returns to zero V and the readout element is biased, in the memory node it is possible to observe a positive charge level and the current shown by the transistor is 6 magnitude order larger than when programmed. A key element that can be observed is the programming/erasing time of 1.7 ns for the characteristic of the nanostructure. Nowadays, the times for flash memory is at least over 3 magnitude order lower than the hybrid structure.

Now that the simulation shows the results for the programming and erasing times, the next step is to fabricate the suspended control gate as well as the MOS transistor.

4.2.4. Fabrication process

The fabrication process for the SGSNM cell, start by considering a substrate based on Si, on top of it, a high quality and thin SiO_2 layer is grown. As the memory node, a monolayer made of silicon nanodots is deposited even sparsely. Each silicon nanodots is isolated between them due to are immersed in a SiO_2 layer. As sacrificial layer, a thick polysilicon layer is used. Finally, as the suspended control gate, an Aluminium layer is deposited. Patterns are performed by using standard photoresist due to the size of the beam being used. To get the shape of the beams, a standard Al wet etchant is considered at 300 K. To release the doubly-clamped beam, a single-step dry etching process is performed. As a result, the control gate is suspended and the device is up to measure. **Figure 17** shows the process above described.

4.3. High-Q resonators

High-Q resonators are one of the main nanodevices to be investigated due to the wide variety of applications that can be used when implemented it. A double-clamped beam as a classical

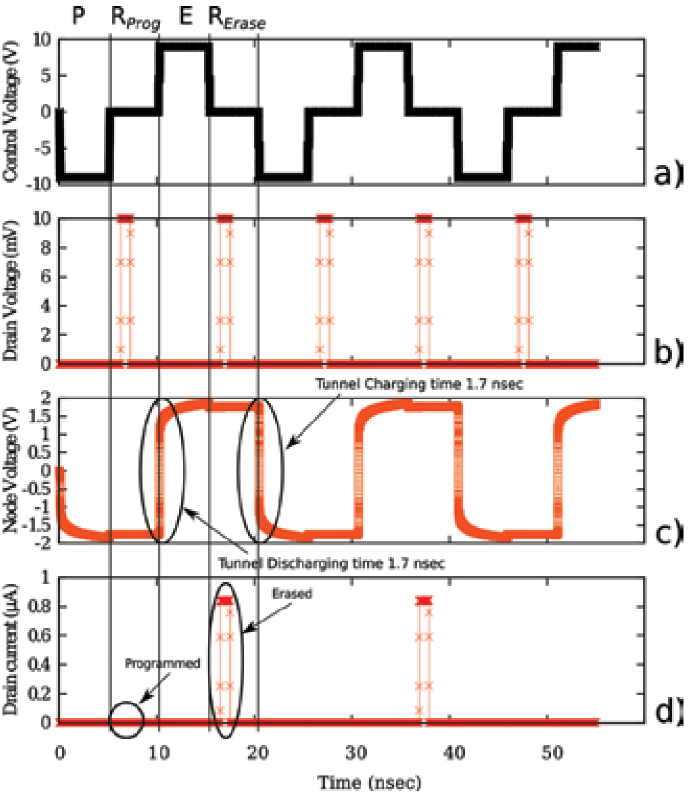


Figure 16. Set of curves obtained from simulating the SGSNM. The first curve defines the piece-wise-linear source that drives the memory operation. While performing a negative voltage, it is possible to see how the electrons are injected in to the memory node. When the PWL source shows zero volts and the readout element is biased, a current peak is observed indicating that the memory node had been programmed. In contrast, when a positive voltage is applied, at the memory node the electrons are retrieved towards the control gate and when the PWL source is zero, the readout element shows a current peak 6 magnitude order larger than when programmed indicating that the memory has been erased.

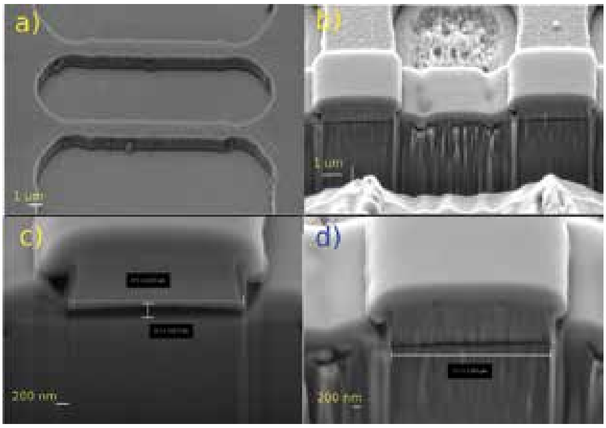


Figure 17. Set of SEM images that shows the doubly-clamped beam suspended. a) Shows the wet etching process result for the Al layer. b) Shows the result for the dry-etching process recipe that displays the beams successfully suspended. c) and d) show a zoom for each beam.

capacitive can be modelled with high accuracy by including non-linear terms. These key characteristics can be strongly related to a spring in which the stiffness can vary according to the electric field applied [30]. The non-linear restoring force F_k can be expressed as:

$$F_k(y) = ky + k_1 y^2 + k_2 y^3 + \cdots O(y^n) \tag{21}$$

The Duffing equation for damping factor is expressed as

$$\frac{d^2 y}{dt^2} + \frac{\gamma}{m}\frac{dy}{dt} + \frac{k}{m} + \alpha y^2 + \beta y^3 = \frac{F_\omega}{m}\cos\omega t \tag{22}$$

where $\alpha = k_1/m$ and $\beta = k_2/m$. A different non-linear models does not consider the second order term due to this term have no impact on the resonant behaviour.

4.3.1. Quality factor

Quality factor is a reference for MEMS/NEMS resonators. It describes the ratio between the energy stored and dissipated defined as:

$$Q_{res} = 2\pi\frac{W_n}{\Delta W} \tag{23}$$

where W_n is the stored energy and ΔW is the energy dissipated each cycle. A fundamental relationship among them can allow to improve for a high quality factor. The maximum energy that can be stored in an electromechanical resonator strongly depends on the vibration mode, the mass of the resonator and the displacement [13].

$$W_n = \frac{\rho}{8} V_r es\, \omega_n^2 x_{nc}^2 \tag{24}$$

According to the material stiffness is related to the quality factor that can be achieved such as poly-diamond that shows a very high Young's modulus. As a matter of fact, the material density modifies the total energy amount that the resonator can drive. Massive resonators that consider extensional or bulk mode resonance present a very high Q-factor.

5. Applications

Hybrid structures have a broad variety of applications being bio-applications key in the development of health services worldwide. In here, we present how the hybrid structures can be applied as bio-sensors.

5.1. Biological applications

The use of nanoelectromechanical systems (NEMS) has had a remarkable impact on different biological areas such as medical, food industry including food safety and analytical. The principle to NEMS development is based on the search for systems that serve as micro-reservoirs, micropumps, valves, sensors and other structures that use biocompatible materials appropriate for chemical or biological molecules release or for their detection. The development of intelligent biomaterials as responsive hydrogels and configurationally imprinted biomimetic polymers (CIBPs) are the preferred materials for biological applications due to high adaptability and compatibility with biological molecules and cells [31]. The use of configurationally CIBPs allows the improvement of molecular recognition systems through the control of chemical functionality and the tridimensional structures.

The constructional designs of devices that operate in an intelligent way, with high sensibility to diverse analytes are capable to control the release of therapeutic or antimicrobial molecules in response to a key biological event allowing it to be used in diverse applications [32]. Medical treatments have innovated in response to the wide variety of pathophysiological conditions that require the development of more effective therapeutic agents and the use of device-integrated biomaterials that can serve as sensors and carriers. NEMS-based devices offer opportunities to address a significant number of unmet medical needs related to dosing, diagnostic and tissue engineering.

Some of the advantages of leading NEMS-based drug delivery in implant/stent have a potential impact if treatment requires local dosing, avoiding the need for injection. In the area of implant/pumps devices, those have the potential to lower total dose due to local administration, avoids the need for injection, permitting a local and systemic parenteral administration. To implant an electronic chip, it could be observed potential to lower total dose due to local administration, the capability of establishing precise timing and control, avoids the need for injection, flexibility of local or systemic parenteral administration depending on the formulation. Finally, the implant/polymer chips show potential to lower total dose due to local administration and avoid the need for injection [33].

In diagnostic applications, the ability to monitor the health status, diseases onset and diseases progression is highly desirable. To develop devices for these applications, it is necessary to know the specific biomarker associated with a health or a disease state. To count with a non-invasive approach to detect and monitor this biomarker and technological capability to discriminate between and among the biomarkers. The development of simple-to-use NEMS-based biosensors could have applications in the identification of major diseases and/or pathogens, rapid diagnosis of exposure and disease and detection of emerging pathogens which could be in parallel for multiple infectious agents, accurate assessment of disease stage and prognosis and a better management of outbreaks and emerging acute and chronic health threats [34]. In the same way, the food industry has developed research focused on food packaging and food safety through the use of NEMS-based biosensors. Nanosensors permit the detection of food-borne contaminants, detection of pathogens and capability to detect and quantify volatile or non-volatile compounds related to quality or natural physiological process [35].

Food contact materials used to monitor the condition of packaged food or the environment surrounding the food have used in both, polymer nanomaterials for food packaging (PNFP) and MEMS/NEMS-based biosensors to create "Intelligent/smart food package". This technology can inform with a visible indicator or other novel systems, the supplier or consumer that foodstuffs are still fresh, or whether the packaging has been breached, kept at the appropriate temperatures throughout the supply chain, or has spoiled [36]. Fresh produce or meats during their maturation or spoilage exhibit odours, colours or other sensory characteristics which can be easily discerned by consumers. However, to determine that the product should be good for the determined period, consumers use information that the producers set, based on a set of idealised assumptions about the way that the food is stored or transported. Some of these assumptions are not real if it is considered that this date may no longer be applicable if this food product was stored above its optimal temperature for an hour, either in a delivery truck or a warm automobile.

Most of the development of these intelligent food packages, look for alternatives to detect small organic molecules that are the result of microbial activities or adulterants, specific gasses, and/or viable foodborne pathogens [31]. The benefits of intelligent packages are related with speed and accuracy with which industries or regulatory agencies can detect the presence of molecular contaminants or adulterants in complex food matrices [31].

Nanoelectromechanical systems used as sensors have a significant impact in the analytical field. Biosensing is a complex task that involves knowledge about the biochemical process in addition to diverse problems related to the nature of the operation medium. Some of the most popular applications for NEMS sensors in analytical chemistry permit the detection of formaldehyde vapour, water-toluene vapour, organics and inorganics, alcohols, H_2, organophosphorous vapour and others (in gas-phase sensing applications). In liquid-phase applications it is possible to analyse acetic acid, aminoethanethiol, retinoid isomers, metal ions and fructose among others. Finally, in biosensing applications it is possible to detect myoglobin, antibody-concentration, liposomes, thiolated single-stranded DNA (ssDNA), thiol modified single-stranded DNA (ssDNA) or the presence of pathogens or their toxins (airborne anthracis spores, Staphylococcus enterotoxin B (SEB), Salmonella typhimurium, airborne virus particles, *Escherichia coli* O^{157}:H^7 and others) [37].

A potential high number of applications in the biosensing or dosing could be proposed by different devices between them hybrids type NEMS-MOS.

6. Conclusions

As the Semiconductor industry has reached an impasse due to the scaling-down process according to Moore's Law for the Metal-Oxide-Technology in use. Alternative technologies are foreseen to allow the development of nanodevices with a broad variety of characteristics such as high switching speed, low power consumption, robust, among others that can overcome the inherent issues for Silicon. A few "exotic materials" appear as good candidates such as Graphene, MoS_2, BN-h, etc. However, the time for the novel technology to be mature will take time. To allow the "exotic materials" to mature, the semiconductor industry needs novel nanostructures capable to overcome a few of the issues that silicon-based technology is facing. As clearly shown, the hybrid nano-structures allow to develop a broad variety of nanodevices such as transistors, memories and sensors. As stated in the chapter, it is demonstrated that hybrid-structures are allowing the emerging technology to become mature to diversify as well as to be reliable as silicon-based technology is.

Acknowledgements

The authors would like to thank to the Nanofab Staff at the University of Southampton, UK. To the Research Council for Science and Technology (CONACyT), to the Universidad Autónoma de Nuevo León and to the Universidad de Guadalajara, México, for the help provided in the development of this research.

Author details

Mario Alberto García-Ramírez[1,2]*, Miguel Angel Bello-Jiménez[3],
María Esther Macías-Rodríguez[4], Barbara Cortese[5], José Trinidad Guillen-Bonilla[1],
Rosa Elvia López-Estopier[3], Juan Carlos Gutiérrez-García[1] and Everardo Vargas-Rodríguez[6]

*Address all correspondence to: seario@gmail.com

1 Electronics and Computer Science Department, Research University Centre for Applied Sciences and Engineering (CUCEI), Universidad de Guadalajara, Guadalajara, Jalisco, México

2 Faculty of Electrical and Mechanical Engineering, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León, México

3 Instituto de Investigación en Comunicación Óptica (IICO), Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico

4 Food Safety Laboratory, Department of Pharmacobiology, Research University Centre for Applied Sciences and Engineering (CUCEI), Universidad de Guadalajara, Guadalajara, Jalisco, México

5 NNL, National Nanotechnology Laboratories of CNR-INFM, Distretto Tecnologico, Universitá del Salento, Lecce, Italy

6 Departamento de Estudios Multidisciplinarios, División de Ingenierías, Universidad de Guanajuato, Yuriria, Guanajuato, México

References

[1] Moore E. Gordon: Cramming more components onto integrated circuits. Electronics. 1969;**38**(8):114

[2] International Technology Roadmap for Semiconductors. 2007. Available from: http://www.itrs.net/Links/2009ITRS/Home2009.html [Accesed: May 20, 2008]

[3] Terrones M, Romo-Herrera JM, Cruz-Silva E, López-Urías F, Muñoz-Sandoval E, Velázquez-Salazar JJ, Terrones H, Bando Y, Golberg D. Pure and doped boron nitride nanotubes. Materials Today. 2007;**10**(5):30-38. DOI: 10.1016/S1369-7021(07)70077-9

[4] Zhang J, Terrones M, Park CR, Mukherjee R, Monthioux M, Koratkar N, Kim YS, Hurt R, Frackowiak E, Enoki T, Chen Y, Chen Y, Bianco A. Carbon science in 2016: Status, challenges and perspectives. Carbon. 2016;**98**:708-732. DOI: 10.1016/j.carbon.2015.11.060. ISSN: 0008-6223

[5] Das S, Robinson JA, Dubey M, Terrones H, Terrones M. Beyond graphene: Progress in novel two-dimensional materials and van der Waals solids. Annual Review of Materials Research. 2015;**45**(1):1-27

[6] Lv R, Robinson JA, Schaak RE, Sun DS,Y, Mallouk TE, Terrones M. Transition metal dichalcogenides and beyond: Synthesis, properties, and applications of single- and few-layer nanosheets. Accounts of Chemical Research. 2015;**48**(1):56-64. DOI: 10.1021/ar5002846

[7] Ionescu AM, Riel H. Tunnel field-effect transistors as energy–efficient electronic switches. Nature. 2011;**479**:329. DOI: 10.1038/nature10679

[8] Butler SZ et al. Progress, challenges, and opportunities in two-dimensional materials beyond graphene. ACS Nano. 2013;**7**(4):2898-2926. DOI: 10.1021/nn400280c

[9] Geim AK, Novoselov KS. The rise of graphene. Nature Materials. **6**:183. DOI: 10.1038/nmat1849

[10] Castro EV et al. Biased bilayer graphene: Semiconductor with a gap tunable by electric field effect. Preprint at http://arxiv.org/abs/cond-mat/0611342. 2006

[11] Radisavljevic B, Radenovic A, Brivio J, Giacometti V, Kis A. Single-layer MoS_2 transistors. Nature Nanotechnology. 2011;**6**:147-150. DOI: 10.1038/nnano.2010.279

[12] Novoselov KS et al. Electric field effect in atomically thin carbon films. Science. 2004;**306**: 666-669

[13] Agache V. Intégration et caractérisation physique de nanostructures pour les technologies de l'information et de la communication. Application au filtrage électromécanique dans la gamme des radiofréquences (0.8-2.4GHz) [PhD dissertation]. IEMN; 2003

[14] Su Y-K. 6.02—Nitride-based LEDs and Superluminescent LEDs. In: Bhattacharya P, Fornari R, Kamimura H, editors. Comprehensive Semiconductor Science and Technology. Amsterdam: Elsevier; 2011. pp. 28-100. DOI: 10.1016/B978-0-44-453153-7.00024-9

[15] There's Plenty of Room at the Bottom, an Invitation to Enter a New Field of Physics. http://www.phy.pku.edu.cn/~qhcao/resources/class/QM/Feynman%27s–Talk.pdf

[16] Toofan M, Toofan J. Chapter 5—A brief review of the cleaning process for electronic device fabrication. In: Kohli R, Mittal KL, editors. Developments in Surface Contamination and Cleaning. Oxford: William Andrew Publishing; 2015. pp. 185-212. DOI: 10.1016/B978-0-323-29961-9.00005-3. ISBN: 9780323299619

[17] Nguyen N-T. Chapter 4—Fabrication technologies. In: Micro and Nano Technologies. Oxford: William Andrew Publishing; 2012. pp. 113-161. DOI: 10.1016/B978-1-4377-3520-8.00004-8. Micromixers (2nd ed.). ISBN: 9781437735208

[18] https://www.samcointl.com/basics–bosch–process–silicon–deep–rie/

[19] Dadgour HF, Banerjee K. Hybrid NEMS-CMOS integrated circuits: A novel strategy for energy-efficient designs. IET Computers and Digital Techniques. 2009;**3**(6):593-608. DOI: 10.1049/iet–cdt.2008.0148

[20] Hybrid Nano-Electro-Mechanical/Integrated Circuit Systems for Sensing and Power Management Applications NEMSIC Project. https://cordis.europa.eu/project/rcn/87279–en.html

[21] García-Ramírez MA, Ghiass AM, Moktadir Z, Tsuchiya Y, Mizuta H. Fabrication and characterisation of a double-clamped beam structure as a control gate for a high-speed non-volatile memory device. Microelectronic Engineering. 2014;**114**:22-25. DOI: 10.1016/j.mee.2013.09.002

[22] Hassani FA, Tsuchiya Y, Mizuta H. In-plane resonant nano-electro-mechanical sensors: A comprehensive study on design, fabrication and characterization challenges. Sensors. 2013;**13**(3):9364-9387

[23] García-Ramírez MA, Tsuchiya Y, Mizuta H. Hybrid numerical analysis of a high-speed non-volatile suspended gate silicon nanodot memory (SGSNM). Journal of Computational Electronics. 2011;**10**(1-2):248-257

[24] Hadjittofis E, Das SC, Zhang GGZ, Heng JYY. Chapter 8—Interfacial phenomena. In: Qiu Y, Chen Y, Zhang GGZ, Yu L, Mantri RV, editors. Developing Solid Oral Dosage Forms. Second edition. Boston: Academic Press; 2017. pp. 225-252. DOI: 10.1016/B978-0-12-802447-8.00008–X. ISBN: 9780128024478

[25] Cadense. https://www.cadence.com/

[26] Comsol Multiphysics. 2007. Available from: https://www.comsol.com/ [Accessed: October 10, 2007]

[27] Coventor Ware. 2007. Available from: https://www.coventor.com/mems-solutions/products/coventorware/ [Accessed: October 2, 2008]

[28] Masuoka F, Momodomi M, Iwata Y, Shirota R. New ultra high density EPROM and flash EEPROM with NAND structure cell. In: 1987 International Electron Devices Meeting IEDM. IEEE; 1987. DOI: 10.1109/IEDM.1987.191485

[29] http://www.designers-guide.org/verilogams/VlogAMS–2.3–pub.pdf

[30] Kaajakari V, Mattila T, Oja A, Seppa H. Nonlinear limits for single-crystal silicon microresonators. Journal of Microelectromechanical Systems. 2004;**13**:715-724

[31] Duncan TV. Applications of nanotechnology in food packaging and food safety: Barrier materials, antimicrobials and sensors. Journal of Colloid and Interface Science. 2011;**363**:1-24

[32] Caldorera-Moore M, Peppas NA. Micro- and nanotechnologies for intelligent and responsive biomaterial-based medical systems. Advanced Drug Delivery Reviews. 2009; **61**:1391-1401

[33] Staples M, Daniel K, Cima MJ, Langer R. Application of micro- and nano-electromechanical devices to drug delivery. Pharmaceutical Research. 2006;**23**(5):847-863

[34] Li Y, Denny P, Ho CM, Montemagno C, Shi W, Qi F, Wu B, Wolinsky L, Wong DT. The oral fluid MEMS/NEMS chip (OFMNC): Diagnostic & translational applications. Advances in Dental Research. 2005;**18**(1):3-5

[35] Sozer N, Kokini JL. Nanotechnology and its applications in the food sector. Trends in Biotechnology. 2009;**27**(2):82-89

[36] Silvestre C, Duraccio D, Cimmino S. Food packaging based on polymer nanomaterials. Progress in Polymer Science. 2011;**36**:1766-1782

[37] Zougagh M, Ríos A. Micro-electromechanical sensors in the analytical field. The Analyst. 2009;**134**:1274-1290

Chapter 6

Comprehensive Analytical Models of Random Variations in Subthreshold MOSFET's High-Frequency Performances

Rawid Banchuin

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72710

Abstract

Subthreshold MOSFET has been adopted in many low power VHF circuits/systems in which their performances are mainly determined by three major high-frequency characteristics of intrinsic subthreshold MOSFET, i.e., gate capacitance, transition frequency, and maximum frequency of oscillation. Unfortunately, the physical level imperfections and variations in manufacturing process of MOSFET cause random variations in MOSFET's electrical characteristics including the aforesaid high-frequency ones which in turn cause the undesired variations in those subthreshold MOSFET-based VHF circuits/systems. As a result, the statistical/variability aware analysis and designing strategies must be adopted for handling these variations where the comprehensive analytical models of variations in those major high-frequency characteristics of subthreshold MOSFET have been found to be beneficial. Therefore, these comprehensive analytical models have been reviewed in this chapter where interesting related issues have also been discussed. Moreover, an improved model of variation in maximum frequency of oscillation has also been proposed.

Keywords: gate capacitance, maximum frequency of oscillation, subthreshold MOSFET, transition frequency, VHF circuits/systems

1. Introduction

Subthreshold MOSFET has been extensively used in many VHF circuits/systems, e.g., wireless microsystems [1], low power receiver [2], low power LNA [3, 4] and RF front-end [5], where performances of these VHF circuits/systems are mainly determined by three major high-frequency characteristics of intrinsic subthreshold MOSFET, i.e., gate capacitance, C_g, transition frequency, f_T, and maximum frequency of oscillation, f_{max}. Clearly, the physical level imperfections and manufacturing process variations of MOSFET, e.g., gate length random fluctuation, line edge roughness, random dopant fluctuation, etc., cause the variations in MOSFET's electrical characteristics, e.g., drain current, I_D and transconductance, g_m, etc. These variations are crucial in the statistical/variability aware analysis and design of MOSFET-based circuits/systems. So, there exist many previous studies on such variations which some of them have also focused on the subthreshold MOSFET [1, 6–12]. Unfortunately, C_g, f_T, and f_{max} have not been considered even though they also exist and greatly affect the high-frequency performances of such MOSFET-based circuits/systems. Therefore, analytical models of variations in those major high-frequency characteristics have been performed [13–17]. In [13], an analytical model of variation in f_T derived as a function of the variation in C_g has been proposed where only strong inversion MOSFET has been focused. However, this model is not comprehensive, as none of any related physical levels variable of the MOSFET has been involved. In [14], the models of variations in C_g and f_T, which are comprehensive as they are in terms of the related MOSFET's physical level variables, have been proposed. Again, only the strong inversion MOSFET has been considered in [14].

According to the aforementioned importance and usage of subthreshold MOSFET in the MOSFET-based VHF circuits/systems, the comprehensive analytical models of variations in C_g, f_T, and f_{max} of subthreshold MOSFET have been proposed [15–17]. Such models have been found to be very accurate as they yield smaller than 10% the average percentages of errors. In this chapter, the revision of these models will be made where some foundations on the subthreshold MOSFET will be briefly given in the subsequent section followed by the revision on models of C_g in Section 3. The models of f_T and f_{max} will, respectively, be reviewed in Sections 4 and 5 where an improved model of variation in f_{max} will also be introduced. Some interesting issues related to these models will be mentioned in Section 6 and the conclusion will be finally drawn in Section 7.

2. Foundations on subthreshold MOSFET

Unlike the strong inversion MOSFET in which I_d is a polynomial function of the gate to source voltage, V_{gs}, I_d of the subthreshold MOSFET is an exponential function of V_{gs} and can be given as follows:

$$I_d = \mu C_{dep}\frac{W}{L}\left(\frac{kT}{q}\right)^2 \exp\left[\frac{V_{gs}-V_t}{nkT/q}\right]\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right] \tag{1}$$

where C_{dep} and n denote the capacitance of the depletion region under the gate area and the subthreshold parameter, respectively.

