Design of 3D protection against electrostatic discharges (ESD) in advanced silicon on insulator FDSOI thin film multilayer technology

Louise de Conti

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **NANO ELECTRONIQUE ET NANO TECHNOLOGIES**

Arrêté ministériel : 25 mai 2016

Présentée par

Louise DE CONTI

Thèse dirigée par **Philippe GALY (STMicroelectronics SA)**
et codirigée par **Sorin CRISTOLOVEANU (IMEP-LAHC) & Maud VINET (CEA LETI)**

Préparée au sein du laboratoire **IMEP-LAHC (Institut de Microélectronique, Electromagnétisme et Photonique - Laboratoire d'Hyperfréquences et de Caractérisation)** dans **l'École Doctorale EEATS (Electronique, Electrotechnique, Automatique et Traitement du Signal)**

Conception de protection 3D contre les décharges électrostatiques (ESD) en technologie silicium avancée sur isolant (FD-SOI) film mince multi couches

Design of 3D protection against electrostatic discharges (ESD) in advanced silicon on insulator FD-SOI thin film multilayer technology

Thèse soutenue publiquement le **2 octobre 2019**,
devant le jury composé de :

Dr. Dionyz POGANY
Professeur à Vienna University of Technology, rapporteur
Dr. Bruno ALLARD
Professeur à l'université de Lyon, rapporteur
Dr. Jean-Pierre COLINGE
Directeur de recherche émérite au CEA LETI Grenoble, examinateur
Dr. Nathalie LABAT
Professeur à l'Université de Bordeaux, examinatrice et présidente du jury
Dr. Philippe GALY
Directeur technique à STMicroelectronics Crolles, directeur de thèse
Dr. Sorin CRISTOLOVEANU
Directeur de recherche émérite CNRS Grenoble, co-directeur de thèse

“What is the use of a house if you haven't got a tolerable planet to put it on?”

Henry David Thoreau

Acknowledgements

"I would rather walk with a friend in the dark,
than alone in the light."
Helen Keller

I would like to thank my supervisors Philippe Galy, Sorin Cristoloveanu and Maud Vinet, for entrusting me with my thesis. They have followed my work all along, suggesting very relevant ideas that have contributed to the final accomplishment of this work. I am particularly grateful to Philippe, who was always patient and available for me. His enthusiasm in front of any idea or scientific work was the vector to find our best results.

I would additionally like to thank the members of the jury and invited members – notably Dionyz Pogany, Bruno Allard, Nathalie Labat and Jean-Pierre Colinge – for their interesting remarks about my work. Also, I want to thank the members of the small jury who evaluated me during the meeting that occurred each year, for the renewal of my PhD contract: Maryline Bawedin and Jean-Pierre Colinge (again).

A special thanks goes to Sotiris Athanasiou, who helped me to understand the ESD topic in the beginning of my PhD, and who was of great support outside working hours during my PhD. I also received a great understanding help from Johan Bourgeat, Ghislain Troussier and Benjamin Viale. I am grateful to Charles-Alexandre Legrand and Blaise Jacquier, who worked with me to perform the ESD measurements, and to Rudy Costanzi, who gave me a formation about the DC measurement test bench. I am thankful to all the ESD team of ST Crolles, who kindly accepted me as a member during the internal international STMicroelectronics ESD workshop.

I do not want to forget the whole MPW team, for the uncountable hours of work to enable the fabrication of wafers. Stéphanie Chouteau, Jocelyne Gimbert and Vincent Dumettier were my most frequent contacts, but many other people put energy on the release of a wafer. Franck Arnaud also was always supporting me and helping me with my designs. I had the occasion to improve them a lot during the timeframe of my PhD, and this is all thanks to the knowledge of Jean-Claude Marin, Bertrand Le-Gratiet, Christian Gardin, Cyril Renard, Charlotte Beylier, Benjamin Dumont and Stéphane Martin.

I want to thank the patent team of STMicroelectronics also: Jean-Marc Tessier, Caroline Maranini and all their colleagues.

I thank Pascale Maillet-Contoz and Annaïck Moreau for taking care of the administrative procedures.

I was presented to a very pleasant group of PhD students with whom I enjoyed eating for lunch, and who – for most of them – became my dear friends. I would like also to thank the engineers I was sharing the open-space with. (I really hope that I did not forget anybody) Thanks to Renan Letiecq, Mohammed Tmimi, Ioanna Kriekouki, Valérian Cincon, Geoffrey Delahaye, Raphael Guillaume, Florian Voineau, David Gaidioz, Florent Torres, Romane

Dumont, Robin Benarrouch, Hanae Zegmout, Mirjana Videnovic-Misic, Hani Malloug, Hani Sherry, Antoine Delmas, Alexia Valery, Hassan El Dirani, Angel De Dios Gonzales Santos, Thibault Despoisse, Zoltan Nemes, Charlelie Pignol, Clément Beauquier, Antoine Le Ravallec, Guillaume Tochou, Alexis Rodrigo Iga Jadue, Thomas Capelli, Andreia Cathelin, Tarun Chawla, Yelda Ozgur, Olivier Jeantet, Thomas Ahrens, Franck Genevaux, Solenne Bergoin, Nour Ben Salem, Veronique Ollagnier and François Maillard, my PhD was much more pleasant. All those people helped me a lot discovering who I was, and what path I want to follow. I wish them all the best for the future.

Thomas Bedecarrats should figure among the list of PhD students who shared their lunch with me, but I had to book for him an entire paragraph, due to his enormous contribution to my work. Not only he was of great support, sharing all my sadness' and joys, but he also truly fed me with his scientific knowledge.

I thank all the readers of this manuscript, yes, also *you*.

I acknowledge all the people who re-use their lunch glass goblet for coffee (instead of wasting a plastic glass) in STMicroelectronics cafeteria.

Last but not least, I am grateful for all the love I received - during this PhD time but also before those three years - from my family, my friends outside work, and from Hector Pharam.

Outline

Dissemination

Journal publication as a first author

L. De Conti, T. Bedecarrats, S. Cristoloveanu, M. Vinet and Ph. Galy, “GDNMOS and GDBIMOS devices for high voltage ESD protection in thin film advanced FD-SOI technology”, Solid State Electronics, vol. 159, pp. 90-98, September 2019.

International conference publications

L. De Conti, T. Bedecarrats, M. Vinet, S. Cristoloveanu and Ph. Galy, “Toward Gated-Diode-BIMOS for thin silicon ESD protection in advanced FD-SOI CMOS technologies”, 2017 IEEE International Conference on IC Design and Technology (ICICDT), Austin TX USA, May 2017.

L. De Conti, S. Cristoloveanu, M. Vinet and Ph. Galy, “GDNMOS and GDBIMOS devices for ESD protection in 28nm thin film UTBB FD-SOI technology”, EUROSOI-ULIS 2018: Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon, Granada, March 2018.

L. De Conti, S. Cristoloveanu, M. Vinet and Ph. Galy, “Thin-film FD-SOI BIMOS topologies for ESD protection”, IRPS 2019: IEEE International Reliability Physics Symposium, Monterey CA USA, April 2019.

Ph. Galy, L. De Conti, G. Delahaye, L. Anghel, M. Vinet, S. Cristoloveanu, “Topology and design investigation on thin film silicon BIMOS device for ESD protection in FD-SOI technology”, ESREF, Toulouse, 2019, *IN PRESS*.

Patents

Ph. Galy and L. De Conti, “ESD boost solution for MOS/BIMOS in thin silicon film & hybrid bulk for FD-SOI advanced technology” or “Dispositif électronique de protection contre les décharges électrostatiques”, STMicroelectronics reference 17-GR1-0545, patent filed on December the 13th, 2017.

T. Bedecarrats, L. De Conti and Ph. Galy, “Non homogeneous silicide for thin-film carrier transport control for example to build an ESD high or low voltage protection in UTBB FD-SOI technology” or “Circuit électronique”, STMicroelectronics reference 18-GR1-0233, patent filed on June the 29th, 2018.

L. De Conti and Ph. Galy, “Transistor body contact, in particular for a matrix arrangement” or “Prise de contact substrat pour transistor, destiné en particulier à un arrangement matriciel”, STMicroelectronics reference 17-GR1-0713, patent filed July the 3rd, 2018.

L. De Conti and Ph. Galy, “3D ESD protection in FD-SOI with silicon continuity thanks to compact hybrid direct merged thin film” or “Circuit électronique”, STMicroelectronics reference 18-GR1-0531, patent filed on May the 9th, 2019.

Ph. Galy and L. De Conti, “In-situ coupled bias resistance in thin silicon film and/or hybrid bulk for FD-SOI ultra compact ESD device protection improvement”, STMicroelectronics reference 18-GR1-0667, patent filed on May the 9th, 2019.

Chapter 1: Introduction

"Tell me and I will forget
Show me and I will remember
Involve me and I will understand."
Confucius

The Integrated Circuits (IC) using MOSFET (Metal Oxide Semiconductor Field Effect Transistor) [1] devices have drastically evolved those last decades, thanks to the components' miniaturization (the semiconductor industry has grown along with the Moore's law). One of the main reasons for this miniaturization is the cost reduction of the integrated circuits that goes along with the increase in the number of components for the same silicon area used. The other main reason is the performance of devices and circuits. In order to enable the MOSFET scaling, some technological parameters had to be modified. For example, the gate oxide that was initially made of SiO_2 is now composed of a high-k dielectric layer in addition to a SiO_2 layer, in order to increase the gate capacitance without shrinking the gate oxide thickness too much. Implantations have been introduced, for instance the LDD implantation (Lightly Doped Drain) that was used to limit the lateral electric field; the epitaxy to raise the drain and the source allowed to reduce the access resistance (this list of modifications is not exhaustive).

However, when pursuing the shrinking, physical limitations in the planar technology were so strong that the only way to continue the scaling was to change the MOSFET architecture drastically. One of the solutions is the Silicon-On-Insulator (SOI) technology [2]. Historically MOS transistors were built on top of silicon wafer. This is the bulk technology. However transistors with small gate length are subject to short channel effects such as DIBL (Drain Induced Barrier Lowering) [3], surface scattering [4], velocity saturation [5], impact ionization [6], Hot Carrier Injection (HCI) [7] and Gate Induced Drain Leakage (GIDL) [8] [9]. Some short-channel effects are reduced in the FD-SOI (Fully Depleted Silicon On Insulator) technology. FD-SOI transistors are built on the thin-film (silicon layer) that is on top of a Buried OXide (BOX), therefore they are insulated from the substrate (Figure 1).

Another way of improving the performance of dies (than shrinking the size of its components) would be to stack them in three dimensions (3D). This allows to increase the density of transistors per surface, but also to reduce the metallic interconnections' delays. A classical solution consists in stacking two processed wafers on top of each other; it is called the TSV (Through Silicon Via) method. But the alignment cannot be done with a high precision. For a 3D stack with a precision of the pitch of a transistor, the best is to process transistors on top of each other directly. This is the technique proposed by the Coolcube™ project. The problem is that the temperature used to create the top transistors will harm the bottom transistors. Thus, the temperature of fabrication of MOS transistor should be decreased in order to be able to use this 3D technique. Some very advanced work is done on "cold" CMOS (Complementary Metal Oxide Semiconductor) technology in CEA [10]. Other 3D

implementations could also be imagined. This field is relatively new, and many opportunities and solutions can still be discovered.

Now that scalability is at its limits, we have entered into the More than Moore era. The market of semiconductor industry is forecasted to grow thanks to the Internet of Things (IoT). New technological revolutions will be made in embedded non-volatile memories, hybrid devices with merged logic and memory functionalities, monolithic three-dimensional integration and heterogeneous integration techniques [11].

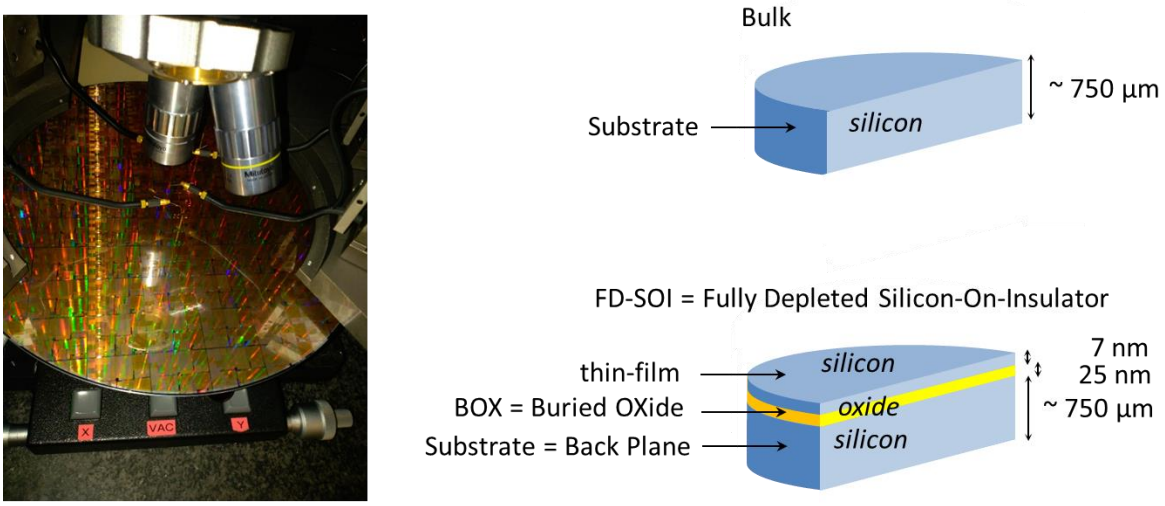


Figure 1: Left: Wafer under measurement. Right: Wafer in bulk and FD-SOI technology.

Because of the size of the components, all those technologies - regardless of whether they are planar or 3D - are extremely sensitive to electrostatic discharges (ESD) and need to be protected. Indeed, ESD events can happen at any time (during the fabrication of the circuit, or during its lifetime) and involve huge currents and voltages. In this manuscript some solutions will be presented.

I. Presentation of the technology

1. Introduction to the FD-SOI technology

MOSFET transistors were initially fabricated on top of bulk silicon wafers. Historically SOI technology was introduced for spatial industry requirements since transistors were less sensitive to radiations. Nevertheless, FD-SOI technology presents other advantages. It enabled to shrink devices further, mainly because it was reducing short channel effects. It also suppressed other parasitics that could affect negatively the devices' behavior. Indeed, the BOX is here to isolate the active device from the others and from the parasitic elements that are in the substrate. No current is flowing in the substrate from one transistor to the others, and the latch-up phenomenon with the parasitic SCR (Silicon Controlled Rectifier) is relaxed. Other parasitic elements of the substrate - that negatively impact the performance of transistors - are suppressed by the presence of the BOX: some parasitic capacitances (between the implants), parasitic bipolar transistors and different paths of leakage currents. One of the only remaining parasitic element is the capacitance of the BOX itself, but it can be considered as an opportunity to better control the device, by using the back-plane as a back gate. The electrostatic control is also better on FD-SOI devices thanks to the junctions that are less deep.

The fabrication of SOI wafer that is standardly used in the industry is via the Smart-Cut technique used by SOITEC [12]. At first there are two bulk wafers A and B. Wafer A is oxidized to create an insulating layer. The thickness of this insulating layer will correspond to the one of the BOX of the future SOI wafer. Then the Smart-Cut ion implantation is performed in order to form an in-depth weak layer (where the wafer will be cut afterward). The depth of this hydrogen implantation corresponds to the thickness of the silicon film above the BOX. After some cleaning, wafer A and B are bonded. The smart cut is performed, and the bonded wafer is cut at the weak layer position of implanted ions. Wafer "B" is becoming the SOI wafer after annealing and CMP (Chemical Mechanical Polishing). Wafer "A" is recycled and can be used to become a new wafer A or B for another SOI wafer fabrication (Figure 2).

The thickness of the silicon film on top of the BOX is managed by adjusting the implantation energy. When the film is sufficiently thick, the depleted region under the gate in a NMOS does not reach the BOX, and the transistor is partially depleted; it is PDSOI (Partially Depleted SOI). When it is less than 50 nm, the silicon film is called thin-film, and the transistor is fully depleted (FD-SOI).

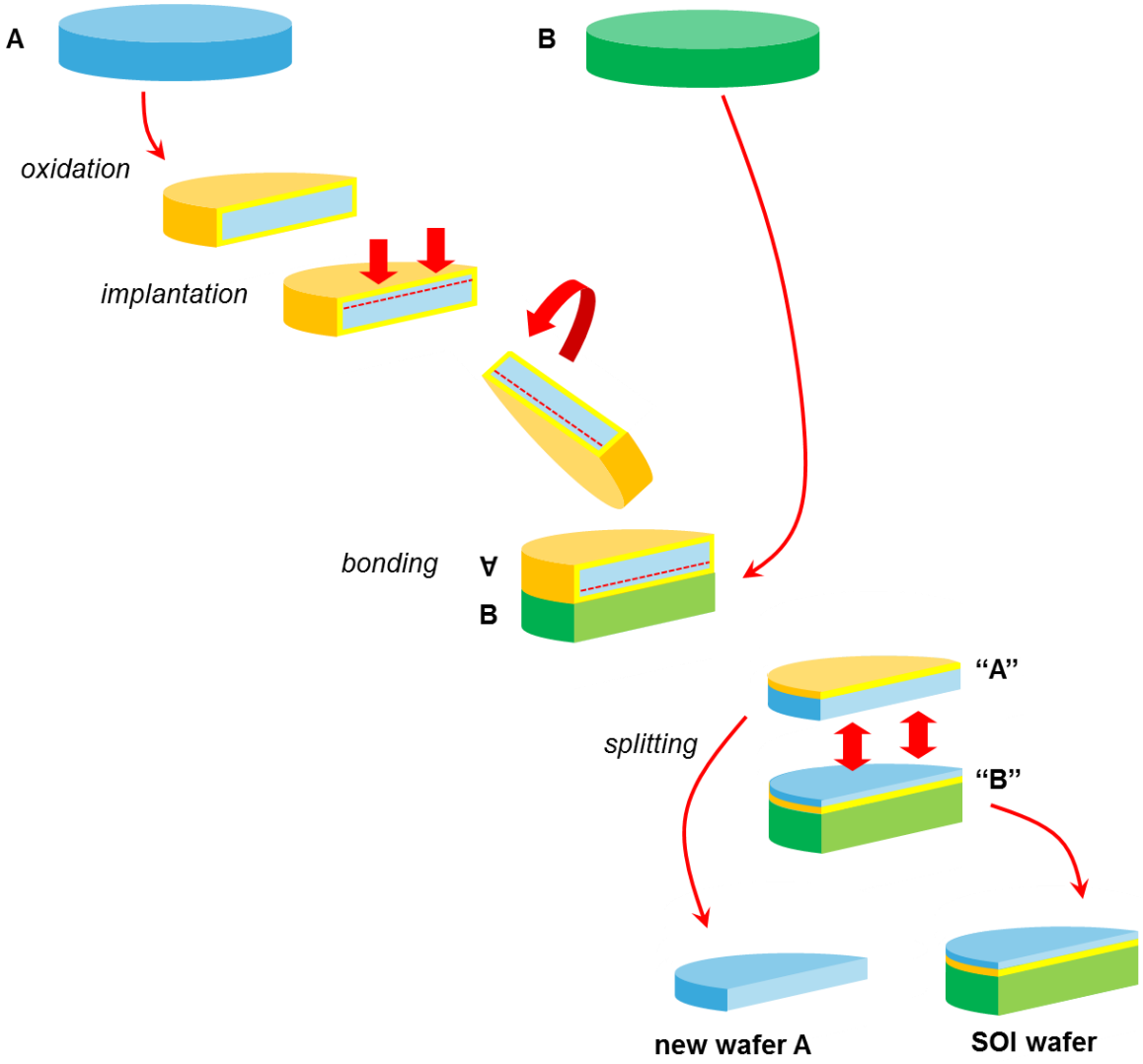


Figure 2: Schematic of the fabrication process of SOI wafers using the Smart-Cut™ process. Adapted from [13].

2. MOSFET in the 28 nm UTBB FD-SOI technology

The MOSFET in thin-film FD-SOI is a four-terminal device: the gate, the back-plane - or substrate (which can also be considered as a back gate) -, the source and the drain. It is mainly used as an electrical switch. Indeed, by controlling the gates' voltages, the transistor can let the current pass between the source and the drain (in logic it is called a "1" and the transistor is considered ON) or block it (then it is a "0" and the transistor is OFF). The transistors can be of type N (then the carriers flowing from one side of the channel to the other are electrons) or P-type (the carriers are holes). Here are the different features of the MOS transistors in the 28 nm FD-SOI Ultra-Thin-Body and BOX (UTBB) technology node: two types of back-plane (NWELL and PWELL doping) make it possible to get two different threshold voltages for each type of MOSFET. If the doping is the same type in the back-plane and in the channel, it is a LVT (Low Threshold Voltage) transistor, else it is a RVT (Regular Threshold Voltage) transistor (Figure 3). The structure of the RVT N-MOS FD-SOI transistor is described in Figure 4. In the NRVT MOS, the channel of the transistor is left undoped (Pint doping). The source and the drain are epitaxially raised for reducing the access resistance, and their doping is N^+. The lightly doped extensions of the source and the drain (N-LDD doping) are here to limit the lateral electric field (indeed, the reduction of the lateral field is less abrupt with extensions, because it is staggered all along the extension) in order to attenuate hot carrier degradation. The goal of the spacer is to limit the prolongation of the lightly doped extensions under the gate. The metallic gate is made with polysilicon and titanium nitride. This material has been chosen for its work function, because it is a mid-gap material allowing to have equilibrate threshold voltages between the NMOS and the PMOS [14]. In the 28 nm FD-SOI technology that is used in STMicroelectronics [15] [16] [17] [18], the channel is very thin (T_{Si} = 7 nm) for a better electrostatic control of the gates on the channel. The BOX is 25 nm thick. Under the BOX, the back-plane is P-doped (PWELL doping). The NISO doping (or deep NWELL) is used to isolate this PWELL doping from the rest of the substrate. On top of source, drain and gate, salicidation is realized to reduce the access resistance, and then metallic connections are made. The transistor is insulated from others thanks to oxide trenches named STI (Shallow Trench Isolation).

When a positive bias is applied on the gate (V_G), a conductive N-channel (made of electrons) takes place in the silicon channel under the gate oxide between the source and the drain, and a vertical electric field is created in the oxide. If a positive bias is also applied on the drain (V_D), a lateral electric field is generated, and the carriers in the conductive channel can move from source to drain; this leads to a current I_D. Polarizing the substrate can have some interests for example in case of threshold voltage tuning. The different modes of the transistor operation are seen on the I_D vs V_G curve for a fixed V_D (Figure 5), for example $V_D = V_{D\_USE} > V_{DSAT}$:

- The transistor is blocked when V_G = 0 V ($I_D = I_{OFF}$)
- It is in weak inversion when $V_G < V_{TH}$
- It is in strong inversion when $V_G > V_{TH}$ (usually I_{ON} is selected at $V_G = V_{D\_USE}$).

V_{TH} is the threshold voltage delimiting the two states (weak and strong inversion). The Subthreshold Slope (SS) is extracted from the curve slope below V_{TH} while plotting I_D vs V_G in logarithmic scale:

$$SS = \frac{\partial I_D}{\partial V_G}$$

Its theoretical limit is 60 mV/decade at 300 K. I_{OFF} is undesirable because it corresponds to a leakage current increasing power consumption when transistors and circuits are in standby mode.

If the thickness of the gate oxide (Tox) is reduced, then the gate control on the channel is enhanced, but the possibility that carriers cross the potential barrier of the gate oxide by tunnel effect is increased. The resulting leakage current between the channel and the gate - but also between the gate and the source or drain junctions (because of the underlap) - is then stronger. Silicon oxide is replaced by high-k dielectrics in order to minimize this effect; indeed, the thickness of the oxide can be larger for a same value of capacitance of the gate oxide. That is why the term of EOT (equivalent oxide thickness) is employed. An interfacial layer of SiO_2 is necessary to be able to deposit HfO_2 above the silicon thin-film. GO1 (thin oxide) RVT transistors have a V_{D_USE} = V_{DD} of 1 V and GO2 (thick oxide) transistors have a V_{DD} of 1.8 V. The typical threshold voltage of a GO1 RVT transistor is about 0.4 V.

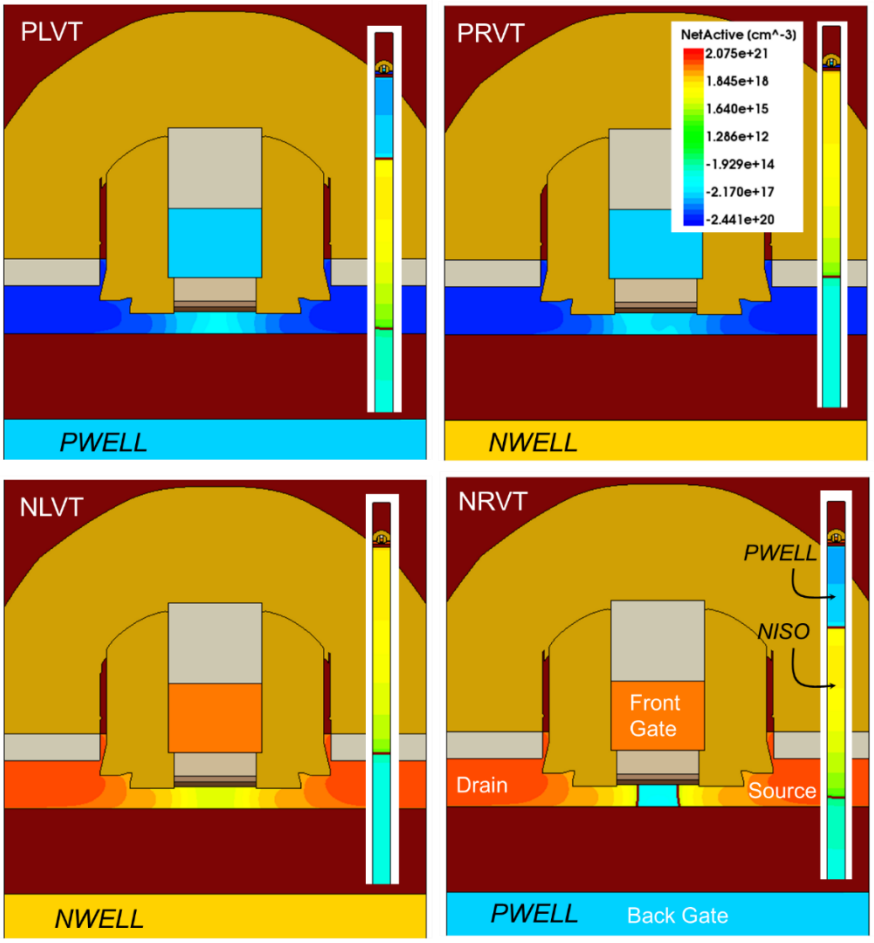


Figure 3: Cross section of PLVT, PRVT, NLVT and NRVT transistors in UTBB FD-SOI technology. Insights: doping scale and structures with the substrate doping.