By using Eq. (1) and keeping in mind that $g_m = dI_d/dV_{gs}$, g_m of subthreshold MOSFET can be given by

$$g_m = \frac{\mu}{n}C_{dep}\frac{W}{L}\left(\frac{kT}{q}\right)^2 \exp\left[\frac{V_{gs}-V_t}{nkT/q}\right]\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right] \tag{2}$$

3. Variation in gate capacitance (C_g)

Before reviewing the models of variation in C_g of subthreshold MOSFET, it is worthy to introduce the mathematical expression of C_g as it is the mathematical basis of such models. Here, C_g which can be defined as the total capacitance seen by looking in to the gate terminal of the MOSFET as shown in **Figure 1**, can be given in terms of the gate charge, Q_g as [15]

$$C_g = \frac{dQ_g}{dV_{gs}} \tag{3}$$

where

$$Q_g = \frac{\mu W^2 L C_{ox}^2}{I_d} \int_0^{Vgs-Vt} \left(V_{gs} - V_c - V_t\right)^2 dV_c - Q_{B,max} \tag{4}$$

It is noted that $Q_{B,max}$ stands for the maximum bulk charge [15]. By using Eq. (1), Q_g of the subthreshold MOSFET can be found as

$$Q_g = \frac{\left[\frac{WL^2 C_{ox}^2}{C_{dep}(kT/q)^2}\right]\left(V_{gs} - V_t\right)^3}{3\left[1 - \exp\left[-\frac{V_{ds}}{kT/q}\right]\right]\exp\left[\frac{q}{nkT}\left(V_{gs} - V_t\right)\right]} - Q_{B,max} \tag{5}$$

As a result, the expression of C_g can be obtained by using Eqs. (1) and (5) as follows

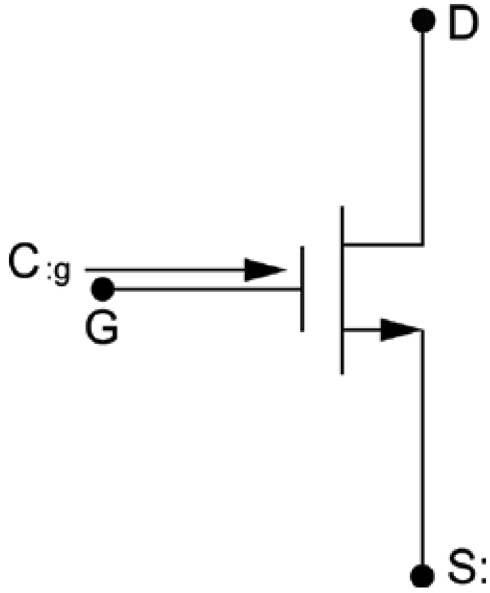


Figure 1. The conceptual definition of C_g (referenced to N-type MOSFET).

$$C_g = \frac{1}{3}\left[\frac{WL^2C_{ox}^2}{C_{dep}(kT/q)^2}\right]\left[3(V_{gs}-V_t)^2 - \frac{q}{nkT}(V_{gs}-V_t)^3\right]\exp\left[-\frac{q}{nkT}(V_{gs}-V_t)\right] \tag{6}$$

By taking the physical level imperfections and manufacturing process variations of MOSFET into account, random variations in MOSFET's parameters such as V_t, W, L, etc., denoted by ΔV_t, ΔW, ΔL, and so on existed. These variations yield the randomly varied C_g i.e. $C_g(\Delta V_t, \Delta W, \Delta L, \ldots)$ [15]. Thus, the variations in C_g, ΔC_g can be mathematically defined as [15]

$$\Delta C_g \overset{\Delta}{=} C_g(\Delta V_t, \Delta W, \Delta L, \ldots) - C_g \tag{7}$$

where C_g stands for the nominal gate capacitance in this context.

With this mathematical definition and the fact that ΔV_t is the most influential in subthreshold MOSFET [18], the following comprehensive analytical expression of ΔC_g has been proposed in [15]

$$\begin{gathered}\Delta C_g = 2\left[\sqrt{\frac{W}{C_{dep}}}\frac{LC_{ox}}{kT/q}\right]^2\left[\exp\left[-\frac{V_{ds}}{kT/q}\right]-1\right]^{-1} \\ \left[V_{gs}-V_{FB}-\phi_s-N_{eff}W_{dep}\right]\left[V_t-V_{FB}-\phi_s-N_{eff}W_{dep}\right]\end{gathered} \tag{8}$$

where N_{eff}, V_{FB}, W_{dep}, and ϕ_s denote the effective values of the substrate doping concentration $N_{sub}(x)$, the flat band voltage, depletion width, and surface potential, respectively. Moreover, N_{eff} can be obtained by weight averaging of $N_{sub}(x)$ as [15]

$$N_{eff} = 3\int_0^{W_{dep}} N_{sub}(x)\left(1-\frac{x}{W_{dep}}\right)^2\frac{dx}{W_{dep}} \tag{9}$$

As ΔC_g is a random variable, it is necessary to derive its statistical parameters for completing the comprehensive analytical modeling. Among various statistical parameters, the variance has been chosen as it determines the spread of the variation in a convenient manner. Based on the traditional analytical model of statistical variation in MOSFET's parameter [19], the variances of ΔC_g, $Var[\Delta C_g]$ can be analytically obtained as follows [15]

$$Var[\Delta C_g] = \frac{8q4N_{eff}W_{dep}WL}{\varepsilon_0^2k^2T^2C_{dep}^2}\left[\exp\left[-\frac{V_{ds}}{kT/q}\right]-1\right]^{-2}\left[V_{gs}-V_{FB}-\phi_s-N_{eff}W_{dep}\right]^2 \tag{10}$$

where ε_0 stands for the permittivity of free space. At this point, it can be seen that the comprehensive analytical model of ΔC_g proposed in [15] is composed of Eqs. (8) and (10) where the latter has been derived based on the former. In [15], $(Var[\Delta C_g])^{0.5}$ calculated by using the proposed model has been compared to its 65 nm CMOS technology-based benchmarks obtained by using the Monte Carlo simulation for verification where strong agreements between the model-based $(Var[\Delta C_g])^{0.5}$ and the benchmark have been found. The average deviation from the benchmark obtained from the entire range of V_{gs} used for simulation given by 0–100 mV has been found to be 9.42565 and 8.91039% for N-type and P-type MOSFET-based comparisons, respectively [15].

Later, an improved model of ΔC_g has been proposed in [16] where the physical level differences between N-type and P-type MOSFETs, e.g., carrier type, etc., has also been taken into account. Such model is composed of the following equations

$$\begin{aligned}\Delta C_{gN} = {} & 2\left[\sqrt{\frac{W}{C_{dep}}\frac{LC_{ox}}{kT/q}}\right]^2\left[\exp\left[-\frac{V_{ds}}{kT/q}\right]-1\right]^{-1}\left[V_{gs}-V_{FB}-2\phi_F-C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_a\left(2\phi_F+V_{sb}\right)}\right] \\ & \times\left[V_t-V_{FB}-2\phi_F-C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_a\left(2\phi_F+V_{sb}\right)}\right]\end{aligned} \tag{11}$$

$$\begin{aligned}\Delta C_{gP} = {} & 2\left[\sqrt{\frac{W}{C_{dep}}\frac{LC_{ox}}{kT/q}}\right]^2\left[\exp\left[-\frac{V_{ds}}{kT/q}\right]-1\right]^{-1}\left[V_{gs}-V_{FB}+\left|2\phi_F\right|+C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_d\left(\left|2\phi_F\right|-V_{sb}\right)}\right] \\ & \times\left[V_t-V_{FB}+\left|2\phi_F\right|+C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_d\left(\left|2\phi_F\right|-V_{sb}\right)})\right]\end{aligned} \tag{12}$$

$$\begin{aligned}Var\left[\Delta C_{gN}\right] = {} & \frac{12q^6N_{eff}W_{dep}WL^3}{C_{dep}^2}\left(\frac{C_{ox}}{kT}\right)^4\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{-2} \\ & \frac{\left[V_{gs}-V_{FB}-2\phi_F-C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_a\left(2\phi_F+V_{sb}\right)}\right]^2}{V_t^{-1}\left[V_{FB}+2\phi_F+C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_a\left(2\phi_F+V_{sb}\right)}\right]}\end{aligned} \tag{13}$$

$$\begin{aligned}Var\left[\Delta C_{gP}\right] = {} & \frac{12q^6N_{eff}W_{dep}WL^3}{C_{dep}^2}\left(\frac{C_{ox}}{kT}\right)^4\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{-2} \\ & \frac{\left[V_{gs}-V_{FB}+\left|2\phi_F\right|+C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_d\left(\left|2\phi_F\right|-V_{sb}\right)}\right]^2}{V_t^{-1}\left[V_{FB}-\left|2\phi_F\right|-C_{ox}^{-1}\sqrt{2q\varepsilon_{S_i}N_d\left(\left|2\phi_F\right|-V_{sb}\right)}\right]}\end{aligned} \tag{14}$$

where ΔC_{gN} and ΔC_{gP} are ΔC_g of N-type and P-type MOSFETs, respectively. Moreover, N_a, N_d, V_{sb}, and ϕ_F denote acceptor doping density, donor doping density, source to body voltage, and Fermi potential, respectively [16]. Also, it is noted that Eqs. (13) and (14) have been, respectively, derived by using Eqs. (11) and (12) based on the up-to-date analytical model of statistical variation in MOSFET's parameter [20] instead of the traditional one.

In [16], a verification similar to that of [15] has been made, i.e., $(Var[\Delta C_{gN}])^{0.5}$ and $(Var[\Delta C_{gP}])^{0.5}$ have been, respectively, compared with their 65 nm CMOS technology-based benchmarks. Both $(Var[\Delta C_{gN}])^{0.5}$ and $(Var[\Delta C_{gP}])^{0.5}$ have been calculated by using the proposed model, and the benchmarks have been obtained from the Monte Carlo simulation. The comparison results have been redrawn here in **Figures 2** and **3** where strong agreements with their benchmarks of the model-based $(Var[\Delta C_{gN}])^{0.5}$ and $(Var[\Delta C_{gP}])^{0.5}$ can be seen for the whole range of V_{gs}. The average deviations determined from such range have been found to be 8.45033 and 6.53211%, respectively [16], which are lower than those of the previous model proposed in [15]. Therefore, the model proposed in [16] has also been found to be more accurate than its predecessor

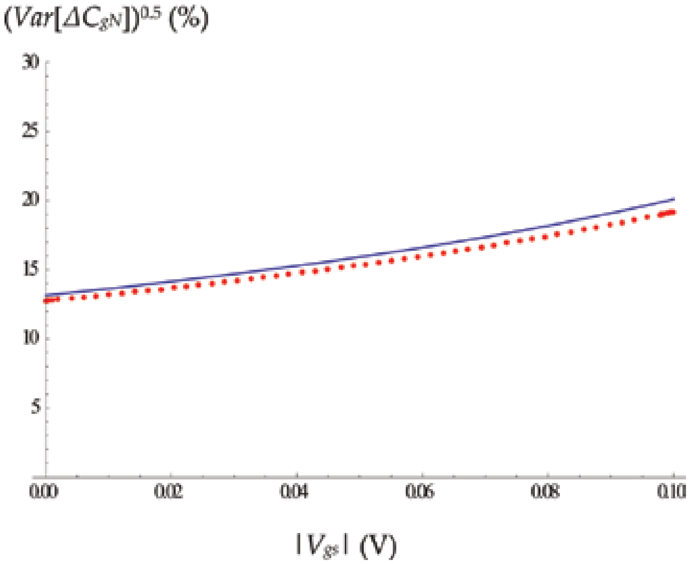


Figure 2. Comparative plot of the model-based $(Var[\Delta C_{gN}])^{0.5}$ (line) and the Monte Carlo simulation-based $(Var[\Delta C_{gN}])^{0.5}$ (dotted) with respect to V_{gs} [16].

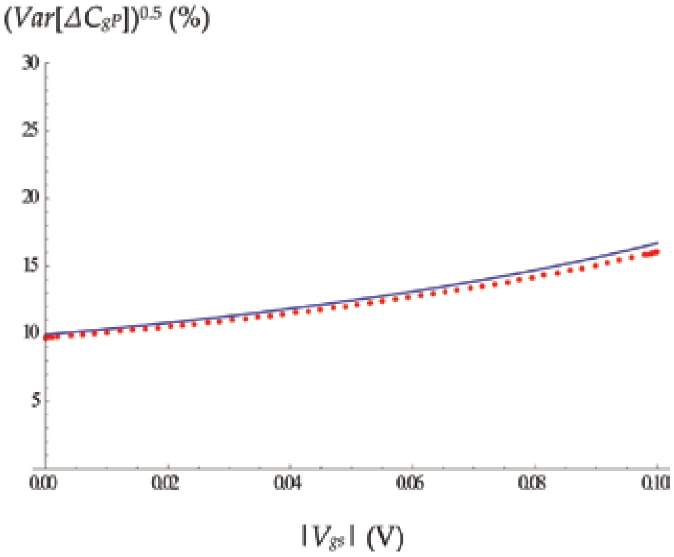


Figure 3. Comparative plot of the model-based $(Var[\Delta C_{gP}])^{0.5}$ (line) and the Monte Carlo simulation-based $(Var[\Delta C_{gP}])^{0.5}$ (dotted) with respect to V_{gs} [16].

[15] apart from being more detailed as the physical level differences between N-type and P-type MOSFETs have also been taken into account.

4. Variation in transition frequency (f_T)

Apart from that of ΔC_g, the comprehensive analytical model of variation in f_T of subthreshold MOSFET, Δf_T has also been proposed in [16]. Before reviewing such model, it is worthy to show the definition of f_T and its comprehensive analytical expression derived in [16]. According to [21], f_T can be defined as the frequency at which the small-signal current gain of the device drops to unity, while the source and drain terminals are held at ground and can be related to C_g by the following equation [13]

$$f_T = \frac{g_m}{2\pi C_g} \tag{15}$$

By using Eqs. (2) and (6), the following comprehensive analytical expression of f_T can be obtained [16]

$$f_T = \frac{3}{2}\left[\frac{\mu C_{dep}^2 (kT/q)^3}{2n\pi L^3 C_{ox}^2}\right]\left[1 - \exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^2 \left[\frac{\exp\left[\frac{2q}{nkT}\left(V_{gs} - V_t\right)\right]}{3\left(V_{gs} - V_t\right)^2 - \frac{q}{nkT}\left(V_{gs} - V_t\right)^3}\right] \tag{16}$$

Similar to ΔC_g, Δf_T can be mathematically defined as [16]

$$\Delta f_T \triangleq f_T(\Delta V_t, \Delta W, \Delta L, \ldots) - f_T \tag{17}$$

where f_T stands for the nominal transition frequency in this context.

By also keeping in mind that ΔV_t is the most influential, the following comprehensive analytical expression of Δf_T has been proposed in [16] where the aforesaid physical level differences between N-type and P-type MOSFETs have also been taken into account.

$$\Delta f_{TN} = \frac{\mu C_{dep}^2 (kT/q)^3 \left[1 - \exp\left[-\frac{V_{ds}}{(kT/q)}\right]\right]^2 \left(V_{FB} + 2\phi_F + C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_a\left(2\phi_F + V_{sb}\right)} - V_t\right)}{\pi n L^3 C_{ox}^2 \left(V_{gs} - V_{FB} - 2\phi_F - C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_a\left(2\phi_F + V_{sb}\right)}\right)^3} \tag{18}$$

$$\Delta f_{TP} = \frac{\mu C_{dep}^2 (kT/q)^3 \left[1 - \exp\left[-\frac{V_{ds}}{(kT/q)}\right]\right]^2 \left(V_{FB} - |2\phi_F| - C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_d\left(|2\phi_F| - V_{sb}\right)} - V_t\right)^{-1}}{\pi n L^3 C_{ox}^2 \left(V_{gs} - V_{FB} + |2\phi_F| + C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_d\left(|2\phi_F| - V_{sb}\right)}\right)^3} \tag{19}$$

It is noted that Δf_{TN} and Δf_{TP} are Δf_T of N-type and P-type MOSFETs, respectively. By also using the up-to-date analytical model of statistical variation in MOSFET's parameter, we have [16]

$$Var\left[\Delta f_{TN}\right] = \frac{\mu^2 C_{dep}^4 (kT)^6 q^{-4} N_{eff} W_{dep} \left[1 - \exp\left[-\frac{Vds}{kT/q}\right]\right]^4 V_t^{-1} \left(V_{FB} + 2\phi_F + C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_a \left(2\phi_F + V_{sb}\right)}\right)}{3\pi^2 n^2 W L^7 C_{ox}^6 \left(V_{gs} - V_{FB} - 2\phi_F - C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_a \left(2\phi_F + V_{sb}\right)}\right)^6} \tag{20}$$

$$Var\left[\Delta f_{TP}\right] = \frac{\mu^2 C_{dep}^4 (kT)^6 q^{-4} N_{eff} W_{dep} \left[1 - \exp\left[-\frac{Vds}{kT/q}\right]\right]^4 V_t^{-1} \left(V_{FB} - \left|2\phi_F\right| - C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_d \left(\left|2\phi_F\right| - V_{sb}\right)}\right)}{3\pi^2 n^2 W L^7 C_{ox}^6 \left(V_{gs} - V_{FB} + \left|2\phi_F\right| + C_{ox}^{-1}\sqrt{2q\varepsilon_{Si} N_d \left(\left|2\phi_F\right| - V_{sb}\right)}\right)^6} \tag{21}$$

At this point, it can be stated that the comprehensive analytical model of Δf_T proposed in [16] is composed of Eqs. (18), (19), (20), and (21). For verification, $(Var[\Delta f_{TN}])^{0.5}$ and $(Var[\Delta f_{TP}])^{0.5}$ calculated by using the proposed model have also been compared with their corresponding 65 nm CMOS technology-based benchmarks obtained from the Monte Carlo simulation. The results have been redrawn here in **Figures 4** and **5** where strong agreements to the benchmarks of the model-based $(Var[\Delta f_{TN}])^{0.5}$ and $(Var[\Delta f_{TP}])^{0.5}$ can be observed. The average deviations have been found to be 8.22947 and 6.25104%, respectively [16]. Moreover, it has been proposed in [16] that there exists a very strong statistical relationship between ΔC_g and Δf_T of any certain subthreshold MOSFET as it has been found by using the proposed model that the magnitude of the statistical correlation coefficient of ΔC_g and Δf_T is unity for both N-type and P-type devices.

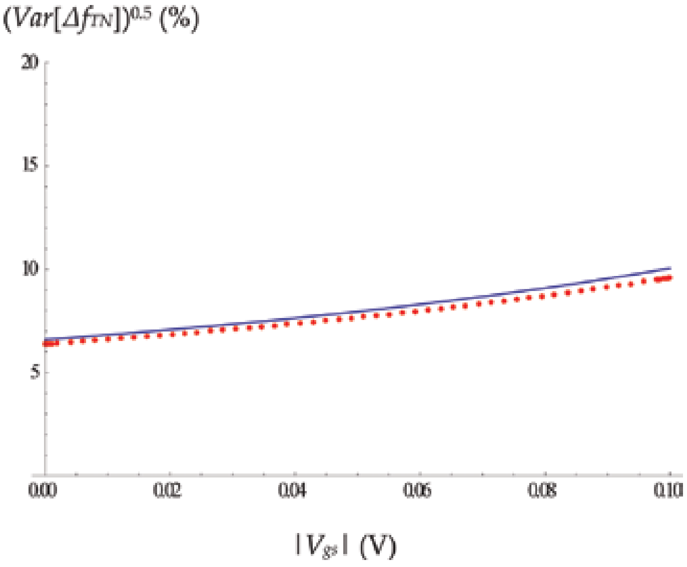


Figure 4. Comparative plot of the model-based $(Var[\Delta f_{TN}])^{0.5}$ (line) and the Monte Carlo simulation-based $(Var[\Delta f_{TN}])^{0.5}$ (dotted) with respect to V_{gs} [16].

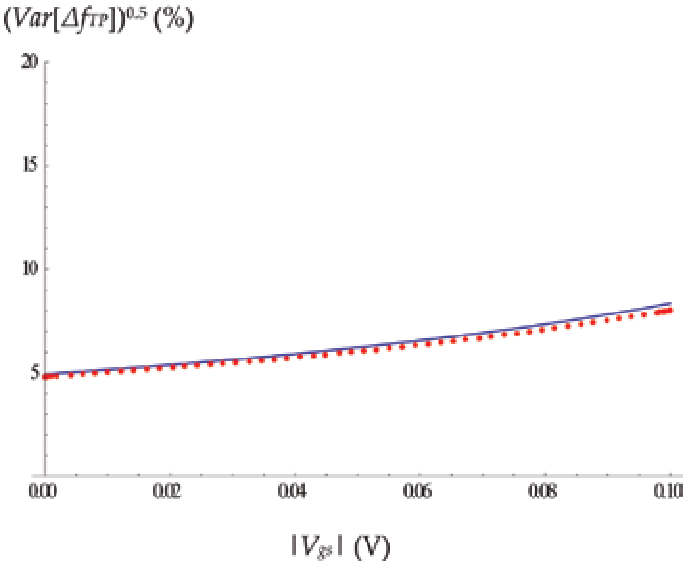


Figure 5. Comparative plot of the model-based $(Var[\Delta f_{TP}])^{0.5}$ (line) and the Monte Carlo simulation-based $(Var[\Delta f_{TP}])^{0.5}$ (dotted) with respect to V_{gs} [16].

5. Variation in maximum frequency of oscillation (f_{max})

Before reviewing the model of variation in f_{max} of subthreshold MOSFET, it is worthy to introduce its definition and mathematical expression. The f_{max}, which takes the effect of the resistance of gate metallization into account, can be defined as the frequency at which the power gain of MOSFET becomes unity. Such gate metallization belonged to the extrinsic part of MOSFET. According to [17], f_{max} can be given under an assumption that C_g is equally divided between drain and source by

$$f_{max} = \frac{1}{4\pi C_g}\sqrt{\frac{2g_m}{R_g}} \tag{22}$$

where R_g stands for the resistance of gate metallization [17].

By substituting g_m and C_g as respectively given by Eqs. (2) and (6) into Eq. (22), we have

$$f_{max} = \frac{\sqrt{\frac{2\mu}{n}\left[\exp\left[-\frac{V_{gs}-V}{nkT/q}\right]\right]^{-1}\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]}}{\frac{4\pi}{3}\sqrt{\frac{W}{C_{dep}R_g}}\left[\frac{L^{2.5}C_{ox}^2}{(kT/q)}\right]\left[3\left(V_{gs}-V_t\right)^2-\frac{\left(V_{gs}-V_t\right)^3}{nkT/q}\right]} \tag{23}$$

Similar to the other variations, Δf_{max} can be mathematically defined as [17]

$$\Delta f_{max} \overset{\Delta}{=} f_{max}(\Delta V_t, \Delta W, \Delta L, \ldots) - f_{max} \tag{24}$$

where f_{max} stands for the nominal maximum frequency of oscillation in this context.

In [17], the comprehensive analytical model of Δf_{max} have been proposed. Such model is composed of the following equations.

$$\begin{aligned}\Delta f_{max} = {} & \frac{1}{\sqrt{2}\pi}\left(\frac{\mu}{nR_g}\right)^{\frac{1}{2}}\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{\frac{1}{2}}\left(\frac{kT}{q}\right)\exp\left[\frac{V_{gs}-V_t}{2nkT/q}\right]\left[\left(\frac{C_{dep}W}{L}\right)^{\frac{1}{2}}\right. \\ & +\left[1-\exp\left[-\tfrac{V_{ds}}{kT/q}\right]\right]^{-1}\left(\tfrac{WL}{C_{dep}}\right)^{\frac{3}{2}}\left(\tfrac{C_{ox}}{kT/q}\right)^{2}\times\left[V_{gs}-V_{FB}-\phi_s-N_{eff}W_{dep}\right] \\ & \left.\times\left[V_t-V_{FB}-\phi_s-N_{eff}W_{dep}\right]\right]\end{aligned} \tag{25}$$

$$\begin{aligned}Var[\Delta f_{max}] = {} & \frac{\mu q 4 N_{eff} W_{dep} W^2}{\pi^2 n C_{dep} R_g \varepsilon_0^2 k^2 T^2}\left[\exp\left[-\frac{V_{ds}}{kT/q}\right]-1\right]^{-1}\exp\left[\frac{V_{gs}-V_t}{nkT/q}\right]\left(\frac{kT}{q}\right)^2 \\ & \left[V_{gs}-V_{FB}-\phi_s-N_{eff}W_{dep}\right]^2\end{aligned} \tag{26}$$

It is noted that Eq. (25) has been derived by also keeping in mind that ΔVt is the most dominant. Moreover, Eq. (26) has been formulated based on Eq. (25) and the traditional model of statistical variation in MOSFET's parameter. The model-based $(Var[\Delta f_{max}])^{0.5}$ has been compared with its 65 nm CMOS technology-based benchmarks obtained by the Monte Carlo simulation for verification. The strong agreements between the model-based $(Var[\Delta f_{max}])^{0.5}$ and the benchmark can be observed from the whole simulated range of V_{gs} given by 0–100 mV. The average deviation has been found to be 9.17682 and 8.51743% for N-type and P-type subthreshold MOSFETs, respectively, [17].

Unfortunately, the model proposed in [17] did not take the physical level differences between N-type and P-type MOSFETs into account. By taking such physical level differences into consideration, we have

$$\begin{aligned}\Delta f_{maxN} = {} & \frac{1}{\sqrt{2}\pi}\left(\frac{\mu}{nR_g}\right)^{\frac{1}{2}}\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{\frac{1}{2}}\left(\frac{kT}{q}\right)\exp\left[\frac{V_{gs}-V_t}{2nkT/q}\right]\left[\left(\frac{C_{dep}W}{L}\right)^{\frac{1}{2}}\right. \\ & +\left[1-\exp\left[-\tfrac{V_{ds}}{kT/q}\right]\right]^{-1}\left(\tfrac{WL}{C_{dep}}\right)^{\frac{3}{2}}\left(\tfrac{C_{ox}}{kT/q}\right)^{2}\times\left[V_{gs}-V_{FB}-2\phi_F-C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_a(2\phi_F+V_{sb})}\right] \\ & \left.\times\left[V_t-V_{FB}-2\phi_F-C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_a(2\phi_F+V_{sb})}\right]\right]\end{aligned} \tag{27}$$

$$\begin{aligned}\Delta f_{maxP} = {} & \frac{1}{\sqrt{2}\pi}\left(\frac{\mu}{nR_g}\right)^{\frac{1}{2}}\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{\frac{1}{2}}\left(\frac{kT}{q}\right)\exp\left[\frac{V_{gs}-V_t}{2nkT/q}\right]\left[\left(\frac{C_{dep}W}{L}\right)^{\frac{1}{2}}\right. \\ & +\left[1-\exp\left[-\tfrac{V_{ds}}{kT/q}\right]\right]^{-1}\left(\tfrac{WL}{C_{dep}}\right)^{\frac{3}{2}}\left(\tfrac{C_{ox}}{kT/q}\right)^{2}\times\left[V_{gs}-V_{FB}+|2\phi_F|+C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_d(|2\phi_F|-V_{sb})}\right] \\ & \left.\times\left[V_t-V_{FB}+|2\phi_F|+C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_d(|2\phi_F|-V_{sb})}\right]\right]\end{aligned} \tag{28}$$

where Δf_{maxN} and Δf_{maxP} are Δf_{max} of N-type and P-type MOSFETs, respectively. By using the up-to-date analytical model of statistical variation in MOSFET's parameter, we have

$$\begin{aligned} Var\left[\Delta f_{maxN}\right] = & \frac{3q^2 N_{eff} W_{dep} W^{-3} L^{-1}\left(\mu/nR_g\right)\left(kT/q\right)^2}{2\pi^2 V_t^{-1}\left[V_{FB}+2\phi_F+C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_a\left(2\phi_F+V_{sb}\right)}\right]}\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right] \\ & \left[\exp\left[\frac{V_{gs}-V_t}{2nkT/q}\right]\right]^2 \times \left[\left(\frac{C_{dep}W}{L}\right)^{\frac{1}{2}}+\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{-1}\left(\frac{WL}{C_{dep}}\right)^{\frac{3}{2}}\left(\frac{C_{ox}}{kT/q}\right)^2 \right. \\ & \left. \times\left[V_{gs}-V_{FB}-2\phi_F-C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_a\left(2\phi_F+V_{sb}\right)}\right]\right] \end{aligned} \tag{29}$$

$$\begin{aligned} Var\left[\Delta f_{maxP}\right] = & \frac{3q^2 N_{eff} W_{dep} W^{-3} L^{-1}\left(\mu/nR_g\right)\left(kT/q\right)^2}{2\pi^2 V_t^{-1}\left[V_{FB}+2\phi_F+C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_a\left(2\phi_F+V_{sb}\right)}\right]}\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right] \\ & \left[\exp\left[\frac{V_{gs}-V_t}{2nkT/q}\right]\right]^2 \times \left[\left(\frac{C_{dep}W}{L}\right)^{\frac{1}{2}}+\left[1-\exp\left[-\frac{V_{ds}}{kT/q}\right]\right]^{-1}\left(\frac{WL}{C_{dep}}\right)^{\frac{3}{2}}\left(\frac{C_{ox}}{kT/q}\right)^2 \right. \\ & \left. \times\left[V_{gs}-V_{FB}+\left|2\phi_F\right|+C_{ox}^{-1}\sqrt{2q\varepsilon_{Si}N_d\left(\left|2\phi_F\right|-V_{sb}\right)}\right]\right] \end{aligned} \tag{30}$$

At this point, it can be seen that the improved model of Δf_{max} is composed of Eqs. (27), (28), (29), and (30). For verification, the model-based $(Var[\Delta f_{maxN}])^{0.5}$ and $(Var[\Delta f_{maxP}])^{0.5}$ have been compared with their corresponding 65 nm CMOS technology-based benchmarks obtained by

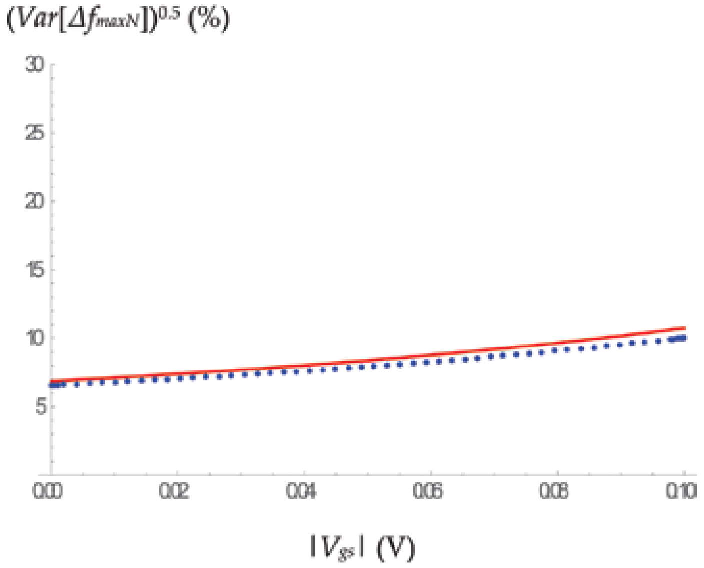


Figure 6. Comparative plot of the model-based $(Var[\Delta f_{maxN}])^{0.5}$ (line) and the Monte Carlo simulation-based $(Var[\Delta f_{maxN}])^{0.5}$ (dotted) with respect to V_{gs}.

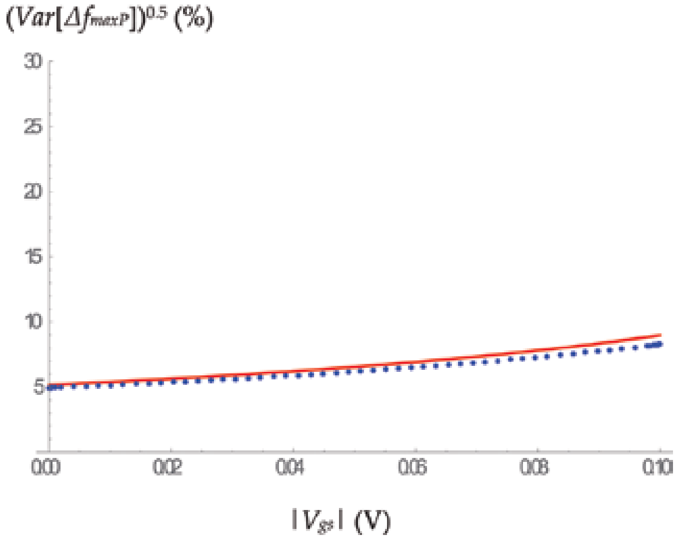


Figure 7. Comparative plot of the model-based $(Var[\Delta f_{maxP}])^{0.5}$ (line) and the Monte Carlo simulation-based $(Var[\Delta f_{maxP}])^{0.5}$ (dotted) with respect to V_{gs}.

using the Monte Carlo simulation. The results are as shown in **Figures 6** and **7** where strong agreements to the benchmarks of the model-based $(Var[\Delta f_{maxN}])^{0.5}$ and $(Var[\Delta f_{maxP}])^{0.5}$ can be observed. The average deviations from the benchmarks have been found to be 6.11788 and 5.85574% for $(Var[\Delta f_{maxN}])^{0.5}$ and $(Var[\Delta f_{maxP}])^{0.5}$, respectively, which are lower than those of the model proposed in [17]. Therefore, our improved model Δf_{max} is also more accurate than the previous one apart from being more detailed as the physical level differences between N-type and P-type MOSFETs have also been taken into account.

Before proceeding further, it should be mentioned here that C_g has more severe variations compared to the other high-frequency characteristics and the P-type subthreshold MOSFET is more robust than the N-type as can be seen from **Figures 2–7**. Moreover, it can be implied that there exists a strong correlation between Δf_{max} and Δf_T as f_{max} is related to f_T by Eq. (31). An implication of strong correlation between Δf_{max} and ΔC_g can be similarly obtained by observing Eq. (22) that is given as

$$f_{max} = \frac{f_T}{\sqrt{2 g_m R_g}} \tag{31}$$

6. Some interesting issues

6.1. Statistical/variability aware design trade-offs

For the optimum statistical/variability aware design of any MOSFET-based VHF circuit, ΔC_g, Δf_T, and Δf_{max} must be minimized. It has been found from Eqs. (13), (14), (20), (21), (29), and (30) that $Var[\Delta C_g] \propto L^3$, $Var[\Delta f_T] \propto L^{-7}$ and $Var[\Delta f_{max}] \propto L^{-1}$ for both types of MOSFET. Therefore, it can be seen that shrinking L can reduce ΔC_g of the subthreshold MOSFET of any type with the increasing Δf_T and Δf_{max} as penalties. Moreover, we have also found that $Var\left[\Delta C_g\right] \propto T^{-2}$, $Var\left[\Delta f_T\right] \propto T^6$, and $Var\left[\Delta f_{max}\right] \propto T^2$. This means that we can reduce Δf_T and Δf_{max} by lowering T with higher ΔC_g as a cost. These design trade-offs must be taken into account in the statistical/variability aware design of any subthreshold MOSFET-based VHF circuits/systems.

6.2. Variation in any high-frequency parameter

Occasionally, determining the variation in other high-frequency parameters apart from C_g, f_T, and f_{max} e.g., bandwidth, f_{BW}, etc., has been found to be necessary. The determination of variation in f_{BW} as a function of Δf_T has been shown in [16]. In general, let any high-frequency parameter of the subthreshold MOSFET be P, the amount of its variation, ΔP, can be determined given the amounts of ΔC_g, Δf_T, and Δf_{max} if P depends on C_g, f_T, and f_{max}. It is noted that the amounts of ΔC_g, Δf_T, and Δf_{max} can be predetermined by using the reviewed comprehensive analytical models. Mathematically, ΔP can be expressed in terms of ΔC_g, Δf_T, and Δf_{max} as follows

$$\Delta P = \left(\frac{\partial P}{\partial C_g}\right)\Delta C_g + \left(\frac{\partial P}{\partial f_T}\right)\Delta f_T + \left(\frac{\partial P}{\partial f_{max}}\right)\Delta f_{max} \tag{32}$$

Therefore, the variance of ΔP, $Var[\Delta P]$ can be given by keeping the aforementioned strong statistical relationships among ΔC_g, Δf_T, and Δf_{max} in mind as follows

$$\begin{aligned} Var[\Delta, P] = \; & \left(\frac{\partial P}{\partial C_g}\right)^2 Var\left[\Delta_{Cg}\right] + \left(\frac{\partial P}{\partial f_T}\right)^2 Var\left[\Delta f_T\right] + \left(\frac{\partial P}{\partial f_{max}}\right)^2 Var\left[\Delta f_{max}\right] \\ & + 2\left(\frac{\partial P}{\partial C_g}\right)\left(\frac{\partial P}{\partial f_T}\right)\sqrt{Var\left[\Delta_{Cg}\right]Var\left[\Delta f_T\right]} + 2\left(\frac{\partial P}{\partial C_g}\right)\left(\frac{\partial P}{\partial f_{max}}\right)\sqrt{Var\left[\Delta_{Cg}\right]Var\left[\Delta f_{max}\right]} \\ & + 2\left(\frac{\partial P}{\partial f_T}\right) \times \left(\frac{\partial P}{\partial f_{max}}\right)\sqrt{Var\left[\Delta f_T\right]Var\left[\Delta f_{max}\right]} \end{aligned} \tag{33}$$

Noted also that the $Var[\Delta C_g]$, $Var[\Delta f_T]$, and $Var[\Delta f_{max}]$ can be known by applying those reviewed models.

6.3. High-frequency parameter mismatches

The amount of mismatches in C_g, f_T, and f_{max} of multiple subthreshold MOSFETs can be determined by applying those reviewed comprehensive analytical models of ΔC_g, Δf_T, and Δf_{max} even though they are dedicated to a single device. As an illustration, the mismatches in C_g, f_T, and f_{max} of two deterministically identical subthreshold MOSFETs, i.e., M1 and M2, will be determined. Traditionally, the magnitude of mismatch can be measured by using its variance [22]. Let the mismatches in C_g, f_T, and f_{max} of M1 and M2 be denoted by ΔC_{g12}, Δf_{T12}, and Δf_{max12}, respectively, their variances, i.e., $Var[\Delta C_{g12}]$, $Var[\Delta f_{T12}]$, and $Var[\Delta f_{max12}]$, can be respectively related to $Var[\Delta C_g]$, $Var[\Delta f_T]$, and $Var[\Delta f_{max}]$ of M1 and M2, which can be determined by using those reviewed models, via the following equations

$$Var\left[\Delta C_{g12}\right] = Var\left[\Delta C_{g1}\right] + Var\left[\Delta C_{g2}\right] - 2\rho_{\Delta C_{g1}\Delta C_{g2}}\sqrt{Var\left[\Delta C_{g1}\right]Var\left[\Delta C_{g2}\right]} \tag{34}$$

$$Var\left[\Delta f_{T12}\right] = Var\left[\Delta f_{T1}\right] + Var\left[\Delta f_{T2}\right] - 2\rho_{\Delta C_{g1}\Delta C_{g2}}\sqrt{Var\left[\Delta f_{T1}\right]Var\left[\Delta f_{T2}\right]} \tag{35}$$

$$Var\left[\Delta f_{max12}\right] = Var\left[\Delta f_{max1}\right] + Var\left[\Delta f_{max2}\right] - 2\rho_{\Delta C_{g1}\Delta C_{g2}}\sqrt{Var\left[\Delta f_{max1}\right]Var\left[\Delta f_{max2}\right]} \tag{36}$$

It is noted that ΔC_{gi}, Δf_{Ti}, Δf_{maxi}, $Var[\Delta C_{gi}]$, $Var[\Delta f_{Ti}]$, and $Var[\Delta f_{maxi}]$, respectively, denote ΔC_g, Δf_T, Δf_{max}, $Var[\Delta C_g]$, $Var[\Delta f_T]$, and $Var[\Delta f_{max}]$ of Mi where $\{i\} = \{1, 2\}$. Moreover, ρ_{XY} stands for the correlation coefficient of X and Y where $\{X\} = \{\Delta C_{g1}, \Delta f_{T1}, \Delta f_{max1}\}$ and $\{Y\} = \{\Delta C_{g2}, \Delta f_{T2}, \Delta f_{max2}\}$. For closely spaced MOSFETs with positive correlation, ρ_{XY}can be given by 1 as the statistical correlation between closely spaced devices is very strong [22]. As a result, the mismatches are maximized. If the negative correlation is assumed on the other hand, ρ_{XY} become -1 and the mismatches are minimized [16]. For distanced devices, we have, $\rho_{XY} = 0$ as the correlation is very weak and can be neglected.

If we assume that both M1 and M2 are statistically identical, we have $Var[\Delta C_{g1}] = Var[\Delta C_{g2}] = Var[\Delta C_g]$, $Var[\Delta f_{T1}] = Var[\Delta f_{T2}] = Var[\Delta f_T]$, and $Var[\Delta f_{max1}] = Var[\Delta f_{max2}] = Var[\Delta f_{max}]$. Thus, Eqs. (34), (35), and (36) become

$$Var\left[\Delta C_{g12}\right] = 2Var\left[\Delta C_g\right]\left(1 - \rho_{\Delta C_{g1}\Delta C_{g2}}\right) \tag{37}$$

$$Var\left[\Delta f_{T12}\right] = 2Var\left[\Delta f_{T1}\right]\left(1 - \rho_{\Delta f_{T1}\Delta f_{T2}}\right) \tag{38}$$

$$Var\left[\Delta f_{max12}\right] = 2Var\left[\Delta f_{max1}\right]\left(1 - \rho_{\Delta f_{max1}\Delta f_{max2}}\right) \tag{39}$$

From these equations, it can be seen that $Var[\Delta C_{g12}]$, $Var[\Delta f_{T12}]$, and $Var[\Delta f_{max12}]$ can all be approximately given by 0 if those statistically identical devices are closely spaced and positively correlated as all ρ_{XY}'s are given by 1. This implies that the high-frequency parameter mismatches of statistically identical, closely spaced, and positively correlated subthreshold MOSFETs can be neglected.

6.4. Variation in any VHF circuit/system

By using the reviewed models, the variation in the crucial parameter of any subthreshold MOSFET-based VHF circuit/system can be analytically formulated. As a case study, the subthreshold MOSFET-based Wu current-reuse active inductor proposed in [1] will be considered. This active inductor can be depicted as shown in **Figure 8**. According to [1], the inductance, l, of this active inductor can be given by

$$l = \frac{C_{g1}}{g_{m1}g_{m2}} \tag{40}$$

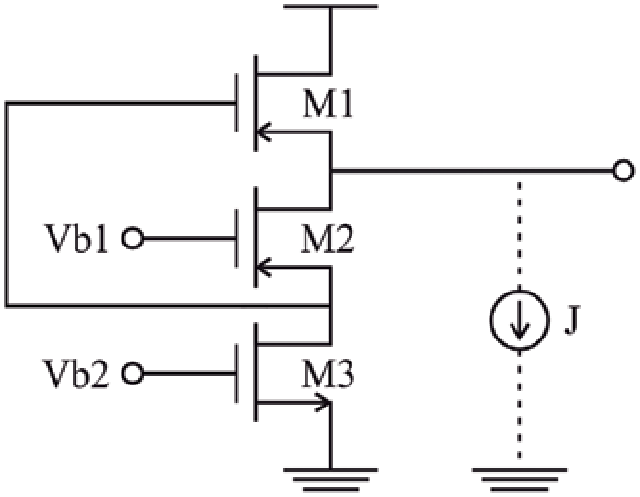


Figure 8. Wu current-reuse active inductor [1].

where C_{g1}, g_{m1}, and g_{m2} are gate capacitance of M1, transconductance of M1, and transconductance of M2, respectively.

By using Eq. (40), the variation in l, Δl due to the variation in C_{g1}, ΔC_{g1} can be immediately given by [16]

$$\Delta l = \frac{\Delta C_{g1}}{g_{m1} g_{m2}} \tag{41}$$

Therefore, we have the following relationship between the variances of Δl and ΔC_{g1}

$$Var[\Delta l] = \frac{Var\left[\Delta C_{g1}\right]}{g_{m1} g_{m2}} \tag{42}$$

It is noted that $Var[\Delta C_{g1}]$ can be determined by using those reviewed models. It can also be seen that $Var[\Delta l] \propto Var\left[\Delta C_{g1}\right]$and $Var[\Delta l] \propto 1/g_{m1} g_{m2}$ [16]. Therefore, it is far more convenient to minimize Δl by reducing g_{m1} and g_{m2} as they are electronically controllable unlike ΔC_{g1}, which must be minimized at the physical level by lowering L as stated above.

6.5. Reduced computational effort simulation

If we let the key parameter of any subthreshold MOSFET-based VHF circuit/system with M MOSFETs under consideration be Z, its variance, $Var[Z]$, which is the desired statistical/variability aware simulation result, can be given by.

$$
\begin{aligned}
Var[Z] = {} & \sum_{i=1}^{M}\left[\left(S^{Z}_{C_{gi}}\right)^{2}\sigma^{2}_{\Delta C_{gi}} + \left(S^{Z}_{f_{Ti}}\right)^{2}\sigma^{2}_{\Delta f_{Ti}} + \left(S^{Z}_{f_{maxi}}\right)^{2}\sigma^{2}_{\Delta f_{maxi}}\right] \\
& + \sum_{\substack{i=1 \\ i\neq j}}^{M}\sum_{j=1}^{M}\left[\left(S^{Z}_{C_{gi}}\right)\left(S^{Z}_{C_{gj}}\right)\rho_{\Delta C_{gi}\Delta C_{gj}}\sqrt{\sigma^{2}_{\Delta C_{gi}}}\sqrt{\sigma^{2}_{\Delta C_{gj}}} + \left(S^{Z}_{f_{Ti}}\right)\left(S^{Z}_{f_{Tj}}\right)\rho_{\Delta f_{Ti}\Delta f_{Tj}}\sqrt{\sigma^{2}_{\Delta f_{Ti}}}\sqrt{\sigma^{2}_{\Delta f_{Tj}}}\right. \\
& \left. + \left(S^{Z}_{f_{maxi}}\right)\left(S^{Z}_{f_{maxj}}\right)\rho_{\Delta f_{maxi}\Delta f_{maxj}}\sqrt{\sigma^{2}_{\Delta f_{maxi}}}\sqrt{\sigma^{2}_{\Delta f_{maxj}}}\right] \\
& + 2\sum_{i=1}^{M}\sum_{j=1}^{M}\left[\left(S^{Z}_{C_{gi}}\right)\left(S^{Z}_{f_{Ti}}\right)\rho_{\Delta C_{gi}\Delta f_{Tj}}\sqrt{\sigma^{2}_{\Delta C_{gi}}}\sqrt{\sigma^{2}_{\Delta f_{Tj}}} + \left(S^{Z}_{C_{gi}}\right)\left(S^{Z}_{f_{maxj}}\right)\rho_{\Delta C_{gi}\Delta f_{maxj}}\right. \\
& \left. \sqrt{\sigma^{2}_{\Delta C_{gi}}}\sqrt{\sigma^{2}_{\Delta f_{maxj}}} + \left(S^{Z}_{f_{Ti}}\right)\left(S^{Z}_{f_{maxj}}\right)\rho_{\Delta f_{Ti}\Delta f_{maxj}}\sqrt{\sigma^{2}_{\Delta f_{Ti}}}\sqrt{\sigma^{2}_{\Delta f_{maxj}}}\right]
\end{aligned}
\tag{43}
$$

It is noted that the magnitude of ρ_{XY}, where $\{X\} = \{\Delta C_{gi}, \Delta f_{Ti}, \Delta f_{maxi}\}$, $\{Y\} = \{\Delta C_{gj}, \Delta f_{Tj}, \Delta f_{maxj}\}$, and the subscripts i and j refers to the arbitrary ith and jth MOSFET, respectively, in this scenario, approaches 1 when $i = j$ as it determines the correlation of the same device. Moreover, $S^{Z}_{C_{gi}}\left(S^{Z}_{C_{gj}}\right)$, $S^{Z}_{f_{Ti}}\left(S^{Z}_{f_{Tj}}\right)$, and $S^{Z}_{f_{maxi}}\left(S^{Z}_{f_{maxj}}\right)$ denote the sensitivity of Z to C_g, f_T, and f_{max} of ith (jth) MOSFET, respectively. By using Eq. (43) and the reviewed comprehensive analytical models for predetermining all $Var[X]$'s and $Var[Y]$'s, $Var[Z]$ can be numerically determined in a reduced computational effort manner as those sensitivities can be obtained by using the sensitivity analysis [23], which required much less computational effort compared to the conventional Monte Carlo simulation. This is because the circuit/system of interest is needed to be solved only once for obtaining the sensitivities and then $Var[Z]$ can be immediately determined unlike the Monte Carlo simulation that requires numerous runs in order to reach the similar outcome [16]. Therefore, much of the computational effort can be significantly reduced.

7. Conclusion

In this chapter, the comprehensive analytical models of ΔC_g, Δf_T, and Δf_{max} of subthreshold MOSFET, which serves as the basis of many VHF circuits/systems, have been reviewed. Interesting issues related to these models i.e., statistical/variability aware design trade-offs of subthreshold MOSFET-based VHF circuit/system; determination of variation in any high-frequency parameter and mismatch in C_g, f_T, and f_{max}; determination of variation in any subthreshold MOSFET-based VHF circuit/system; and the computationally efficient statistical/variability aware simulation with sensitivity analysis have been discussed. Moreover, a modified version of the comprehensive analytical model of Δf_{max} has also been proposed. This revised model has been found to be more accurate and detailed than the previous one.