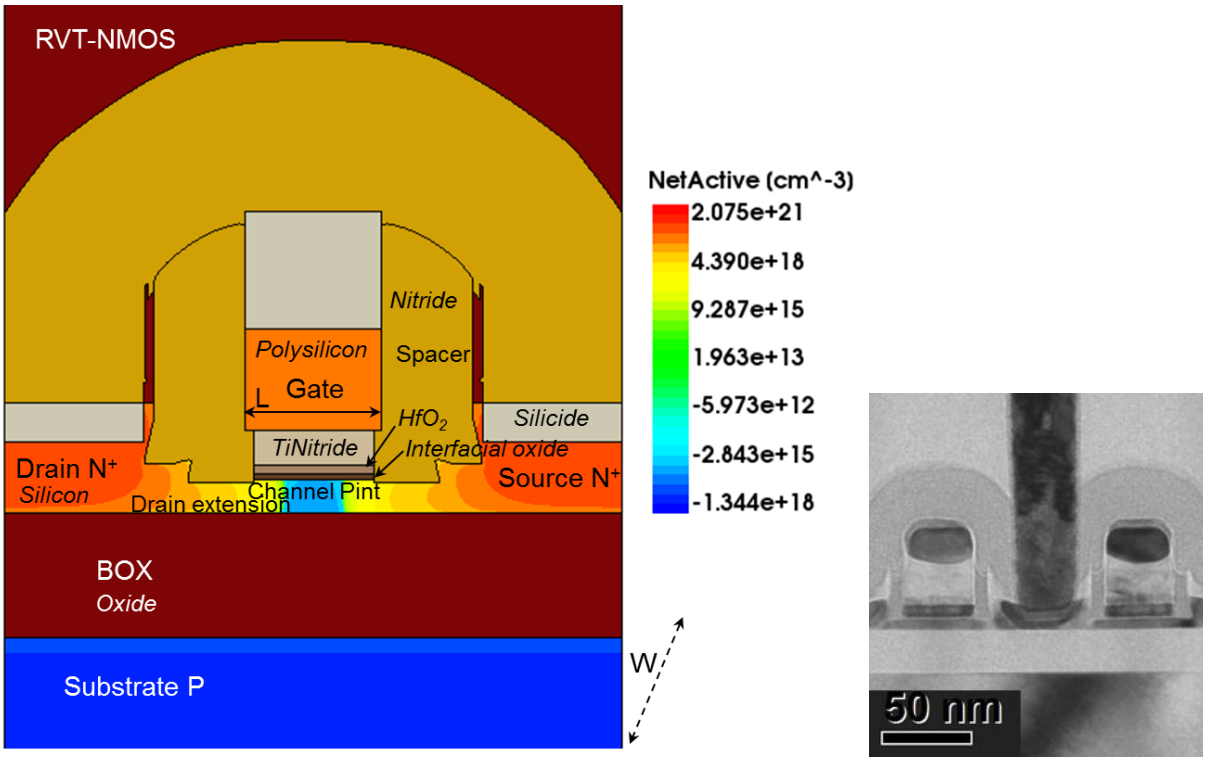


Figure 4: Left: Cross section of a N-MOS FD-SOI transistor. Right: TEM (Transmission Electron Microscopy) view of a multi-finger NMOS transistor.

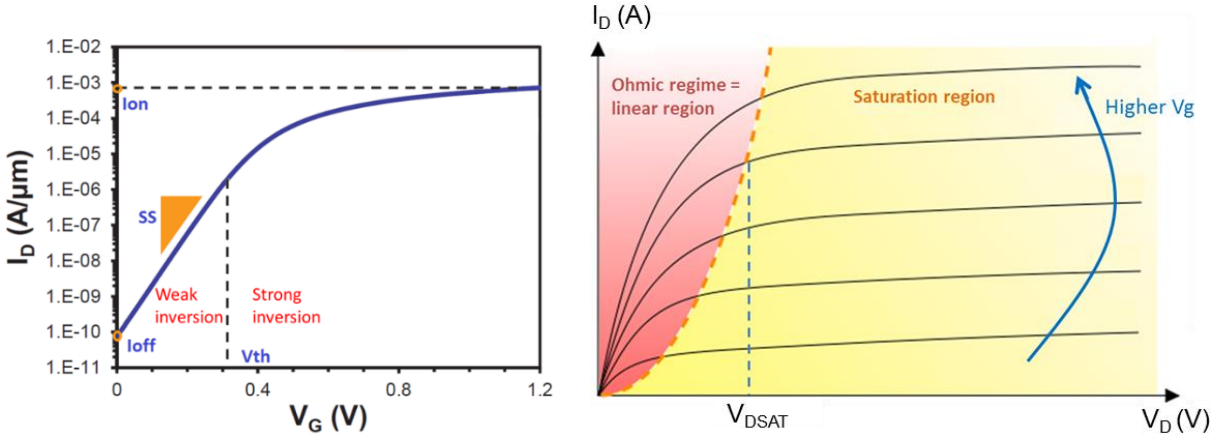


Figure 5: Left: Typical I_D vs V_G curve (at fixed V_D = 0.9 V). Right: Typical I_D vs V_D curve (for various V_G). Adapted from [19].

II. The electrostatic discharge

1. Definition of ESD and importance of ESD protections

An electrostatic discharge (ESD) is the sudden flow of current between two electrically charged objects ("The rapid, spontaneous transfer of electrostatic charge induced by a high electrostatic field." [20]). In the everyday life, one can experience it - for example - by feeling a spark when touching the door of a car or walking on a carpet. Another very common example of discharge is a lightning bolt. ESDs can be caused by triboelectricity, induction or direct conduction, and happen all the time, involving different orders of magnitude of current and voltage. A human has to be charged to 3 kV minimum in order to be able to feel the ESD, he has to be charged to 5 kV to hear it and to 10 kV to see a spark [21]. In the industry, electronic circuits are supposed to undergo ESDs up to 4 kV at component level, and multiple amperes in a few nanoseconds, which may already be destructive if the circuit has not been protected. As a matter of fact, more than 50% of all failures are attributed to ESD and EOS (Electrical OverStress), ESD being a subset of EOS (Figure 6) [22]. "Electrical overstress (EOS) is any electrical stress that exceeds any of the specified absolute maximum ratings (AMR) of a product and causes it to fail (reversibly or irreversibly, immediately or delayed)." [23]

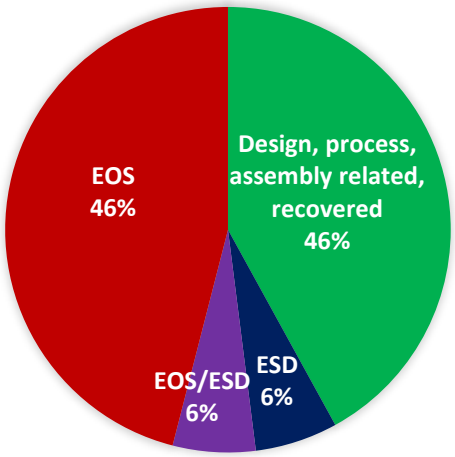


Figure 6: Distribution of failure models in silicon ICs. Adapted from [22].

An ESD event can damage the oxides [24], junctions, metals [25], and the plastic of the package of the circuit, as well as produce hot carrier degradation and melting (Figure 7). For example, the discharge can provoke an increase in drain voltage on a transistor, therefore increasing the electrical field between the drain and the gate. When this field is too high the gate oxide is damaged [26] [27], which induces leakage current in the MOSFET, or even a short circuit between the drain and the gate. In the channel, discharges increase the temperature [28]. The semiconductor can reach fusion and conductive filaments [29] can be created between the source and the drain. Due to the high voltages, impact ionization occurs, therefore hot carriers [7] are generated and injected into the oxides, damaging them and increasing the leakage in the MOSFET. A critical value of electron trap density injected into the oxide leads to gate oxide breakdown [30] [31]. In the metallic vias and interconnections, a discharge can cause electro-migration or simply increase the temperature by Joule effect, and some melting can occur, thus cutting the line. Metal filaments can also intersect junctions in the devices [32]. As a consequence, soft damage (aging) or even hard damage (immediate breakdown) happens [33]. As a matter of fact, there is a need in protecting integrated circuits from ESDs in the industry.

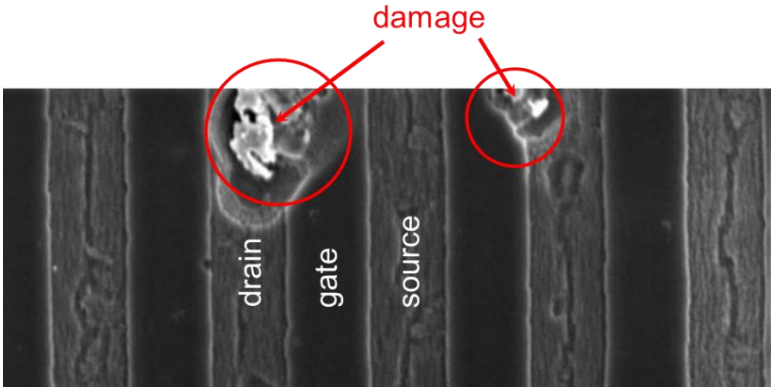


Figure 7: NMOS failure due to CDM (this term will be explained later). Information about this type of failure (the tongue effect) can be found in [34].

2. ESD stress standards

An infinite type of discharges is possible with large energy deposition, and it can be very expensive to develop methodologies to protect ICs from ESDs. That is why the ESDA (ESD Association) has been created to establish standards and share the improvements and developments between industrials and researchers. In this framework, three main models of ESD are used in microelectronics:

- The Human Body Model (HBM) describes a discharge from a charged human to the device.

- The Machine Model (MM) is used to mimic a discharge from a charged machine touching the device.

- The Charged Device Model (CDM) is used when the component is charged itself and the discharge is transferred to another object.

Those different configurations are represented each by an RLC equivalent circuit; thus, a second order differential equation can describe the ESDs.

Those models correspond to component-level test methods, useful for estimating the robustness of circuits in a manufacturing environment, while system-level ESD testing (or Gun testing) is relevant for final product testing. To predict the ESD performance under system-level stress condition, the Human Metal Model (HMM) [35] [36] has been proposed by the ESDA. Cable Discharge Event (CDE) [37] [38] [39] should also be taken into account for the qualification of products. This manuscript only deals with component-level tests so system-level tests are only mentioned for information purpose.

a. Human Body Model

For the HBM [40], the charge transfer happens when a charged person is touching the integrated circuit. The current is flowing from the touched pin of the chip to another one that is grounded. All the other pins are supposed to be left floating. The equivalent electrical circuit of this HBM discharge is depicted in Figure 8. The contact of the person touching the chip is considered of being a resistor of 1.5 kΩ (this is why R_{ESD} = 1.5 kΩ). C_{ESD} = 100 pF is representing the body capacitance of the person with respect to the ground. The impedance of the Device Under Test (DUT) is neglected (hypothesis of a short circuit: the model is valid when considering that the ESD protection device – which is the DUT – is ON because the ESD event was detected), as well as the parasitic capacitances of the tester (that is subject to a maximum number of tested pins). The inductance of the tester (used to test the device as if it was undergoing an HBM discharge) is considered (L_{ESD} = 7.5 µH). C_{ESD} is charged to a specific ESD voltage, then the contact is established and the discharge happens. The intensity of the discharge depends on this pre-charge voltage, considered as V_{HBM} = 1 kV for RF (radio frequency) applications and 4 kV for more challenging environment such as military applications. The ESDA is constantly making those voltage requirements evolve because it is more and more difficult to provide on-chip ESD protection devices that are able to stand

such high voltages. More and more protection devices are characterized and approved for being able to stand V_{HBM} = 500 V and less [41]. In real manufacturing environment, the pre-charge voltage is limited by preventive actions such as operators wearing wristbands and foot straps to dissipate the discharge, equipments being grounded, humidity control, etc.

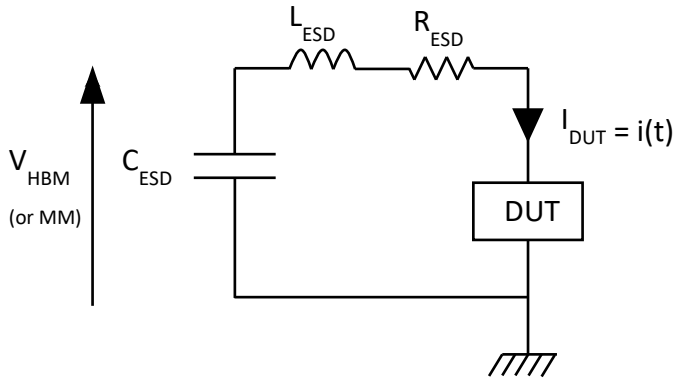


Figure 8: HBM and MM models and associated parameters.

The second order differential equation associated to the equivalent circuit is:

$$\frac{d^2i(t)}{dt^2} + \frac{R_{ESD}}{L_{ESD}} \cdot \frac{di(t)}{dt} + \frac{1}{L_{ESD} \cdot C_{ESD}} \cdot i(t) = 0$$

R_{ESD} is large enough, so:

$$\Delta = \left(\frac{R_{ESD}}{L_{ESD}}\right)^2 - \frac{4}{L_{ESD} \cdot C_{ESD}} > 0$$

Therefore, the discharge current can be calculated this way:

$$I_{HBM}(t) = \frac{V_{HBM} \cdot C_{ESD} \cdot {w_o}^2}{2 \cdot \sqrt{\lambda^2 - {w_o}^2}} \cdot \left(e^{\left(-\lambda + \sqrt{\lambda^2 - {w_o}^2}\right) \cdot t} - e^{\left(-\lambda - \sqrt{\lambda^2 - {w_o}^2}\right) \cdot t}\right)$$

with

$$\lambda = \frac{R_{ESD}}{2 \cdot L_{ESD}}$$

and

$${w_o}^2 = \frac{1}{L_{ESD} \cdot C_{ESD}}$$

Or put differently:

$$I_{HBM}(t) = \frac{V_{HBM} \cdot C_{ESD} \cdot w_o^2}{\sqrt{\lambda^2 - {w_o}^2}} \cdot e^{-\lambda t} \cdot \sinh\left(\sqrt{\lambda^2 - w_o^2} \cdot t\right)$$

Simplifications can be made because $\lambda >> w_o$. Therefore:

$$I_{HBM}(t) = \frac{V_{HBM}}{R_{ESD}} \cdot \left(1 - e^{\frac{-R_{ESD}}{L_{ESD}} \cdot t}\right) \cdot e^{\frac{-t}{R_{ESD} \cdot C_{ESD}}}$$

Transient waveforms are shown in Figure 9. There is no current oscillation because R_{ESD} is large (damped system).

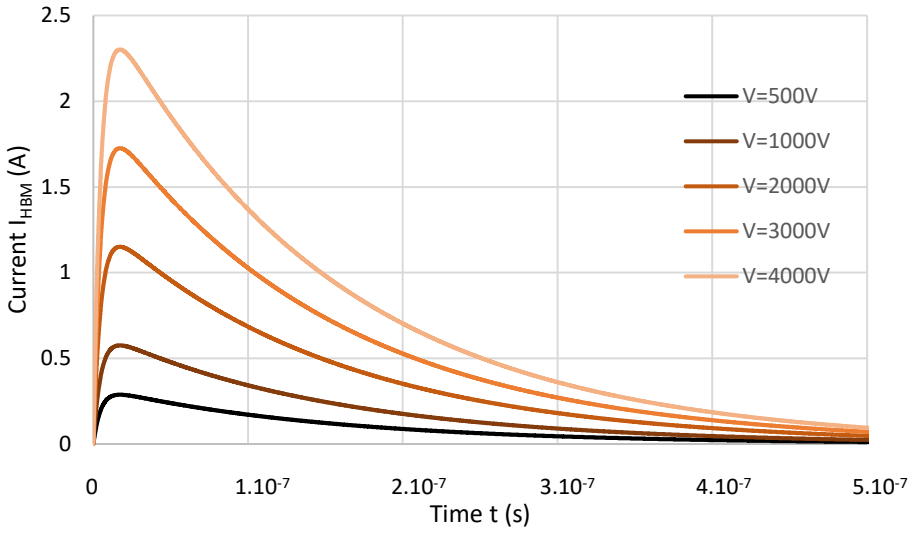


Figure 9: Waveshapes for HBM model for several voltage values of V_{HBM}.

b. Machine Model

For the MM [42], the same conditions are applied (same equivalent circuit as for HBM). This time, since it is a machine touching the IC and not a human, R_{ESD} is way smaller: R_{ESD} = 10 Ω for simulating a metallic contact. The typical V_{MM} is situated between 100 V and 500 V. C_{ESD} = 200 pF and L_{ESD} is considered between 0.5 µH and 2.5 µH.

Since R_{ESD} is small, Δ < 0 and the discharge current can be calculated as:

$$I_{MM}(t) = \frac{V_{MM} \cdot C_{ESD} \cdot w_o^2}{\sqrt{w_o^2 - \lambda^2}} \cdot e^{-\lambda t} \cdot \sin\left(\sqrt{w_o^2 - \lambda^2} \cdot t\right)$$

Simplifications with $w_o >> \lambda$ yield:

$$I_{MM}(t) = V_{MM} \cdot \sqrt{\frac{C_{ESD}}{L_{ESD}}} \cdot e^{-\frac{R_{ESD}}{2 \cdot L_{ESD}} \cdot t} \cdot \sin\left(\frac{1}{\sqrt{L_{ESD} \cdot C_{ESD}}} \cdot t\right)$$

As a result, the waveforms present oscillations. Those oscillations are smoothed after a hundred of nanoseconds (Figure 10). Because of the small R_{ESD}, they are strongly dependent on the tester, therefore the tests are not really reproducible. The MM is no longer considered as a qualification standard: it has been assessed that meeting the HBM and CDM industry standard was sufficient for the component manufacturer [43].

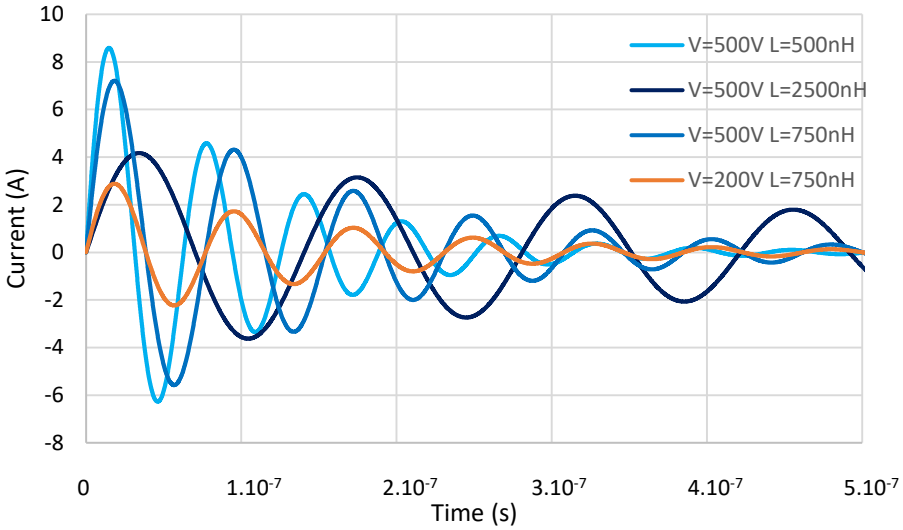


Figure 10: Waveshapes for MM model for several values of V_{MM} and L_{ESD}.

c. Charged Device Model

The CDM [44] is for the case when the charged device is discharging itself to another object. The current flows from everywhere in the circuit toward a single stressed pin which is grounded. It is the most difficult surge to protect the circuit for, because protections have to be placed within the core of the integrated circuit in addition to all the pins. To be CDM qualified, a particular test bench is required, and the DUT is charged thanks to either Direct Charging Method or Field-Induced Charging Method. In this manuscript however, the same method of measurement is used to simulate a CDM discharge and an HBM one (see the section about TLP and VF-TLP measurement).

The equivalent electrical circuit of a CDM is shown in Figure 11. The pre-charge voltage V_{CDM} is taken between 125 V and 1 kV. C_{ESD} corresponds to the capacitance of the circuit to be protected and can vary between 4 pF and 50 pF [45] [46]. L_{ESD} is considered between 2.5 nH and 6.5 nH. R_{ESD} = 10 Ω like for MM. This small R_{ESD} leads to Δ < 0. Therefore:

$$I_{CDM}(t) = \frac{V_{CDM}}{L_{ESD} \cdot w} \cdot e^{-\lambda t} \cdot \sin(wt)$$

with

$$w = \sqrt{w_o^2 - \lambda^2}$$

As a result, the waveforms present oscillations, like for the MM, but they are smoothed after only a few nanoseconds (Figure 12).

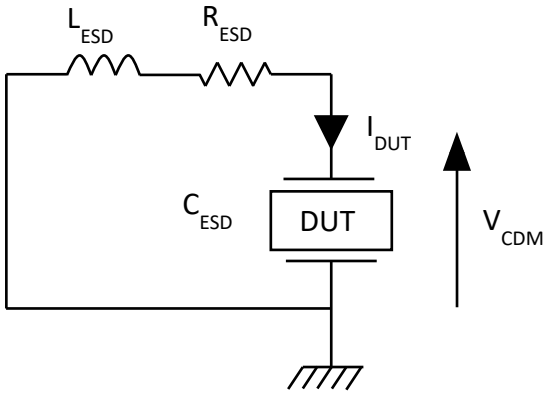


Figure 11: CDM model.

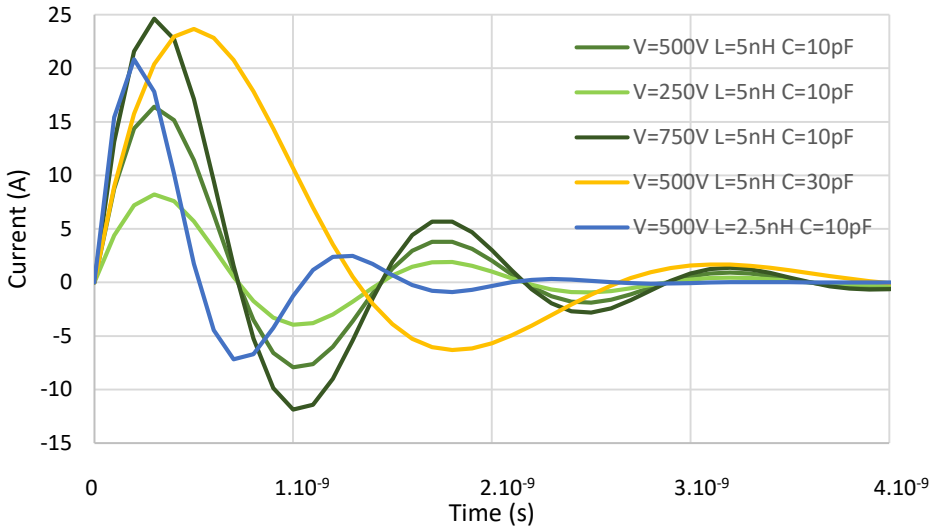
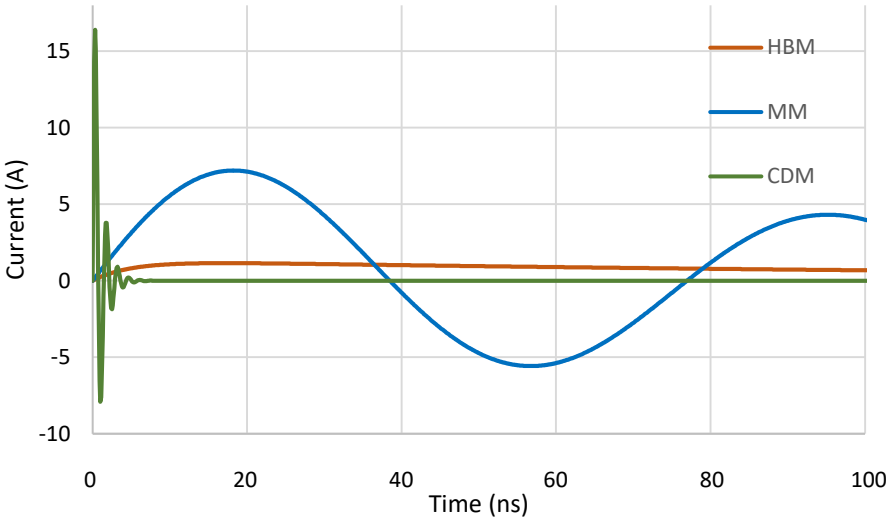


Figure 12: Waveshapes for CDM model for several values of V_{CDM}, L_{ESD} and C_{ESD}.

Figure 13 compares the HBM, MM and CDM discharges. What can be remembered is that a typical HBM discharge is delivering a peak current of 1 A (this corresponds approximately to the 2 kV level of stress), and its duration is about 100 ns long with a rise time of 10 ns. The CDM stress is approximately 1 ns long with 100 ps of rise time. Despite its short duration, the CDM surge can be quite destructive because it involves very high currents (tens of amperes).



	R_{ESD} (Ω)	L_{ESD} (nH)	C_{ESD} (pF)	V (V)
HBM	1500	7500	100	2000
MM	10	750	200	500
CDM	10	5	10	500

Figure 13: ESD waveforms comparison. Table: RLC parameters used for each ESD event model.

3. ESD design window

In order to protect circuits from electrostatic discharges, there are two types of complementary strategies. The first one consists in preventing discharges to happen, thanks to the management of the fabrication environment of the circuit: grounding all the surfaces that can touch the components to be fabricated, using antistatic surfaces and coatings, controlling the humidity in the air and so on. However, this is not sufficient and some discharges still arise. The second strategy aims at deviating the discharges on the circuit so that they do not affect the components of the core, thanks to dedicated protection devices.

The role of the ESD protection is to evacuate a sufficient amount of current while limiting the voltage at the terminals of the protected region, in case of ESD event, so that this destructive current does not pass by the operating part of the IC. The ESD protection should be transparent in the normal IC operating mode; this means that at the operating voltage V_{DD} its leakage current I_{leak} is as low as possible, and the ESD structure is not active. Its parasitic capacitance is low in order to maintain the integrity of rapid signals [47]. The ideal anode current versus anode voltage curve (Figure 14) corresponds to a device that is normally OFF but able to switch abruptly, at a given trigger voltage, in ON mode. This trigger voltage is called V_{T1}. If the curve features a snap-back, the holding voltage, V_H, is the smallest voltage applied on the device while it is in ON state; and when the device is conductive, it has a resistivity of R_{ON}. Without snap-back, the I-V curve is close to a straight line with a slope $1/R_{ON}$ starting at V_{T1}. I_{T2} and V_{T2} correspond to the failure current and voltage, respectively. The ESD protection should activate before the components of the integrated circuit suffer from breakdown, at the voltage V_{BD}. Therefore, there is a design window, establishing that the ESD protection is ON (*i.e.* low impedance) for a certain voltage range $[(V_{DD}+10\%) – (V_{BD}-10\%)]$ only. If V_H is too close from V_{DD}, there is a risk of Latch Up (LU); it means that once the protection device is activated, it stays ON even if the discharge energy is evacuated and the circuit is back to its normal operational mode. This is why a margin of 10% has to be taken. The design window depends on the IC to be protected. The ESD protection should be efficient and robust: it does not break before having sufficiently protected the IC. This means that its failure current I_{T2} is the highest possible. In fact, I_{T2} should exceed the value of the peak current of the HBM discharge corresponding to the required norm (for example 1.2 A for a 2 kV HBM). The triggering should be very fast, because ESDs happen over a short time (1 – 100 ns).

The protection device should not only be compliant with the design window in quasi-static, but also with other measurement techniques. Indeed, some transient overvoltage can appear before the protection reaches its quasi-static behavior. Such an overvoltage can be a cause of failure.