Author details

Rawid Banchuin

Address all correspondence to: rawid.ban@siam.edu

Graduated School of Information Technology and Faculty of Engineering,
Siam University, Thailand

References

[1] Yushi Z, Yuan F. Subthreshold CMOS active inductor with applications to low-power injection-locked oscillators for passive wireless microsystems. In: Proceedings of the IEEE International Midwest Symposium on Circuits and System (MWSCAS ′10); August 1–4, 2010. Seatle: IEEE; 2010. pp. 885-888

[2] Perumana BG, Mukhopadhyay R, Chakraborty S, Lee C-H, Laskar J. A low power fully monolithic subthreshold CMOS receiver with integrated LO generation for 2.4 GHz wireless PAN application. IEEE Journal of Solid-State Circuits. 2008;**43**:2229-2238. DOI: 10.1109/JSSC.2008.2004330

[3] Perumana BG, Chakraborty S, Lee C-H, Laskar J. A fully monolithic 260-μW, 1-GHz subthreshold low noise amplifier. IEEE Microwave and Wireless Components Letters. 2005;**15**:428-430. DOI: 10.1109/LMWC.2005.850563

[4] Lee H, Mohammadi S. A 3 GHz subthreshold CMOS low noise amplifier. In: Proceedings of the IEEE Radio Frequency Integrated Circuits Symposium (RFIC′06); June 10–13, 2006. San Francisco: IEEE; 2006. pp. 494-497

[5] Kim S, Choi J, Lee J, Koo B, Kim C, Eum N, Yu H, Jung H. A subthreshold CMOS front-end design for low-power band-III T-DMB/DAB recievers. ETRI Journal. 2011;**33**:969-972. DOI: 10.4218/etrij.11.0211.0055

[6] Masuda H, Kida T, Ohkawa S. Comprehensive matching characterization of analog CMOS circuits. IEICE Transaction on Fundamentals of Electronics, Communications and Computer Sciences. 2009;**E92-A**:966-975. DOI: 10.1587/transfun.E92.A.966

[7] Banchuin R. Process induced random variation models of nanoscale MOS performance: Efficient tool for the nanoscale regime analog/mixed signal CMOS statistical/variability aware design. In: Proceedings of the International Conference on Information and Electronics Engineering (ICIEE ′11); May 28–29, 2011. Bangkok: IACSIT Press; 2011. pp. 6-12

[8] Banchuin R. Complete circuit level random variation models of nanoscale MOS performance. International Journal of Information and Electronic Engineering. 2011;**1**:9-15. DOI: 10.7763/IJIEE.2011.V1.2

[9] Weifeng L, Lingling S. Modeling of current mismatch induced by random dopant fluctuation. Journal of Semiconductors. 2011;**32**:084003-1-084003-5. DOI: 10.1088/1674-4926/32/8/084003

[10] Papatanasiou K. A designer's approach to device mismatch: Theory, modeling, simulation techniques, scripting, applications and examples. Analog Integrated Circuits and Signal Processing. 2006;**48**:95-106. DOI: 10.1007/s10470-066-5367-2

[11] Rao R, Srivastava A, Blaauw D, Sylvester D. Statistical analysis of subthreshold leakage current for VLSI circuits. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2004;**12**:131-139. DOI: 10.1109/TVLSI.2003.821549

[12] Forti F, Wright ME. Measurement of MOS current mismatch in the weak inversion region. IEEE Journal of Solid-State Circuits. 1994;**29**:138-142. DOI: 10.1109/4.272119

[13] Kim H-S, Chung C, Lim J, Park K, Oh H, Kang H-K. Characterization and modeling of RF-performance (f_T) fluctuation in MOSFETs. IEEE Electron Device Letters. 2009;**30**:855-857. DOI: 10.1109/LED.2009.2023826

[14] Banchuin R. Novel complete probabilistic models of random variation in high frequency performance of nanoscale MOSFET. Journal of Electrical and Computer Engineering. 2013;**2013**:1-10. DOI: 10.1155/2013/189436

[15] Banchuin R, Chaisricharoen R. Analytical analysis and modelling of variation in gate capacitance of subthreshold MOSFET. In: Proceedings of the Joint International Conference Information and Communication Technology, Electronic and Electrical Engineering (JICTEE '14); March 5–8, 2014. Chiang Rai: IEEE. 2014. pp. 1-4

[16] Banchuin R. Analysis and comprehensive analytical modeling of statistical variations in subthreshold MOSFET's high frequency characteristics. Advances in Electrical and Electronic Engineering. 2014;**12**:47-57. DOI: 10.15598/aeee.v12i1.909

[17] Banchuin R, Chaisricharoen R. Analytical analysis and modelling of variation in maximum frequency of oscillation of subthreshold MOSFET. In: Proceedings of the Joint International Conference Information and Communication Technology, Electronic and Electrical Engineering (JICTEE '14); March 5–8, 2014. Chiang Rai: IEEE; 2014. pp. 1-4

[18] Kwong J, Chandrakasan AP. Advances in ultra-low-voltage design. IEEE Solid State Circuits Magazines. 2008;**13**:20-27. DOI: 10.1109/N-SSC.2008.4785819

[19] Pelgrom MJM, Duinmaijer ACJ, Welbers APG. Matching properties of MOS transistors. IEEE Journal of Solid-State Circuits. 1989;**24**:1433-1439. DOI: 10.1109/JSSC.1989.572629

[20] Takeuchi K, Nishida A, Hiramoto T. Random fluctuations in scaled MOS devices. In: Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD '09); September 9–11, 2009. San Diego: IEEE; 2009. pp. 1-7

[21] Razavi B. Design of Analog CMOS Integrated Circuits. Boston: McGraw-Hill; 2001. p. 684

[22] Cathignol A, Mennillo S, Bordez S, Vendrame L, Ghibaudo G. Spacing impact on MOSFET mismatch. In: Proceedings of the IEEE International Conference on Microelectronic Test Structure (ICMTS '08); March 24–27, 2008. Edinburgh: IEEE; 2009. pp. 90-94

[23] Cijan G, Tuma T, Burmen A. Modeling and simulation of MOS transistor mismatch. In: Proceedings of the Eurosim Congress on Modeling and Simulation (EUROSIM '07); September 9–13, 2007. Ljubljana: SLOSIM; 2007. pp. 1-8

Section 3

Applications in the Present and Future Eras

Chapter 7

6T CMOS SRAM Stability in Nanoelectronic Era: From Metrics to Built-in Monitoring

Bartomeu Alorda, Gabriel Torrens and Sebastia Bota

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73539

Abstract

The digital technology in the nanoelectronic era is based on intensive data processing and battery-based devices. As a consequence, the need for larger and energy-efficient circuits with large embedded memories is growing rapidly in current system-on-chip (SoC). In this context, where embedded SRAM yield dominate the overall SoC yield, the memory sensitivity to process variation and aging effects has aggressively increased. In addition, long-term aging effects introduce extra variability reducing the failure-free period. Therefore, although stability metrics are used intensively in the circuit design phases, more accurate and non-invasive methodologies must be proposed to observe the stability metric for high reliability systems. This chapter reviews the most extended memory cell stability metrics and evaluates the feasibility of tracking SRAM cell reliability evolution implementing a detailed bit-cell stability characterization measurement. The memory performance degradation observation is focused on estimating the threshold voltage (V_{th}) drift caused by process variation and reliability mechanisms. A novel SRAM stability degradation measurement architecture is proposed to be included in modern memory designs with minimal hardware intrusion. The new architecture may extend the failure-free period by introducing adaptable circuits depending on the measured memory stability parameter.

Keywords: SRAM reliability, process variability, memory cell stability margins, lifetime monitoring

1. Introduction

CMOS technology has been adopted by the digital IC market for a wide range of applications from high-performance computing and graphics to mobile applications, wearable electronics and IoT applications. Technology scaling has been constantly evolving offering new opportunities to adapt each technology node to new challenging applications. Modern multi-core system trends result in a significant percentage of the total die area being dedicated to memory blocks. As larger densities of static memories are embedded inside complex SoC designs, analyzing memory reliability becomes more critical, as it may be an important source of the overall system error rate. For instance, the contribution of the SRAM parameter variability dominates the overall circuit parameter characteristics, including leakage and yield [1]. In addition, a deep knowledge and analysis about the SRAM cell noise margin and the impact of physical parameters variation is therefore becoming a must in modern CMOS designs.

The IC technologies have been constantly and aggressively scaled down due to efficient computation requirements. The critical dimension reduction in poly and diffusion features entails an increase in statistical physical parameters variation in the transistor parameters: threshold voltage (V_{th}), channel length, and mobility [2]. In this sense, embedded SRAM circuits are becoming more vulnerable because memory cells are scaled near the minimum available size in each technology node and the power supply is reduced. In this scenario, memory failures are drastically increasing due to higher device parameter variability, more defect density and new reliability mechanisms [3]. This has a direct impact on many parameters like SRAM performance, bit density, V_{DDmin}, leakage, dynamic power reduction, yield and failure probability. In addition, new reliability mechanisms may produce changes in the initial statistical parameter variability depending on user workload application, boosting the emergence of failures in field like bias temperature instability [4, 5].

The initial memory cell parameter variability profile due to fabrication process is defined using a combination of metrics. The most used ones are: the cell stability metrics (noise margins), the functional access time, the power consumption profile and the minimum V_{DD}. Due to the reliability degradation mechanisms, mainly due to V_{th} drift, the initial profile may change dramatically while the circuit is on field, increasing the functional failure probability and/or lowering the circuit performance profile.

Traditionally, on field circuit reliability effects have been minimized using several techniques at the design step. The first common methodology to reduce vulnerability of memory cells is based on introducing some reliability safety margins by design, in addition to the variability guard bands needed to overcome process variation issues. These margins lead to some cost in terms of performance, consumption or area. A second mitigation alternative has been proposed in the literature based on including operational assistance circuits, like read and write assist circuits introduced in memories to assure fault-free operations [5–7]. In more recent approaches, adaptive solutions are also proposed to mitigate BTI effects recovering the V_{th} drift [8]. These approaches involve memory modification to include additional adaptive circuits that in some scenarios have demonstrated to contribute to increase the functional failure probability [6, 7]. Therefore, in some applications, it may be important to periodically monitor the profile changes to detect which memory cells are likely to fail in the near future, and try to take some decision to avoid the failure [9]. The objective is to reduce the failure in time rate, and to improve the overall system reliability while remaining compatible with assist techniques or improved memory cell designs.

The next sections will review the conventional and novel SRAM noise metrics proposed in the literature and their suitability as observable parameters to estimate the threshold voltage (V_{th}) drift of 6T-based SRAM cells. The stability metrics will be analyzed and compared keeping in mind their suitability to be used in an implementable built-in monitor architecture. It is well known that the V_{th} variability is caused by process variation and reliability mechanisms but the implementation of a direct V_{th} measurement built-in monitor without affecting the memory array performance is difficult due to the need of internal memory-array node accessibility. Therefore, this approach analyses the stability metrics defined in the literature and proposes a built-in monitor architecture taking profit of their feasibility to measure and track the evolution of the memory cell effects due to reliability mechanism by observing the effect of the stability margin drift caused by the V_{th} drift.

2. Static random allocate memory

A typical SRAM is designed as a memory-cell matrix organized in N rows and M columns, see **Figure 1**. The SRAM performs three operations: Hold, Read and Write. The hold operation consists in storing the cell values and remains unaltered while the memory is powered on. The read operation accesses to a specific memory cell to read-out the value stored without destroy it. Finally, the write operation updates the stored value in a concrete memory cell changing the previous value.

During an operation, the row and column decoders translate the memory address into an internal cell matrix position. The row address identifies only one row (shadowed row in **Figure 1**) during a read/write access. The column address selects which specific cell from the selected row is actually read-out or write-in (dark cell in **Figure 1**). Finally, the read/write circuits perform the read/write operation to the selected memory cell.

While the memory cell addressed by the row and column decoder is the "selected" cell because both decoders point out the cell, the rest of the row cells are the "Half-Selected" cells, because only the row decoder is pointing out them but the read/write circuits have not access to those memory cells. In each read/write operation, there are M-1 half-selected cells for each selected cell. The presence of half-selected cells is important to understand why read operation is considered during the cell design as the weakest operation in terms of memory stability [10].

2.1. Memory cell with six transistors

The conventional 6T SRAM memory cell is formed by two cross-coupled CMOS inverters connected to the complementary bit-lines through two pass transistors. **Figure 2** shows this well-known memory cell and the main signals to perform read/write operations. Following

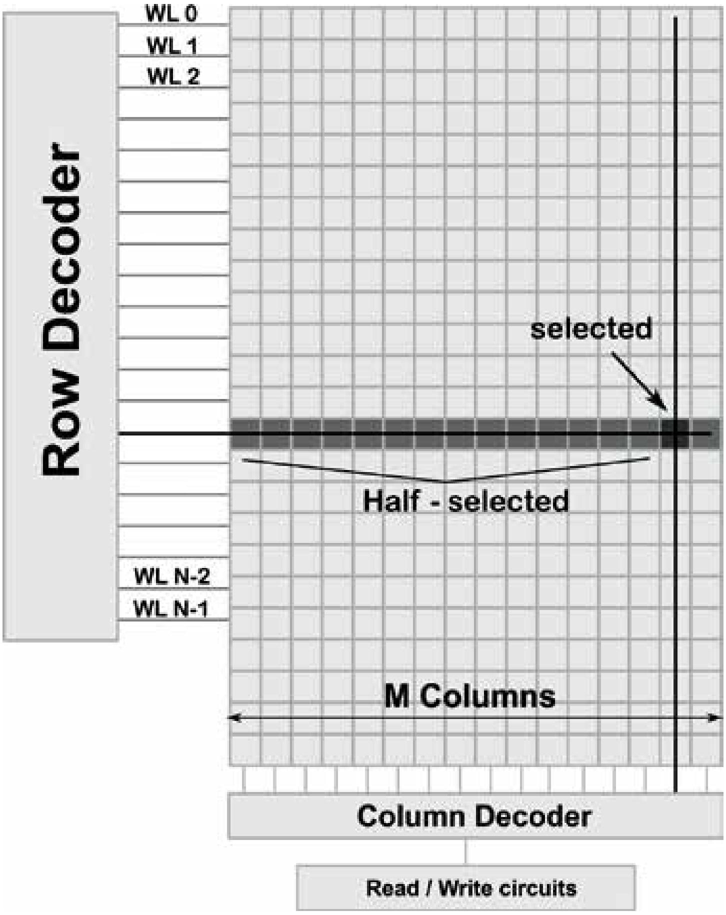


Figure 1. Typical SRAM internal organization with the main parts.

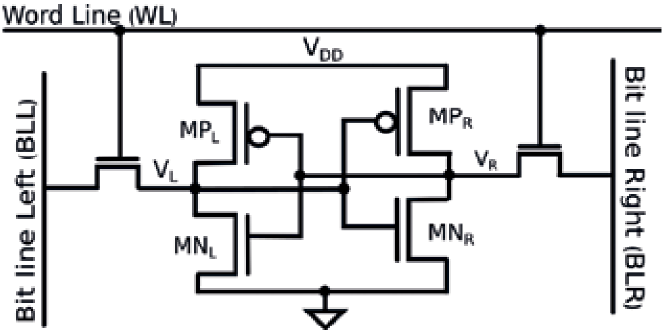


Figure 2. The six CMOS transistors SRAM cell schematic (6T).

the matrix distribution showed in **Figure 1**, all cells in the same row share the word line (WL in **Figure 2**) signal that is connected to the corresponding output of the row decoder. In the same way, all cells in the same column share the bit-lines (BLL and BLR in **Figure 2**) forming the column signals to the read/write circuits, see **Figure 1**.

The access transistors have their gate node connected to the WL to open or close the connection of internal cell nodes (V_L and V_R in **Figure 2**) to the bit-lines (BLL and BLR respectively in **Figure 2**). So, bit-lines act as input/output nodes carrying the data from the selected cell to the read circuits in a read operation, or from write circuits to the selected cell in a write operation.

During the hold period the memory cell maintains a stable value due to the feedback reinforcement of cross-coupled inverters. The WL signal remain low, the BLL and BLR signals are high (are pre-charged waiting for the next operation) and the memory cell has their internals nodes disconnected from the bit-lines.

A read operation is performed connecting the internal memory nodes to both bit-lines precharged to high value. The internal node (V_L or V_R) at low value discharges the connected bit-line (BLL or BLR) though the voltage divider formed by the access transistor and the pull-down transistor (MN_L or MN_R in **Figure 2**). The read circuit senses and amplifies the difference between both bit-lines and the read-out value is latched.

A write operation starts when the write circuits set up the bit-lines with the adequate complementary value to write (BLL with the data value, and BLR with the complementary data value or vice versa). Then, the WL connects the selected memory cell to the bit-lines and the external values force the update of the stored value. In this case, the new value is written though the voltage divider formed by the access transistor and the pull-up transistors (MP_L or MP_R in **Figure 2**). Finally, the memory cell is disconnected from the bit-lines and the new value is stored.

When the memory is performing a write operation on a selected memory cell, there are half-selected memory cells that operate like in a read operation. These cells share the same word-line than the cell which is actually being written, for this reason, their internal nodes are connected to the bit-lines, which sense the cell stored value as in a read operation. In this situation, the cell is in its worst-case cell stability mode as it is reported in [11, 12]. In general, the read operation is more critical that write operation, and the presence of half-selected cells has motivated that the read vulnerability is guaranteed with bigger guard bands in exchange for writability degradation.

3. Memory stability metrics

Stability has been used for years as a useful metric to optimize the design of SRAM cells and predict the effect of parameter variation. Cell stability has been traditionally obtained by computing the noise margins for each memory operation. The noise margins represent the quantification of the cell ability to tolerate a certain presence of noise (in terms of current or voltage). This section will introduce the proposed noise margins considering the kind of nodes involved in the measurement: Internal cell nodes or External cell nodes. In both cases, noise margins will be organized in terms of the measured electrical variable (voltage, current or digital value) and the operation performed (read or write).

3.1. Stability metrics defined on internal cell nodes

The stability defined from the noise metrics on internal cell nodes tries to analyze the impact of voltage or current noise presence on the internal nodes and the maximum range tolerated by the cell. These metrics are widely used for their ability to be implemented in computer simulations at the design phases. It is, therefore, a metric based on the internal nodes ability to tolerate noise in the form of voltage or current.

3.1.1. Based on voltage transfer characteristics

The popular definition for the cell noise margin is obtained using the voltage transfer curves (VTC) considering both read and write operations.

3.1.1.1. Read static noise margin

The read operation is the weakest situation because the cell transistors must be stronger enough to discharge the pre-charged bit-line without flipping its value stored. In a read operation, the memory cell is connected to the bit-lines and the internal nodes are disturbed. The node (V_L or V_R) at low voltage value must remain at this value to maintain the stored value in the cell, while the bit-line is discharged through the pull-down transistor. The static noise margin (SNM) quantifies the maximum amount of voltage noise that can be tolerated at the cross-inverters output nodes without flipping the cell. In the case of a read operation, **Figure 3** shows the node values setup and the noise voltage sources (V_n) to introduce the disturbance simulating a DC sweep between 0 and V_{DD}. The transistor MN_L is trying to maintain the V_L node as low as possible discharging the BLL. The effect of V_n introduces an extra voltage step that produces, if high enough, the loss of the stored value. The maximum extra voltage tolerated by the cell previous to lose the data is defined as the read static noise margin (RSNM).

The graphical method to determine the RSNM uses the static voltage transfer characteristics of the SRAM cell inverters. **Figure 4** superposes the voltage transfer characteristic (VTC) of one cell inverter to the inverse VTC of the other cell inverter. The resulting two-lobed graph is called a "butterfly" curve and is used to determine the RSNM. Its value is defined as the side length of the largest square that can be fitted inside the lobes of the "butterfly" curve [13]. **Figure 4** shows the hold/read operation dependence of the "butterfly" curves. **Figure 4(b)** shows how read operation reduces the noise margin due to the internal nodes connection to the bit-lines.

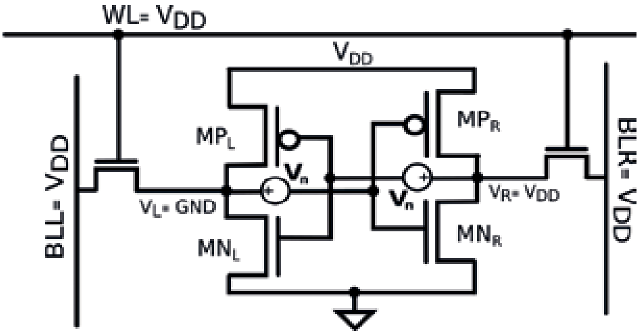


Figure 3. The setup for the read static noise margin definition.

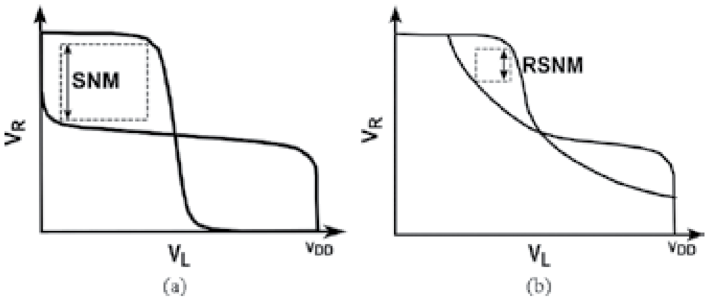


Figure 4. The VTCs of 6T CMOS based memory cell during (a) hold and (b) read access and graphical noise margin representation.

The memory cell designers try to maximize the RSNM value in order to obtain an optimum stability profile during read operation. To maximize the RSNM, the pull-down transistors width (MN_L and MN_R) must be set higher than the access transistors width. The size relationship between the pull-down and access transistors is called cell ratio (CR) and its value is usually designed to be higher than 1.

3.1.1.2. *Write noise margin*

During a write operation the stability is defined considering that the objective of the write operation is to force a new value into the cell, so break the cell stability. In that case, **Figure 5(a)** shows the cell setup considered to measure the write noise margin (WNM). The case where the new value is equal to the stored value is not considered because the cell does not change the internal values. When the cell is written and the value must be updated to the opposite value, both sides of the cross-coupled inverters (V_L and V_R) are confronted to two different situations. The first one, the cell side where internal node is at low value and the bit-line is at high value (V_L and BLL in **Figure 5(a)**), the transistor involved is the pull-down transistor (MN_L). In the second case, the cell side where internal node is at high and the bit-line is at low value (V_R and BLR in **Figure 5(a)**), the transistor involved is the pull-up transistor (MP_R).

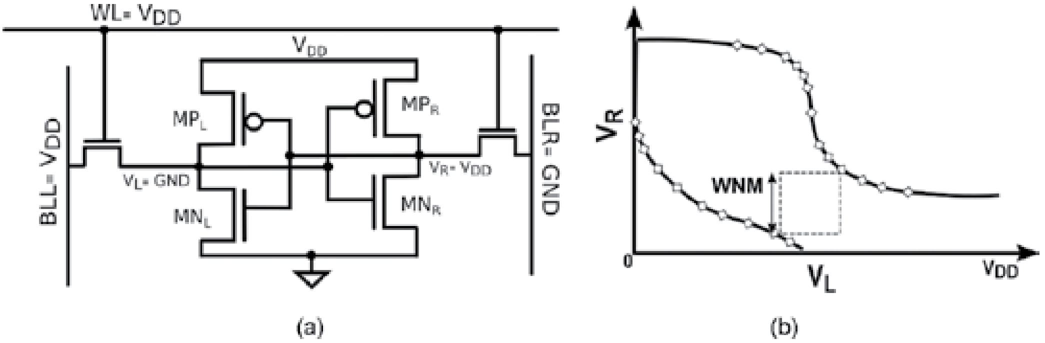


Figure 5. (a) The 6T CMOS based memory cell setup for the write static noise margin and (b) the VTCs during write access and graphical WNM representation.

Therefore, the write margin will be measured by the side of the smallest square embedded between the read and the write VTC measured from the same memory cell at the lower half of the read curves, past the trip point. **Figure 5(b)** shows the graphical representation of WNM.

In this case, the higher WNM is, the lower the writability of the memory cell results. Therefore, memory-cell designers try to slightly reduce the WNM value to obtain an optimum stability profile during write operation without affect the read operation. In order to reduce the WNM, the pull-up transistors width (MP_L and MP_R) must be established lower than the access transistors width. The size relationship between the pull-up and access transistors is called pull-up ratio (PR) and its value is usually designed to be slightly lower than 1.

3.1.2. Based on N-curve

Alternative SRAM noise metrics can be characterized using the N-curve [14]. In this case, the read and write noise margins are defined using the N-curve trip points showed in **Figure 6(b)** as A, B and C. The trip points are obtained using the setup showed in **Figure 6(a)**. The N-curve represents the current (I_{LX}) injected to the internal grounded node when the voltage source (V_{LX}) is swept from 0 to V_{DD}. A pair of current and voltage components defines the read/write noise margins. The N-curve values between trip points A and B, see **Figure 6(b)**, define the read metrics: the static voltage noise margin (SVNM), as the maximum DC voltage tolerable at the internal node previous to flip the memory cell content, and the static current noise margin (SINM), as the maximum DC current value that can be injected in the memory cell before its content changes.

The N-curve values between trip points B and C, see **Figure 6(b)**, define the write metrics: the write trip voltage (WTV), as the DC voltage drop needed to flip the memory cell content, and the write trip current (WTI), as the amount of DC current injected in the memory cell to change its content.

3.2. Stability metrics defined on external cell nodes

The main drawback of stability margin metrics defined on internal cell nodes is that they overestimate read failures and underestimate write failures since it assumes an infinitely long operating duration. However, those parameters are easy to simulate and have a graphical interpretation.

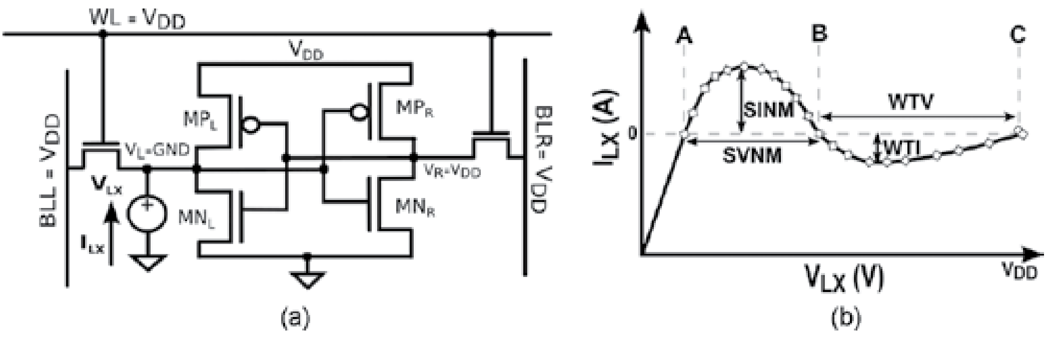


Figure 6. (a) The N-curve measurement setup and (b) the stability parameters defined.

Other metric alternatives are based on direct access to external memory cell nodes: power supply, word-line, and bit-line nodes [15]. In addition, bit-line current or digital stored value measurements are proposed to characterize the stability with less memory array intrusions [16].

3.2.1. Based on a bit-line current observation

These methodologies measure the bit-line current variation while adjusting voltages of bit-lines, word-lines or cell power supply node to obtain read and write stability data.