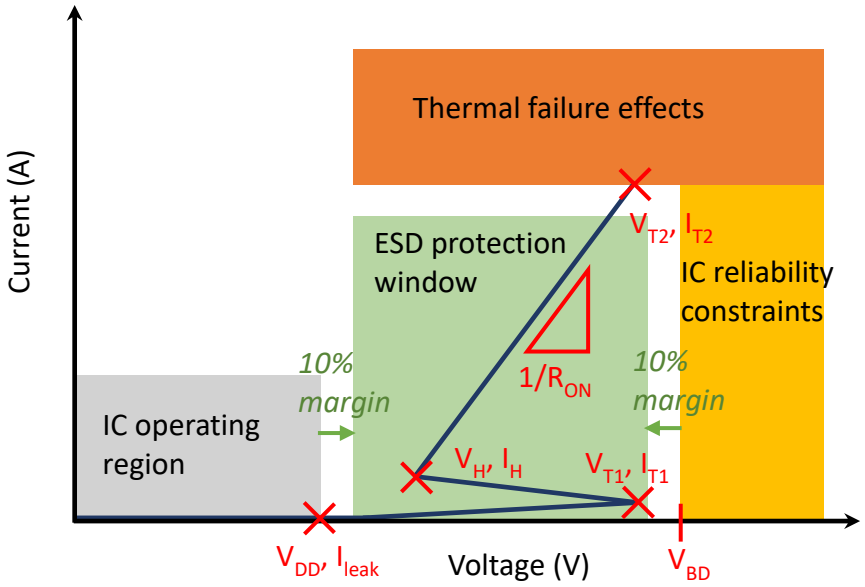


Figure 14: ESD design window.

Nowadays, it is more and more difficult to make ESD protections fit the design window, because the difference between the breakdown voltage and the operating voltage is decreasing [48] [49] [50] [51]. Another difficulty concerns the silicon footprint of the protections. In order to have a high failure current, protection devices need to be designed with a big width, so that the huge current of the discharge can flow through a sufficient section of silicon. In 28 nm FD-SOI technology, the silicon surface dedicated for ESD protections occupies 30% of the I/O ring in average (I/O means Input Output pad, it is the cell that allows the interface between the logic inside the chip and external system components. The I/O ring is a ring of I/Os situated around the chip. Pads are regions on which an external connection is possible). Pressure is put on designers to decrease this surface and this explains also why the voltage requirements for HBM and CDM are regularly updated.

4. Protection strategies

Protection strategies are implemented in order to mutualize the protections, so that the number of ESD protection devices to be used is limited. Indeed, a high number of ESD protections negatively impacts the chip in terms of area, capacitance and leakage current, and does not contribute to the functionality of the circuit.

For addressing HBM and MM surges (that come from outside of the circuit), the primary protection network is used. Protections are placed at every entry of the circuit (I/O pads, power supplies like V_{DD}, and the ground pad Gnd). ESD events have a polarity: the surge can be positive or negative (the sign of the current is changing). By symmetry, only the positive surges need to be calculated. However, when implementing the protections, the ESD qualification should be done for both polarities of ESDs. The protections can be unidirectional (for addressing only positive surges) or bidirectional (for positive and negative surges). Bidirectional protections usually have a larger area footprint than unidirectional ones. They must protect the core of the circuit from positive or negative ESDs for all combinations of pads (a discharge from Gnd toward V_{DD}, V_{DD} toward Gnd, Gnd toward I/O_1, I/O_1 toward Gnd, V_{DD} toward I/O_1, I/O_1 toward V_{DD}, I/O_1 toward I/O_2, I/O_2 toward I/O_1, etc.). Therefore, when dealing with ESDs, all the pads of the circuit are considered floating. During an ESD event, if several pads are in contact with the charged human or machine, multiple paths of conduction for the discharge are taken in parallel. A trigger circuit can be used to help the protections to trigger. Two categories of protection strategies can be described: the local and the global (with the remote and the distributed strategies).

For CDM, it is complex to forecast the ESD path in the circuit since the discharge comes from the circuit itself. The typical protection elements for addressing CDM discharges are: resistors (they help to increase the total resistance of an uncontrolled ESD current path), diodes and Grounded-Gate NMOS (GGNMOS) devices. The protective elements have to be small and localized at strategical places in the IC. This is the secondary network protection. It is designed to efficiently turn ON the protections during a fast – CDM - event. This second network is complementary to the first network. The localization of the second network protections depends on the circuit, the design, the substrate, the metal connections and the package.

In this section, the mentioned ESD protection strategies are depicting ESD protection networks inside one power domain. When dealing with multi-power circuits, additional protection blocks are necessary, but it is not the scope of this manuscript.

a. Local protection strategy

The local strategy consists in having bidirectional ESD protections between each pad (I/O pad or power) and the ground pad (Figure 15). Those protections are typically designed in each pad, which is why the pad area can be very consequent, all the more so as bidirectional protections take a lot of silicon area. The protection device situated between the power net (V_{DD}) and the ground (Gnd pad) is called a power clamp and is bidirectional. The worse possible path for the discharge is if it happens between a V_{DD} pad and an I/O pad, because two protections in series are involved instead of one. The advantage of this strategy is that it does not require complex placement rules.

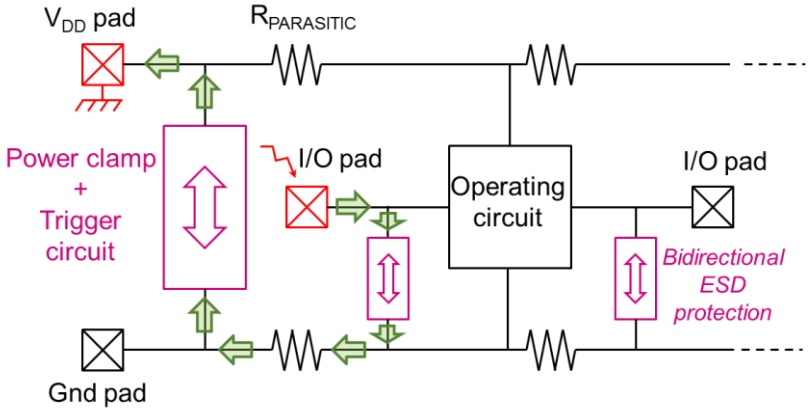


Figure 15: Local ESD protection strategy. One bidirectional ESD protection is placed in each pad. If a discharge arises from one I/O pad to V_{DD} pad for example, its current will follow the path drawn with the green arrows. The pink arrows depicted in the ESD block are here to show the direction of the current when there is a discharge. (In the local strategy the protection devices are all bidirectional, this is why the current can flow in both directions.) When the circuit is in normal operating condition the protection devices are OFF.

b. Remote protection strategy

The remote ESD protection strategy (described in Figure 16) allows to save some silicon area with respect to the local strategy [52]. It consists in placing unidirectional protections at each I/O pads. The power clamps (protections placed between V_{DD} and Gnd nodes) and their trigger circuit (of great dimension) are placed along the I/O ring such as to limit the parasitic resistance between two clamps. When an ESD event arises between two pads, the maximum voltage during the ESD event has to be smaller than the smallest breaking voltage that exists between those two pads. For example, the protections are here to prevent a critical voltage V_{CRIT} to be reached across the structures of the operational circuit between the I/O pad and the ground. Considering that only one path is turned ON, the discharge flows from the I/O pad toward Gnd pad with the path depicted in Figure 16, and the following inequality has to be respected:

$$V_{PROTECTION} + V_{CLAMP} + R_{PARASITIC} \cdot I_{ESD} < V_{CRIT}$$

where $V_{PROTECTION}$ and V_{CLAMP} are the trigger voltage of the unidirectional protection and the ESD clamp respectively; $R_{PARASITIC}$ is the resistance value of all the metal rails involved in the path, and I_{ESD} is the ESD discharge current peak value. This example shows how important it is to make sure that the distance between two clamps is not too long.

The power clamps are typically made with huge NMOS transistors, their trigger circuit being a RC circuit that is plugged to their gate; a reverse diode is plugged in parallel between V_{DD} and Gnd to insure the bidirectionality of the protection. The unidirectional ESD protections are usually made with diodes.

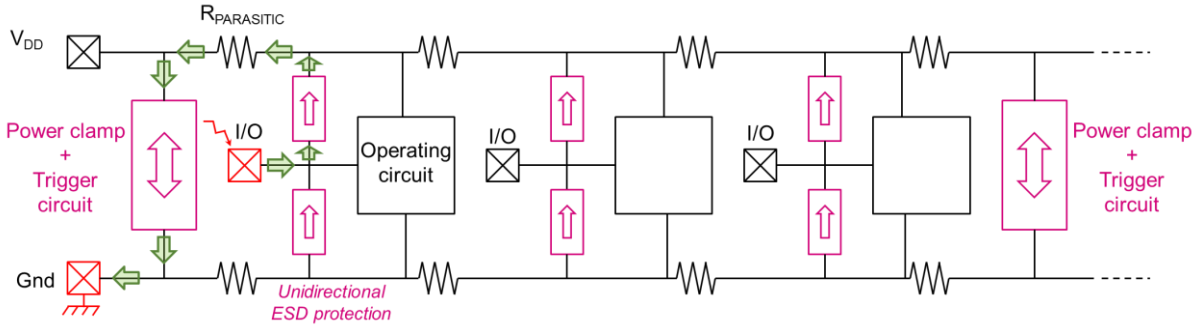


Figure 16: Remote ESD protection strategy. One unidirectional ESD protection is placed in each I/O pad, and bidirectional protections are regularly placed between the V_{DD} and Gnd rails. If a discharge arises from the closest I/O pad to the power clamp toward Gnd pad for example, its current will follow the path drawn with the green arrows.

c. Distributed protection strategy

In the distributed ESD protection strategy, the trigger circuit is decentralized from the central clamp, and small bidirectional clamps are distributed in every I/O cell (Figure 17) [53]. A rail Trigger is connecting all the gates of the distributed NMOS protections. The remote Trigger Circuits (TC) are connected to the rails Boost, Trigger and Gnd; they have to be placed regularly along the I/O ring.

Figure 18 is depicting what happens if there is an ESD surge from one I/O pad to the second. At first the ESD event has to be detected. A fraction of the current is flowing from the ESD entry pad to the rail Boost (this current is negligible compared to the current flowing through the rail V_{DD}). This will activate the trigger circuit, which will be maintained ON while the ESD event is occurring. The activated trigger circuit opens all the bidirectional protection that are distributed all over the chip thanks to the rail Trigger. As a result, the main ESD current can see several different paths to flow until the second I/O pad through the rail V_{DD}. It is because of this multitude of paths that those bidirectional clamps are allowed to be designed with small dimensions.

This strategy enables to relax the distance constraint due to the parasitic resistances of the rails, with respect to the remote strategy. Indeed, the evacuation of current is generalized into all the I/O clamp protections, and a specific path is avoided.

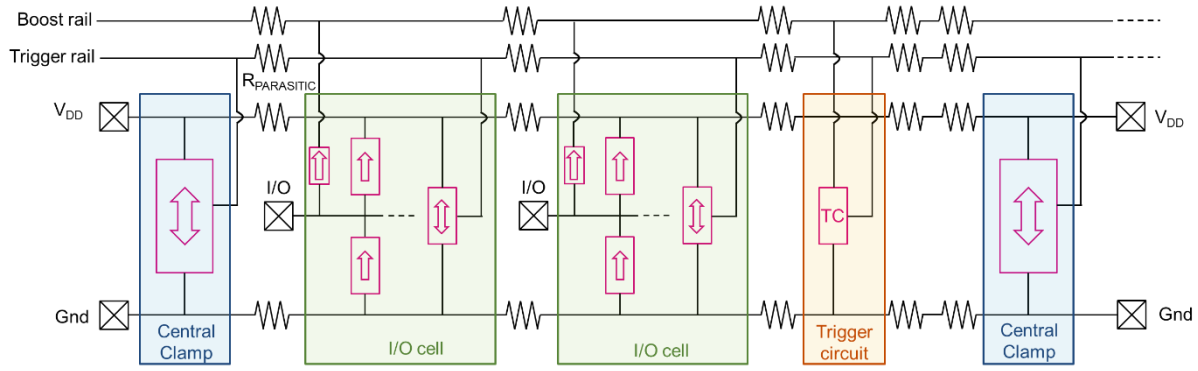


Figure 17: Distributed ESD protection strategy. The operating part of the circuit is not represented.

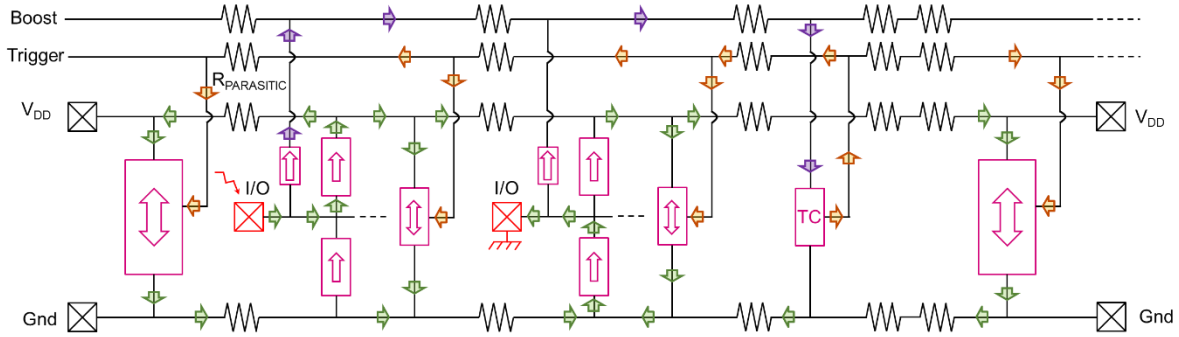


Figure 18: The different conduction currents of the distributed ESD protection strategy are represented in case of an ESD surge from one I/O pad to the second. The purple arrows are for the "boost" signal propagation. The orange arrows are for the "trigger" signal propagation, and the green arrows are for the ESD current main flow through the protection devices. Adapted from [21].

III. Context of study and tooling

1. Measurements

a. TLP measurements

Classical DC (Direct Current) information provide some insights about the behavior of the ESD device. But in order to realize a characterization that is closer to the HBM model, a better characterization technique is the Transmission Line Pulse (TLP) [54] [55] [56]. The advantage of this method is to be able to obtain I-V behaviors with high currents while limiting self-heating effects that can affect the device.

Several variants of TLP characterization test benches are available, but the principle remains the same. It operates as following: a square waveform pulse (of 100 ns of duration for example) is provided to the DUT by a previously charged coaxial transmission line (Figure 19). The resulting square current and voltage pulses across the device are measured. An average is made on the value of the current and the voltage waveforms (typically 30% of the pulse duration is selected). These values represent one point in the I-V curve of the device (Figure 20). After this test, the leakage current in the device at a given voltage (for example the leakage current at $V=V_{DD}$) is measured thanks to a non-destructive DC measurement, in order to verify that the device is not degraded. If no degradation is observed then the amplitude of the incoming pulse is increased in order to obtain the next I-V data point. The whole I-V curve is plotted thanks to this method. TLP measurements are stopped when the DUT is considered damaged. A stopping criterion can correspond for example to a maximum DC leakage current of 1 µA. The last I-V point that was measured before breakdown is considered as being the failure current I_{T2} and voltage V_{T2} of the device. Note that with this technique, all I-V points are thermodynamically uncorrelated.

The duration of the pulse is a function of the length of the coaxial cable and the propagation speed of the waves inside the cable. The pulse is sufficiently narrow to prevent self-heating in the DUT, but sufficiently wide so that its capacitances and inductances are stabilized. Few transmission lines are actually available in order to choose the duration of the pulse for each TLP setup. Few filters are also available for selecting the rise time of the pulse. A typical TLP pulse width is often chosen as 100 ns with a rising and falling time of 10 ns, in order to provide an energy comparable to the one of an HBM discharge. Indeed, a good correlation has been reported between the TLP failure current of devices and the HBM failure current, provided that the pulse duration was about 100 ns, since this duration leads to a similar dissipated energy in the device [57].

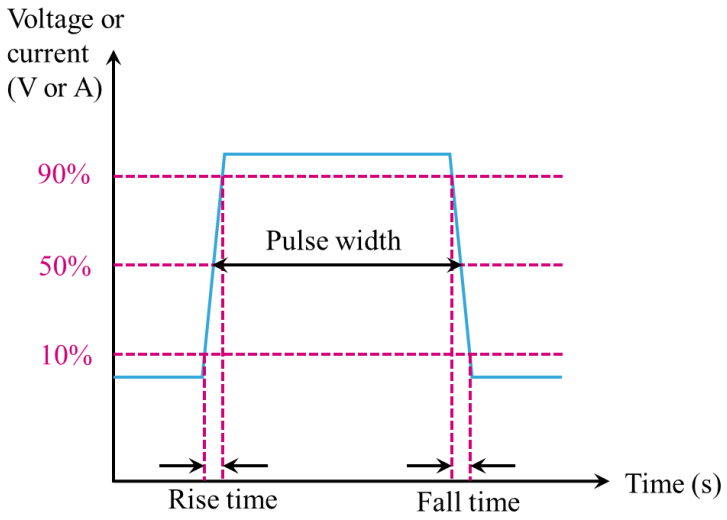


Figure 19: Square pulse for TLP characterization. Adapted from [21].

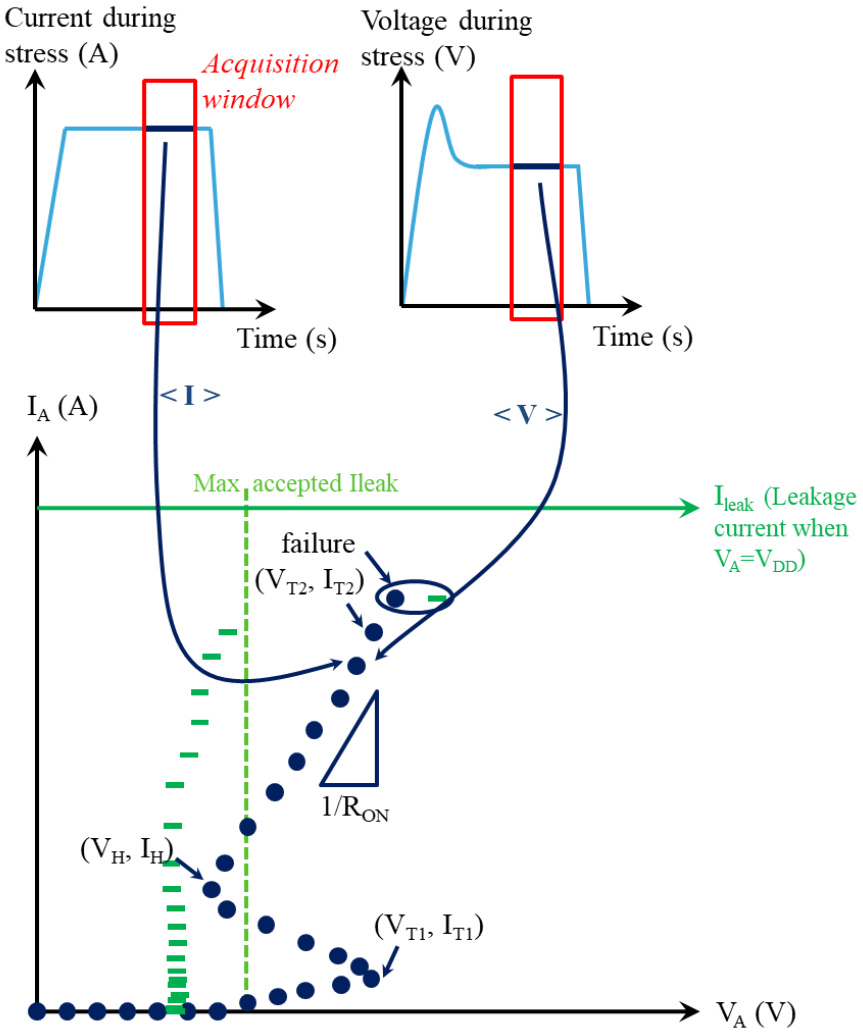


Figure 20: Top: Principle of TLP measurement. Square pulses are sent, then the voltage and current are measured to draw the I-V characteristic. Bottom: Typical I-V curve of an ESD protection device, with its leakage (green scale).

The average that is made on the value of the current and the voltage waveforms (in order to get an I-V point of the TLP curve) is obtained from an acquisition time window that is situated after the transient phenomena in the waveforms. An example of well-known transient phenomenon is the overshoot. It is the voltage peak that can be seen on the voltage waveform measured across the device (Figure 20). The choice of this acquisition time window has a significant impact on the resulting TLP I-V curve [58].

All our TLP measurements were performed at room temperature on several dies with a TLP test bench that uses Time Domain Reception (TDR) and 4-point Kelvin probes (the real impedance of the device is measured instead of the whole setup impedance). 10 ns rise time and 100 ns pulse duration were selected. Each point of the TLP curve corresponds to the average of the voltage and current measured in the time interval between 70 ns and 90 ns. After each I-V point is measured, a non-destructive DC measurement is performed to get the leakage current in the device at 1 V. TLP measurements are stopped when this leakage increases by more than 500% with respect to the first leakage current that was measured before the DUT had undergone pulses.

b. VF-TLP measurements

The Very Fast TLP (VF-TLP) method is similar to the TLP characterization technique, except that the pulses have a much shorter width, that is comparable to the CDM duration. A typical VF-TLP pulse duration is 1 ns with a rising and falling time of 250 ps. The aim of this method is to get information about the behavior of ESD protection structures and their ability to protect the circuit against a fast discharge such as the CDM. However, no correlation is possible between CDM and VF-TLP [59]. Indeed, when a CDM event occurs, the circuit itself represents the carriers' reservoir that discharges into one pad. Therefore, it leads to build-up voltage drops across devices that can reach their breakdown voltage. In the VF-TLP characterization, the waveform is applied between the two pads of the ESD device, which is limiting the voltage at its terminals.

All our VF-TLP measurements were performed at room temperature on several dies. The pulse duration was 1 ns with a rise time of 100 ps (except in the chapter 3, where the rise time was modified to observe its impact on the curves). After each I-V point is measured, a non-destructive DC measurement is performed to evaluate the leakage current in the device at 1 V.

c. DC measurements

DC (quasi-static) measurements that are presented in this manuscript consist in raising progressively the voltage on the device while monitoring the resulting current. The measurement time is sufficiently long to be considered infinite (all transient phenomena in the devices have time to fade). It provides information about the leakage current that is expected to flow from the device for each supposed nominal power voltage V_{DD}. All our DC measurements were performed at room temperature on several dies.

2. TCAD as a predictive tool of investigation

TCAD (Technology Computer Aided Design) Synopsys Sentaurus™ tool [60] is ideal for simulating a single device, in order to understand the phenomena involved in the device electrical behavior, or to compare a set of devices without having to wait for measurement results (the fabrication of devices can be very time consuming). First the structure has to be generated: the different regions along with their materials are defined, the doping layers are placed as well as the electrodes (where electrical and thermal conditions will be applied). Then the device is meshed: it is discretized onto a non-uniform grid of nodes. Currents, voltages and other physical parameters can then be calculated at each node of the meshing. A trade-off has to be found for the number of nodes, between convergence and accuracy.

Our devices were meshed in 3D in a process compliant way. The simulations have been beforehand calibrated thanks to a standard NMOS structure (like the one in Figure 4). Its quasi-static I_D-V_G curve has been adjusted with the doping levels, the gate work function and other parameters, in order to obtain the same V_{TH} as in the measurements. The gate stack was not simulated for reducing computation time; only a SiO_2 layer was created, and a gate work function was selected (such as obtaining a similar V_{TH} in the standard NMOS transistor). The metal interconnections were not simulated. All the simulations were done with 3D TCAD.

The aim of our simulations is not to get the exact ESD parameters such as V_{T1} and V_H, but to be able to compare the different structures together, to understand better the phenomena provoking the different behaviors, and to have a trend.

a. Setup of the TCAD simulations

For a more extensive description of the TCAD setup, please read the Appendix 1.

A set of physical device equations that describes the carriers' distribution and conduction mechanisms have to be specified for the computation. The system of coupled equations to be solved in our TCAD simulations at each node is the following:

- **Poisson equation**

The Poisson equation is used to compute the electrostatic potential for a given charge distribution.

- **Continuity equations**

The charge conservation principle allows to obtain an equation describing the time and space evolution of the charge concentrations. The time variation of the number of electrons and holes depends on the currents due to the spatial movement of carriers, and on their generation and recombination rates.

- **Transport equations**

The total current of carriers is the addition of a conduction current due to an electric field and a diffusion current due to a gradient of carriers (drift-diffusion model).

- **Contact equations**

All contacts (electrodes) on semiconductors are ohmic, subject to charge neutrality and equilibrium. When all contacts of a device are biased to the same voltage, the device is in equilibrium, and the electrons' and holes' densities are described by a constant Fermi potential. In our simulations, the gate work function (used in the contact equations) has been chosen thanks to the calibration of the TCAD curves with a reference NMOS device.

- **Circuit equations**

The connectivity of the circuit is considered. Some external resistances, current and voltage sources, or other additional external devices can be added (lumped elements).

- **Heat equations**

Most of the simulations done in this work do not take into account the temperature effects, unless they are explicitly labeled as electro-thermal simulations. For electro-thermal simulations, the self-heating of the device is taken into account by adding the lattice temperature equations to the previous set of equations. A thermal electrode (thermode) is defined in order to apply a temperature boundary condition. In our electro-thermal simulations, room temperature is applied on an electrode situated under the bulk region situated below the BOX of the device. When electro-thermal ACS simulations were performed, the simulations were starting with a device temperature of 300 K and were stopped with the hot spot of the device reaching 800 K. We did not perform simulations to investigate the effect of an external temperature on our devices (in harsh environment for example).

Physical models are selected to be included in the numerical resolution of the previous equations. In our simulations, the default model parameters were left unchanged. The following models were used:

- **Boltzmann statistics**

Electrons' and holes' densities were computed from the electrons' and holes' quasi-Fermi potentials using the Maxwell-Boltzmann instead of the Fermi-Dirac statistics.

- **Semiconductor band structure**

A silicon band gap narrowing model was included to determine the intrinsic carrier concentration. Band gap narrowing occurs when the semi-conductor is degenerate (with high doping or with high current injections). By default, band gap narrowing is active in all TCAD simulations. Different band gap models are available, and the Slotboom model has been chosen [60].

- **Generation-recombination models**

In our TCAD simulations, the SRH model (Shockley–Read–Hall recombination with doping-dependent lifetime) was used along with Auger recombination model, avalanche van Overstraeten-de Man model (electron–hole pair generation by impact ionization) and non-local path band-to-band tunneling model.

- **Mobility models**

Free carriers in the material gain momentum from electric fields and lose momentum through scattering with perturbations to the spatial periodicity of the lattice potential. Phonon scattering, doping-dependent mobility degradation, mobility degradation at interfaces (with transverse field dependence) and mobility degradation due to high field saturation were considered in our TCAD simulations. The different mobility contributions are combined by Matthiessen's rule.

b. Average Current Slope and Average Voltage Slope

For the simulation of the device, TLP method would be lengthy to reproduce, and the simulation would take a too long time before having results. Therefore, another technique is used to simulate the I-V curve: the Average Current Slope (ACS) technique (Figure 21). The device is subject to a current ramp. The rise time of the ACS is 100 ns, in order to mimic a TLP test for HBM. The voltage is initialized at 0 V and then let free to change. As a result, an I-V response of the device is obtained. It is very dependent on the slope of the current, which is adjusted thanks to the rise time of the ACS and the maximum amplitude of the current. For having the best possible equivalence to the HBM TLP measurements, the amplitude of the current ramp is chosen knowing that the failure current in all our thin-film devices is situated around 0.1 A. Therefore, in our simulations the devices are subject to a current ramp that goes from 0 to 0.1 A.

The main advantage of ACS is that only one simulation is needed for obtaining an I-V curve, while TLP requires as many simulations as the number of points that are present in an I-V curve. It has been shown [61] that the ACS simulation with 100 ns duration could be a good first approximation before the complete TLP measurements. Because of the short duration of the ACS event, the carriers' flow and the trigger mechanisms are conserved. As a consequence, the ACS is a valid fast method to address a full characterization in one run.

Another advantage of the ACS is that the current is progressively raised and all the points of the I-V curve are correlated, contrary to the TLP I-V curve. Therefore, some effects - such as the overvoltage in the TLP voltage waveform - are taken into account. The overvoltage is not part of the TLP I-V curve since the data points come from an average done after the stabilization of the square waveforms, which leads sometimes to an I-V TLP curve that perfectly fits to the ESD design window but a device that is still unable to stand the intended ESD stress.

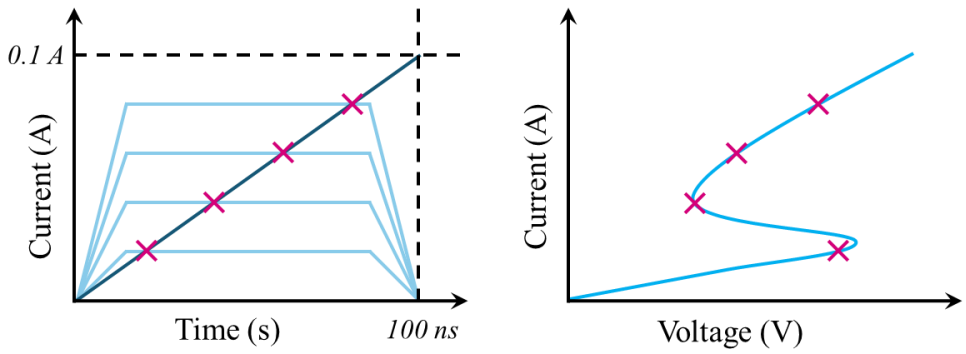


Figure 21: Principle of the ACS method: equivalence of TLP and of current ramp (left), and obtained I-V characteristics of the device (right). Adapted from [62].

For a rise time of the duration of a TLP HBM pulse, the ACS assessment is particularly designed for the study of the triggering of the device. However, as for any simulation, the deductions drawn from an ACS analysis must be carefully evaluated and used only as an indication of the general tendencies, which have to be verified with measurements.

In our ACS simulations, all the structures have 10 µm width. For simplification purpose, only one finger was simulated.

The Average Voltage Slope (AVS) simulation consists in simulating the device with a voltage ramp. The rise time of the AVS is chosen as 1 ms in order to simulate a DC voltage sweep measurement. The voltage ramp is chosen to start at 0 V and stop at 5 V. The resulting current flowing through the device versus the voltage builds a DC I-V curve. The stopping voltage is not important because even if it changes the slope of the ramp, the time scale is sufficiently long so that only the quasi-static behavior of the device is reached anyway. The aim of the AVS is to determine the leakage current that a device would have for a given V_{DD} that can be chosen among all the values of the voltage ramp. Note that the eventual snap-back of the device cannot be seen, instead a very sharp current rise is observed. This is because the device voltage is forced by the simulation to always increase.

IV. ESD Protection devices

Hybrid technology allows removing the BOX layer in order to get bulk devices along with thin-film devices on a FD-SOI wafer [63] [64]. ESD devices are typically designed in the bulk, because their performance is significantly reduced in the thin-film (due to its extremely thin thickness) [65] [66]. The main bulk ESD devices that are currently used in the 28 nm FD-SOI UTBB technology node are diodes, MOSSWIs (MOS Switches), GGNMOS, BIMOS (Bipolar-MOS effect device) and SCR. Those devices can be adapted to be designed in the thin-film. Also, other devices can be used as a protection, as long as they fill the requirements of the ESD window. In this section, “classical” and more innovative devices will be presented.

1. Diode

A diode is basically a P/N junction. The difference of energy band level between the N-doped and the P-doped regions constitutes a potential barrier that prevents the carriers to flow. When the diode is forward biased, the barrier potential accross the junction reduces. When the turn-on voltage is reached, current flows across the diode from the P-doped to the N-doped region. When the (anode) voltage is too high, the diode undergoes self-heating with the strong injection regime, and the mobility of carriers is degraded, thus leading to a saturation in the I-V characteristic. The saturation in the diode I-V curve can also come from series resistances. High currents can be achieved before the breakdown of the diode. When the diode is reverse-biased, the external voltage enhances the barrier potential. The diode is blocked until the avalanche voltage of the junction is reached, with a sufficiently high electric field. When reverse-biased, the diode cannot achieve high currents before breakdown.

For an ESD application, the diode is reverse-biased with respect to the core circuit, and forward-biased for the ESD surge; it is a unidirectional component. By placing several diodes in series, the threshold voltage of the protection can be modified, but in the meantime, it increases the leakage current, the silicon area and parasitic effects (for example a parasitic SCR can be built if dimensions and doping levels are not well selected).

The gated diode [67] is a type of diode where a metal gate with high-k dielectrics is placed between the N^+ and P^+ regions. Thin-film gated diode model [68] also exist (for addressing a CDM issue for example). In the 28 nm FD-SOI technology, the gated diode doping layers are P^+/Pint/N^+ (Figure 22). The channel is Pint doped (it means that it is an intrinsic semiconductor), which minimizes the leakage of the diode.

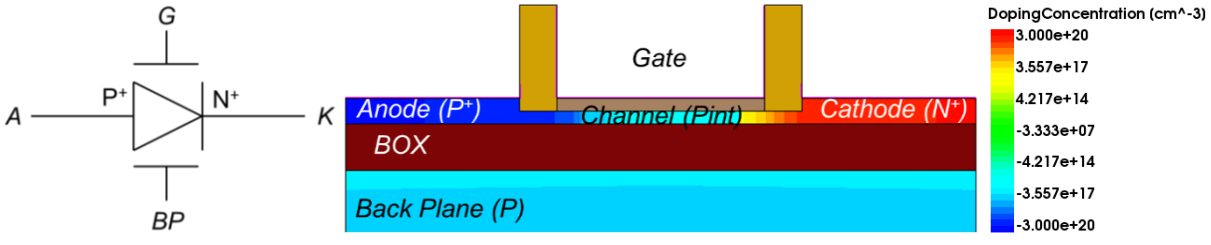


Figure 22: Schematic of a diode (left), and TCAD view of a thin-film gated diode.

2. Protection devices built from NMOS devices

The NMOS transistor is a bidirectional component, therefore it can be used as a central clamp. The drain of the NMOS is connected to the contact to be protected.

a. MOS switch

A possibility to use it for an ESD application is to ground the gate when there is no ESD, and to bias the gate so that the transistor is ON when an ESD event occurs. It is then called MOSSWI for MOS switch. A circuit is needed (for example an RC high-pass filter, like in Figure 23) to detect ESDs and pilot the gate voltage. Care is taken about the time constant of this local trigger circuit. The bidirectional nature of the MOSSWI depends on the architecture of its trigger circuit. A reverse biased diode can be added in parallel to ensure the bidirectionality of the system. The conduction in a bulk MOSSWI mostly occurs near the surface, therefore it is possible to design it in the thin-film of the FD-SOI wafer without reducing too much its robustness (its failure current I_{T2}). The surface conduction of the bulk MOSSWI is also the reason of its great dimensions. Indeed, it has to be designed with a large width for being able to stand high currents. The bulk MOSSWI is easily controlled and does not present latch up risks, that is why it is the most commonly used protection device (along with a diode) as a central clamp in digital CMOS circuits. The thin-film MOSSWI has less leakage than in the bulk, and back-plane biasing (as a back gate) can be used in the triggering mechanism.

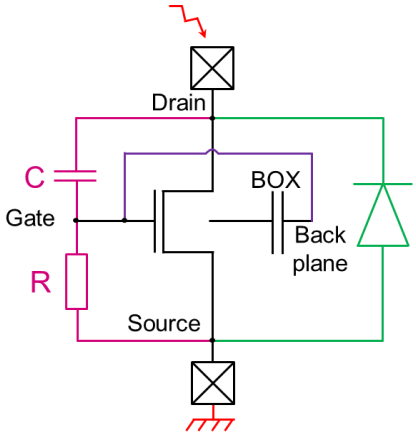


Figure 23: Schematic of a thin-film MOSSWI. A RC high-pass filter - the trigger circuit - (in pink) is used to control the gate: when the frequency of the signal is low (for a normal use), the gate is grounded through the resistor, and the MOS conduction is blocked. When there is a fast signal (for example an ESD), the gate node voltage increases through capacitive coupling. Thanks to this voltage on the gate, the MOSFET is ON and current can flow between the drain and the source. The back-plane is used as a back gate (it is connected to the front gate in violet), in order to reduce the threshold voltage. A reverse diode is implemented (in green) for addressing negative ESDs.

b. Grounded Gate NMOS

Another possibility is to use the NMOS as a Grounded Gate NMOS (GGNMOS). The gate and the source are plugged together (Figure 24), so the MOS effect is blocked, and it is the NPN parasitic Lateral Bipolar Junction Transistor (LBJT) of the structure that matters [69]. The source of the transistor corresponds to the emitter, the substrate corresponds to the base, and the drain coincides with the collector. When a positive ESD arises, hot carriers are generated in the drain/channel junction by impact ionization. The hole current goes toward the body, thus increasing the body-source potential V_{bs} (through the parasitic resistor R_{WELL} of the substrate in bulk GGNMOS). Due to this potential, the source injects electrons toward the drain, thus fully activating the parasitic bipolar transistor (with a snap-back on its I_D-V_D characteristics). When a negative ESD arises, the gate is biased (since it is connected to the source, which undergoes the ESD surge), thus turning ON the GGNMOS. Bulk GGNMOS benefits from volumic conduction, this is why thin-film GGNMOS (surface conduction) have a lower robustness than bulk ones. However, they trigger at a lower voltage [66]; as a consequence, thin-film GGNMOS are mostly used as CDM protections. The grounded-gate PMOS is not used because holes have a low mobility, therefore its performances as an ESD protection are reduced with respect to the grounded-gate NMOS. Silicide is typically removed from drains and sources in order to prevent multi-triggering (zig-zag I-V curve due to "sequential" triggering of fingers in a multi-finger device) due to the multiple fingers [70].

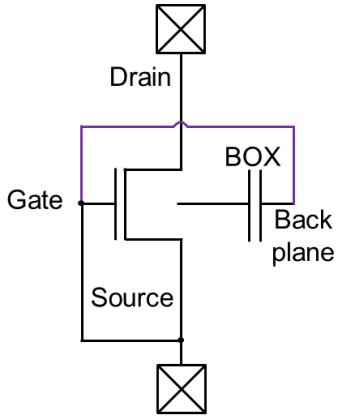


Figure 24: Schematic of a thin-film GGNMOS: gates are grounded.

c. Bipolar MOS

A BIMOS (Bipolar MOS effect) is an N-type MOS transistor on which a P^+ body contact is placed in order to have access to the base of the NPN parasitic LBJT (which is the channel of the MOSFET). The BIMOS device can be designed in bulk [71] [72] [73] [74] [75] or in SOI [76] [77] [78]. For ESD applications [74] this contact and the gate are both plugged to the same external polysilicon resistor (Figure 25).

The BIMOS triggers both dynamically and in static mode. When an ESD arises, there is impact ionization at the junction between the channel and the drain which turns on the bipolar transistor formed by the drain (collector), the source (emitter) and the channel (base), like in the GGNMOS. Then, since the base is also plugged to the gate of the transistor, the MOSFET part is activated and an inversion layer is created. The current is conducted through MOSFET subthreshold operation and superimposed to the bipolar current. The higher the value of the polysilicon resistor, the smaller the trigger voltage of the BIMOS. Indeed, a higher value of resistor is transformed into a higher voltage at its terminal, and since the resistor is also connected to the gate of the BIMOS, the voltage on the gate is higher. The BIMOS device has better performances than the GGNMOS one, since the external resistor increases V_{bs} further. V_{bs} is the base-emitter voltage of the parasitic NPN bipolar transistor, and it is also the body-source voltage of the MOSFET. Hence, increasing V_{bs} helps the parasitic LBJT to trigger earlier, but also it reduces the threshold voltage of the MOSFET thanks to body effect. The BIMOS can be triggered dynamically also thanks to its parasitic capacitances between the drain and the gate, and the drain and the bulk. Provided that the ESD surge has a short rise time, when the drain voltage raises, it will increase the gate and bulk voltages through parasitic capacitances, thus helping the BIMOS system to activate. The BIMOS can also address negative ESD surges. The gate and body terminals are then biased through the external resistor, and this turns ON the BIMOS.

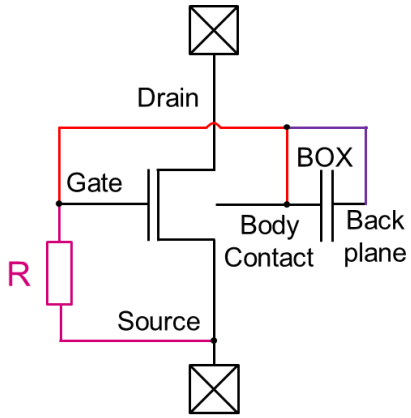


Figure 25: Schematic of a thin-film BIMOS structure.

3. Protection devices built from the SCR

a. Silicon Controlled Rectifier

A Silicon Controlled Rectifier (SCR) [79] [62], also called thyristor, is a P/N/P/N doped structure. It can be seen as two merged bipolar transistors NPN and PNP that are looped (Figure 26). The SCR is a unidirectional component. A bidirectional component - the triac - is obtained by connecting two SCRs in a head to tail way (Figure 27).

The I_A-V_A curve of the SCR is similar to the one of the GGNMOS (with a snap-back). The strength of the SCR is its low dynamic resistance R_{ON} [80]. Indeed, when it triggers, a very high current can flow in it. The working principle of the SCR is the loop of bipolar transistors that allows them to amplify the current. When the bipolar transistor gains are verifying the condition:

$$\beta_{NPN} \cdot \beta_{PNP} \geq 1$$

then the collector current of the PNP transistor increases, which rises the base current of the NPN transistor, thus the collector current in the NPN transistor enlarges, leading to the augmentation of the base current of the PNP transistor, which closes the loop with the raise of the collector current of the PNP transistor. The gain of the transistors depends on the doping levels and on the dimensions of the structure.

If the base of the two bipolar transistors are floating, the difference of voltage between the SCR anode and cathode should be higher than the avalanche voltage of the P-N junction in order to activate the SCR. The trigger voltage of the SCR can be lowered if a high potential is applied to the basis of the NPN transistor (which is the P-doped trigger of the SCR: G_P), or if a low potential is applied to the base of the PNP transistor (N-doped trigger G_N). This is the goal of the trigger circuit [81] [82] [83].

Since the current conduction is volumic in the SCR, this device is used in the bulk. However, some efforts are made to design it on top of the SOI wafer [84] [85]. The Z^2-FET (Zero subthreshold swing and Zero impact ionization FET) [86] [87] [88] and GDNMOS (Gated Diode NMOS) devices result from SCR integration trial in the thin-film SOI, however those devices are very different from the SCR. Other very innovative devices are emerging, like the Z^3-FET (Zero gate, Zero swing slope and Zero impact ionization FET) [89] for example.

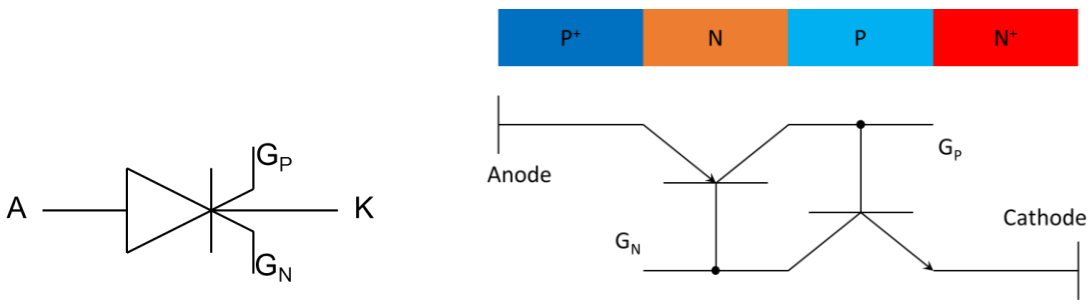


Figure 26: Schematics of a SCR structure. A stands for anode, K for cathode; G_N and G_P are the N and P-doped trigger of the SCR, respectively.

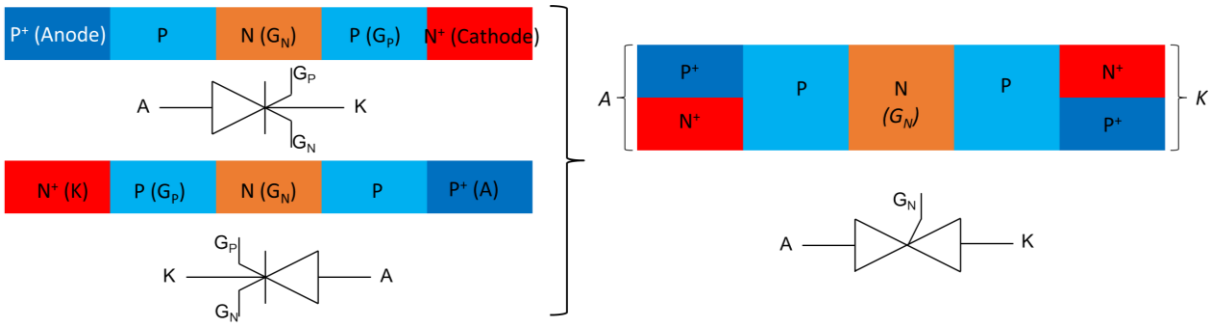


Figure 27: Schematics of a triac (left): it is comprised of two SCRs (right).

b. Zero subthreshold swing and Zero impact ionization FET

The Z^2-FET (Zero subthreshold swing and Zero impact ionization FET) structure is depicted in Figure 28. It is controlled by biasing the front and the back gate, the aim being to modulate injection barriers for the carriers [90] [91] [92]. Super-coupling is not a problem in the Z^2-FET structure because the population of electrons and holes are separated laterally [93]. It can also be seen as a partially gated diode. The Z^2-FET device is only mentioned in this section for information purpose, it is not part of the scope of this thesis.

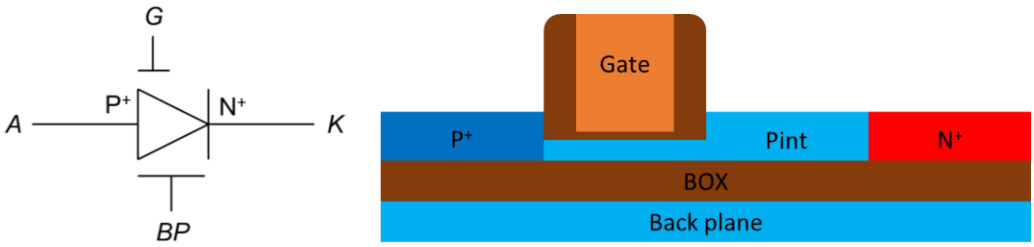


Figure 28: Schematic (left) and cross-section (right) of a Z^2-FET structure.

c. Gated Diode NMOS

The GDNMOS [94] is a gated diode that shares its cathode with the drain of a NMOS transistor. Theoretically the device could be considered also as a SCR since it is a P/N/P/N doped structure (Figure 29). This device is explained more extensively in chapter 2.

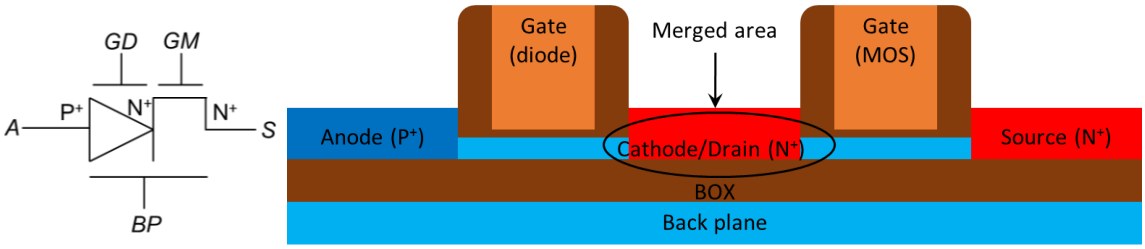


Figure 29: Schematic (left) and cross-section (right) of a GDNMOS structure.

d. Beta-Matrix architecture

The beta-matrix [95] [96] [97] is an architecture of ESD protection that behaves like a network of triacs. It can be considered as a protection device as much as a global protection strategy. In Figure 30, it can be observed that each triac will provide a preferred path for the ESD between the VDD GND and I/O pad for each I/O of the circuit. This beta matrix can be used with a floating trigger G_N for high voltage application. It has to be connected to a trigger circuit (to control G_N) in order to fit other ESD design windows.

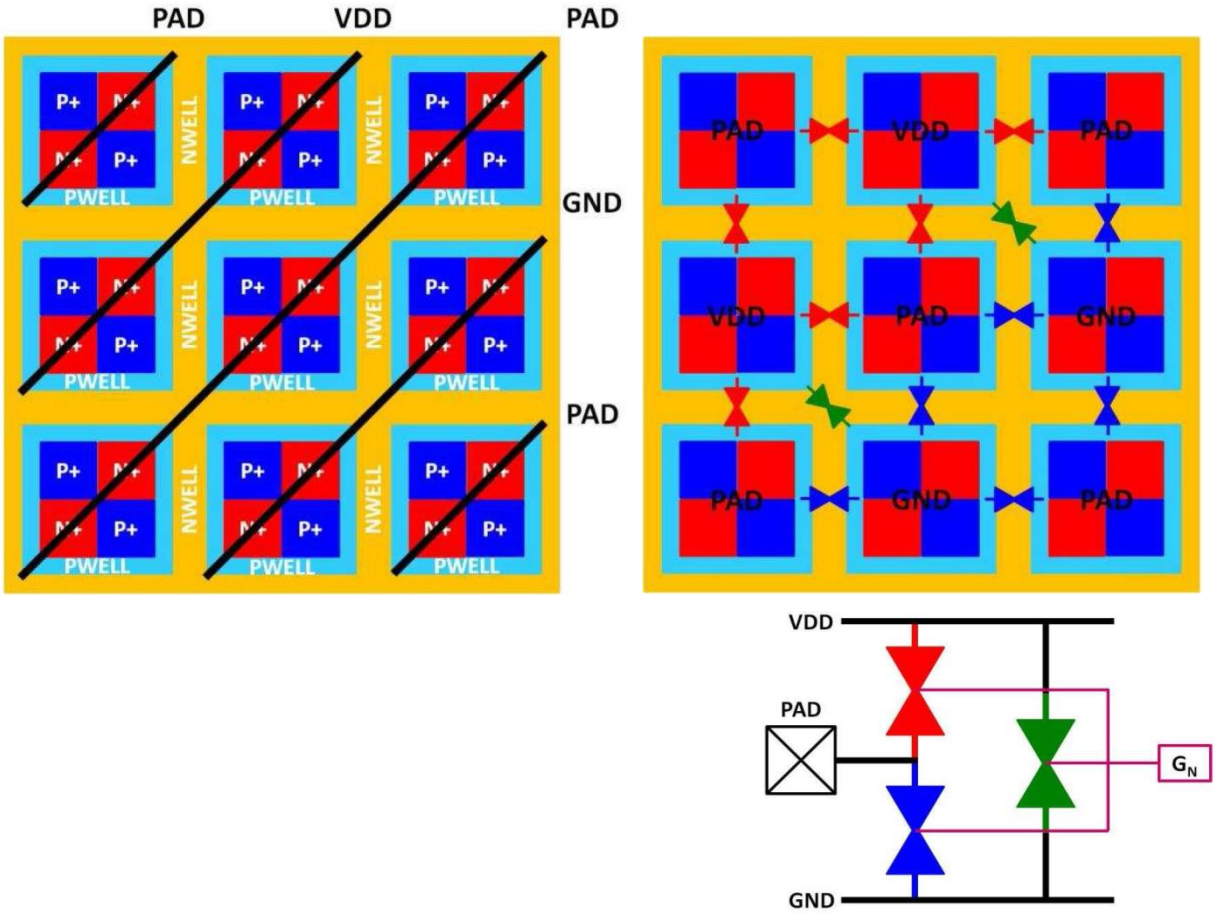


Figure 30: Beta-matrix structure. Top left: Top view of the structure (P^+ doping in dark blue, N^+ in red, Nwell in yellow and Pwell in light blue) with its metallic connections (thick black lines that represent connections lines to some I/O PAD, to VDD or to GND). Top right: the different triacs are represented on the beta-matrix; in red the triacs that link VDD and PAD, in blue the ones between PAD and GND, and in green the ones between VDD and GND. Bottom: electric schematic of the beta-matrix (which explains why it is a global protection strategy) [79].

V. Objectives

The work presented in this manuscript is focused on the “design of 3D protections against ESD in advanced FD-SOI thin-film multilayer technologies”.

In Chapter 1, an overview of the 28 nm UTBB (ultra-thin Body and BOX) FD-SOI technology is given. Then ESD are defined and the importance of protecting circuits against ESD is highlighted. The three main models of ESD discharges are explained, the expectations for a good electrical behavior of ESD protections are set, and strategies of placement of ESD protections in the circuit are described. The state-of-the-art of existing ESD protection devices of interest for this PhD research is provided. Measurement and simulation tools for investigating the electrical behavior of the devices are introduced.

Chapter 2 is dedicated to 1D ESD protections in the thin-film. 1D is thought as the current flowing in one direction mainly. The aim is to improve existing thin-film devices and better understand them. This paves the way to 2D ESD protections. Indeed, the first step is to master thin-film protections and understand triggering phenomena, before studying 2D protections. The study of classical ESD devices in thin-film also enables to build ESD protections for 3D technologies such as Coolcube^{TM}, where the upper part will consist in thin silicon layers. Another information to mention is that the BOX is usually removed from the FD-SOI wafer to obtain what is called a “hybrid” region where ESD protections are built. The goal is to place protections in the bulk part of the wafer for benefiting from volumic conduction of current. Therefore, building ESD protections in the thin-film of FD-SOI wafers is already quite innovative anyways. The whole challenge of this chapter is to provide such protections built in the thin-film silicon that can fit different ESD design windows and are able to stand the maximum possible discharge current.

Matrices are introduced in Chapter 3, which deals with 2D ESD protections. Those devices take advantage of the current conduction in two directions. This also facilitates imagining 3D matrix of devices. We dedicated our study to the BIMOS device. At first, a new topology of BIMOS is introduced: the BIMOS dot, where the body contact is shifted inside the gate. Then a matrix of BIMOS dot devices is proposed and compared to a matrix of classical BIMOS devices. In a second part, different known 1D BIMOS topologies are compared to a matrix of BIMOS, where the body contact is situated around the matrix of sources and drains. The aim is to verify if the technology is mature enough to allow 2D structures to obtain better electrical performances than in 1D.