3.2.1.1. Supply read retention voltage (SRRV)

The SRRV metric based on the observation of I_{BL} estimates the read margin based on the power supply swept. The I_{BL} is monitored to determine when the memory cell losses their ability to remain unaltered and changes the stored value.

Figure 7(a) shows the cell setup values when the bit-lines are set to pre-charged value. The word-line is ramped up until the sudden transition of I_{BL} appears to remain at a low current value. **Figure 7(b)** represents graphically the evolution of I_{BL} versus power supply voltage. When the current drops from its maximum value, the SRRV is defined as the maximum power supply voltage drop to produce a successful read operation. Therefore, SRRV is obtained from the difference between the nominal power supply voltage and the minimum power supply voltage to disturb the stored value.

3.2.1.2. Word-line read retention voltage (WRRV)

The WRRV metric is based on the observation of I_{BL} and estimates the read ability of the cell when the word-line is swept above V_{DD}. The I_{BL} is monitored to determine when the memory cell changes the stored value losing their ability to perform a non-destructive read operation.

Figure 8(a) shows the cell setup values when the bit-lines are pre-charged to V_{DD}. The WL is ramped up until a sudden transition of I_{BL} appears. **Figure 8(b)** represents graphically the evolution of I_{BL} for values of word-line voltage above V_{DD}. When the current drops from its maximum value, the WRRV is defined as the difference between the maximum word-line voltage and the nominal power supply voltage.

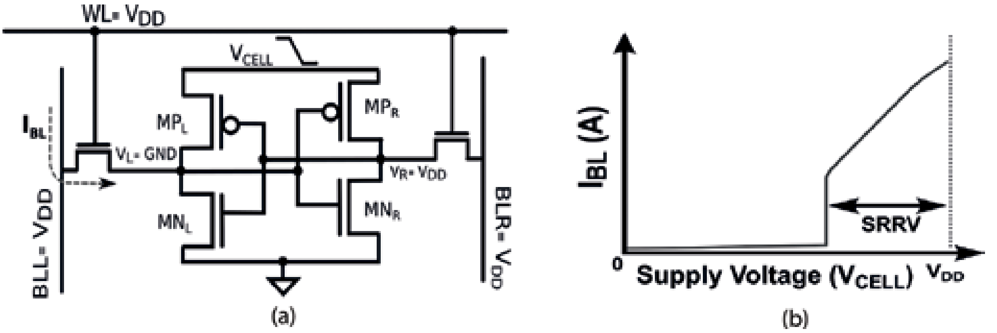


Figure 7. (a) The setup for the supply read retention voltage observation and (b) the graphical SRRV definition from current-voltage transfer curves.

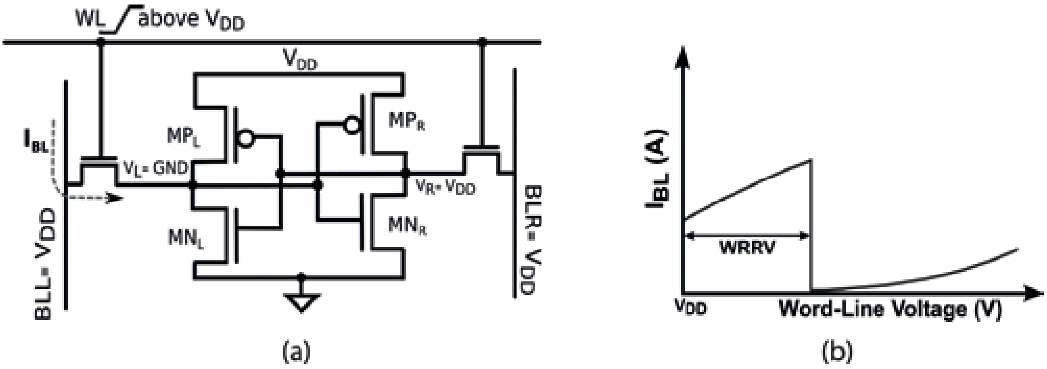


Figure 8. (a) The setup for the word-line read retention voltage observation and (b) the graphical WRRV definition from current-voltage transfer curves.

3.2.1.3. Bit-line write trip voltage (BWTV)

The BWTV estimates the cell writability as the maximum bit-line voltage tolerated by the BLR node (see **Figure 9(a)**) able to flip the cell value during a write cycle.

Figure 9(a) shows the cell setup to perform the margin measurement when it is initialized to store a '0' (V_L retains the '0' and V_R the '1'). The word-line (WL) and the left bit-line (BLL) are biased to V_{DD}, while the right bit-line (BLR) is ramped low from V_{DD}. The current measured on BLL node (I_{BL}) is monitored expecting a sudden drop (see **Figure 9(b)**). When this condition occurs, it indicates a successful write operation and defines the lower bit-line voltage tolerable by the cell. **Figure 9(b)** shows the bit-line current waveform and graphically shows the noise margin.

3.2.1.4. Word-line write trip voltage (WWTV)

The WWTV metric based on the observation of I_{BL} estimates the write margin based on word-line sweep. The I_{BL} is monitored to determine when the memory cell changes the stored value.

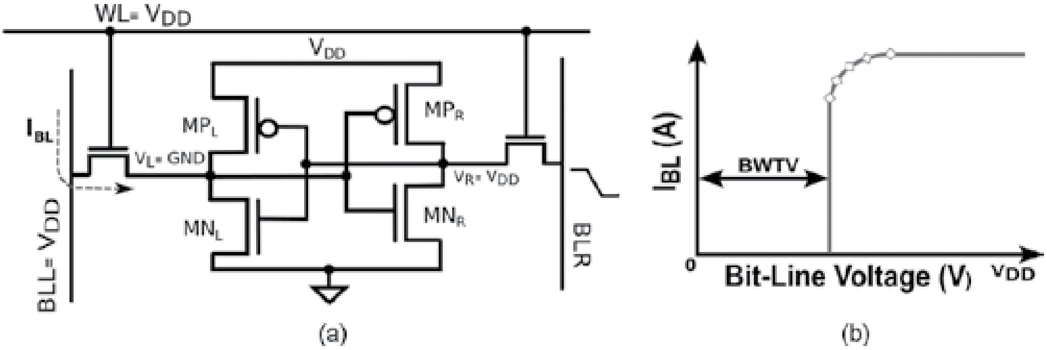


Figure 9. (a) The setup for the bit-line write trip voltage observation and (b) the graphical BWTV definition from current-voltage transfer curves.

Figure 10(a) shows the cell setup values when the bit-lines are set with inverted values. The WL is ramped up until the sudden transition of I_{BL} appears. **Figure 10(b)** represents graphically the evolution of I_{BL} for different values of word-line voltage. When the current drops from its maximum value, the WWTV is defined as the maximum word-line voltage drop to produce a successful write operation. Therefore, WWTV is obtained from the difference between the nominal power supply voltage and the minimum word-line voltage to change the stored value.

3.2.2. Based on stored value observation

The current based noise metrics requires analogue measurements from bit-lines requiring memory cell array modifications. To overcome these requirements, another metric is proposed in [17] with minimal redesign requirements. It is based on word-line voltage sweep and requires read/write memory operations because the stored value is the observation parameter.

3.2.2.1. Maximum word-line voltage margin (MWLV)

The MWLV estimates the writability margin finding the minimum word-line voltage level to produce an effective write on a specific cell [17].

For each value in the word-line voltage level the stored value is read to determine when the memory cell changes the value. **Figure 11(a)** shows the cell setup values when the bit-lines are set with inverted values. The V_{WL} is ramped up until the new value is written. **Figure 10(b)** represents graphically the evolution of stored value for different values of word-line voltage. When the write operation is successful, the MWLV is defined as the maximum word-line voltage drop to produce a successful write operation. Therefore, MWLV is obtained from the difference between the nominal power supply voltage and the minimum word-line voltage to change the stored value.

This technique was proposed in previous works [11] to improve the read/write stability with minimal SRAM circuit modifications. The reduction of word-line voltage during read/write

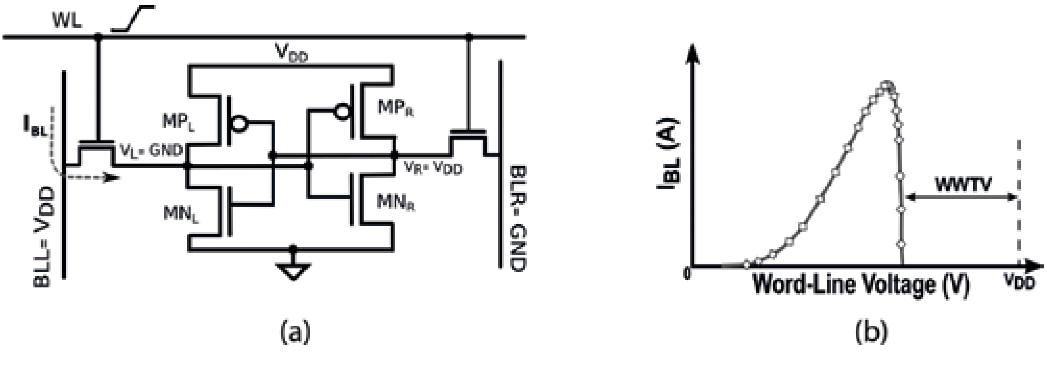


Figure 10. (a) The setup for the word-line write trip voltage observation and (b) the graphical WWTV definition from current-voltage transfer curves.

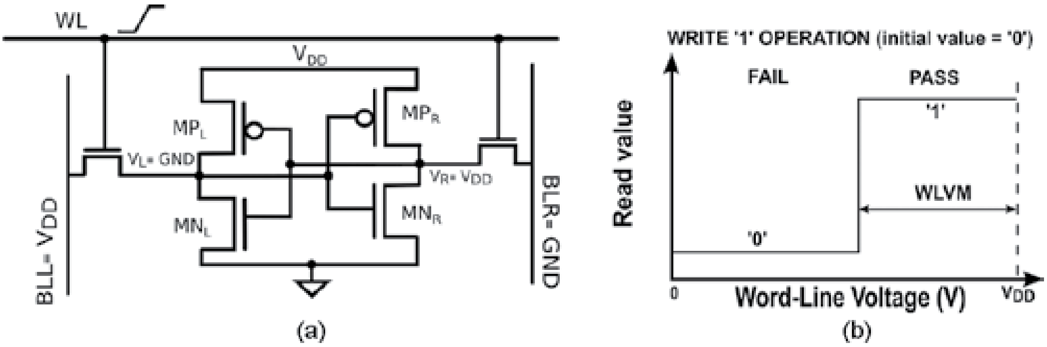


Figure 11. (a) The setup for the maximum word-line voltage margin observation and (b) the graphical MWLV definition from digital-voltage transfer curves.

operation increases cell stability during operations because the internal nodes of half-selected cells are connected to the bit-lines through a weaker connection, and thus the memory cell becomes more stable.

Finally, it is important to note that the MWLV metric is similar to WWTV metric because in both cases the word-line voltage is used to disturb the write operation. In this sense, the voltage conditions showed in **Figures 9(a)** and **10(a)** are equivalent. The main difference consists in how the behavior of the cell is monitored. In WWTV the I_{BL} current is used to detect a successful write on cell, while the word-line voltage is ramped up. Therefore, the WWTV

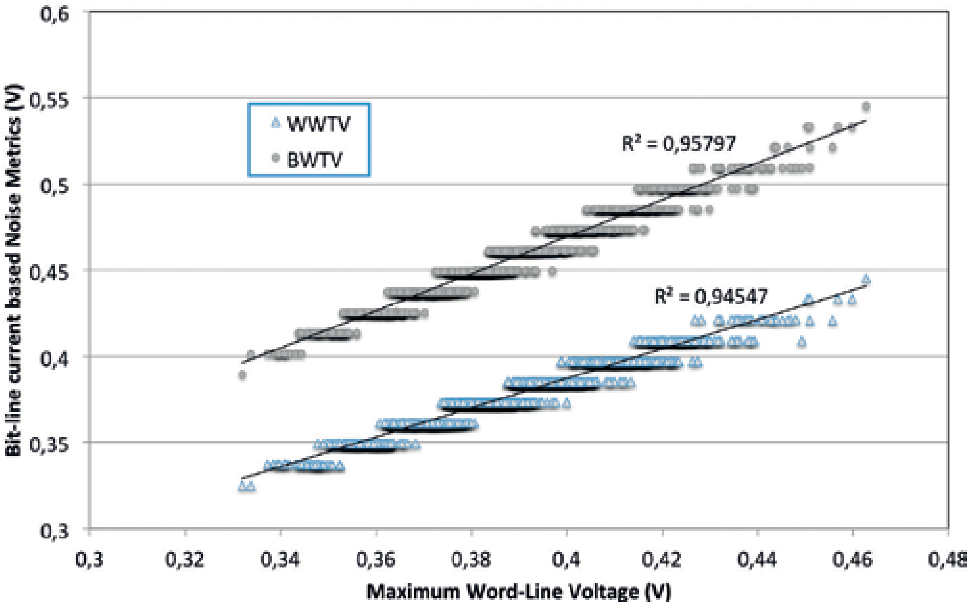


Figure 12. High grade of correlation between MWLV and WWTV/BWTV using Monte Carlo simulations with 65 nm CMOS technology.

considers the DC behavior response. By contrast, in MWLV a successful write operation is observed reading the stored value after a regular write operation at lower word-line voltage. That is, the cell memory is operated using memory read/write operation at regular designed timings. Therefore, the MWLV metric analyses the transient behavior because timing and dynamic features of write operation are included.

The existence of several metrics combining different methodologies to monitor the cell stability requires a deep analysis. In this sense, intensive Monte-Carlo simulations considering process variation on a commercial 65 nm CMOS technology have been performed to determine the correlations between current based writability margins and the digital based MWLV metric. **Figure 12** shows the linear correlation obtained between BWTV, WWTV and MWLV metrics. A linear correlation between metrics with a coefficient near to 0.95 is archived in both cases. This result suggests a remarkable equivalence between the different metrics and highlights the opportunity for freely selecting the most adequate methodology.

4. Defining a built-in stability monitor

Detecting SRAM performance shifts due to parameter variation and BTI involves sensing SRAM cell and peripheral circuit degradation. In this work, we center our attention on the sensing process and on the long-term variability effects on the memory cell margins, which depend on the threshold voltage shift of all NMOS and PMOS devices. It is well known that the fabrication processes in nanometer era introduce parameter variability, which translates in functionality effects at the device level.

The process variability has an impact on the noise margins, showed in **Figure 13**, where corner and Monte-Carlo analysis results show a high variability in the read static noise margin.

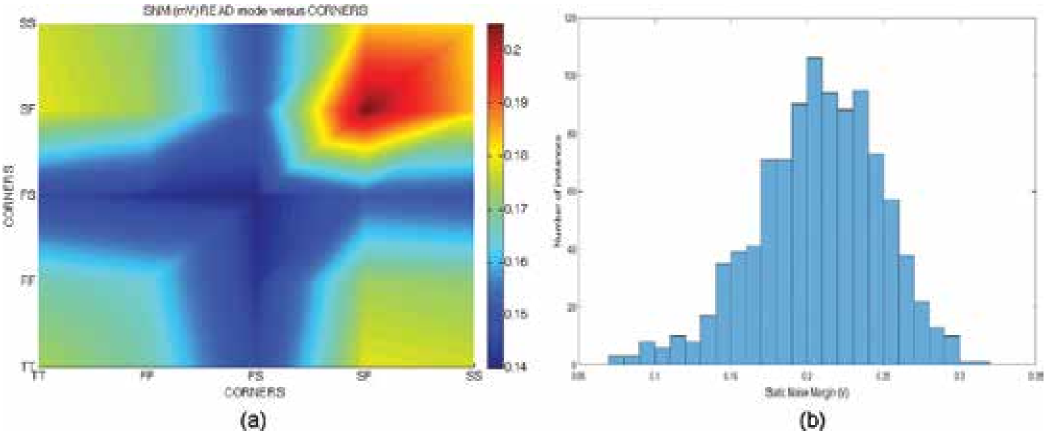


Figure 13. Process variability on read static noise metric obtained from (a) corners and (b) Monte-Carlo analysis with a commercial 65 nm CMOS technology.

The corner analysis assigns the slow-fast corner (SF in **Figure 13(a)**) as the most stable one that increases the stability in 17.1% from typical-typical corner (TT in **Figure 13(a)**), while the weakest is the fast-slow corner (FS in **Figure 13(a)**) and decrease the stability in 25% from TT corner. The maximum variability of RSNM is in the range of 60 mV.

The RSNM histogram, showed in **Figure 13(b)**, has been obtained from a 1000 iterations Monte-Carlo analysis considering process variation with a 65 nm CMOS technology. The variability spread is 240 mV with the mean value of 207.4 mV, and the standard deviation of 40.8 mV. Similar process variability impact may be observed using other noise margins.

In addition to the process variability, the use of the device may introduce extra parameter variability due to wearout/aging mechanisms. The influence of V_{th} variability on SRAM write margin metrics has been reported in **Figure 14(b)** [15, 17], where different write metrics (write noise margin, bitline write trip voltage, wordline write trip voltage and maximum wordline voltage) are explored considering V_{th} deviations due to wearout/aging effects.

Figure 14(a) shows the V_{th} variability impact on SRAM read noise margins (read static noise margin, supply read retention voltage, wordline read retention voltage). In the case of read noise margins, the variability behavior shows that the V_{th} drift decrease the read stability.

In the case of write noise margins, the write stability increases, this is, the memory cells are more easily written. Consequently, BWTV, WWTV and MWLV ramp up with V_{th}. According to their metric definitions, it means that the bit-line or word-line voltage can be lower previous to produce a fail in a write operation. However, WNM decreases with V_{th}, pointing out that the memory cell reduces its ability to tolerate noise during a write operation, because it is weaker against write processes. Therefore, although the four write metric curves showed in **Figure 14(b)** evolve in different directions, the meaning is equivalent in all of them: the memory cell is more stable during write operations allowing easy write operations with the increment of V_{th} value.

Finally, external variables may influence on metric values. To illustrate this impact, **Figure 15** reports the impact of power supply reduction on RSNM considering corners analysis (**Figure 15(a)**). The temperature decreases the noise margin during read operations. This temperature effect

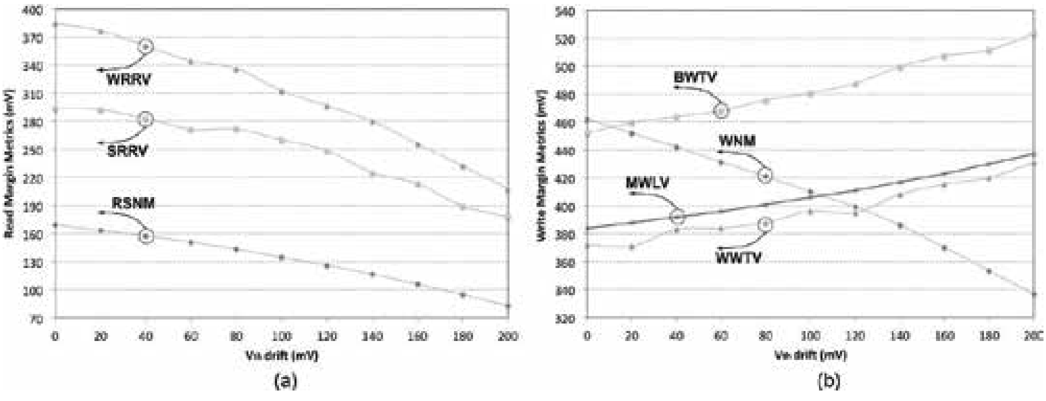


Figure 14. V_{th} drift impact on (a) read and (b) write noise margin definitions using a commercial 65 nm CMOS technology.

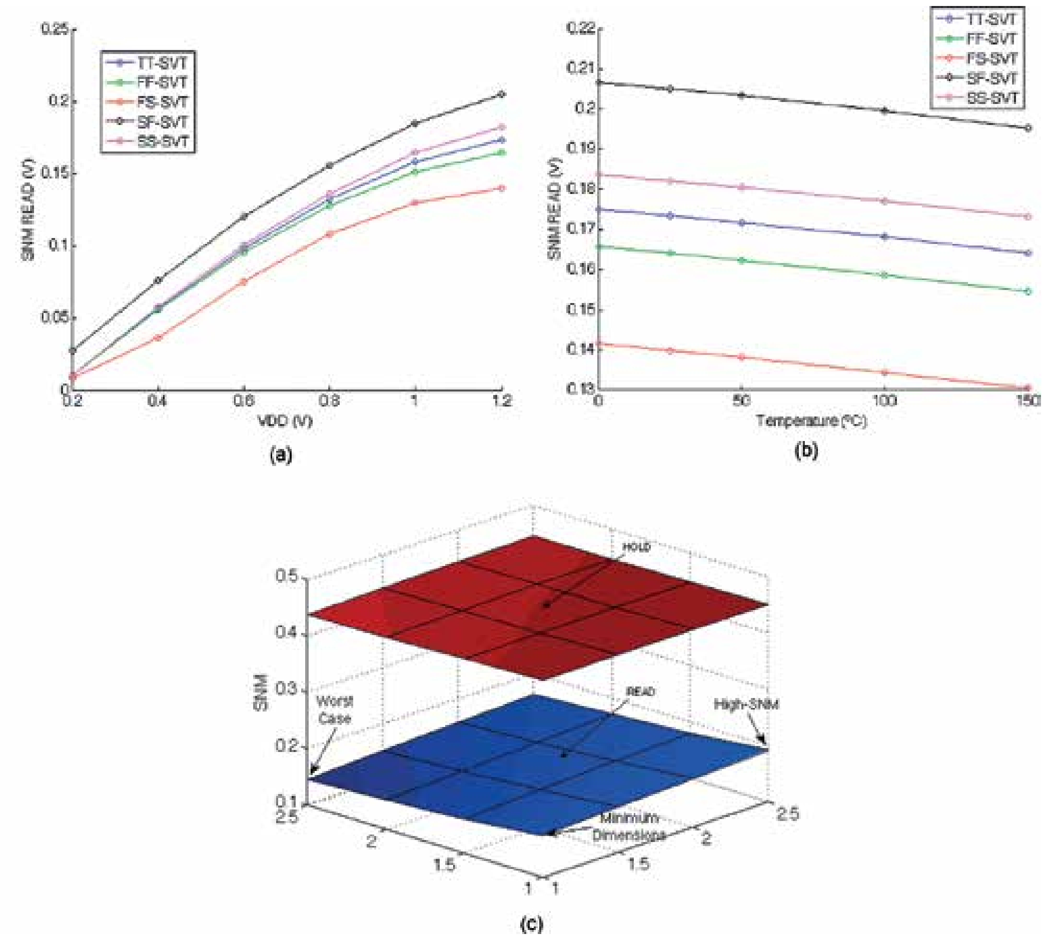


Figure 15. Impact of (a) power supply, (b) temperature and (c) operation and transistor width on read static noise metric using a commercial 65 nm CMOS technology.

considering corners analysis is showed in **Figure 15(b)**. And finally, the functional state (hold or read) of the circuit also is reported to influence the noise margin. **Figure 15(c)** warms about this effect representing the static noise margin of a memory cell in hold and read operation.

4.1. Stability metrics: from simulation to implementation

Although, the described static stability margin definitions are proposed to help designers during the pre-silicon step, the large number of devices in a memory array, and the increasing variability of technology processes, difficult the development of accurate models to simulate random effects in critical design parameters. Therefore, post-production or livelong memory measurements strategies are becoming important issues in modern system designs with high number of memory instances per chip.

Those popular metrics are suitable for simulation estimation but are too difficult to measure on real circuits. In this sense, DC read/write margins measurements were proposed using similar simulation methodologies. Direct access to internal storage nodes was implemented in [18] using large analogue switch networks circuits to connect internal cell nodes to external voltage sources and current monitoring circuits. Although, this methodology may achieve higher accuracy in SRAM failure analysis than simulation, its main drawback is the memory array redesigning efforts and hardware complexity required to perform voltage/current DC sweeps. In addition, the stability results and memory performance may also be affected negatively.

To decrease the intrusion on the memory array layout, the metrics based on bit-line current measurements have been proposed for large memory arrays [15]. Direct bit-line current measurement has been proposed in the literature to characterize noise metrics [15] or aging effect [19] in large memory instances with less memory cell array modifications. These approaches measure bit-line current variation, while adjusting bit-lines, word-lines and cell supply voltages to obtain writability data. Therefore, it is necessary to analyze the voltages and currents to obtain stability margins. Even though bit-line based metrics report good dependence with V_{th} drift, their implementation is resource demanding mainly in terms of SRAM redesign and area overhead. In this sense, several previous works have proposed an SRAM array schema designed for large-scale memories to perform bit-line current measurements [15]. The hardware requirements of [15] are: independent cell supply, cell ground, N-well bias, and P-well bias used for voltage adjusting conditions. Furthermore, column read/write circuitry must be shut off while a complex switch network for direct bit-line access enables to measure bit-lines current. The total area overhead is estimated to be around 20% [15]. Hence, the costs in terms of hardware redesign are elevated. In addition, these direct bit-line measurement methodologies may accelerate transistor aging because, during its measurement, the memory cell is forced to work at non-nominal DC voltage or above the nominal values as in the WRRV metric. In addition, DC currents values flowing through the devices may increase the faulty probability due to electro-migration effects.

Finally, the digital based metric is proposed in [16, 17] reducing the hardware requirements, and the needs of memory redesign, while maintaining the capabilities to estimate the write margin. Apart from not requiring the redesign of memory array, the memory cell operates at nominal values, i.e. Accesses time schedule and voltage/current levels. **Table 1** compares the features of the different noise margin metrics included in this work.