Finally, new 3D designs are proposed and described in Chapter 4. They consist in stacking one device above and another one under the BOX. Those devices can be merged by opening the BOX in selected regions, in order to benefit from 3D conduction current. A great care has been taken so that the designs can directly be fabricated using the actual FD-SOI technology of STMicroelectronics.

Chapter 2: ESD thin film devices

"In theory, there is no difference between theory and practice.
In practice, there is."
Attributed to Jan L. A. van de Snepscheut

To optimize chips and to be more competitive for the CMOS technology, the 3D technology is investigated. An example is the Coolcube™ project [10] [98] [99]. The idea is to achieve a monolithic integration by stacking vertically layers of devices (Figure 31). This allows to align the upper layer with a precision of the size of one transistor only. The process flow of the interconnections is the same as for standard planar tungsten vias. The contacts are made through oxide with a higher depth, in order to have enough space for the second layer of devices. There are many challenges to increase the yield of devices sequentially stacked. Indeed, the first layer of devices would not stand the use of a classical temperature of process (1200°C) when fabricating the second layer, due to all the interconnections that could melt or the doping layers that could expand. Therefore, the second layer has to be processed at a low temperature (500°C) in order not to destroy the first layer. Some work is done in the CEA Grenoble to ensure that "cold" transistors [100] [101] [102] – processed at low temperature – would still have the required electrical characteristics. Devices as in Figure 31 were not studied in this manuscript.

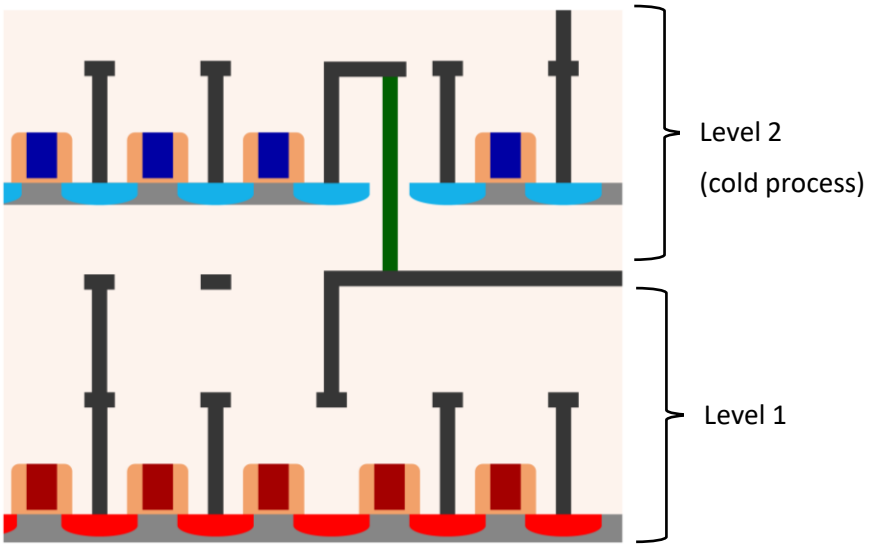


Figure 31: Layer of "cold" transistors above a layer of "hot" transistors, to illustrate the principle of Coolcube™.

In order to be able to stack layers above one another, it is very important to ensure that each layer is protected against ESDs. Traditionally, ESD protections in FD-SOI are made in the hybrid section of the wafer, for benefiting from the volumic conduction of the bulk. But to protect an upper layer of the circuit, there would be no choice but to design ESD protections in the thin-film.

Other generic advantages of implementing ESD protections in the thin-film are: (i) the possibility of using the back-plane as a back gate; (ii) less parasitic elements (junctions, capacitor, bipolar, …) for better control and silicon area saving; (iii) there is no need to open the BOX, so process steps are potentially easier and less margin with neighboring components is required; and (iv) the difference between the triggering mechanisms in bulk and thin-film offers additional possibilities of implementation in order to shift the electrical behavior of the protections.

ESD protection in the thin-film is the first step toward protecting emerging technologies, or other 3D monolithic integrated circuits. It leads to new ESD challenges. Therefore, the aim of this chapter will be to understand better some protections in the thin-film, and to improve them. The study will be done thanks to 3D TCAD and to the characterization of fabricated devices.

I. ESD boost solution for MOSFET and BIMOS

1. Context

The BIMOS and the GGNMOS transistors are among the most used ESD protections in the industry. When they are designed in the thin-film, the parasitic capacitances play an important role in their triggering. Since the transistor gate and body contact are connected together, the capacitance between the drain and the channel (or body contact) - called C_{DB} - and the capacitance between the drain and the gate - C_{DG} - help raising the voltage on the gate of the transistor with the increase in voltage on the drain, where the ESD arises (Figure 32). Therefore, increasing the value of the parasitic capacitances decreases the trigger voltage of the BIMOS.

C_{DB} is the capacitance due to the N^+-Pint junction of the drain and the channel. C_{DG} is C_{DG_spacer} the capacitance of the spacers that lie between the drain and the gate, in parallel with C_{DG_ox} the capacitance of the thin oxide of the gate where there is an overlap of the drain doping under the gate.

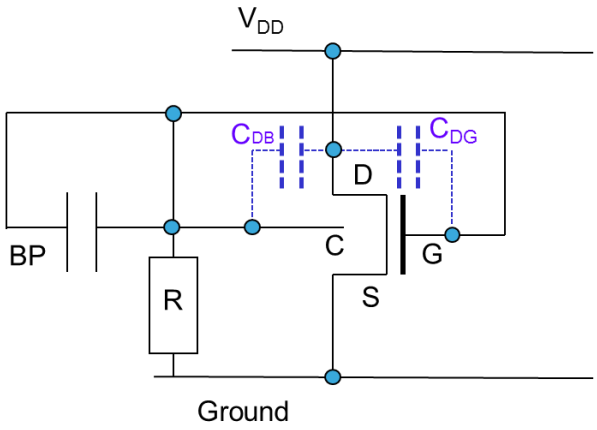


Figure 32: Schematic of a BIMOS transistor used as an ESD protection for the V_{DD} node. D stands for drain, G for gate, S for source, C for body contact and BP for back gate (back-plane). The resistor is external to the BIMOS while C_{DB} and C_{DG} are intrinsic capacitances.

Those parasitic capacitances drastically decrease for advanced technology. Indeed, C_{DB} is very low in thin silicon film compared to bulk structure, since the junction between the drain and the channel is only 7 nm thick (the size of the thin-film). In addition, C_{DG} is very low in stair epi (the spacers have the shape of stairs) compared to the MOSFET structure of the previous node. This is because $C_{DG\_spacer} = C_{DG0} + C_{DG1}$ with $C_{DG0} > C_{DG1}$ (Figure 33).

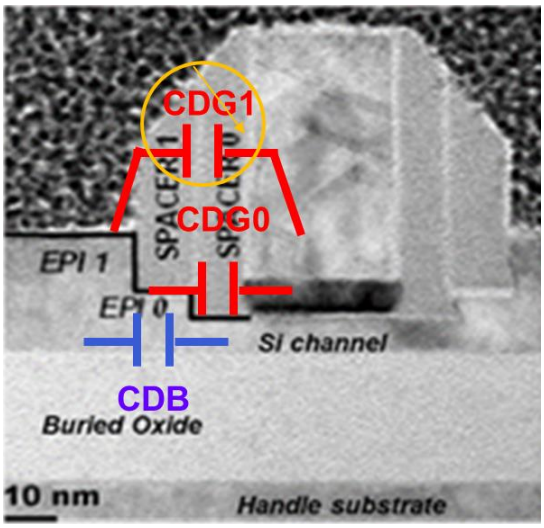


Figure 33: MOS transistor with stair epitaxy.

A solution can be to increase the value of the external polysilicon resistor that is plugged to the gate, in order to compensate for the shift in trigger voltage of the structure and keep the same ESD performance. Nevertheless, increasing the value of the resistor leads to an ultra-low AVS threshold voltage (below 1 V), because all the leakage current flowing in the channel is transformed in a non-negligible voltage through the resistor. This could cause a risk of latch up. Therefore, there is a real need in increasing the parasitic capacitance between the node drain and the node gate.

2. Analysis

We propose to use a boost in capacitance, by extending the "active" drain (in cadence virtuoso layouts, the "active" is represented by the green layer "RX" and corresponds to the thin silicon film) and the front gate. Four BIMOS devices are compared (Figure 34 and Figure 35): a BIMOS with one finger ("BIMOS_1finger"), with two fingers ("BIMOS_2fingers"), and two BIMOS with one finger but an extended drain for having a capacitive boost. They have the same silicon footprint than the BIMOS_2fingers. One of them is called "BIMOS_capaboost_float" because it has connected gates and a floating active, which leads to one additional parasitic capacitance compared to the BIMOS_1finger. The other one is called "BIMOS_capaboost"; both the extended active and the additional gate are connected, for having two additional parasitic capacitances with respect to the BIMOS_1finger.

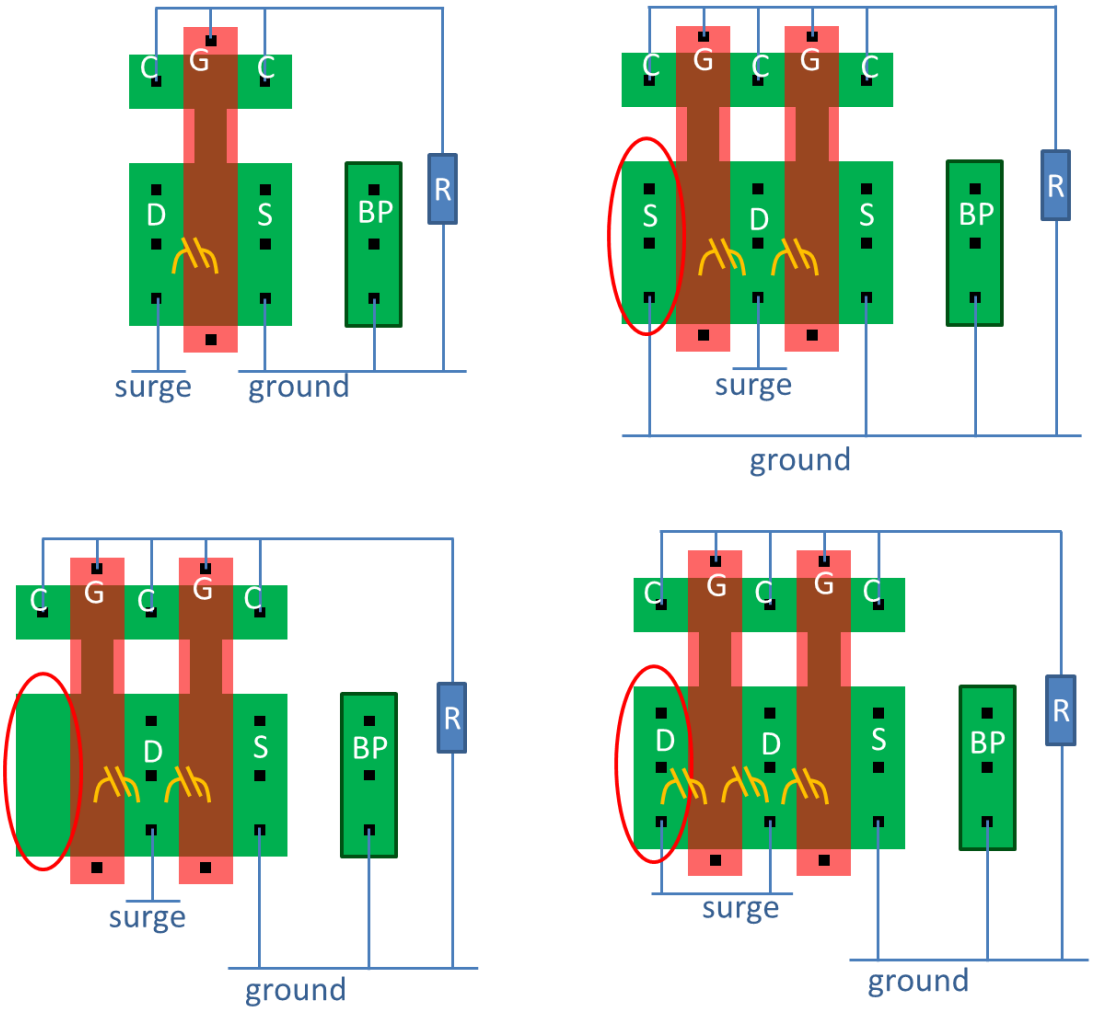


Figure 34: Top views of BIMOS and their connectivity. Top left: BIMOS_1finger; Top right: BIMOS_2fingers; Bottom left: BIMOS_capaboost_float; Bottom right: BIMOS_capaboost. The colors correspond to cadence virtuoso layers, like for Figure 35 (active in green, gate stack in red). The parasitic capacitances are displayed in yellow. Note the connectivity of the additional finger (encircled in red).

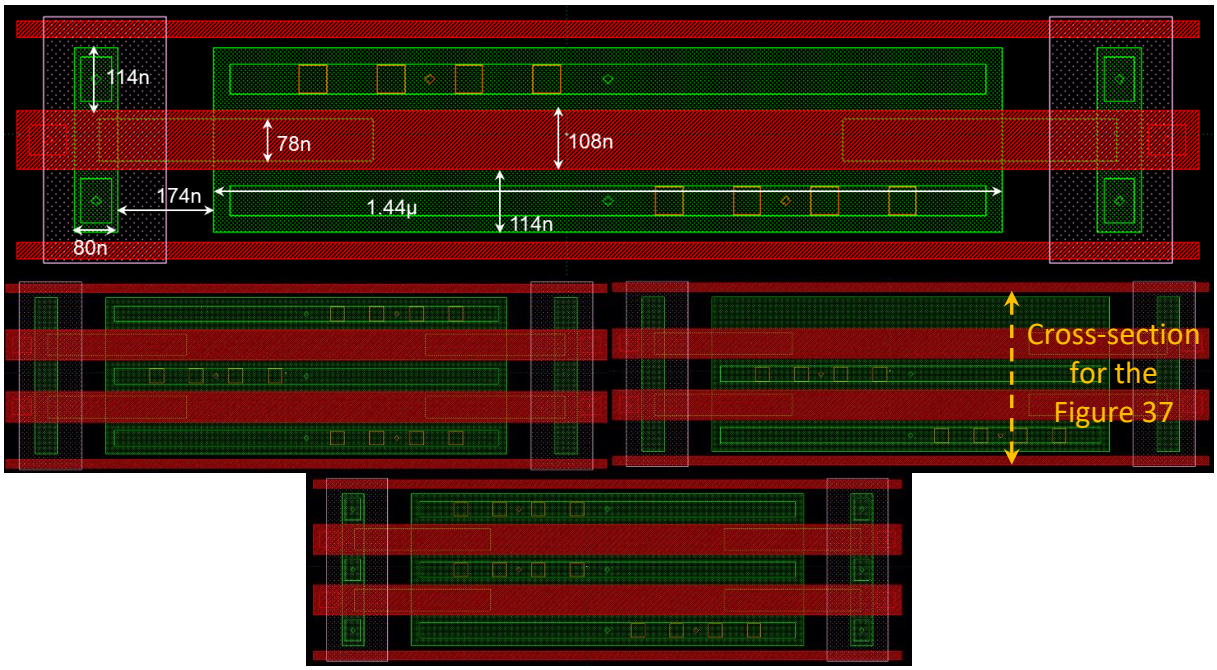


Figure 35: Layouts. Top: BIMOS_1finger and its dimensions; Middle left: BIMOS_2fingers; Middle right: BIMOS_capaboost_float; Bottom: BIMOS_capaboost. The orange squares show where the contacts are situated; there is one metal line for connecting the drain (left of each device) and one for the source (right of each device) but it is not displayed.

Electro-thermal ACS simulations were performed on devices with a width (finger length) of 10 µm. All the simulations start with a device temperature of 300 K and are stopped when the hot spot of the device reaches 800 K (Figure 36). It can be seen that the hot spot is near the drain, in the channel (Figure 37, Figure 38 and Figure 39). In Figure 37, devices are plotted when the drain current is 10 mA; the hot spot corresponds to the red region in the BIMOS_1finger device and to the green region in the BIMOS_2fingers and BIMOS_capaboost devices. In Figure 39, which is obtained from a cross-section in Figure 37, the hot spot location can be found in the X axis when looking at the maximum of the curve in terms of temperature. The Figure 38 is analogous to the Figure 39 except that the temperature in the devices is plotted for a drain current of 1 mA instead of 10 mA.

It is normal that for a given drain current, the maximal temperature inside the BIMOS_1finger is higher than the one inside the BIMOS_2fingers (Figure 38 and Figure 39, the light blue curve - corresponding to the BIMOS_1finger - has a higher maximal temperature than the dark blue curve – corresponding to the BIMOS_2fingers). Indeed, the second finger of the BIMOS_2fingers can also help evacuating some current so the current in each finger is relaxed, thus lowering the maximal temperature. The temperature is not equally distributed in each finger (Figure 38 and Figure 39, the dark blue curve corresponding to the BIMOS_2fingers exhibits two maxima – one for each finger –, but one of them corresponds to a higher temperature than the other) due to meshing differences. In real devices, process differences will favor the current to flow into one finger preferentially to the other.

The BIMOS_capaboost has only one finger of conduction of current, like the BIMOS_1finger. Though, the temperature inside its conductive channel (the hot spot) is less than in the BIMOS_1finger (lower maximal temperature in the case of the red curve with respect to the light blue curve in Figure 38 and Figure 39) since the extended active can allow a more uniform temperature spread in the device. Also, thanks to the parasitic capacitances of the BIMOS_capaboost, the device triggers earlier, and as a result, at 10 mA the BIMOS_capaboost has a similar drain voltage as the BIMOS_2fingers, and a lower drain voltage than the BIMOS_1finger (Figure 40). At a given current, the temperature is lower in a device that has a lower voltage. This explains that the maximal temperature is similar in the BIMOS_2fingers and in the BIMOS_capaboost at 10 mA (Figure 39). As a matter of comparison, the maximal temperature is higher in the BIMOS_capaboost than in the BIMOS_2fingers at 1 mA (Figure 38), with the shift in drain voltage (Figure 40).

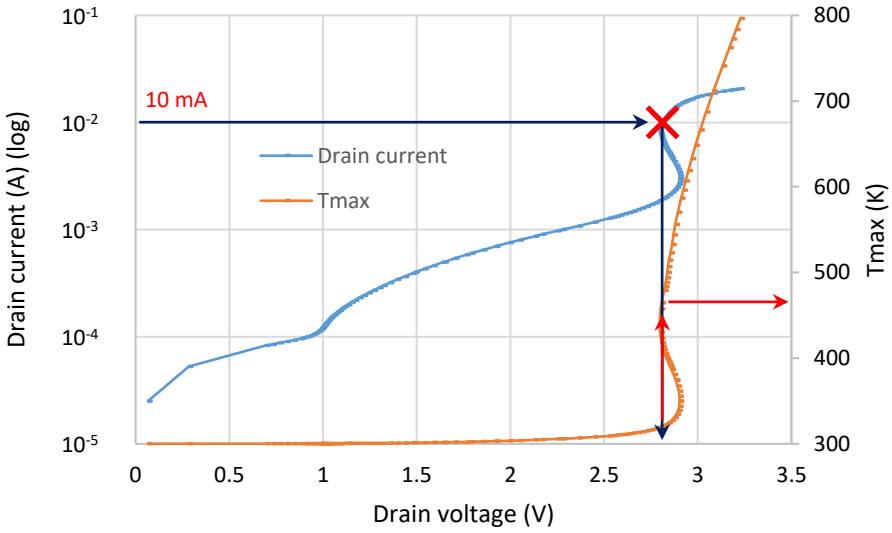


Figure 36: Electro-thermal ACS TCAD simulation. Thermal response of the BIMOS_capaboost. Figure 37 is obtained by plotting the device when the current is 10 mA, which corresponds to the emphasized point on this graph.

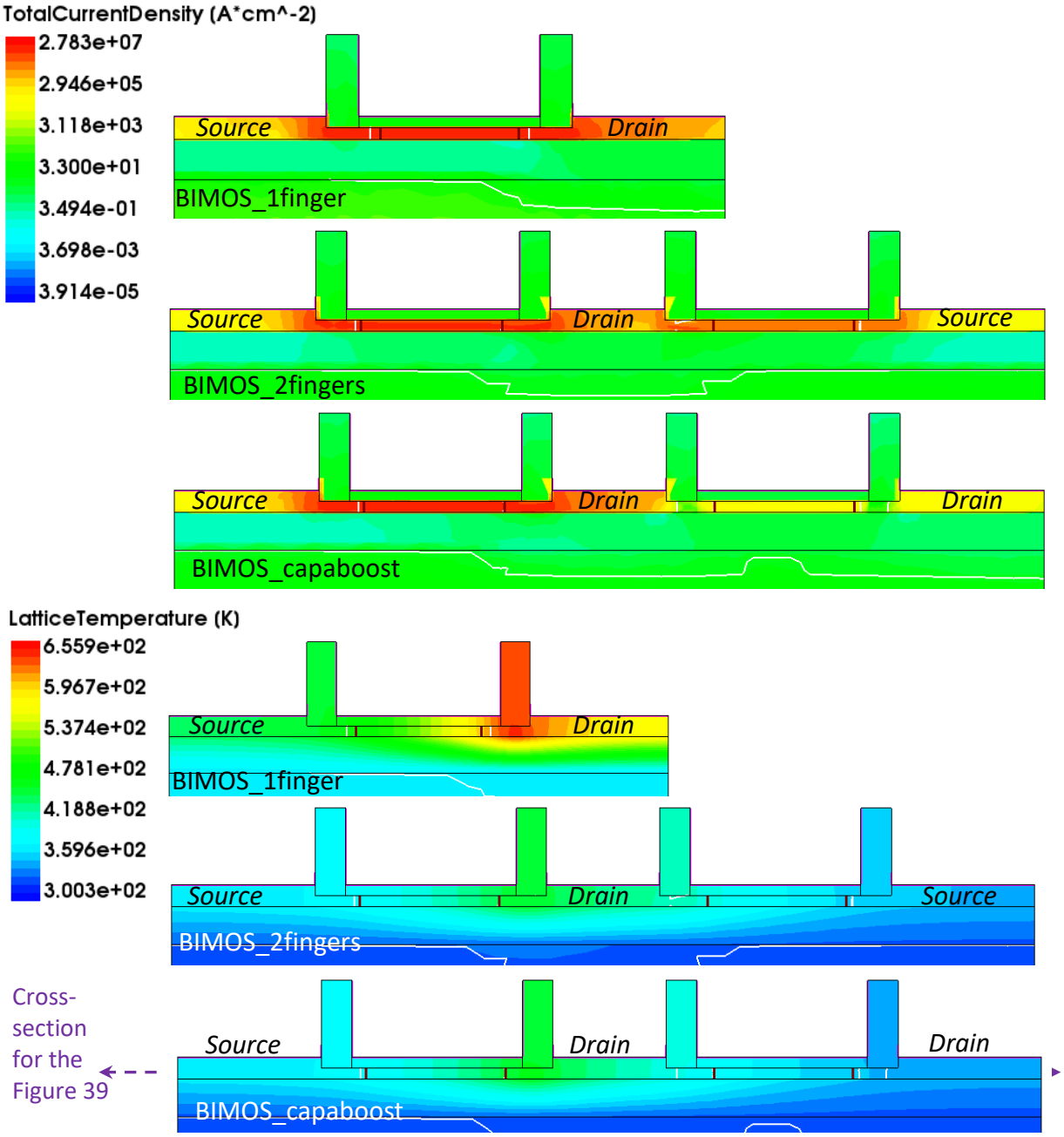


Figure 37: Electro-thermal TCAD simulation. Current density and temperature in various devices at 10 mA.

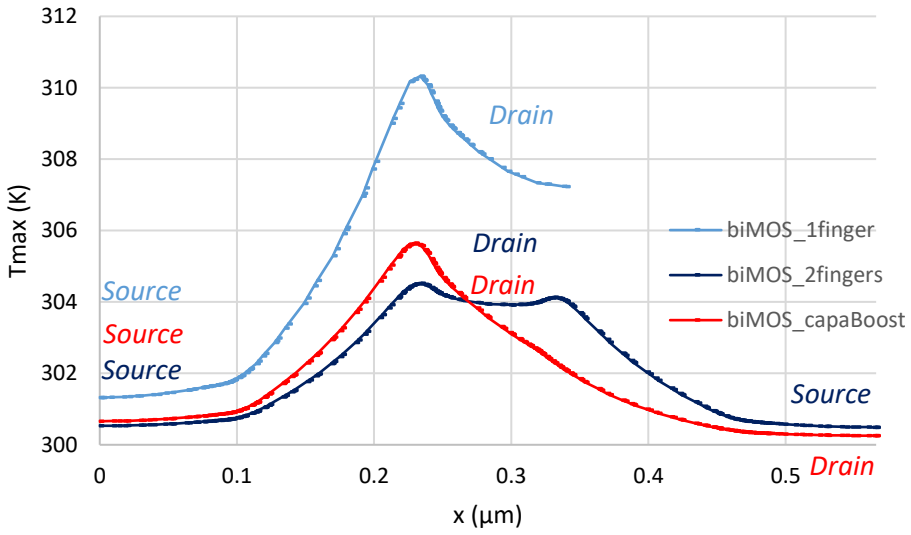


Figure 38: Electro-thermal ACS TCAD simulation. Temperature comparison (Tmax) inside devices at 1 mA.

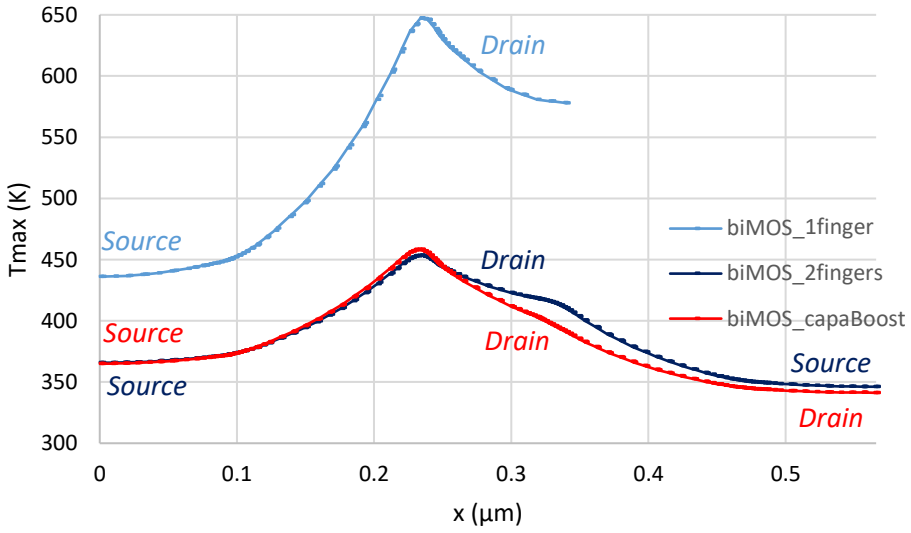


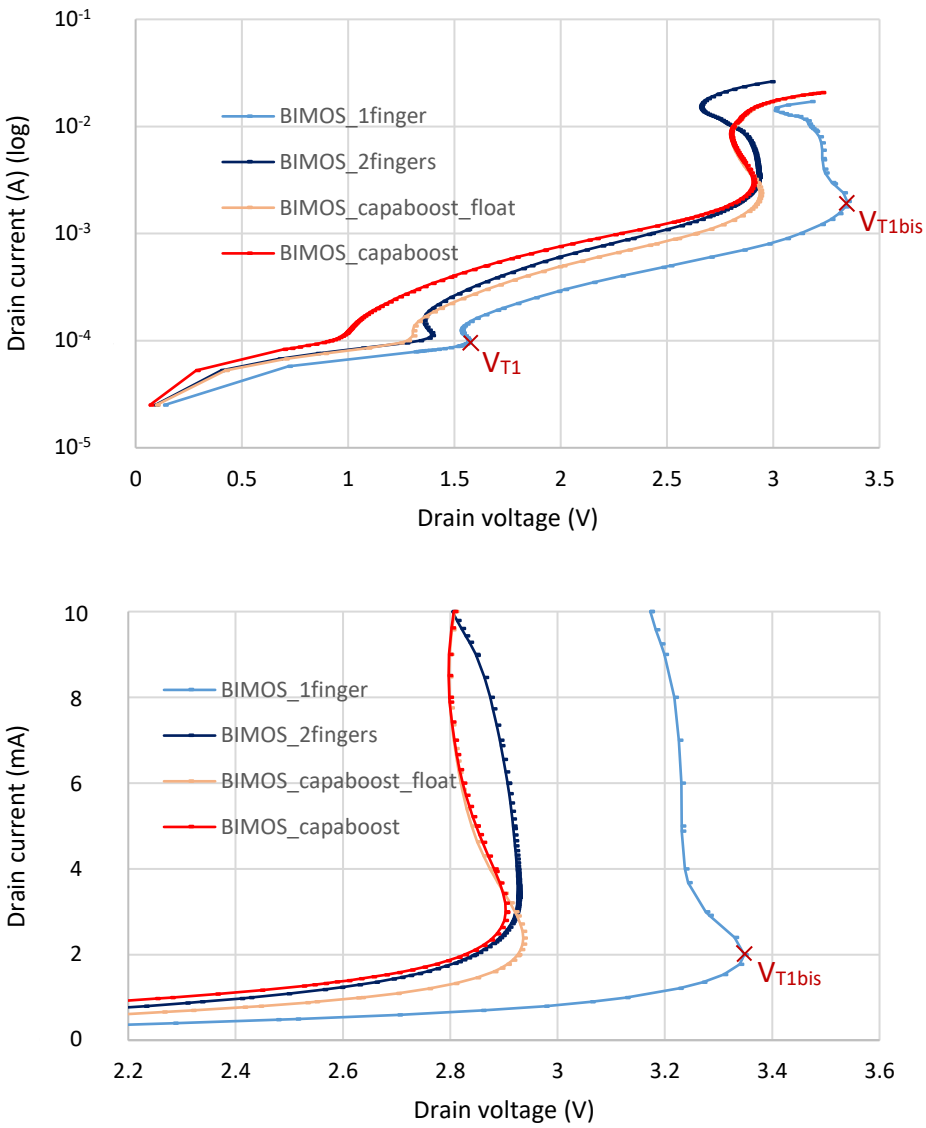
Figure 39: Electro-thermal ACS TCAD simulation. Temperature comparison (Tmax) inside devices at 10 mA. This figure is obtained thanks to the cross-section shown in Figure 37.

Figure 40 compares the four devices with an ACS. Two snap-backs can be observed. The first one occurs at V_{T1} because it is the real trigger voltage of the structure, when the structure starts to conduct significantly more current. The second one happens at V_{T1bis}, and when performing TLP measurements it is this snap-back that is observed. Indeed, the TLP workbench cannot capture currents that are smaller than 5 mA. The capacitive boost is a phenomenon that can be observed near V_{T1} and not near V_{T1bis} on the ACS, because the change in capacitance matters at short time scale ($\tau = RC$). In an ACS, a ramp of current is performed in 100 ns, therefore only the smallest currents are influenced by a change in capacitance, because they are caught at a short time.

The TLP curve would not show any difference of trigger voltage between the devices (because it would show the V_{T1bis} only). However, the ESD protection that has the lowest V_{T1} would better stand an ESD event than the other protections that are apparently similar on the TLP curves but that actually exhibit a higher V_{T1}. There are two reasons for this.

First, the overshoot of the protection that has the highest V_{T1} would be higher than the others. This is because the overshoot is the voltage phenomenon that happens when the voltage pulse is established, during the rise time, so when the time is very short. Therefore, it corresponds to what is observed at small current in the ACS (for a short time of simulation, when the device did not have the time to reach its equilibrium). The overshoot is typically not taken into account in the TLP curves, even though a too intense overshoot can damage the protection. The only way to capture the overshoots of the TLP measurement is to look at the waveforms of the voltage versus time. Another technique to be able to get the small-time scale effects on devices is to perform a VF-TLP characterization.

The second reason is that the protection with the lowest V_{T1} will be more likely to trigger even when a low energy ESD is occurring. By low energy ESD we consider what could happen when the ESD protection network of devices experiences a small abnormal current. For example, during an ESD event, the protection device triggers and most of the charge is evacuated. When the ESD event is over, the protection becomes OFF. The small residual current I_ε - due to charges that did not have time to be evacuated before the protection closed - is problematic. Indeed, even if it is small, it is transformed into a high voltage at the terminals of the ESD protection, since the protection acts like a highly resistive element (because it is closed). This high voltage can be very destructive for the components to be protected (the ESD protection did not manage to play its role of voltage clamp and failed to protect the functional circuit). This phenomenon of high voltage occurring right after the protection closes can be caught on the voltage versus time waveforms of the TLP curve, by looking at what happens after the voltage pulse, for times greater than 100 ns. Another example of low energy ESD is when a small current I_0 arrives on the I/O ring and spreads into the different rails, so it is divided into smaller currents. A small I/O ring - that does not have a lot of branches - will not have problems, since each small current will be still high enough to trigger the protection devices. But in the case of a bigger I/O ring with a lot of branches, the current I_0 will be divided into too small currents I_ε, so the protections will not be triggered. This current I_ε will be transformed into a high voltage through the high impedance of the closed protection. As a consequence, the circuit will experience the breakdown of its components. A smaller V_{T1} in the ACS curve means that even if the protection is still not able to evacuate a huge amount of charges yet, it starts to conduct a small current for a smaller voltage than the protection that exhibits a higher V_{T1}. So even if a small current I_ε arises, the protection will let it flow since it will be already a little opened. Therefore, the problem of initiating a high voltage is avoided.



	V_{T1} (V)	V_{T1bis} (V)
BIMOS_1finger	1.6	3.3
BIMOS_2fingers	1.4	2.9
BIMOS_capaboost_float	1.3	2.9
BIMOS_capaboost	1.0	2.9

Figure 40: Electro-thermal ACS TCAD simulation. Comparison of the BIMOS_1finger, BIMOS_2fingers, BIMOS_capaboost_float and BIMOS_capaboost devices. Bottom: zoom on the V_{T1bis} of the curves. Table: extraction of V_{T1} and V_{T1bis} in the devices.

Thanks to its additional parasitic capacitances, the BIMOS_capaboost triggers earlier than the BIMOS_1finger and the BIMOS_2fingers, because - as expected - its V_{T1} is smaller (Figure 40). The BIMOS_2fingers and BIMOS_capaboost_float have a similar V_{T1} since they both have a similar amount of parasitic capacitances, and they trigger before the BIMOS_1finger. The BIMOS_1finger, BIMOS_capaboost and BIMOS_capaboost_float have

similar R_{ON} because they only have one finger of conduction; it can be seen in Figure 40 where the curves between V_{T1} and V_{T1bis} are parallel. The R_{ON} of the BIMOS_2fingers is smaller than the one of the other devices since it has a second finger of conduction, and this is why the slope between V_{T1} and V_{T1bis} of the BIMOS_2fingers is a bit steeper than the others.

Figure 41 shows that the external polysilicon resistor can still be used to control the trigger voltage, in addition to using the capa-boost technique.

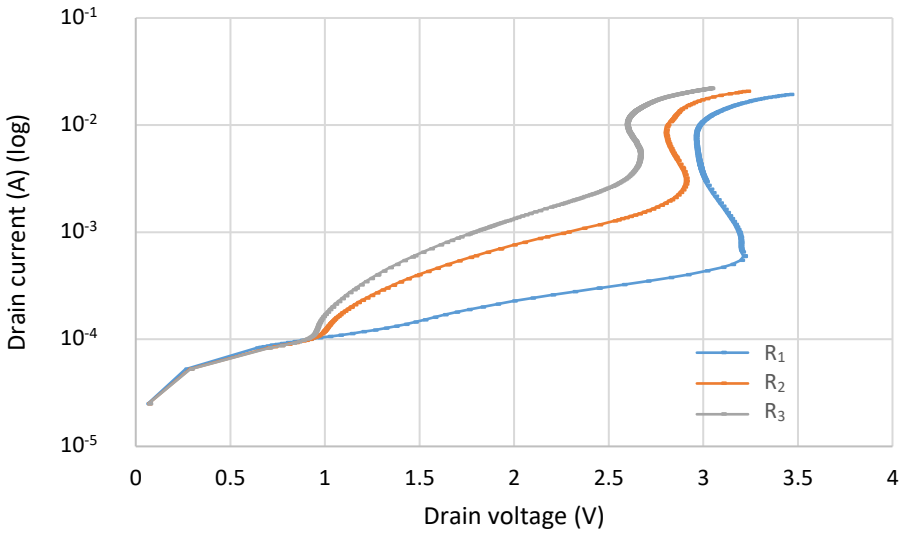


Figure 41: Electro-thermal ACS TCAD simulation. BIMOS_capaboost, with different values of external resistor ($R_1 < R_2 < R_3$).

A finger can also be added in a MOSFET structure for increasing the number of parasitic capacitances. The ACS behavior is typically the same as for the BIMOS device, *i.e.* with the V_{T1} drop thanks to the capacitance boost. When performing an AVS however, there is no difference between a MOS_1finger, a MOS_2fingers and a MOS_capaboost, since the AVS is performed over a long time (1 ms), which makes all the capacitive phenomena vanish. On the AVS curve (Figure 42) there is a difference of trigger voltage between the BIMOS_1finger, BIMOS_2fingers and BIMOS_capaboost, thanks to the body contact being connected to the gate in the BIMOS configuration.

The leakage current (at low voltage) is higher for the BIMOS_2finger and the BIMOS_capaboost than for the BIMOS_1finger, because those devices have two fingers that can allow a leakage current between the anode and the Body Contact, through the junction; it is as if their width was twice larger. The trigger voltage of the BIMOS_2fingers is the smallest, due to all the leakage currents that flow through the Body Contact being transformed via the external resistor into a voltage on the gates. The BIMOS_capaboost has a higher trigger voltage, indeed the current is less likely to flow in the finger with the other extremity of the active that is also at the highest potential (drain connection) than in the other finger. The BIMOS_1finger has the highest trigger voltage since it has less leakage.

An interesting hypothesis about the leakage current at low voltage is that it is partly (maybe mostly) constituted by the displacement current. In the AVS curves, dV/dt is 5.10^3 $V.s^{-1}$ (because the voltage sweep is performed in 1 ms and goes from 0 to 5 V), therefore there is a non-negligible displacement current CdV/dt of the order of 1.10^{-10} A (if we take typical values of parasitic capacitances), which gives the lower boundary of the current that can be simulated. This means that the static leakage can even be lower than the simulated leakage current. It also means that the values of the parasitic capacitances of the BIMOS could be extracted from the AVS curve (Figure 42), and that the BIMOS_2fingers and BIMOS_capaboost have a higher leakage current at low drain voltage than the BIMOS with one finger because they have a higher parasitic capacitance.

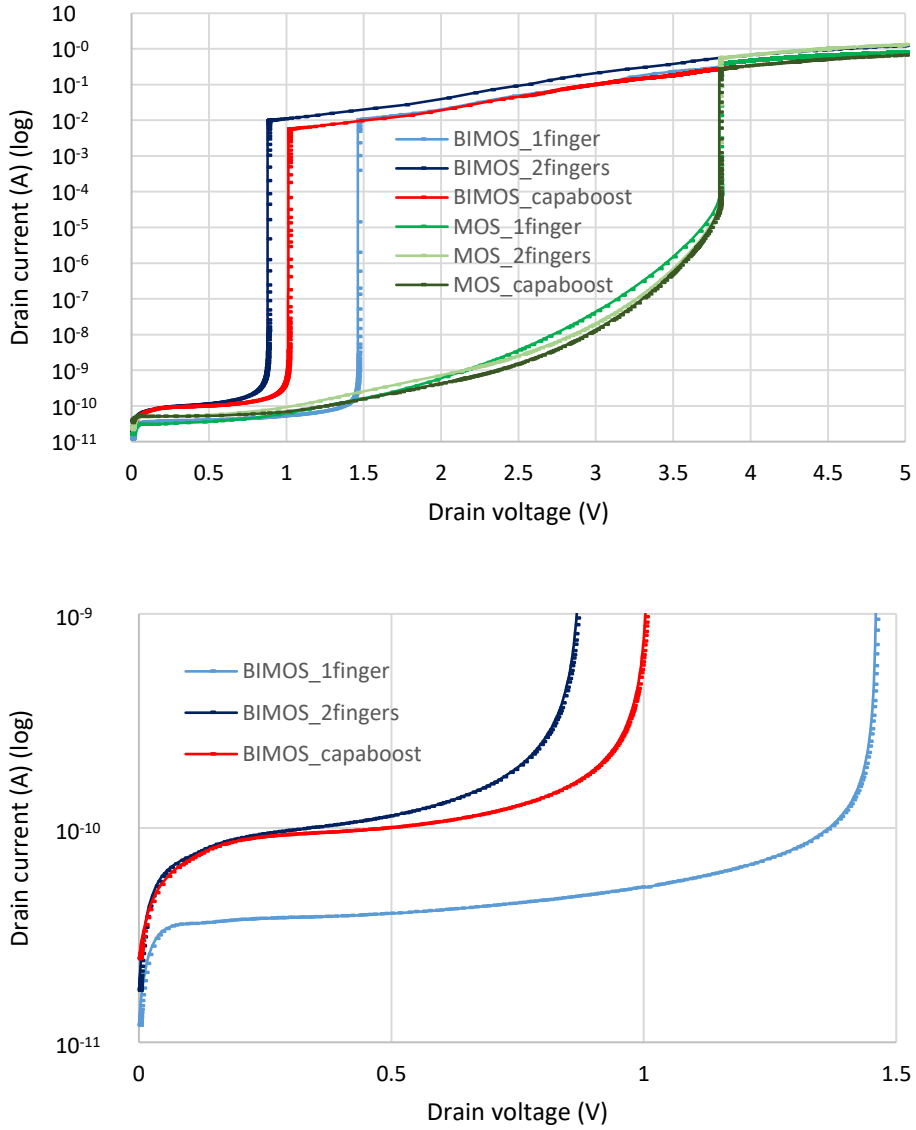


Figure 42: Electro-thermal AVS TCAD simulation of NMOS and BIMOS devices with a channel of 112 nm. Bottom: zoom on the BIMOS curves.

One can notice that the ACS and AVS results are not consistent: the trigger voltage of the ACS curves is not always the same as in the AVS curves. This remark holds for the whole manuscript; even if the ACS and AVS trigger voltage of the BIMOS_1finger is almost the same (Figure 43), it is only a coincidence, because it does not rely at all on the same mechanisms. Other devices exhibit a much larger difference of trigger voltage between their ACS and the AVS curves, like the GDNMOS device (Figure 86, Figure 95 and Figure 107 for example).

- In the case of the ACS, the phenomenon is transient (rapid). The BIMOS activates thanks to the RC circuit built by its parasitic capacitances (between the drain and the gate) and the external resistor (connected to the gate), as explained in this section.

- In the case of AVS, we observe the quasi-static behavior of the BIMOS device. An extensive work has been performed by Thomas Bedecarrats [103] to understand the BIMOS in DC. A summary of some of his work is provided in appendix, to explain the four identified regions of the AVS curve (Figure 43). To put it in a nutshell, the BIMOS triggers because of band to band tunneling and impact ionization currents.

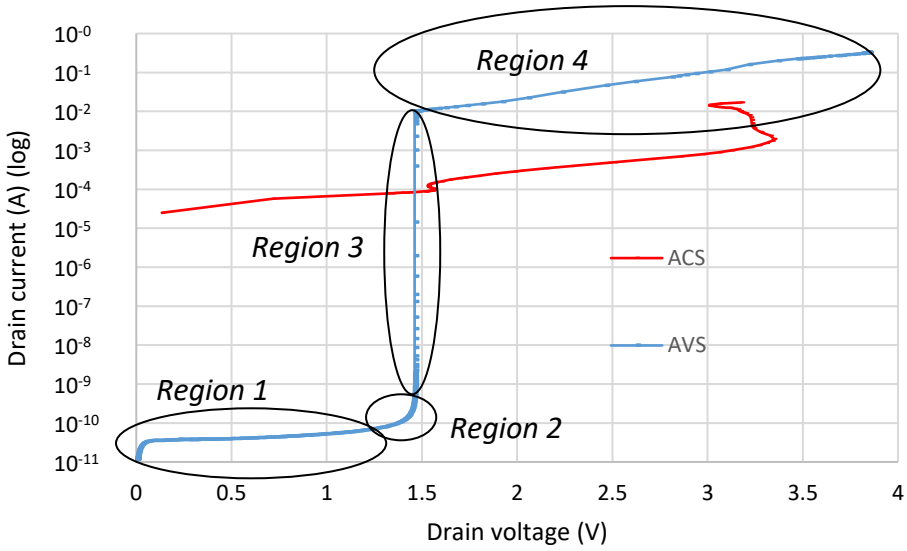


Figure 43: Electro-thermal TCAD simulations of the BIMOS_1finger device with a channel of 112 nm. Comparison of the ACS and the AVS behavior of the device. Four regions of the AVS curve are identified.

To conclude this first part of the chapter 2, the capa-boost solution is effective to reduce the trigger voltage of the protection devices, therefore increasing their effectiveness. It can be used on the BIMOS as well as on other devices (for example on a MOSFET). The reasonable use of the “capa-boost” effect still has to be investigated for RF requirements, even if some techniques can be used to manage the capacitances of the ESD protections [47].

II. GDxMOS device for high and low-voltage ESD protection

Previous studies were conducted on the GDNMOS (Gated Diode merged NMOS) device [104], [105], [106]. It is an ESD protection device where a gated diode is merged with an NMOS transistor (Figure 44). The anode and the cathode of the device correspond to the anode (P^+) of the diode and the source (N^+) of the NMOS, respectively. The shared region is located in the cathode of the diode, which is also the drain of the MOSFET. The GDNMOS device is designed in ultra-thin film SOI, the interest of this structure lying in the fact that, in principle, a SCR-like structure (with PNPN junctions) can be formed.

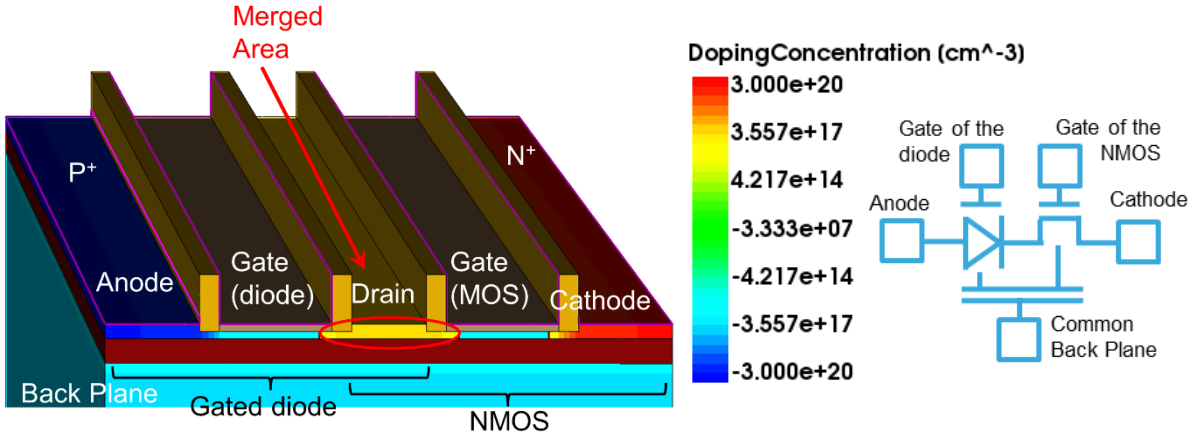


Figure 44: TCAD view (left) and schematic (right) of a GDNMOS structure.

In this section, the GDNMOS will be investigated further, and a new device - the GDBIMOS (Gated Diode merged BIMOS) - will be presented and compared to the GDNMOS. This device is composed of a gated diode merged with a BIMOS transistor. We named this family of devices GDxMOS, "X" being replaced by a "N" or a "BI" to get a GDNMOS or a GDBIMOS. In this study, the GDBIMOS has two body contacts that can be seen in the top view presented in Figure 45.

GDxMOS were fabricated using the 28 nm node ultra-thin film UTBB FD-SOI with metal gate and high-k dielectrics CMOS technology (Figure 46). All the measured structures have 100 µm total width (10 fingers of 10 µm). For simplification purpose, only one finger was simulated. In the anode-to-cathode direction, the minimum dimensions allowed by the design rules have been used when possible. The anode current and voltage were measured and simulated for a high number of variants with different connectivity conditions on the terminals. The devices are reconfigurable and promising for high and low-voltage ESD protection applications.

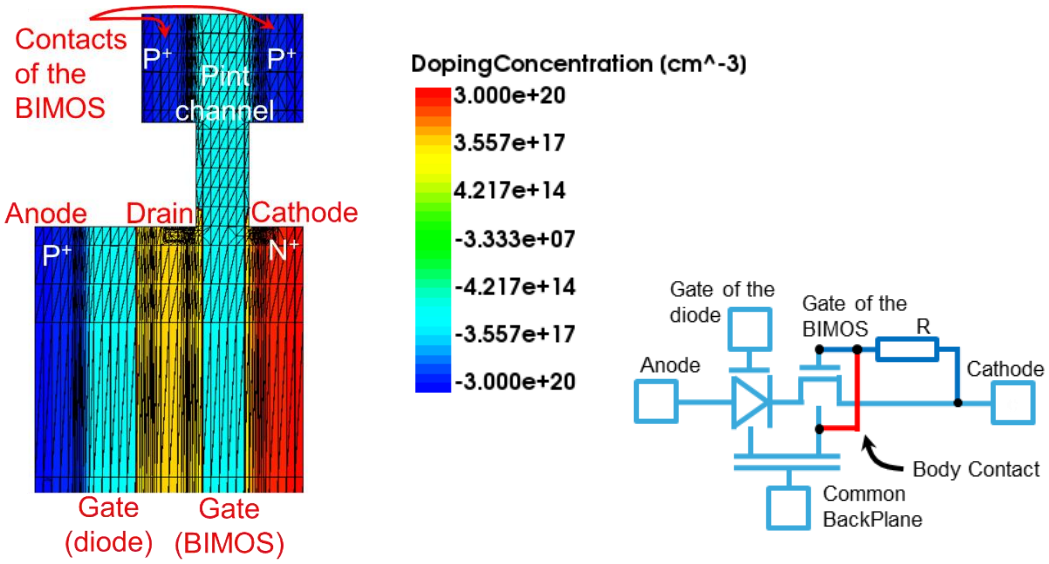


Figure 45: Top view (left) and schematic (right) of a GDBIMOS structure.

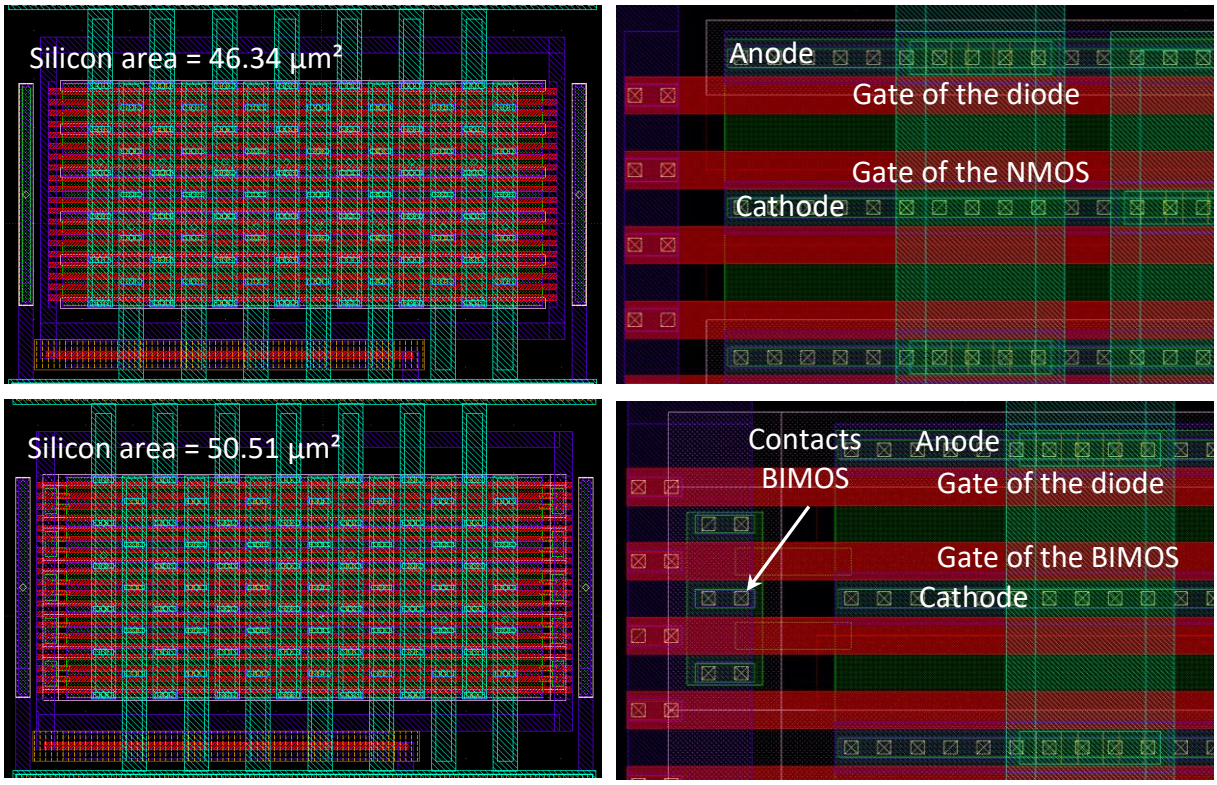


Figure 46: Typical layout of the GDxMOS structures (zoom on the right). Top: GDNMOS. Bottom: GDBIMOS.