MWLV is measured reducing the word-line voltage peak value during a write operation, without alterations on memory cell performance, and requiring minimal memory overhead. In fact, it is obtained performing only a sequence of read/write operations on the target cell with different word-line voltages. Despite of that, the authors do not focus the attention only on MWLV noise margin, but also on some of the preciously introduced metrics, discussing their feasibility for lifetime aging effects monitoring. Next section will propose a built-in monitor approach feasible for noise metrics based on bit-line current or digital value observation.

4.2. Built-in monitor proposal

The built-in aging monitor approach based on noise margin measurements is feasible for different noise metric search. **Figure 16** shows the monitor schema adding the yellow blocks

		RSNM	*WNM*	*SVNM*	*SINM*	*WTV*	*WTI*	*SRRV*	*WRRV*	*BWTV*	*WWTV*	*MWLV*
ΔWM / ΔVth	*SYM*	-0.431	-0.613	-0.009	0.051	0.141	-0.075	-0.598	-0.884	0.350	0.291	0.263
	ASYM	-0.446	-0.311	1.111	0.175	-1.012	-0.018	0.853	0.379	0.234	0.291	0.329
Lineal coefficient (R^2)	*SYM*	0.987	0.988	0.999	0.983	0.996	0.997	0.947	0.985	0.986	0.966	0.989
	ASYM	0.999	0.999	0.998	0.999	0.999	0.921	0.92	0.986	0.993	0.966	0.993
Variable observed		V	V	V	I	V	I	I	I	I	I	D
Implementation		DC	DC	DC	DC	DC	DC	DC	DC	DC	DC	TRAN
Cell Array redesign		high	high	high	high	high	high	medium	high	medium	high	low
Area Overhead Needs		high	high	high	high	high	high	medium	medium	medium	medium	low
Access to cell nodes		Y	Y	Y	Y	Y	Y	N (V_{DD}, BL)	N (WL, BL)	N (BL)	N (WL, BL)	N

Table 1. Stability metrics comparison for lifetime monitoring.

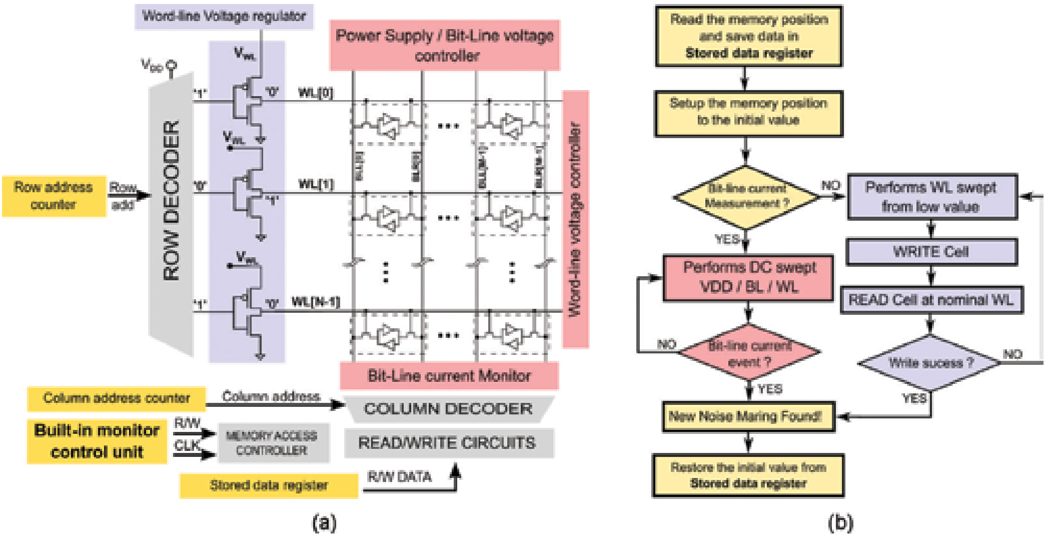

Word-line Voltage regulator
Power Supply / Bit-Line voltage controller
V_{DD}
V_{WL}
'1'
'0'
WL[0]
WL[1]
WL[N-1]
Row address counter
Row add
ROW DECODER
BL[0]
BL[0]
BL[M-1]
BL[M-1]
Word-line voltage controller
Bit-Line current Monitor
Column address counter
Column address
COLUMN DECODER
Built-in monitor control unit
R/W
CLK
MEMORY ACCESS CONTROLLER
READ/WRITE CIRCUITS
Stored data register
R/W DATA
(a)
Read the memory position and save data in Stored data register
Setup the memory position to the initial value
Bit-line current Measurement ?
NO
YES
Performs WL swept from low value
WRITE Cell
READ Cell at nominal WL
Performs DC swept VDD / BL / WL
Bit-line current event ?
Write sucess ?
New Noise Maring Found!
Restore the initial value from Stored data register
(b)

Figure 16. (a) The lifetime NM monitor schematic basics and (b) the NM search algorithm proposal for the built-in monitor schema approach.

needed in all noise margins implementations, the red blocks needed in direct bit-line current measurements and the blue blocks needed only for MWLV metric.

The Built-in monitor schema is supposed to perform the noise metric search in field, such as between activity periods. Therefore, the memory array will be disconnected from external signals to run the monitor algorithms. The common elements needed are the row and column counters (proposed to store the current memory address and run sequential accesses), and the stored data register (that saves temporarily the previous data read from current memory address), because the noise margin search algorithms are destructive (DC voltage swept or write operations).

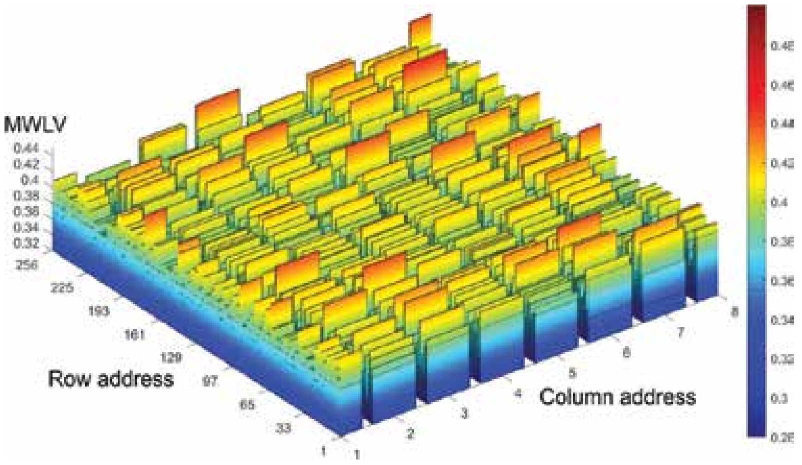


Figure 17. Spatial MWLV margin distribution due to process variability considering an 8 × 256 memory cell array using a commercial 65 nm CMOS technology.

The direct bit-line current-based noise metrics have the following requirements: the bit-line current monitor, the power supply/bit-line voltage controller and the word-line voltage controller. The voltage controllers are introduced to generate the desired DC voltage swept on the internal memory array nodes. The bit-line current monitor is able to detect the value sudden drop signaling the new noise margin.

The MWLV noise margin has less hardware requirements because the circuit changes are mainly centered on controlling the word-line voltage peak. The built-in approach proposes a word-line voltage controllability implementation based on using as isolated power supply node of all last row decoder gates. A feasible digitally controlled word-line regulator was reported in [17].

The build-in control unit implements the search algorithm depending on the noise metric implemented. A feasible algorithm proposal is showed in **Figure 17** highlighting the functions related with each metric methodology. The direct bit-line current measurement is based on the implantation reported in [15], while the MWLV metric is based on the design reported in [17]. **Figure 17** shows a 3D representation of the MWLV values measured from a 256 × 8 bytes memory implemented using a 65 nm CMOS technology [17] showing the suitability of this built-in approach.

Finally, it is important to note that the proposed search algorithm may be applied at any time during the normal lifetime of the memory. To determine the degradation evolution is not necessary to perform a whole exploration and the NM value may be estimated using a random address evaluation.

In addition, a novel online bit-line current measurement strategy has been recently proposed by [19] to measure aging effects on memory cell PMOS devices that the interest to perform online aging estimations is increasing in importance and is a challenging topic.

5. Conclusions

The runtime Reliability monitoring challenge in 6T CMOS SRAM has been addressed. The post-silicon stability profile has been highlighted as an observable signature to extract performance degradation due to reliability mechanisms. The different noise margins are identified as a suitable metric considering SRAM design with extra reliability guard bands. In addition, the write margins are suitable for designs oriented to guarantee read operation in exchange for writability degradation.

Author details

Bartomeu Alorda*, Gabriel Torrens and Sebastia Bota

*Address all correspondence to: tomeu.alorda@uib.eu

Electronic Systems Group, Physics Department, University of the Balearic Islands, Palma, Spain

References

[1] Vishvakarma SK, Reniwal BS, Sharma V, Kushwah CB, Dwivedi D. Nanoscale memory design for efficient computation: Trends, challenges and opportunity. In: Proceedings of the IEEE International Symposium on Nanoelectronic and Information Systems; 2015. pp. 29-34. DOI: 10.1109/iNIS.2015.58

[2] Mann RW, Hook TB, Nguyen DPT, Calhoun BH. Nonrandom device mismatch consideration in nanoscale SRAM. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. Jul. 2012;**20**(7):1211-1220. DOI: 10.1109/TVLSI.2011.2158863

[3] Khan S, Hamdioui S. Trends and challenges of SRAM reliability in the nano-scale era. In: Proceedings of the Inter. Conf. on Design & Technology of Integrated Systems in Nanoscale Era; 2010. DOI: 10.1109/DTIS.2010.5487565

[4] Lin JC, Oates AS, Yu CH. Time dependent Vccmin degradation of SRAM fabricated with high-k gate dielectrics. In: Proceedings of the IEEE International Reliability Physics Symposium; 2007. pp. 439-444. DOI: 10.1109/RELPHY.2007.369930

[5] Pilo H, Barwin C, Braceras G, Browning C, Lamphier S, Towler F. An SRAM design in 65 nm technology node featuring read and write-assist circuits to expand operating voltage. IEEE Journal of Solid-State Circuits. 2007;**42**(4):813-819. DOI: 10.1109/VLSIC.2006.1705289

[6] Karl E, Guo Z, Ng Y-G, Keane J, Bhattacharya U, Zhang K. The impact of assist-circuit design for 22 nm SRAM and beyond. In: Proceedings of the IEEE International Electron Devices Meeting; 2012. pp. 561-564. DOI: 10.1109/IEDM.2012.6479099

[7] Chiu YT, Wang YF, Lee Y-H, Liang YC, Wang TC, Wu SY, Hsieh CC, Wu K. Analysis of the reliability impact on high-k metal gate SRAM with assist-circuit. In: Proceedings of the IEEE International Reliability Physics Symposium; 2014. pp. 4.1-4.4. DOI: 10.1109/IRPS.2014.6861171

[8] Faraji R, Naji HR. Adaptive technique for overcoming performance degradation due to aging on 6T SRAM cells. IEEE Transactions on Device and Materials Reliability. 2014;**14**(4):1031-1040. DOI: 10.1109/TDMR.2014.2360779

[9] Kim W, Chen C-C, Liu T, Cha S, Milor L. Estimation of remaining life using embedded SRAM for wearout parameter extraction. In: Proceedings of the IEEE International Workshop on Advances in Sensors and Interfaces; 2011. DOI: 10.1109/IWASI.2015.7184952

[10] Alorda B, Torrens G, Bota S, Segura J. Adaptive static and dynamic noise margin improvement in minimum-sized 6T-SRAM cells. Microelectronics Reliability. 2014;**54**:2613-2620. DOI: 10.1016/j.microrel.2014.05.009

[11] Alorda B, Torrens G, Bota S, Segura J. Static and dynamic stability improvement strategies for 6T CMOS low-power SRAMs. In: Proceedings of the Design Automation & Test in Europe Conference; 2010. pp. 429-434

[12] Bota S, Torrens G, Alorda B. Critical charge characterization of 6T SRAMs during read mode. In: Proceedings of the IEEE International On-Line Testing Symposium; 2009. pp. 120-125. DOI: 10.1109/IOLTS.2009.5195993

[13] Seevinck E, List FJ, Lohstroh J. Static-noise margin analysis of MOS SRAM cells. IEEE Journal of Solid-State Circuits. 1987;**SC-22**(5):748-754

[14] Wann C, Wong R, Frank DJ, Mann R, Ko S-B, Croce P, Lea D, Hoyniak D, Lee Y-M, Toomey J, Weybright M, Sudijono J. SRAM cell design for stability methodology. In: Proceedings of the IEEE Symposium on VLSI-TSA; 2005. DOI: 10.1109/VTSA.2005.1497065

[15] Guo Z, Carlson A, Pang L-T, Duong KT, King T-J, Nikolic B. Large-scale SRAM variability characterization in 45nm CMOS. IEEE Journal of Solid-State Circuits. 2009;**44**(11):3174-3192. DOI: 10.1109/JSSC.2009.2032698

[16] Alorda B, Carmona C, Torrens G, Bota S. On-line write margin estimator to monitor performance degradation in SRAM cores. In: Proceedings of the International On-Line Testing Symposium; 2016. DOI: 10.1109/IOLTS.2016.7604678

[17] Alorda B, Carmona C, Torrens G, Bota S. An affordable experimental technique for SRAM write margin characterization for nanometer CMOS technologies. Microelectronics Reliability. 2016;**65**:280-288. DOI: 10.1016/j.microrel.2016.07.154

[18] Bhavnagarwala A, Kosonocky S, Chan Y, Stawiasz K, Srinivasan U, Kowalczyk S, Ziegler M. A sub-600 mV fluctuation tolerant 65 nm CMOS SRAM array with dynamic cell biasing. IEEE Journal of Solid-State Circuits. 2008;**43**(4):946-955. DOI: 10.1109/VLSIC.2007.4342773

[19] Ahmed F, Milor L. Online measurement of degradation due to bias temperature instability in SRAMs. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2016;**24**(6):2184-2194. DOI: 10.1109/TVLSI.2015.2500900

Chapter 8

Towards New Generation Power MOSFETs for Automotive Electric Control Units

Kuan W.A. Chee and Tianhong Ye

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70906

Abstract

Power metal-oxide-semiconductor field-effect transistors (MOSFETs) are thought to be highly robust and versatile in high-speed switching applications in power electronics design due to its intrinsic high input impedance and compact size. This chapter concerns the development of a high-performance low voltage rating power MOSFET possessing low on-resistance and excellent avalanche current capability for an automotive electric power steering system (EPS). Using industry-standard Technology Computer-Aided Design (TCAD) tools, the planar- and trench-technology power MOSFETs, have been designed, modeled, simulated and compared. We surveyed and analyzed the specific on-resistance due to the different device structures, and various methods are highlighted and compared so that their benefits can be better understood and adopted. Additionally, the device ruggedness has been investigated and its improvement was evaluated and established for that of the trench MOSFET due to gate corner smoothing.

Keywords: automotive MOSFETs, specific on-resistance, avalanche ruggedness, unclamped inductive switching, silicon carbide power semiconductor, critical breakdown electric field, Technology Computer-Aided Design

1. Introduction

When the first automobiles were invented dating back to 130 years ago, the only expectations were safe operation and durability. Over the years of continual development of the automobile, more and more “bells and whistles” were added, culminating in more innovative features and functions. More recently, driverless cars have become a reality. These features are inevitably empowered by advances in electrical engineering and automation, bringing about the rapid increase in the value of electronics in a car. Particularly, more and more electronic control units (ECUs) have been developed for automobiles and electric vehicles. In certain high-end vehicles, the number of ECUs can be as high as 100 or so. If ECUs are akin to the organs of the car, semiconductor devices are like the cells. The latter we refer especially to those power semiconductor devices that are widely recognized as basic and vital building blocks of electrical and power electronic systems.

Discrete power semiconductors occupy a major share of the ever-increasing revenue from semiconductor devices in the HEV/EV industry over the years, and this is projected to continue beyond 2020 (**Figure 1**) [1]. Specifically, power metal-oxide-semiconductor field-effect transistors (MOSFETs) have gained a lot of popularity due to their simple drive requirements, low on-resistance and fast switching properties. Owing to their high input impedance and energy efficiency excellence in high frequency applications, MOSFETs are the preferred choice to several circuit designers [2]. Notably, power MOSFETs are able to switch high current and voltage levels with enhanced power handling capability in highly efficient power supply circuits and systems [3].

1.1. Power consumption of power MOSFET

One of the key metrics underpinning the performance of the MOSFET is on-resistance (Rdson). High Rdson restricts the maximum current capability; in addition, large power dissipation

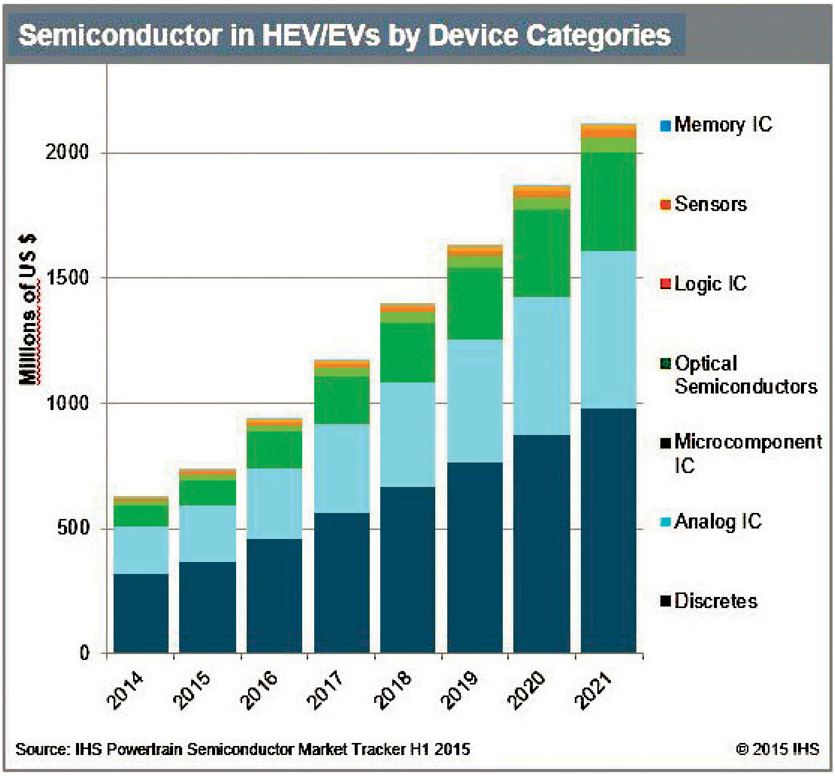


Figure 1. Semiconductors in HEVs/EVs by device categories [1].

($P = V_{dd} \times I_{d,\ ave.} = I_{d,ave.}{}^{2} \times$ Rdson) will lead to unwanted die temperature rise during device operation. It is understood that Rdson is inversely proportional to the cell area for many device technologies Therefore to enable comparison between different designs, e.g. 'trench' versus 'planar' types, a figure-of-merit is introduced called the specific on-resistance, i.e. the product of Rdson and the cell area.

1.2. Ruggedness of power MOSFET

Almost every application circuit has some kind of inductance, not only in the form of load inductance such as solenoids or electric motors, but stray inductances such as wiring and layout inductances.

Figure 2 shows a typical application circuit in an electric power steering system. It can be seen that instantaneous current changes could result from a short circuit in the arm of the H-bridge, a short circuit to the ground or a short circuit to the three-phase motor. When the supply current is rapidly switched off, the changing magnetic field inside the windings induces a back electromotive force. Thus, when dealing with inductive loads in ECUs, a high di/dt commutation rate during switching transitions runs the risk of a surge voltage that may destroy the device [4]. Placing a freewheeling diode anti-parallel to the MOSFET represents one approach to avoid this possible high voltage dump. However, in some applications, for instance gasoline or diesel injection [2], MOSFETs are designed with an intrinsic body diode to withstand this

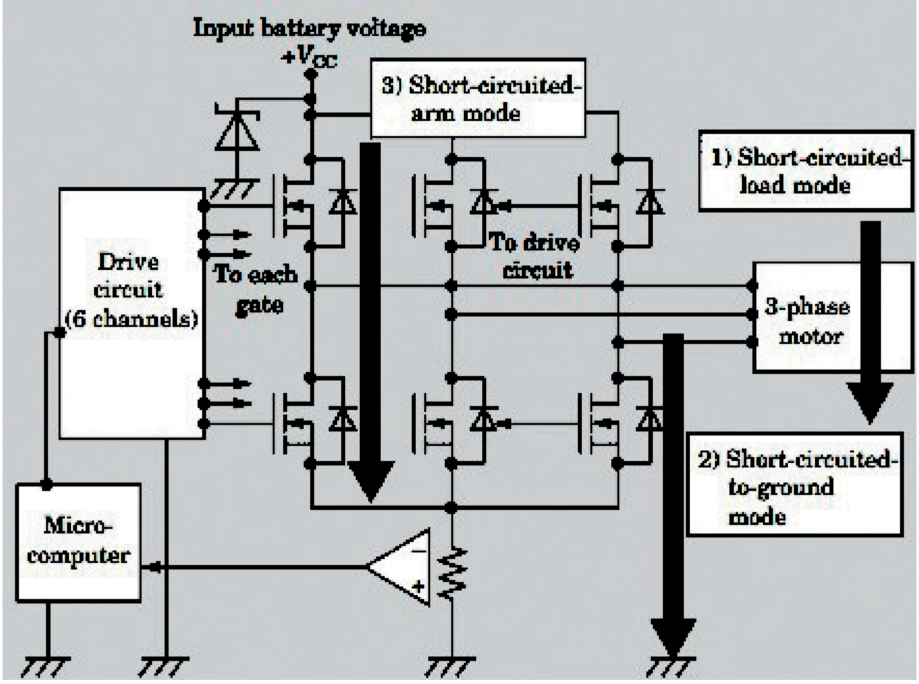

Input battery voltage
+V_{CC}
3) Short-circuited-arm mode
1) Short-circuited-load mode
Drive circuit (6 channels)
To each gate
To drive circuit
3-phase motor
2) Short-circuited-to-ground mode
Micro-computer
−
+

Figure 2. Typical application circuit in electric power steering [2].

possible voltage surge in order to survive any avalanche breakdown threat. Unclamped inductive switching (UIS) is so-called without support of a separate freewheeling diode, and ruggedness is the ability of the MOSFET to resist avalanche failure under UIS conditions. Electron irradiation or platinum doping may also be used for minority carrier lifetime control in the body diode to greatly improve the reverse recovery characteristics.

1.3. Overview of the power MOSFET market segments

According to the QYResearch Group, the global revenue for the discrete power device market in 2016 was valued at $ 7.277 billion, and by the end of 2022 this number was projected to rise to $ 9.135 billion, growing at a compound annual growth rate of slightly above 3.86% between 2016 and 2022 [5]. As aforementioned, the power MOSFET accounts for a significant portion of the total revenue. There are various catalogs of MOSFETs available in the market; the technology used is mainly categorized into the following three types: planar, trench and superjunction. In the low voltage category, besides automotive MOSFETs that form the main focus of this chapter, other power MOSFETs are designed for a range of other applications. Take Infineon for example, they target their commercial power MOSFETs at the following applications [6]:

- DC/DC converters
- 3D printers
- LED lighting
- motor control systems
- solar micro inverters
- battery powered applications, i.e. desktop and notebook
- audio amplifier

Further, Infineon has also developed green and robust packages for their product range, providing the highest current handling capabilities [7]. In the high voltage rating (500–900 V), a very innovative kind of MOSFET has dominated the market, called the superjunction MOSFET, which was originally commercialized by Infineon in 1998 [7]. Normally, the on-resistance is positively related to the voltage rating, which is characteristic of typical high voltage rating devices. This is due to the increase in drift region resistance to support higher voltages. However, thanks to the superjunction MOSFET, this relation does not apply. The most remarkable feature about this kind of MOSFET is the dramatic reduction in on-resistance and switching losses, thus enabling high power density and energy conversion efficiency in high power applications. Finally, the other kind of power MOSFETs is based on the laterally double-diffused short channel structure, or RF LDMOS. Due to its high operating frequency, one typical application of this MOSFET is in telecommunications, for example, in power amplifiers in television systems (especially digital television), radar systems and military communications [8]. Besides a higher gain and linearity, excellent noise-resistant properties and thermal stability are other key advantages of this type of unipolar device [8].

2. Automotive power MOSFET designs

A standard planar MOSFET was designed to meet the performance specifications of the electric power steering circuit. In order to enable better noise resilience, an appropriate threshold voltage (V_{th}) of 3 V was engineered [2]. In the current technology market, the typical supply voltage for the power steering circuit is 42 V [2]. Therefore the designed breakdown voltage of the planar and trench MOSFET should be around 50 V. **Figure 3** shows the structure of the n-channel planar MOSFET including the depletion regions. During forward conduction, electrons flow from the source through the inverted region of the p-well (or n-channel) beneath the gate, then through the JFET region before entering the drift region. Hence, there are four main types of component resistances [9]: (1) source resistance, (2) channel resistance, (3) JFET resistance and (4) drift region resistance, which will be further discussed below. **Figure 4** shows the V_{th} increase with the *p*-well dose. The *p*-well was designed with a dose of 4.5×10^{14} cm^{-2} to meet the V_{th} requirement.