1. GDxMOS as a high voltage protection

In this experiment, the doping in the drain remains conventional. The GDNMOS devices were numbered depending on the biasing conditions (Table 1). Device 1 corresponds to a GDNMOS with grounded front gates. Devices 2 to 5 have a grounded diode gate and their NMOS gate is plugged to an external polysilicon resistor (the value of the resistor increases with the device number *i.e.* $R_1 < R_2 < R_3 < R_4$). Devices 6 to 9 have a grounded NMOS gate and their diode gate is plugged to a resistor. Devices 10 to 13 have both gates tied together and plugged to a resistor. The back gate is grounded for all the devices. The measured GDBIMOS (devices 14 – 18) have all their front gates plugged to a resistor. Devices 14 to 17 have a grounded back gate while device 18 has the back gate plugged to a pad in order to apply different biasing conditions.

device	structure	diode gate	NMOS gate	value of R	Back gate
1	GDNMOS	grounded	grounded	/	grounded
2		grounded	to R	R_1	
3				R_2	
4				R_3	
5				R_4	
6		to R	grounded	R_1	
7				R_2	
8				R_3	
9				R_4	
10		to R	to R	R_1	
11				R_2	
12				R_3	
13				R_4	
14	GDBIMOS	to R	to R	R_1	
15				R_2	
16				R_3	
17				R_4	
18				R_3	pad

Table 1: List of measured devices.

a. ESD robustness measurements

TLP measurements with positive pulses on the anode have been performed on devices 1 – 13, the robustness of which was analyzed. The devices with grounded diode gate (1 - 5) have a robustness issue with an early break (Figure 47 and Figure 48). The reason is the absence of resistor to protect the gate of the diode, which is located just next to the anode. The resistor allows the biasing of the diode gate through the parasitic capacitance C_{AGD} between the anode and the gate of the diode. Therefore, the difference between the anode voltage V_A and the voltage of the diode gate V_{GD} is reduced during ESD event. Else, without a resistor, the potential difference between the anode and the diode grounded gate can reach the breakdown voltage of the oxide that lies between them. An extensive study would be needed in order to find out the exact phenomena that are initiating the breakdown of this oxide. It could be the trap assisted conduction [107], [108].

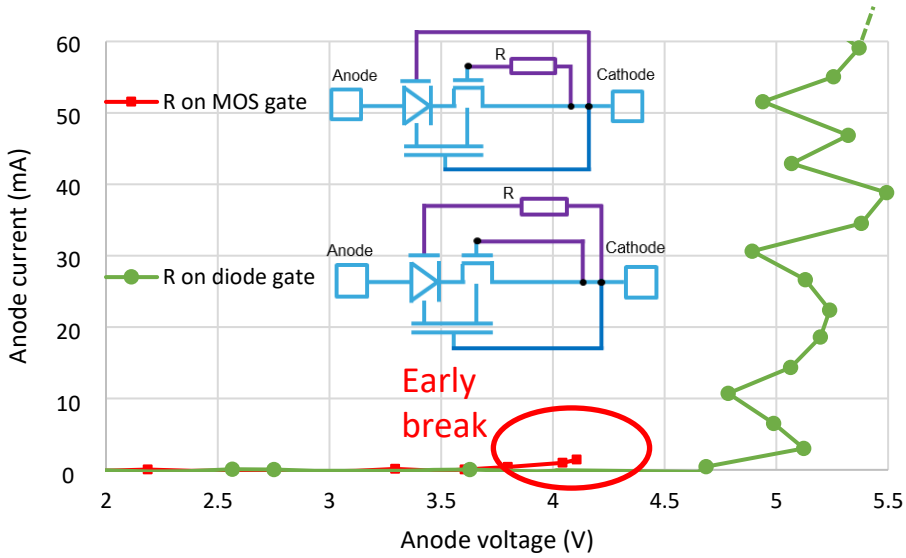


Figure 47: TLP measurements of device 4 (resistor R on NMOS gate) and of device 8 (R on diode gate).

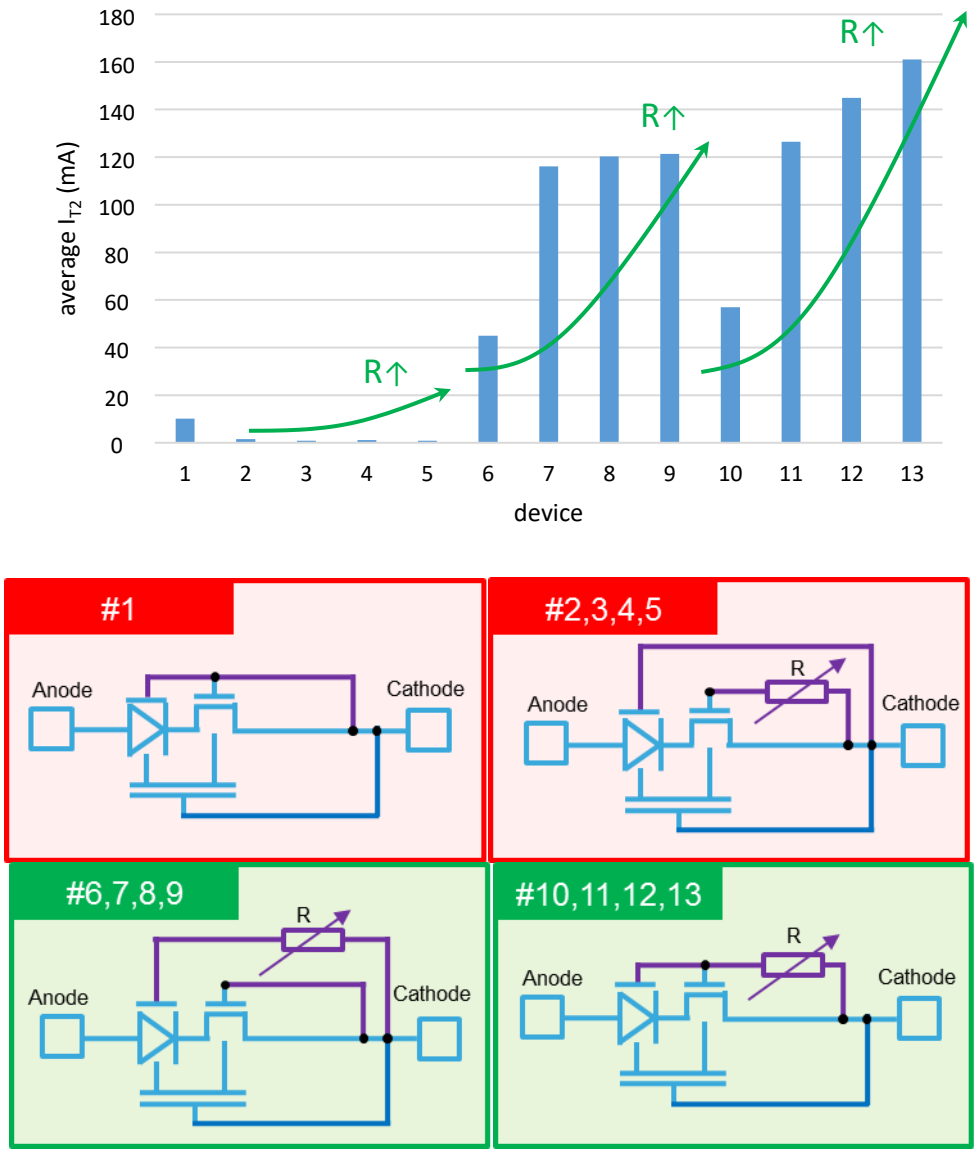


Figure 48: Average failure current I_{T2} for each device, extracted from TLP measurements performed on several dies. The corresponding topologies of structures are shown underneath the graph.

By comparing the I_{T2} of device 6 with device 7, 8 or 9 (Figure 48), it can be seen that the value of the resistor that is plugged to the gate of the diode has to be sufficiently high in order to prevent a too early break. Indeed, the gate voltage can be raised higher with a higher resistance value, thus delaying the failure and enabling a higher current to pass.

The I_{T2} of the device 6 is smaller than the one of device 10 even if the resistor value is the same. This is because device 10 has both gates plugged together and benefits from the parasitic capacitance C_{DGM} between the drain and the gate of the NMOS, in addition to the parasitic capacitance C_{AGD}. Both capacitances help raising the voltage on the gates, therefore, protecting the critical diode gate region.

I_{T2} reaches a maximum value of 0.12 A when only the gate of the diode is connected to a resistor (devices 8 and 9). In the case of both gates plugged into the resistor, I_{T2} can become higher than this value and increases with the resistor value, even at high resistor value (device 11 - 13). The reason is that the I_A vs V_A characteristic is the same for the devices 6 to 9, but the V_{T1} of devices 10 to 13 is lowered for an increased value of the resistance (see Figure 51 and Figure 53 that will be explained in the next section). As a matter of fact, the V_{T2} of the devices is about 5 V (Figure 49). The idea is that, for a fixed V_{T2} and dynamic ON resistance R_{ON}, the smaller the V_{T1}, the higher the I_{T2} (Figure 50). In reality R_{ON} changes from device to device due to the difference in the value of the external resistor R, but the difference in R_{ON} is not sufficient to compensate for the shift of V_{T1}. Therefore, the I_{T2} of devices 11, 12 and 13 is increased above 0.12 A. For the devices 1 to 5 the V_{T2} cannot reach its maximum value because of the issue with the lack of resistor on the gate of the diode.

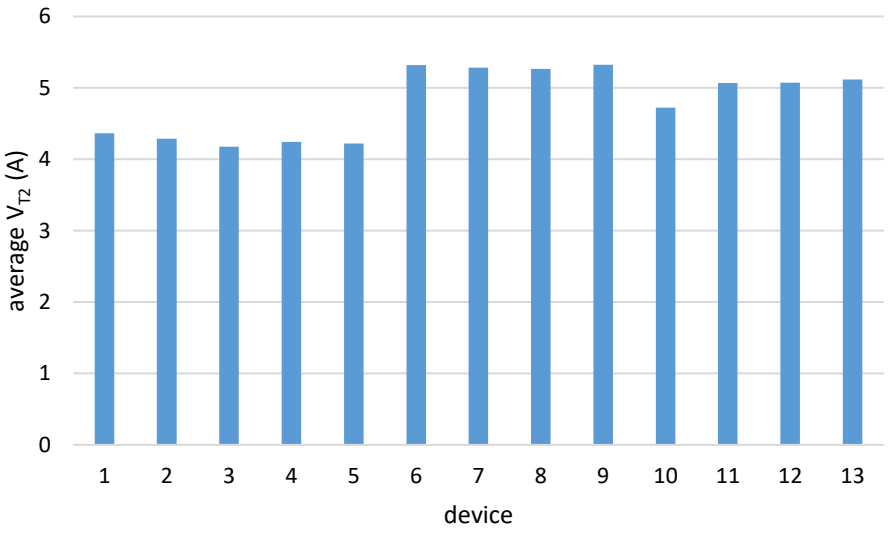


Figure 49: Average failure voltage V_{T2} for each device, extracted from TLP measurements performed on several dies.

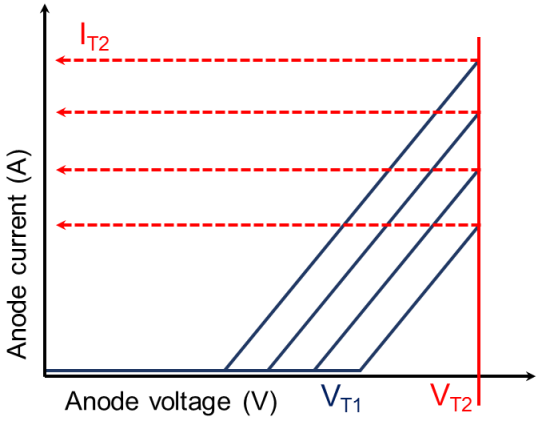


Figure 50: Generic I-V curve explaining why the devices with a smaller V_{T1} reach a higher I_{T2}.

b. Influence of the front gates on the GDNMOS

The ESD behavior of the GDNMOS devices with different biasing conditions on the front gates was investigated.

If the gate of the diode is plugged to a resistor and the gate of the NMOS is tied to the ground, the trigger voltage V_{T1} of the structure does not change with the value of the resistor (Figure 51). In fact, when the anode voltage increases, the parasitic capacitance C_{AGD} causes an increase in V_{GD} through the resistor that is plugged to the gate of the diode. The larger the resistor value, the higher the V_{GD}. A higher diode gate voltage reduces the amount of current (for a given anode voltage) that is flowing through the diode before it triggers. Numerical simulations show that the gate voltage modulates the energy barriers along the channel, which prevents the holes of the anode from diffusing to the cathode of the diode (Figure 52). Thus, with a higher resistor value, the diode is more resistive before triggering. When the potentials are such as the energy bands of the anode and channel tend to coincide, there is no more barrier impeding the injection of the holes from the anode into the channel. At this moment, the anode starts to control the level of the energy bands in the channel at the expense of the gate. When the diode triggers, the voltage of the anode reaches the value for which all the energy bands (in anode, channel and cathode) are almost at the same level, and holes are flowing through the whole diode (from the anode until the cathode). The resistivity of the diode in ON mode does not depend on the value of the resistor that is plugged to the gate, because the number of carriers flowing through the diode is so important that the gate voltage is no longer able to act on the channel. The holes flowing through the diode (from the anode to the cathode, which is also the drain of the MOSFET) are recombined in the drain of the NMOS. Most of the anode potential drops on the drain of the NMOS due to the diode being very conductive. Since the gate of the NMOS is not plugged to the resistor that is connected to the diode gate, the value of this resistor has no impact on the conduction of electrons in the NMOS part of the device.

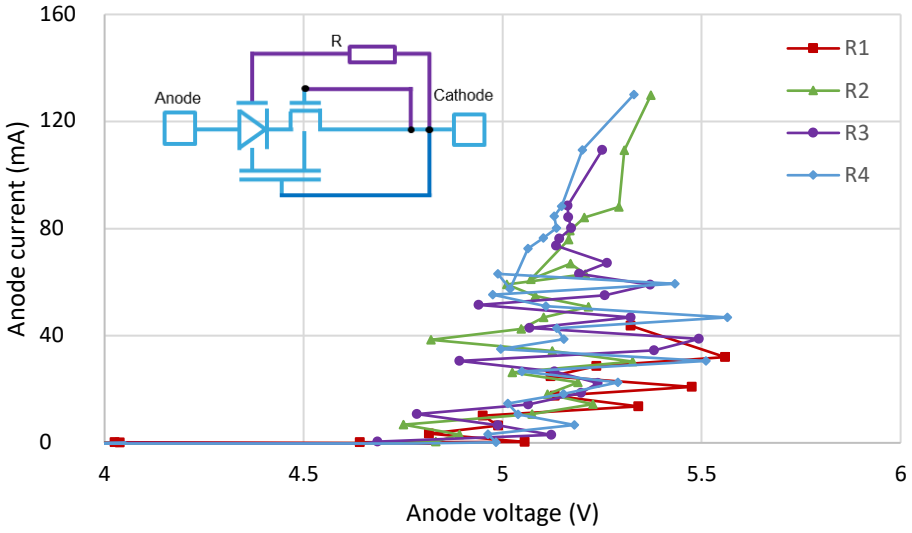


Figure 51: Multi-triggering TLP measurements of a GDNMOS with different values of resistor ($R_1 < R_2 < R_3 < R_4$ corresponding to devices 6 – 9, respectively) plugged to the diode gate. The NMOS gate is tied to the ground.

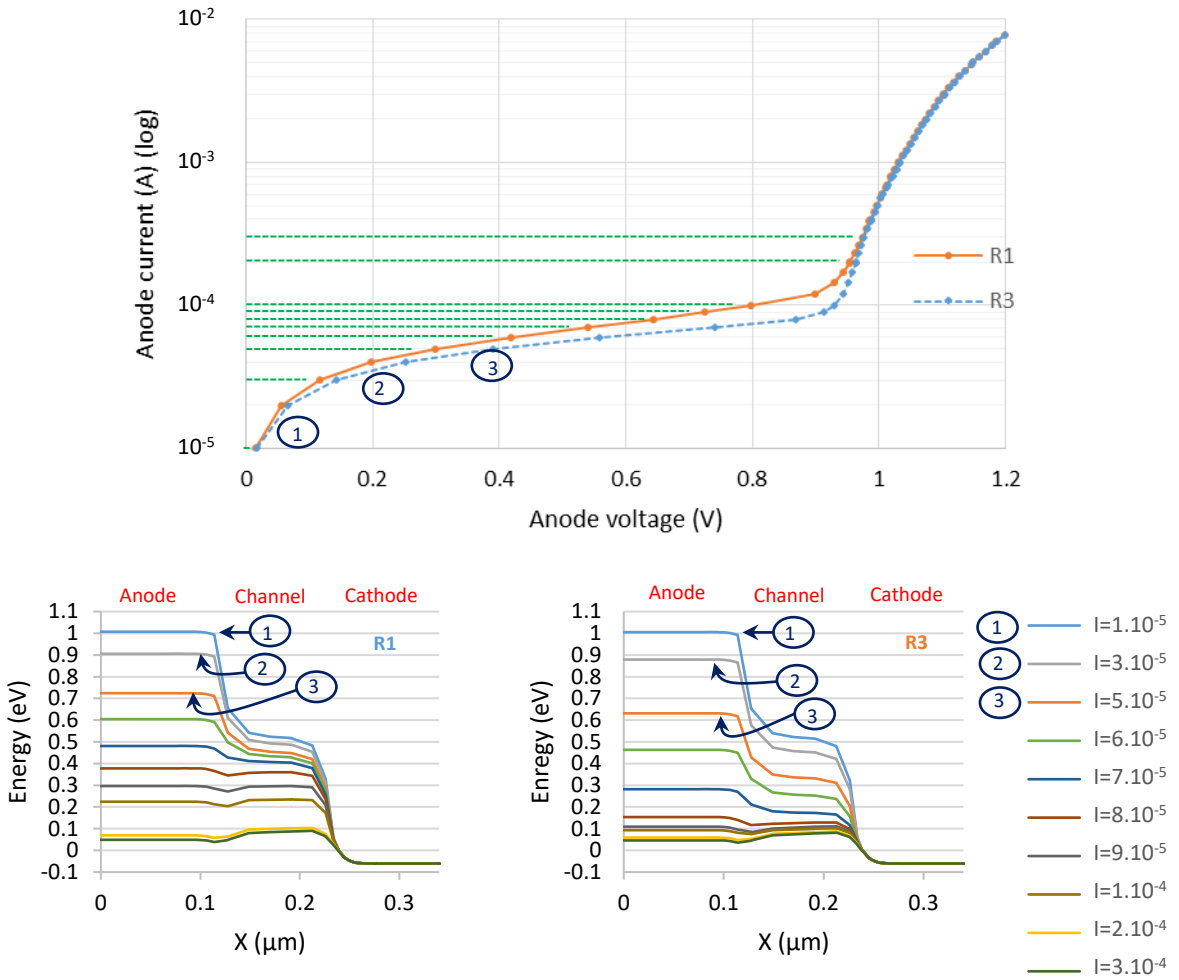


Figure 52: Top: simulated anode current versus anode voltage in two gated diodes with different resistors connected to the diode gate. Bottom: simulated conduction band energy along the gated diode for different anode currents. In the left panel, the value of the resistor R_1 that is plugged to the gate is smaller than on the right (R_3). Note that from the three upper curves (plotted for I=10^{-5} A, 3.10^{-5} A and 5.10^{-5} A) for example, the energy is decreasing faster in the anode and the channel in the case of a higher resistor.

The trigger voltage V_{T1}, measured to be around 4.8 to 5.0 V, is not modified by the value of the resistor plugged to the gate of the diode of the GDNMOS. In other words, we cannot use the resistor value in order to fit the characteristics into a different ESD design window if needed. However, the idea that a resistor on the gate of the diode is helping a gated diode or a GDNMOS to be more resistive before triggering is interesting and can be used to reduce the leakage current. Note in Figure 51, that multi-triggering is observed. The multi-triggering behavior is standard when there are several fingers.

In the case where the gate of the diode and MOSFET are tied together and connected to a resistor, V_{T1} is lower than the V_{T1} of a grounded NMOS gate, for same value of R. Figure 53 shows that a higher resistance value lowers the V_{T1}. This is because the parasitic capacitances of both gates are responsible for raising the voltage on the gates through the resistor. A higher voltage on the gates allows the NMOS part of the device to conduct

current for a lower anode voltage. Here, the V_{T1} (measured at 1 mA) is 4.0 V with R_1, 3.7 V with R_2, 3.3 V with R_3 and 2.7 V with R_4. This significant V_{T1} shift makes it possible to utilize the protection in another ESD design window, by simply adjusting the value of the external polysilicon resistor. Note that the multi-triggering is improved for higher values of resistor.

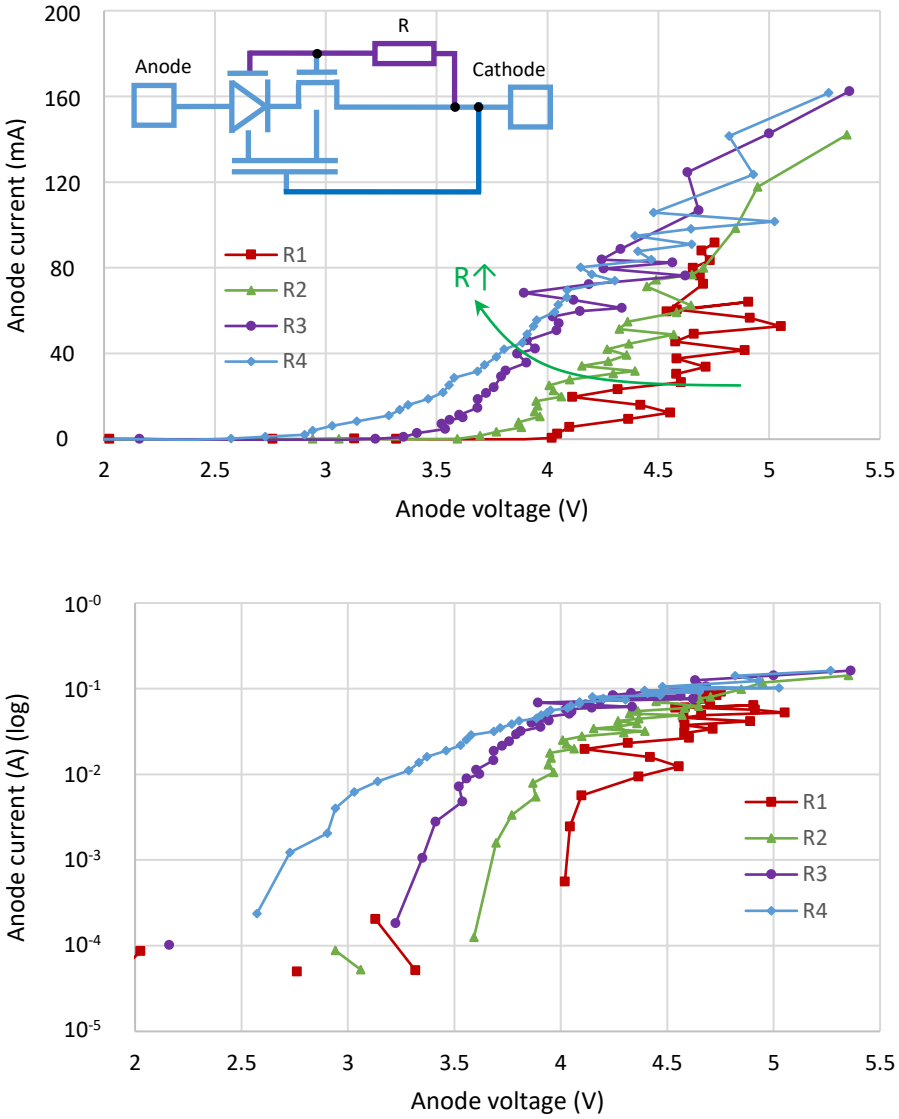


Figure 53: TLP measurements of a GDNMOS with different values of resistor plugged to both of the gates (devices 10 - 13). Top: the current is displayed in linear scale. Bottom: logarithmic scale. 100 µA is the limit of detectability of the TLP equipment. $R_1 < R_2 < R_3 < R_4$.

3D TCAD simulations confirm the V_{T1} shift according to the gates biasing conditions. According to Figure 54, plugging the gate of the diode to a resistor (compared to having grounded gates) does not change the trigger voltage. A resistor on the NMOS gate would reduce the trigger voltage (this situation cannot be observed in the measurements because of the early break). Both gates being plugged to a resistor reduces the trigger voltage further (because all the parasitic capacitances are helping to raise the voltage on the NMOS gate).

Figure 55 shows that the higher the resistor (plugged on both gates), the higher the voltage of the gates and the higher the time constant (RC) while the gate voltage is raised. This increase in gate voltage is provoking the reduction of the trigger voltage (the anode voltage for which the current starts to conduct; there is a direct link between time and current because it is an ACS).

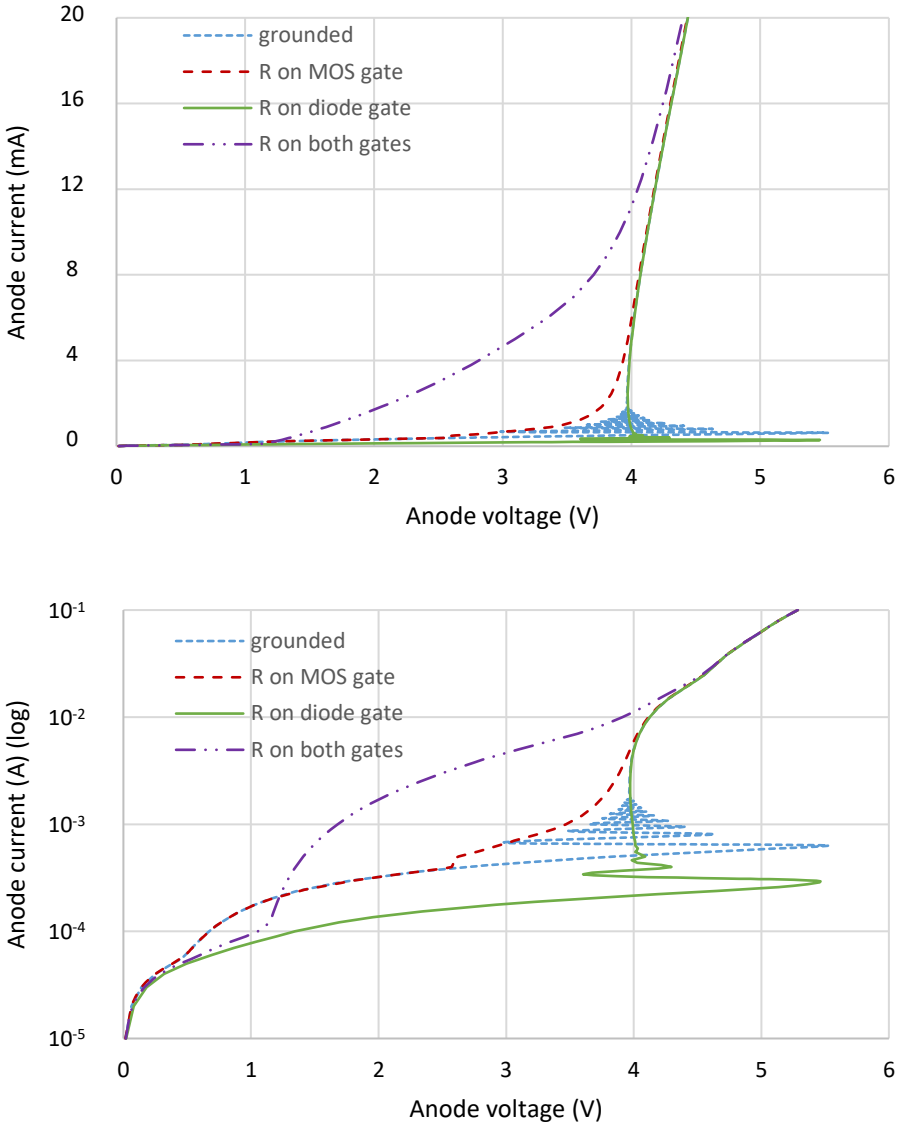


Figure 54: ACS I-V TCAD simulations of GDNMOS with different biasing conditions: both gates grounded (device 1), grounded diode gate and NMOS gate plugged to a resistor (device 4), grounded NMOS gate and diode gate plugged to a resistor (device 8), and both gates tied together and plugged to a resistor (device 12). Top: linear scale; bottom: logarithmic scale.