Nevertheless, it is important to note that for high voltage designs, the drift region resistance is the most significant component, whereas for low voltage designs, channel resistance and source resistance are crucial, in the overall Rdson. As Rdson is negatively correlated to the drift region doping concentration, so is the breakdown voltage (BV), as shown in **Figure 5**. Hence for high voltage rating power MOSFETs, the doping concentration in the drift region should be low enough, which is the reason why the Rdson of high power MOSFETs is typically way larger than that of low power MOSFETs. Fortunately, replacing silicon (Si) with wide bandgap silicon carbide (SiC) would enable a significantly lower drift region resistance [10]. The results for this will be discussed below. Besides, the drift region epitaxial layer thickness (t_{nepi}) also determines Rdson, and **Figure 6** shows that below 5 μm, BV drops dramatically, thereby reflecting the case that avalanche breakdown occurs before the drift region is fully depleted in the off-state. Therefore the optimal t_{nepi} should be slightly larger than 5 μm for the best trade-off between BV and Rdson.

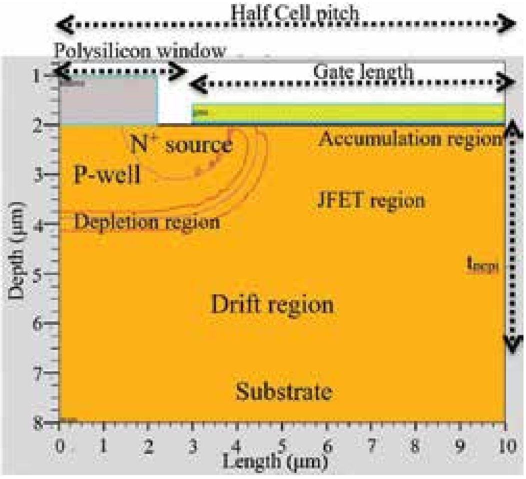


Figure 3. Structure of the planar MOSFET including depletion regions at zero bias [11].

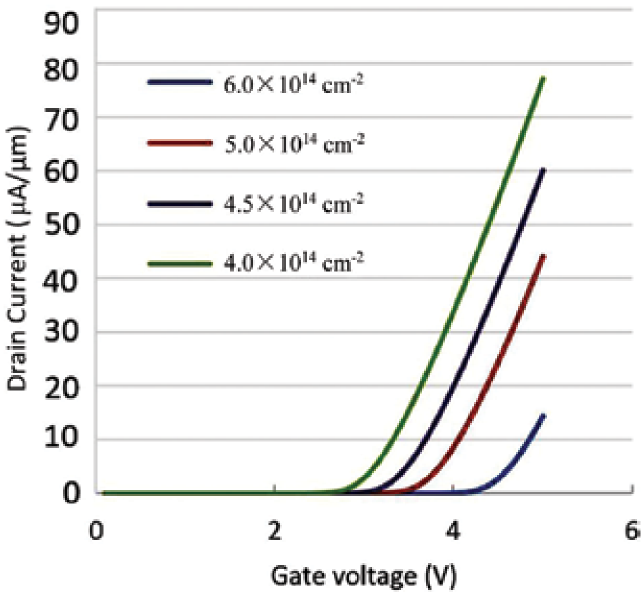


Figure 4. Transfer characteristics at a drain voltage of 0.1 V for various p-well (boron) doses [11].

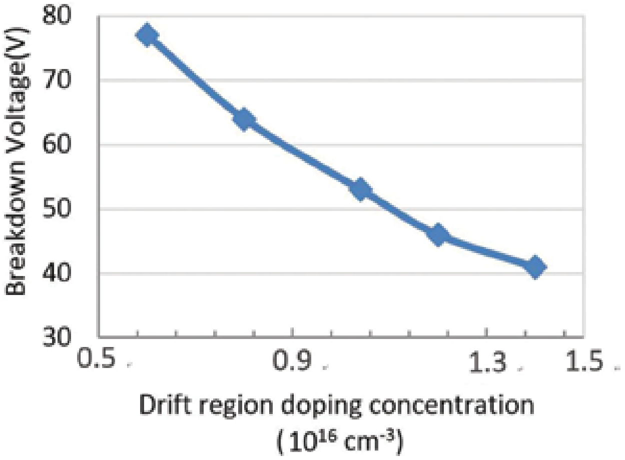


Figure 5. Breakdown voltage as a function of drift region doping concentration [11].

For a half-cell pitch decreasing from 11 to 10 μm, BV increases (**Figure 7**). Below 10 μm, no further increase in BV can occur, owing to a field plate effect that optimizes the electric field distribution at the junction curvature; the electrical field at the junction curvature approximates that of a planar junction. A shorter cell pitch would increase the JFET resistance; therefore the half-cell pitch was chosen to be 10 μm to provide the best trade-off between BV and JFET resistance. The Rdson is 1.56 × 10^4 Ω at a gate bias of 5 V (see **Figure 8**), and with a cell width of 1 μm, the specific on-resistance is 1.56 mΩcm^2.

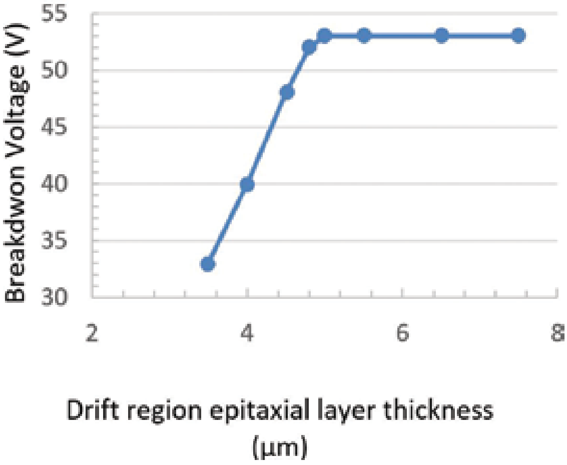


Figure 6. Breakdown voltage as a function of drift region epitaxial layer thickness [11].

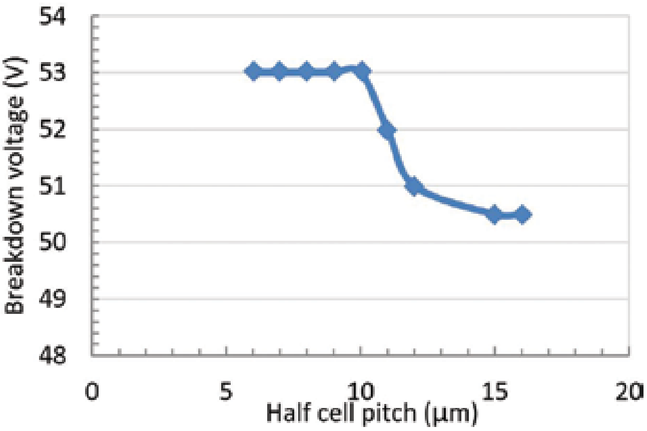


Figure 7. Breakdown voltage as a function of half-cell pitch [11].

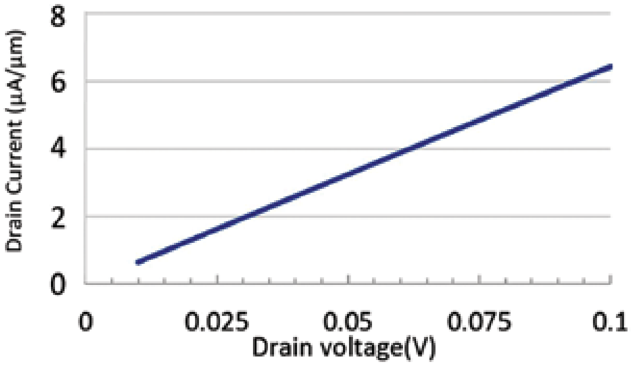


Figure 8. Output characteristics in the linear region of operation at a gate voltage of 5 V [11].

Further, a caveat should be noted that in practice, especially for high voltage devices, BV is limited by the edge termination structure used to control the surface electric field. This is because high voltage planar junctions under reverse bias exhibit significantly lower breakdown voltages than one-dimensional theory predicts due to three-dimensional electric potential line crowding at the junction periphery. Therefore a good edge termination structure is critical to minimize this effect and increase the planar junction BV to near ideal values to maintain the rated BV and reliability of the high voltage power device. When the maximum specified drain to source voltage (or BV) is exceeded when the MOSFET is turned off, the intense surface fields on the field guard rings, beyond the rated design specification, can cause avalanche multiplication, thereby leading to conduction of an overcurrent that damages the device due to excessive power dissipation. This is indicated by the catastrophic damage on the field guard rings of the MOSFET bare die (see **Figures 9** and **10**).

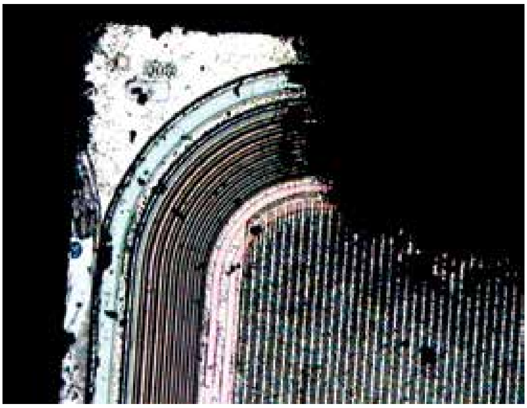


Figure 9. Breakdown damage on field guard rings indicating excessive drain to source voltage.

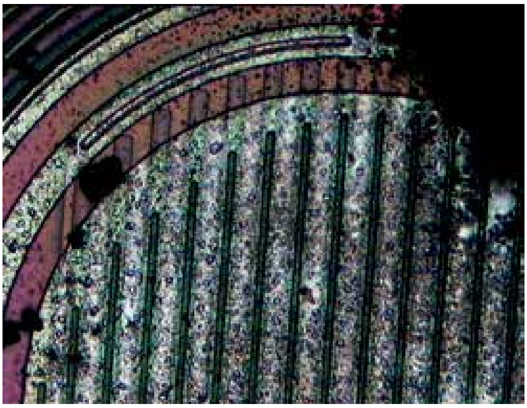


Figure 10. Breakdown damage on field guard rings indicating excessive drain to source voltage (under higher magnification *c.f.* **Figure 9**).

2.1. Design enhancements for low on-resistance

Having high cell densities and large die sizes can achieve lower on-resistances, but concomitantly result in significant gate and output charges, thereby increasing the switching losses. Therefore three main strategies to reduce on-resistance will be illustrated for the planar MOSFET: (1) optimization of gate width-length dimensions; (2) increased doping in the integral JFET region; and (3) adopting wide bandgap SiC as the power semiconductor material. The deep trench design is known to significantly reduce on-resistance owing to a low spreading resistance through the increased accumulation layer, and complete elimination of the JFET resistance.

2.1.1. Gate width-length optimization

Concerning the planar MOSFET, the specific on-resistance of the accumulation layer is positively related, but that of the JFET region is negatively related, to the width-length ratio of the gate electrode [12]. The optimum gate width is *ca.* 3 μm (see **Figure 11**), yielding a specific on-resistance of 1.2 mΩcm^2. Therefore reducing the half-cell pitch from 10 to 6 μm (for a polysilicon window 3 μm in length) reduces the specific on-resistance by 23%. As the gate width reduces below 3 μm (or cell pitch below 6 μm), the specific on-resistance rises sharply due to the short current path in the JFET region, which is pinched off during linear operation. In addition, according to **Figure 7**, when the half-cell pitch is 6 μm, the junction curvature does not lower the BV. Therefore the optimal gate length should be between 5 and 6 μm.

2.1.2. Increased doping in integral JFET region

Figure 12 shows the structure of the power MOSFET with increased doping in the integral JFET region. By increasing the JFET doping concentration (≈ n), the JFET resistivity reduces as $\frac{1}{nq\mu}$ where q is elementary charge and μ is carrier mobility, but the BV also lowers as shown in **Figure 13**. The optimal dose is 2.7 × 10^{15} cm^{-2} for a voltage rating of 50 V; and the optimized specific on-resistance is 1.43 mΩcm^{-2}, representing a reduction by 8.3%.

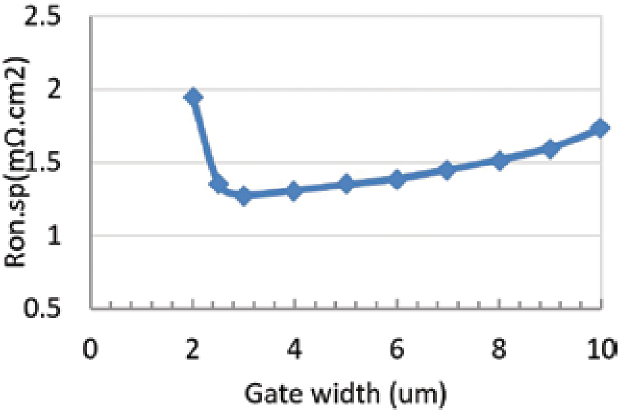


Figure 11. Specific on-resistance versus gate width [11].

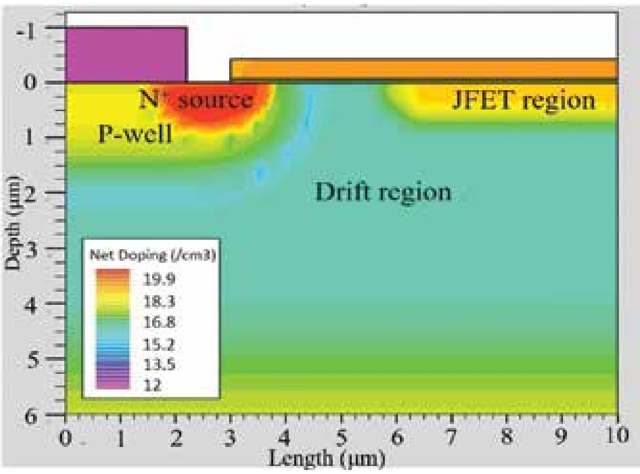


Figure 12. Doping profiles in the power MOSFET with additional dose in the JFET region [11].

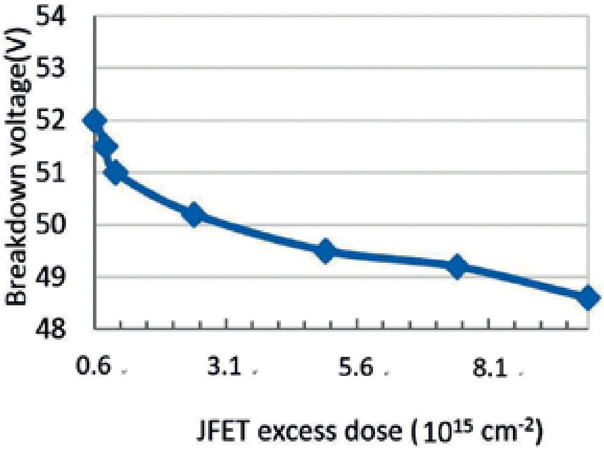


Figure 13. Breakdown voltage as a function of JFET excess dose [11].

2.1.3. Planar MOSFET based on SiC

The structure of a planar gate SiC vertically double-diffused (VD)-MOSFET being modeled is shown in **Figure 14**. The gate oxide thickness is the same as that for the Si planar MOSFET in **Figure 3**. To target a V_{th} of 3 V, the designed doping concentration is 2.6×10^{16} cm^{-3} in the *p*-well.

The critical breakdown electric field of SiC is eight-fold greater than that of Si [13]. Hence, if no reach-through is assumed, in principle a BV up to 411 V can be achieved according to:

$$V_{breakdown} = \frac{E_{critical}{}^{2} \varepsilon_s}{2qN_a} \tag{1}$$

where $E_{critical}$ and ε_s are the critical electric field and dielectric permittivity respectively. N_a is the dopant concentration in the *p*-well, which should be far exceeded by that in the drift region. The minimum *p*-well thickness is governed by the depletion width w_p in the *p*-well at the maximum drain voltage, Va. w_p is 1.46 μm according to:

$$w_n = \sqrt{\frac{2\varepsilon_s V_a}{qN_a}} \tag{2}$$

which in turn sets the minimum *p*-well thickness and channel length. However, this value may be an underestimate neglecting the effects of the junction curvature (see **Figure 15**), as according to the simulations the appropriate channel length is suggested to be at least 1.8 μm for a 50 V device. The *p*-well is designed to confer the blocking voltage capability; hence the dopant concentration in the drift region can be made very high. For example, the drift region dopant concentration may be as high as 10^{18} cm^{-3}, meaning that the depletion width in the drift region becomes extremely small, so that the drift region thickness may successfully be reduced to as thin as 0.3 μm. Hence, the ability to heavily dope and drastically reduce the epilayer thickness of

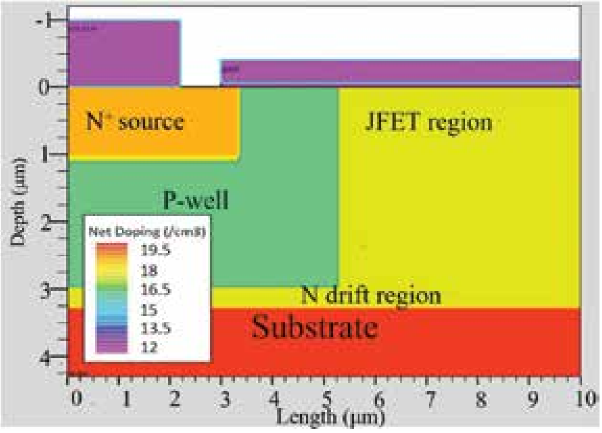


Figure 14. Structure of a planar gate SiC VDMOSFET [11].

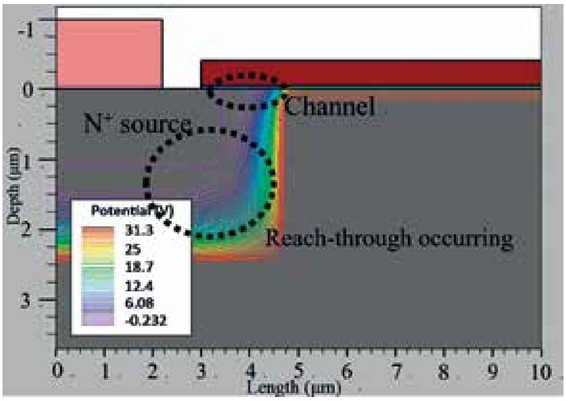


Figure 15. Potential distribution in SiC power MOSFET at a drain voltage of 36 V [11].

the drift region afforded by using SiC as a power semiconductor material mandates an exceedingly low drift region resistance of the planar MOSFET.

Moreover, the JFET region is virtually non-existent because the depletion width is significantly narrower, so that it becomes possible to make the separation between the two *p*-wells very small. As a result, the accumulation layer resistance, which would be significant in the Si planar MOSFET, may also be markedly reduced. The half-cell pitch can shrink to as small as 5.5 μm, yielding a specific on-resistance of 1.08 mΩcm^2, which represents a reduction by 31% compared to that of conventional planar technology in **Figure 3** (1.56 mΩcm^2) or by 10% compared to that of the Si planar MOSFET after gate width optimization (1.2 mΩcm^2).

2.1.4. Trench MOSFET

The cell pitch in the trench design platform can be made very small because there is no JFET region, but is limited by the current fabrication technology. **Figure 16** shows the trench MOSFET structure and current paths at a gate and drain bias of 5 V and 1 V respectively. A half-cell pitch of 2.5 μm is chosen for a typical trench MOSFET, and the gate oxide thickness is 80 nm, the n^+ source junction depth is 0.5 μm, and the dopant concentration in the substrate layer 1 μm thick is 10^{19} cm^{-3}.

To target a V_{th} of 3 V, the *p*-well dopant concentration is 2.6×10^{16} cm^{-3} for this trench design. For a 50 V rating, the designed drift region dopant concentration is 1.4×10^{16} cm^{-3}. **Figure 17** shows the drift region thickness dependence of BV. For an *n*-epilayer thickness below an optimal value of 2.2 μm, the BV reduces dramatically owing to punch-through effects. For this trench MOSFET structure, the specific on-resistance is 0.625 mΩcm^2, which is a reduction by 60% compared to the planar MOSFET (1.56 mΩcm^2).

Therefore, the underpinning reasons for such a low on-resistance of the trench MOSFET can be summarized as follows. By eliminating the intrinsic JFET component in the trench design, the

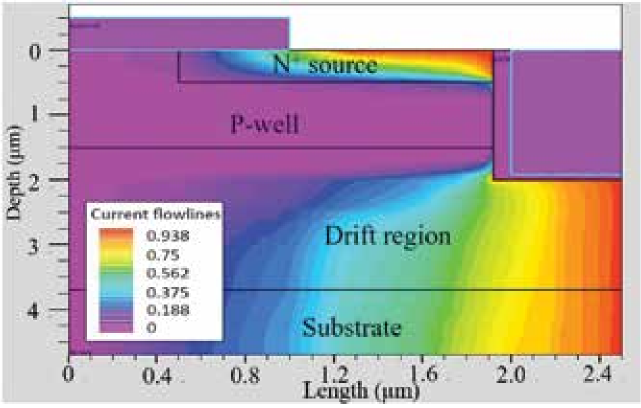


Figure 16. Trench gate power MOSFET structure and current flow lines through the n$^+$ source and n$^+$ substrate of the device with a backside contact. The current density is normalized to the maximum in the device [11].

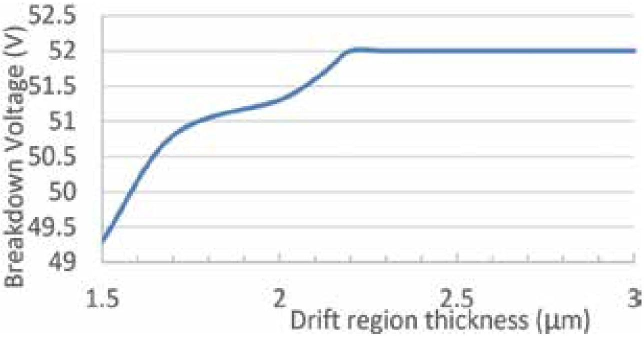


Figure 17. Breakdown voltage versus drift region thickness [11].

cell pitch can be made very small without needing to be concerned about increasing the JFET resistance. In fact, the cell pitch of the designed trench MOSFET is shorter by a factor of 2.5 than that of the planar MOSFET with optimum gate width [11].

3. Avalanche failure of power MOSFET

3.1. Avalanche failure mechanism

Basically, two types of failure modes can be identified in the avalanche condition. One is the active mode, which is caused by the turning on of a parasitic transistor intrinsic in the device through the p-well [14]. During avalanche, the body diode no longer blocks voltage; the electric field in the body diode becomes exceedingly large, above the critical breakdown field magnitudes of Si or SiC, particularly at the junction curvatures. Consequently, the process of impact ionization and avalanche multiplication occurs, thereby leading to a large current flow between the drain and source through the p-well, and power dissipation causes the associated local temperature rise. Due to the positive temperature coefficient of the resistivity of silicon, the p-well resistance (R_{b+}), and in turn, the voltage drop across the p-well (acting as the base-emitter forward bias), will increase. Once this voltage drop exceeds 0.7 V, which is the turn-on voltage of the parasitic BJT, loss of gate control and latch-up occurs, and a hot spot is formed as more current crowds into it, ultimately leading to device destruction due to overcurrent [15]. However, in other cases, avalanche failure is due to a passive mechanism, which essentially arises from a thermal effect [14]. In an avalanche condition, energy stored in the inductor is dissipated in the MOSFET, even in its off state, thereby leading to a local temperature rise within the device. This temperature rise changes the breakdown voltage, which in turn results in significantly larger current flow and increased power dissipation, and eventual thermal runaway; the current percolations through narrow regions due to the positive temperature coefficient of the silicon resistivity bring about secondary breakdown induced by ohmic heating. The secondary breakdown is initiated when the cell temperature reaches a critical value, beyond which the intrinsic carrier concentration exceeds the background doping concentration in the epitaxial layer [16];

and the thermal generation of defects that form current shunts. The avalanche failure site can be optically visualized from burnt marks on the bare die, indicating the occurrence of the hot spots that the current crowd into, eventually causing catastrophic damage.

3.2. Avalanche ruggedness evaluation

Modern day designs are focused on increasing device ruggedness, and thus avalanche testing methods were developed to validate the device avalanche rating. An example of the latter is UIS testing, which is performed using a test circuit like the one shown in **Figure 18**.

The UIS testing procedure is as follows:

1. A gate bias switches on the MOSFET.

2. Current flows through the load (whereas the MOSFET intrinsic resistance can be ignored), and the current increase can be expressed as:

$$I = \frac{VTA}{L} = \frac{VDD \times T}{L} \tag{3}$$

where VDD is the supply voltage, T is the pulse width and L the inductance.

3. When the targeted current is reached, the gate signal is reduced to zero, thereby immediately switching off the MOSFET. However, the current cannot decay abruptly owing to the presence of an inductive load; in fact, the resultant higher voltage exerted on the MOSFET forces the device into avalanche.

4. Avalanche operation is sustained till all the energy stored within the magnetic field due to the inductance is dissipated as heat.

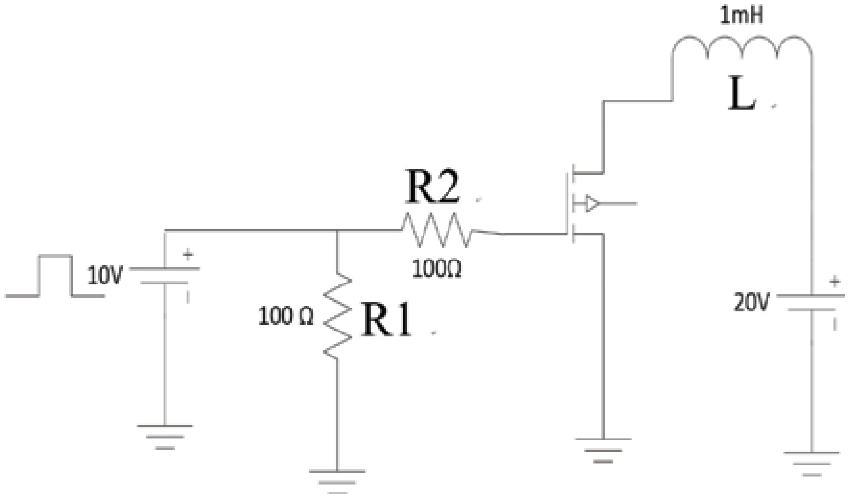


Figure 18. UIS testing circuit [17].

The voltage exerted on the device in the avalanche condition, is not BV but the effective breakdown voltage (BV_{DSS}), which is about 1.3–1.5 fold larger [4]. The avalanche voltage on the inductive load is BV_{DSS} – VDD, and the avalanche duration can be derived from:

$$t_{av} = \frac{I \times L}{\mathrm{BV} - \mathrm{VDD}} \tag{4}$$

Hence, we can compute the single pulse avalanche energy (EAS) from:

$$EAS = \frac{1}{2} I_{av} \mathrm{BV}_{\mathrm{DSS}} \times t_{av} = \frac{1}{2} L \times I^2 \frac{\mathrm{BV}_{\mathrm{DSS}}}{\mathrm{BV}_{\mathrm{DSS}} - \mathrm{VDD}} \tag{5}$$

Since BV_{DSS} is directly proportional to temperature [18], self-heating effects are accounted for in the electro-thermal simulations of the circuit performance. As an example, VDD is 20 V and the inductive load is chosen as 1 mH, and R1 and R2 are both 100 Ω for a typical UIS simulation. The gate signal amplitude is 10 V and pulse width is 2 ms, which turns on the device within the duration when V_{th} is exceeded, but turns off the device otherwise. The 50 V rated MOSFET is designed with a V_{th} of 3 V, and for a die size of 5 mm^2, the waveforms under avalanche operation are shown in **Figure 19**.

The maximum drain current is 40 A, at which instant the gate bias drops below V_{th} so that the MOSFET is turned off and the junction temperature rises sharply from 27 to 123°C within a few nanoseconds as the energy stored in the circuit inductance is dissipated as heat in the device; the drain-source voltage also increases abruptly up to the BV concomitantly with temperature. The peak junction temperature and maximum drain-source voltage occur at the same time because the BV positively correlates with the junction temperature. Subsequently, the device reverts to room temperature after *ca.* 175 μs of avalanche operation, and at which point the drain-source

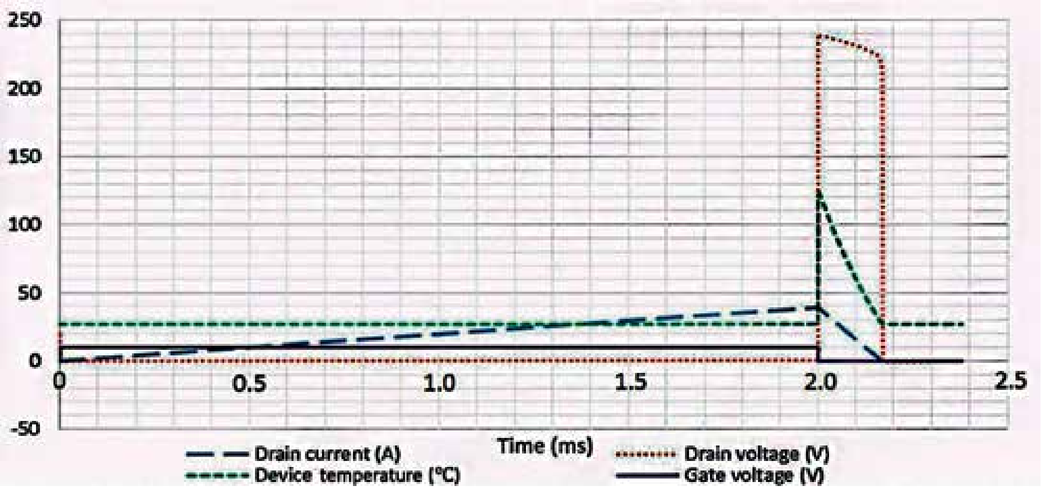


Figure 19. Waveforms under avalanche operation [17].

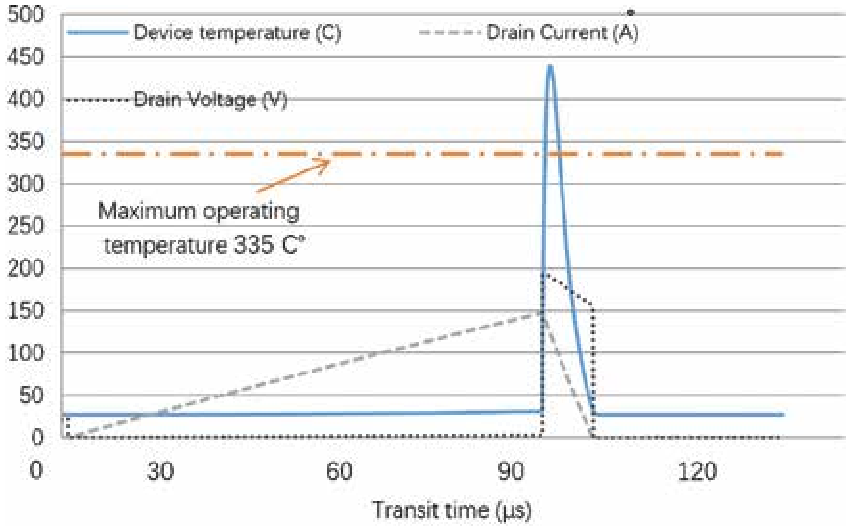


Figure 20. Waveforms under UIS test conditions when avalanche failure is believed to occur. The maximum operating temperature is 335°C [17].

voltage plummets to zero after the excess heat is dissipated into ambient air. According to Eq. (3), the maximum avalanche current is 40 A, which agrees with the simulation result. Also from **Figure 19**, the avalanche time is 0.18 ms, which agrees with calculation using Eq. (4) where BV is 220 V and VDD is 0 V. The EAS is 800 mJ computed according to Eq. (5), and therefore the avalanche power is 4.4×10^4 W (EAS/t_{av}).

3.3. Avalanche performance of planar MOSFET

For the Si device under testing (DUT) to survive under avalanche operation, the device junction temperature cannot exceed 335°C [16]. Otherwise, a large proportion of defects would be thermally generated in the epitaxial layer [16]. As a result, current crowding into a localized hot spot would occur on the chip, melting the aluminum around it and thus destroying the device. Upon optical inspection, the majority of the bare die reveals a catastrophic body diode melt down failure (not shown). **Figure 20** shows avalanche operation when the junction temperature exceeds 335°C. Under this condition, the MOSFET is thought to have failed to survive as the semiconductor approaches intrinsic properties at this high temperature. The lattice temperature profile shown in **Figure 21** illustrates a hot spot at the junction curvature between the *p*-well and the *n*-drift region, where avalanche breakdown occurs. A large amount of current passes through this junction curvature and in the process dissipates substantial power so that the local temperature in this region is the highest.

For a given inductance (0.01 mH), the relationship between the initial junction temperature and maximum avalanche current is shown in **Figure 22**. A linear regression of the data indicates that the maximum initial junction temperature is around 350°C, which closely agrees with the threshold for avalanche failure. For constant inductance, the maximum avalanche current is:

$$I_{av(\max)} \propto T_{jM} - T_{j0} \tag{6}$$

where T_{j0} and T_{jM} are the initial and maximum junction temperatures respectively.

Figure 23 shows the inductive load dependence of $I_{av(\max)}$ at an initial junction temperature of 27°C. As expected, the avalanche current capability becomes weaker as the inductive load increases, and this is because the proportionately large amount of energy stored in the inductance is dissipated in the MOSFET as heat, and risks avalanche failure when the critical lattice temperature is exceeded.

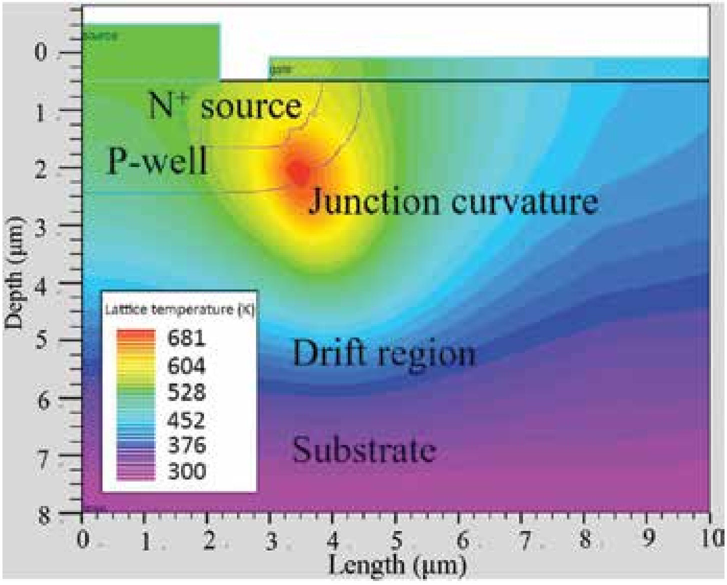


Figure 21. Lattice temperature distribution in the power MOSFET [17].

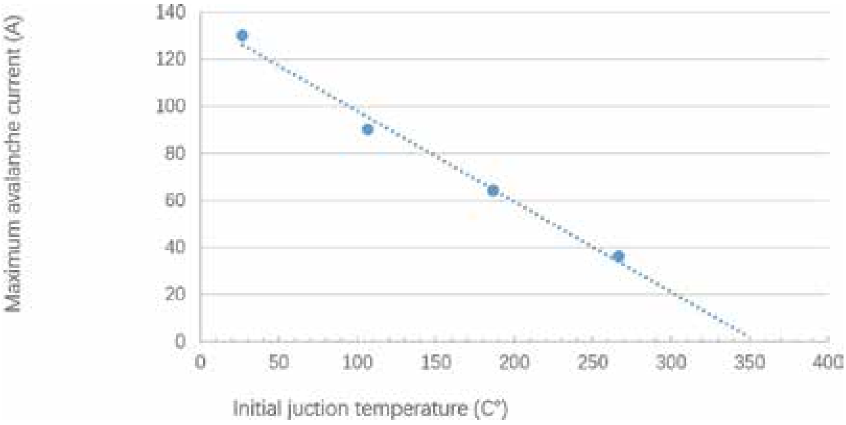


Figure 22. Initial junction temperature and resulting maximum avalanche current [17].

3.4. Avalanche performance of trench MOSFET

The cell pitch of the trench MOSFET can be reduced to 2.5 μm, from the 10 μm of the conventional planar MOSFET. And to maintain the same active area (5 mm^2), the width of the trench MOSFET can also be increased four-fold compared to that of the planar MOSFET. **Figure 24** shows the maximum avalanche current for the planar and trench platforms at an initial junction temperature of 300 K. Clearly, the avalanche current capability of the trench

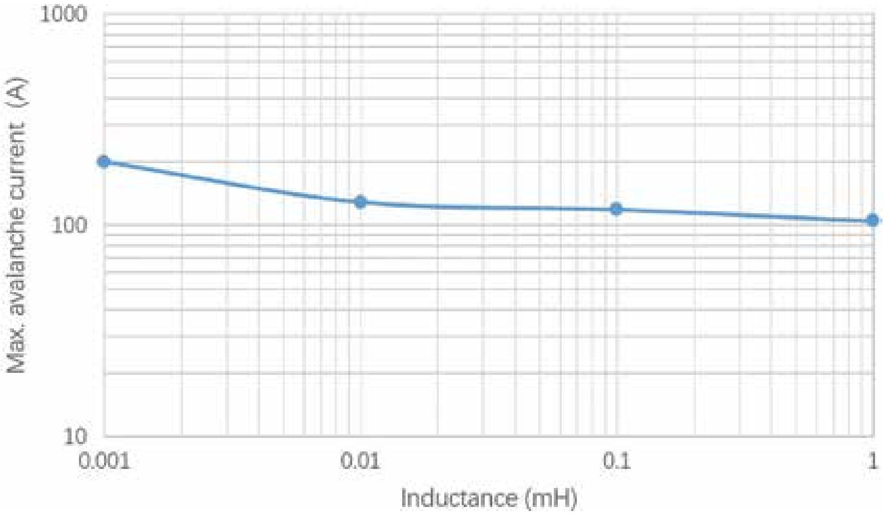


Figure 23. Maximum avalanche current as a function of total inductance [17].

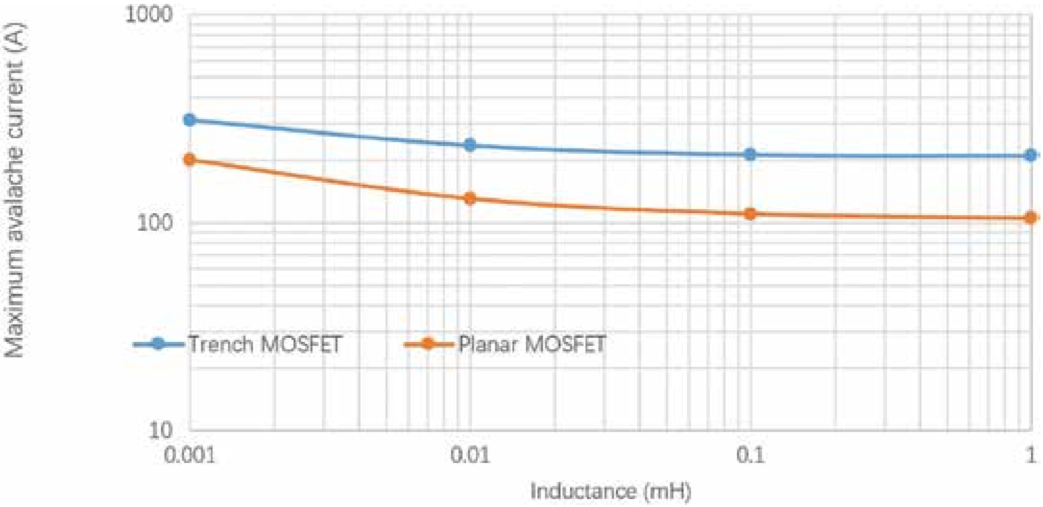


Figure 24. Maximum avalanche current in the planar and trench MOSFETs [17].

variant is 50–100% superior to that of the planar counterpart. **Figure 25** shows that the highest temperature is localized at the planar junction between the *p*-well and the *n*-drift regions, at which point avalanche breakdown occurs.

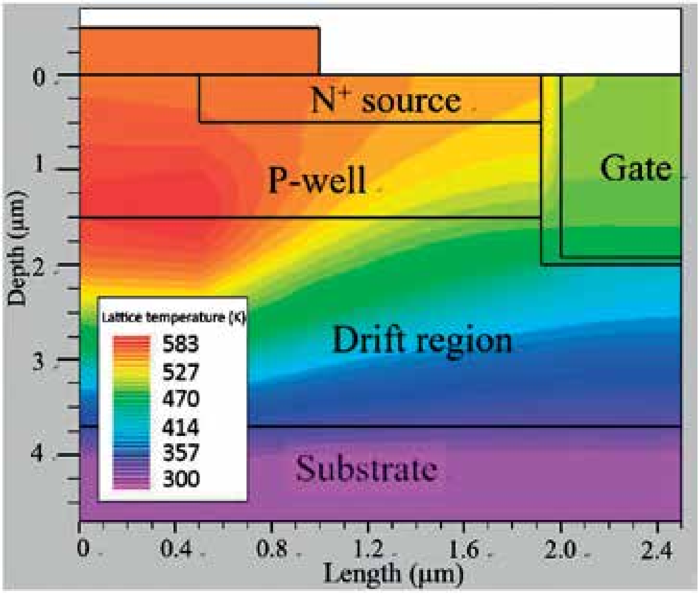


Figure 25. Temperature distribution in the trench MOSFET [17].

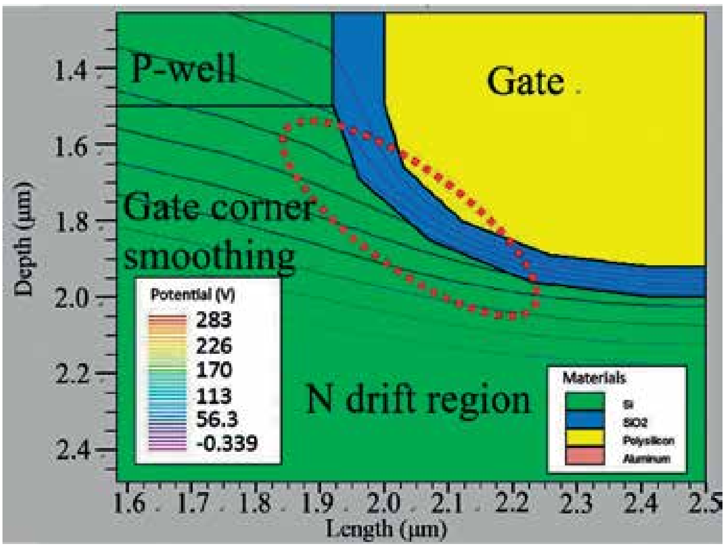


Figure 26. Trench MOSFET with gate corner rounding and potential distribution during avalanche operation [17].

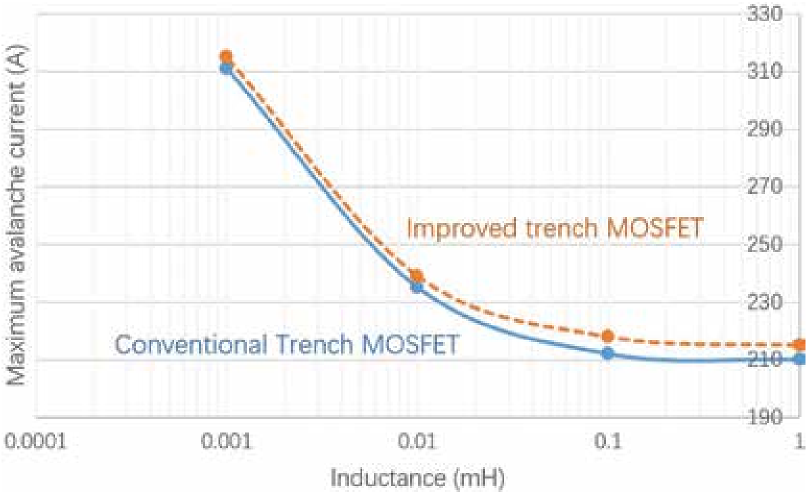


Figure 27. Maximum avalanche current for standard and modified trench MOSFET with rounded gate corners [17].

3.5. Ruggedness improvement of the trench MOSFET

Rounding off the trench gate corners is an approach that can avoid highly intense electric fields under UIS conditions and improve the ruggedness. The resultant potential contours exhibiting less crowding at the edges of the trench gate corner due to the modified design is shown in **Figure 26**. **Figure 27** shows that the maximum avalanche current increases by about 4–10 A per cell using the modified trench gate structure.

4. Conclusions

In this chapter, 50 V rated power MOSFETs based on the planar and trench technologies have been designed, modeled, simulated and compared using industry-standard Technology Computer-Aided Design (TCAD) tools. A survey of some methods to successfully reduce the specific on-resistance has been given. The specific on-resistance can be reduced by 23% through gate width-length optimization of the standard planar Si MOSFET. The increased doping in the JFET region decreases the specific on-resistance by about 8.3% but affects BV. Adopting SiC is more attractive and effective amongst the planar technologies studied where the specific on-resistance can be reduced by *ca.* 31% compared to the planar Si MOSFET. This arises from a shorter cell pitch and heavier doping in the drift region that substantially reduces the drift region resistance. Since the trench MOSFET has no JFET region, optimal design is achievable with a smaller cell pitch. By shrinking the half-cell pitch to 2.5 μm, i.e. reduction by 17% compared to that of the Si planar MOSFET with optimum gate width, the specific on-resistance decreases by more than two-fold. The avalanche ruggedness of the planar and trench MOSFETs has also been evaluated and compared. Experimental microscopy images show notable damage on the die due to avalanche failure. The physical mechanisms that limit the avalanche capability including self-heating effects have been analyzed and taken into account in the electro-thermal modeling and simulations of the circuit performance. The avalanche ruggedness of the trench MOSFET is significantly better compared to that of the planar MOSFET, exhibiting a 50–100% increase in avalanche current capability. For further ruggedness enhancement, the corners of the trench gate may be rounded off to smoothen out the electric field peaks at the edges under UIS conditions. This is expected to increase the maximum avalanche current capability by up to about 3% per cell. However, it may be argued that although the benefit in rounding the trench gate corners scales with the cell density and die size to handle high current levels, it may be outweighed by the additional process costs. Further, due to model simplifications (e.g. one- or two-dimensional finite-element modeling), the simulations investigate the first order effects but do not consider the second or higher order effects. Therefore for more accurate baseline models, the designed edge termination such as the field guard ring structure (see **Figures 9** and **10**), for example, should be taken into account in the simulations and calibrated by the experimental data. Finally, it is crucial to have well designed packaging, such as the bond wires that are imperative to handle high current levels.

Author details

Kuan W.A. Chee[1]* and Tianhong Ye[2,3,4]

*Address all correspondence to: kuan.chee@nottingham.edu.cn

1 Department of Electrical and Electronic Engineering, Faculty of Science and Engineering, The University of Nottingham Ningbo China, Ningbo, Zhejiang, People's Republic of China

2 Department of Electrical and Computer Engineering, Technische Universität München, München, Germany

3 School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore

4 Singapore Design Centre and Headquarter Asia, STMicroelectronics Pte Ltd, Singapore, Singapore

References

[1] Markets and Markets. Power Electronics Market by Material (Silicon, SiC, GaN, Sapphire), Device Type (Discrete, Module, and IC), Vertical (ICT, Consumer Electronics, Power, Industrial, Automotive, and Aerospace and Defense), and Geography - Global Forecast to 2022, Marketsandmarkets.com. 2017 [Online]. Available: http://www.marketsandmarkets.com

[2] Grant D, Gowar J. Power MOSFETS Theory and Applications. New York: Wiley; 1989

[3] STMicroelectronics. Power MOSFET. Available from: http://www.st.com/en/power-transistors/power-mosfets.html [Accessed: 28-07-2017]

[4] Blackburn DL. Turn-off Failure of Power MOSFETs. IEEE Power Electronics Specialists Conference, Puerto Varas, Chile, June 1985

[5] QYResearch Group. Global Discrete Power Device Market 2017 Industry Trends, Sales, Supply, Demand, Analysis & Forecast to 2022; May, Los Angeles, U.S.; 2017

[6] Infineon Technologies AG. Automotive Power Selection Guide. Neubiberg, Germany; 2016

[7] Infineon Technologies AG. Company Presentation; May 2015. p. 6. Retrieved from: http://www.equitystory.com/download/companies/infineon/Presentations/IFX_2015_Q2_en_web.pdf

[8] Erlbacher T. Lateral Power Transistors in Integrated Circuits. Switzerland: Springer International Publishing; 2014. p. 5. ISBN: 978-3-319-00500-3, Chapter 2

[9] Ng JCW, Sin JKO. A low voltage planar power MOSFET with a segmented JFET region. IEEE Transactions on Electron Devices. Aug. 2009;**56**(8):1761-1766

[10] STMicroelectronics. SiC MOSFETs. Available from: http://www.st.com/en/power-transistors/sic-mosfets.html?querycriteria=productId=SC1704 [Accessed: 26-06-2017]

[11] Ye T, Chee KWA. Low on-resistance power MOSFET design for automotive applications. In: IEEE 11th International Conference on ASIC; 2015

[12] Baliga BJ. Fundamentals of Power Semiconductor Devices. New Delhi, India: Springer Science + Business Media, LLC; 2008. p. 351-352

[13] Hull B, Allen S, Zhang Q and Gajewski D. Reliability and stability of SiC power mosfets and next-generation SiC MOSFETs. IEEE Workshop Wide Bandgap Power Devices and Applications (WiPDA), Knoxville, TN, USA ,13-15 Oct. 2014

[14] Murray A, Davis H, Cao J, Spring K, McDonald T. New Power MOSFET Technology with Extreme Ruggedness and Ultra-low RDS(on) Qualified to Q101 for Automotive Applications. International Rectifier, Inc; 2000

[15] Blackburn DL. Power MOSFET failure revisited. Power Electronics Specialists Conference, 19th Annual IEEE, Kyoto, Japan, 11-14 April; 1988

[16] Stoltenburg RR. Boundary of power MOSFET, unclamped inductive-switching (UIS, avalanche current capability). In: APEC IEEE 1989 Conference Proceedings. pp. 359-364

[17] Ye T, Chee KWA. Ruggedness evaluation and design improvement of automotive power MOSFETs. In: 17th International Symposium on Quality Electronic Design (ISQED); 2016

[18] Schleisser D, Ahlers D, Eicher M, Purschel M. Repetitive avalanche of automotive MOSFETs. In: 15th European Conference on Power Electronics and Applications, Lille; 2013. pp. 1-7

Edited by Kim Ho Yeap and Humaira Nisar

In this book, *Complementary Metal Oxide Semiconductor* (CMOS) devices are extensively discussed. The topics encompass the technology advancement in the fabrication process of metal oxide semiconductor field effect transistors or MOSFETs (which are the fundamental building blocks of CMOS devices) and the applications of transistors in the present and future eras. The book is intended to provide information on the latest technology development of CMOS to researchers, physicists, as well as engineers working in the field of semiconductor transistor manufacturing and design.

Published in London, UK

IntechOpen

ISBN 978-1-83881-512-7

9 781838 815127