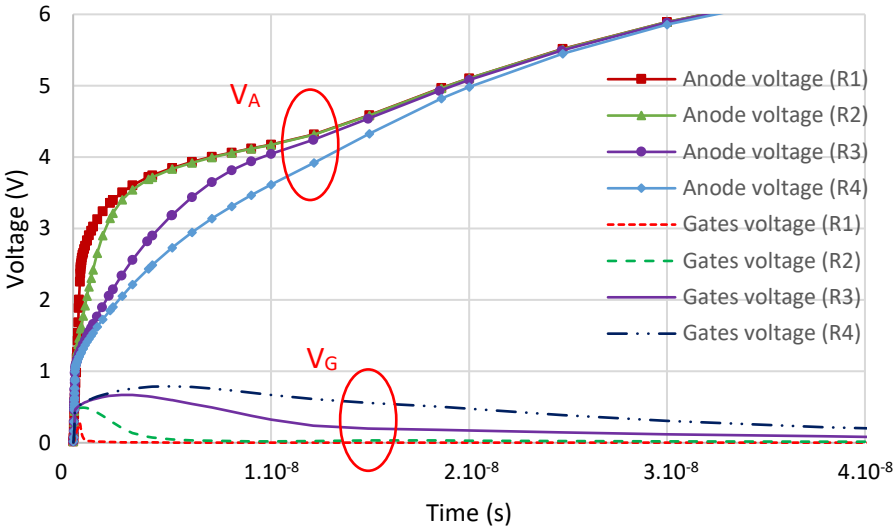


Figure 55: TCAD ACS simulations of a GDNMOS with different values of resistor ($R_1 < R_2 < R_3 < R_4$) plugged to both of the gates (devices 10 - 13). Anode and gates voltage versus time is plotted. Note here that the ACS is a current ramp of 0.1 A in 100 ns, so 10^{-8} seconds correspond to 10 mA.

Note that the oscillations in Figure 54 (and also Figure 99 and Figure 105) are not due to a numeric instability of simulation, nor to difficulties of convergence. In fact, the threshold of V_{T1} is reached (the potential is sufficiently high on the gate to trigger the NMOS; it is the threshold for which current start to flow massively in the device), so the NMOS becomes ON; therefore, the impedance of the device decreases; as a consequence, there is no need of such a high drain voltage as V_{T1} to bias the device and to obtain this current. Therefore, the drain voltage drops. Because of this, the threshold is not reached anymore and the device impedance increases. The drain voltage needs again to increase in order to allow the device to become ON. And so on. Hence, oscillations are observed in ACS, because it is a current ramp with time, so the current is constrained. This phenomenon is not observed in TLP nor in AVS, where the drain current is not forced to increase with time in the device. Some work has been done to exploit such oscillations in order to build spiking neurons [103].

Figure 55 is showing the anode voltage versus time. Since it is an ACS (ramp of current with time), this graph is like a V-I curve instead of an I-V curve (current is proportional to time). The I-V curves from which the data are plotted in Figure 55 are shown in Figure 56. Note that the behavior of the devices in Figure 53 (measurements) is similar to Figure 56 (simulations).

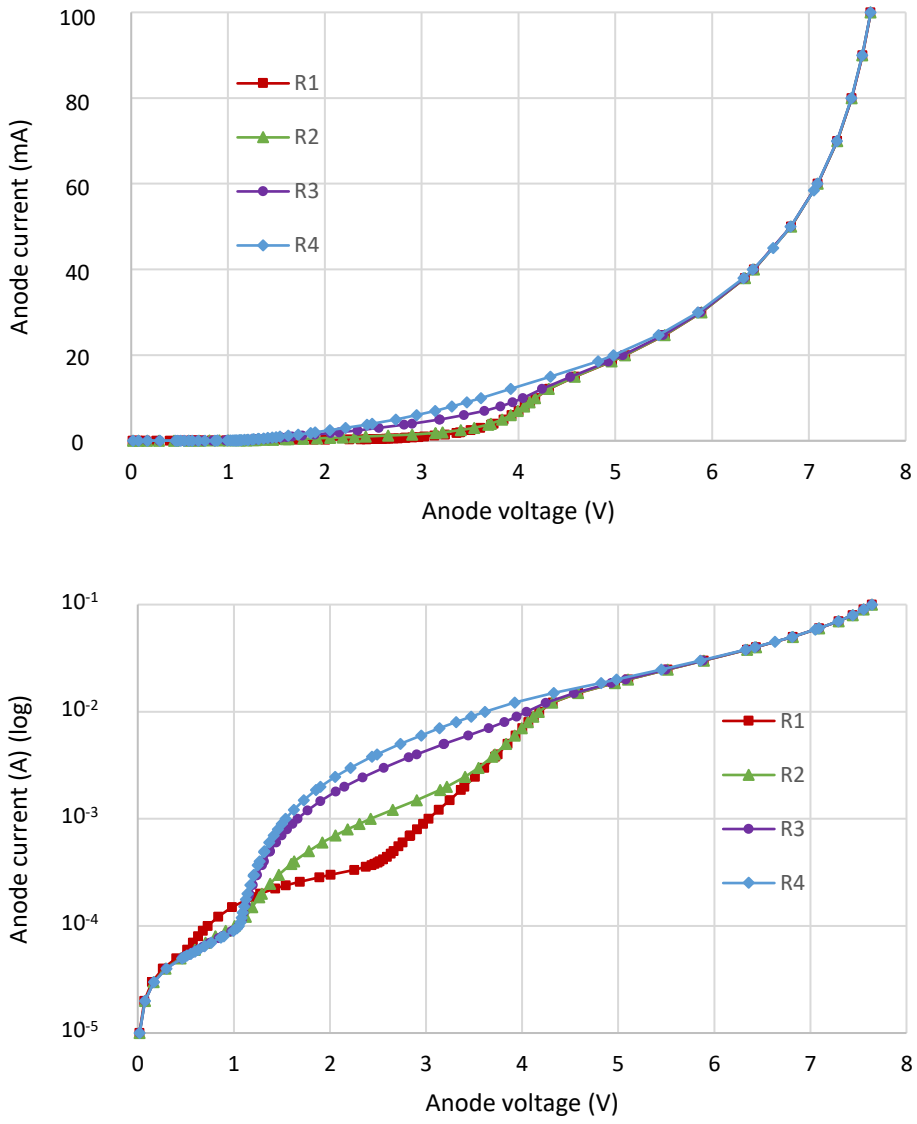


Figure 56: TCAD ACS simulations of a GDNMOS with different values of resistor ($R_1 < R_2 < R_3 < R_4$) plugged to both of the gates (devices 10 - 13). Anode current versus anode voltage is plotted. Top: the current is displayed in linear scale. Bottom: logarithmic scale.

c. Comparison between GDNMOS and GDBIMOS

The GDNMOS devices have the same V_{T1} as the GDBIMOS devices for same value of resistor plugged to both front gates (Figure 57). Due to the multi-triggering events at high voltage, it is difficult to compare experimentally the R_{ON} of these devices. Theoretically, R_{ON} is improved in the GDBIMOS with respect to the GDNMOS, especially if the value of the resistor plugged to its front gates is high. The reason is that the voltage applied on the gates does not decrease to 0 V with time like in GDNMOS (Figure 55 and Figure 58), so this positive gate voltage helps the conduction in the channel of the NMOS part of the GDBIMOS. It is the BIMOS part of the device that maintains the "permanent" positive gate voltage, because it benefits from a positive feedback loop: when there is a current of electrons flowing through the channel of the BIMOS, impact ionization near the drain produces holes that are attracted by the P^+-doped body contact. The hole current in the body contact helps the voltage to be raised on the node of the body contact and gate, through the resistor. With an increased gate voltage, the current in the channel of the BIMOS is enhanced thanks to the MOS effect. This increased current in the channel produces more holes through impact ionization, and so on. It is the parasitic capacitances that provoke the raise in gate voltage as the anode voltage is increasing, therefore turning the device ON, and it is the body contact of the BIMOS that maintains this gate voltage while the device is ON, thus improving the R_{ON}. The I-V curves from which the data are plotted in Figure 58 are shown in Figure 59. Figure 60 compares the curves from Figure 56 (simulations of GDNMOS devices) and Figure 59 (simulations of GDBIMOS devices). Unlike in Figure 57, GDBIMOS devices clearly have a lower R_{ON} than GDNMOS devices. The higher the external resistor value, the lower R_{ON} of the GDBIMOS.

It would be interesting to remove silicide from terminals - anode, drain and source – (Figure 61) to add some resistances in the device, so that fingers all trigger at the same time uniformly. Measurements will then assess if R_{ON} still looks the same in the GDNMOS and GDBIMOS devices, without multi-triggering. With N^+ doping in the drain, the GDNMOS has the behavior of a gated diode in series with a NMOS device. If R_{ON} does not change between the GDNMOS and the GDBIMOS devices, it would mean that it is the diode part that conditions R_{ON} of the device (else, the BIMOS part of the device would help the GDBIMOS to have a lower R_{ON} than in the GDNMOS).

Figure 55 and Figure 58 also provide an interesting information: when the device is ON, the anode voltage is of the order of 4 V and the gate voltage is of the order of 1 V.

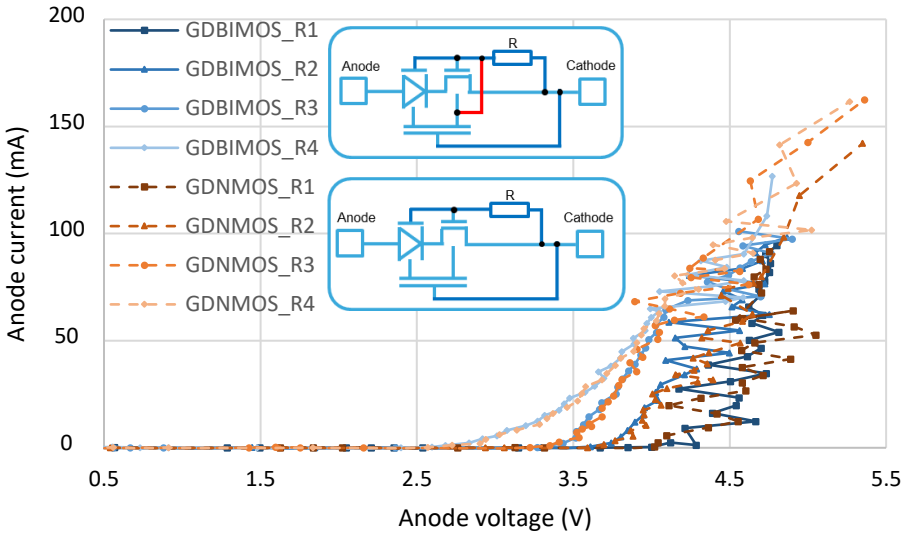


Figure 57: TLP measurements of GDNMOS (devices 10 – 13) and GDBIMOS (devices 14 - 17) with different values of resistor ($R_1 < R_2 < R_3 < R_4$) plugged to both of the gates.

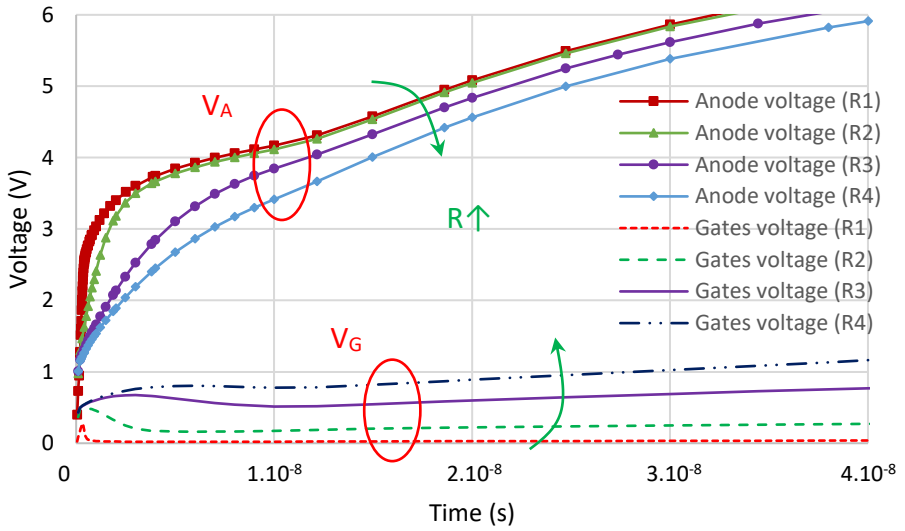


Figure 58: TCAD ACS simulations of a GDBIMOS with different values of resistor ($R_1 < R_2 < R_3 < R_4$) plugged to both of the gates (devices 14 - 17). Anode and gates voltage versus time is plotted. Note the improved R_{ON} and the higher gate voltage with an increase in R.

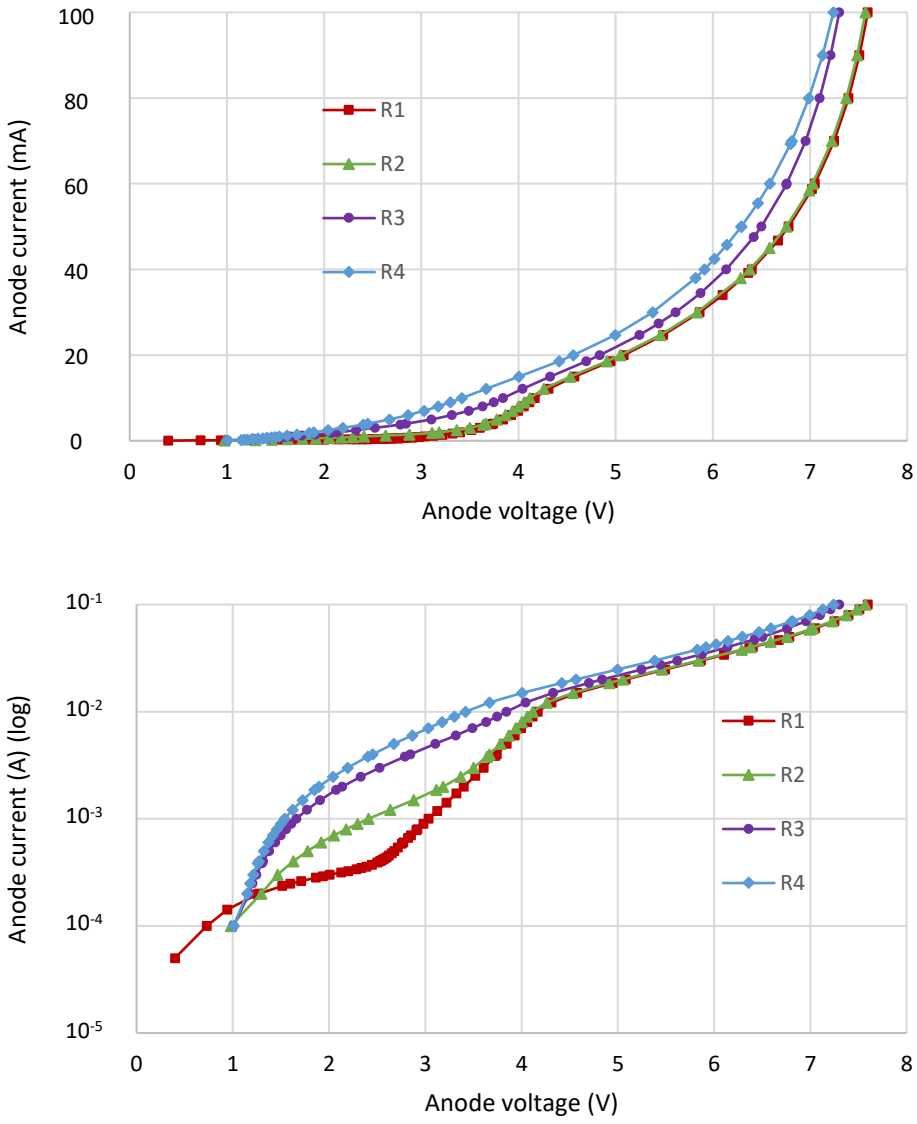


Figure 59: TCAD ACS simulations of a GDBIMOS with different values of resistor ($R_1 < R_2 < R_3 < R_4$) plugged to both of the gates (devices 14 - 17). Anode current versus anode voltage is plotted. Top: the current is displayed in linear scale. Bottom: logarithmic scale.

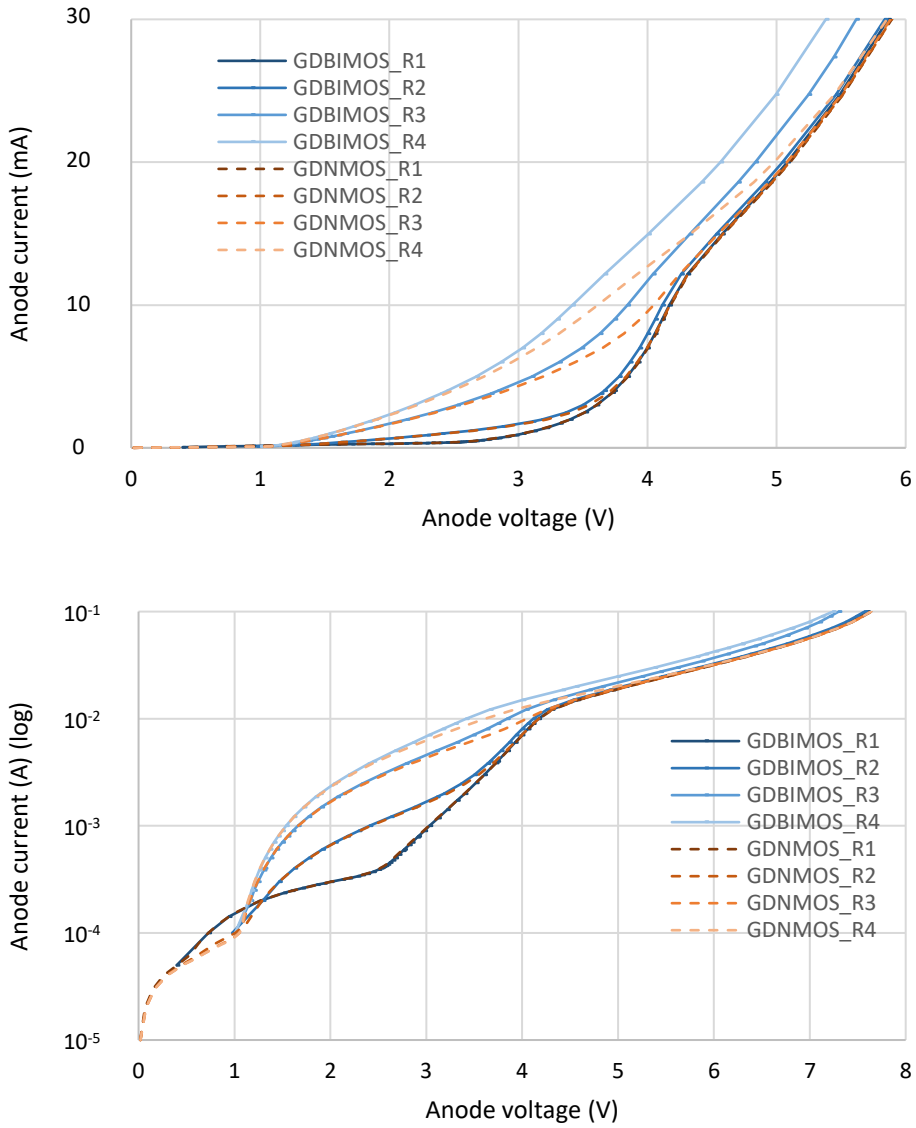


Figure 60: TCAD ACS simulations of GDNMOS (devices 10 – 13) and GDBIMOS (devices 14 - 17) with different values of resistor ($R_1 < R_2 < R_3 < R_4$) plugged to both of the gates. Top: the current is displayed in linear scale. Bottom: logarithmic scale.

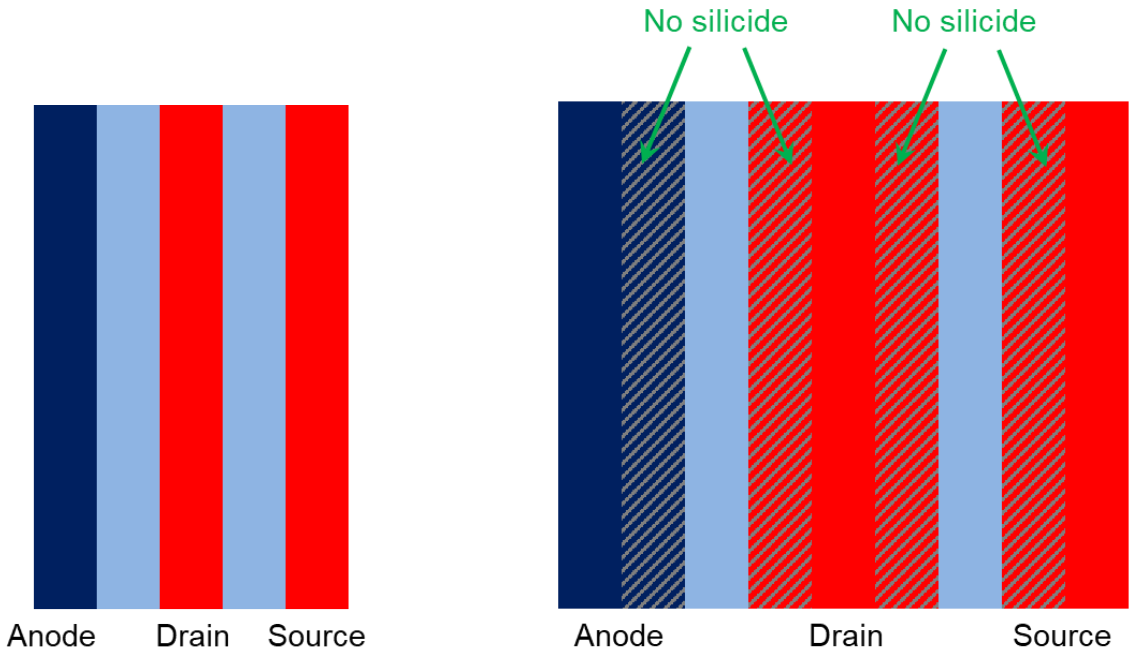


Figure 61: Top view of a GDNMOS device. Left: without silicide removal. Right: with silicide removal, to prevent multi-triggering.

d. Influence of the back gate on the GDBIMOS

The influence of the back gate on the GDBIMOS is investigated using the device 18. We found that increasing the back-plane voltage results in a lower trigger voltage (Figure 62). Therefore, the back-plane is a convenient way to modulate V_{T1} in order to match the desired ESD design window. An increase in leakage for higher positive back-plane bias is also observed (Figure 63). Since the back-plane acts as a back gate, the increase in leakage current reflects the gradual activation of the back-channel MOSFET at the film-BOX interface.

An improved solution of the design would be to connect the back gate to the front gates (that are plugged to the resistor), so that it contributes to (i) a higher current flowing in the structure in ON mode, and (ii) a smaller trigger voltage, while keeping a low leakage before triggering. This elegant solution makes the structure entirely self-biased, without relying on external voltages to control it. In other words, the structure can protect the integrated circuit even if there is no power. In our case, since the buried oxide (BOX) is much thicker than the front gates oxide, the benefits of connecting the back gate to the front gates are less relevant. Indeed, the maximum voltage that arises on the gates does not exceed 1 V (Figure 55 and Figure 58).

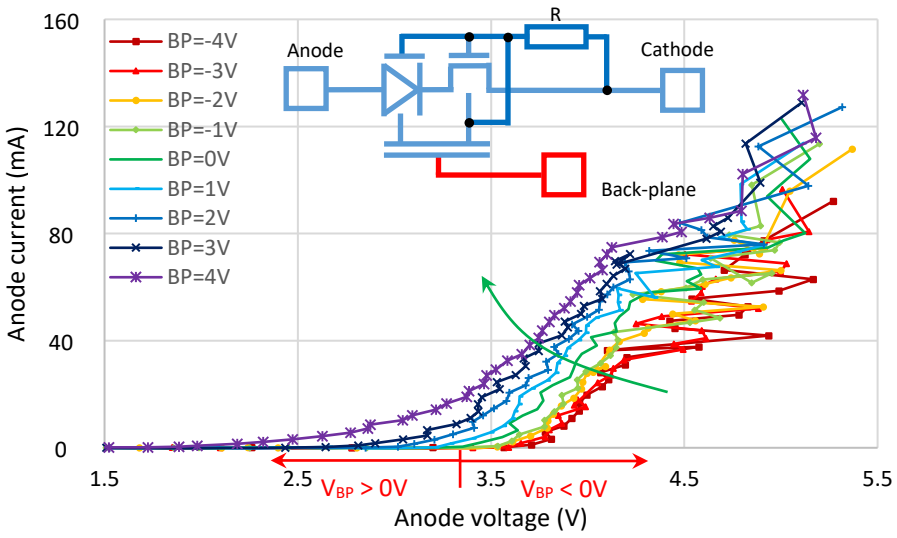


Figure 62: TLP measurements of a GDBIMOS with different back-plane biasing (device 18).

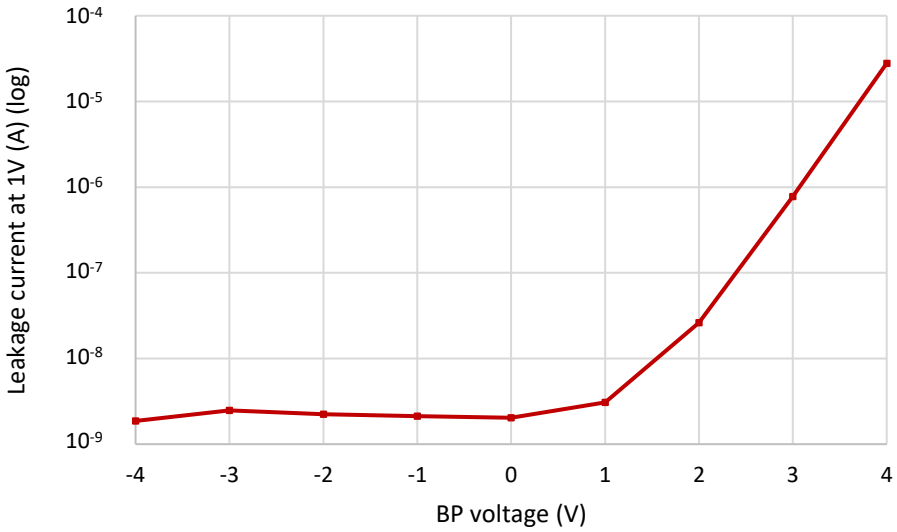


Figure 63: Measured leakage current at anode voltage V_A = 1 V versus back-plane voltage (device 18).

e. Drain connectivities

All the structures mentioned before had a floating drain (their drain was not contacted with tungsten contacts and linked to metallic via). In this section, different devices using the drain contacts will be explored. Devices 19 and 20 are GDNMOS structures; in device 19 the drain is connected to the gate of the diode, while in device 20 it is connected to the anode. Device 21 is a GDBIMOS, with the drain connected to one of the body contacts. For comparison purpose, device 22 is a regular NMOS transistor (Figure 64).

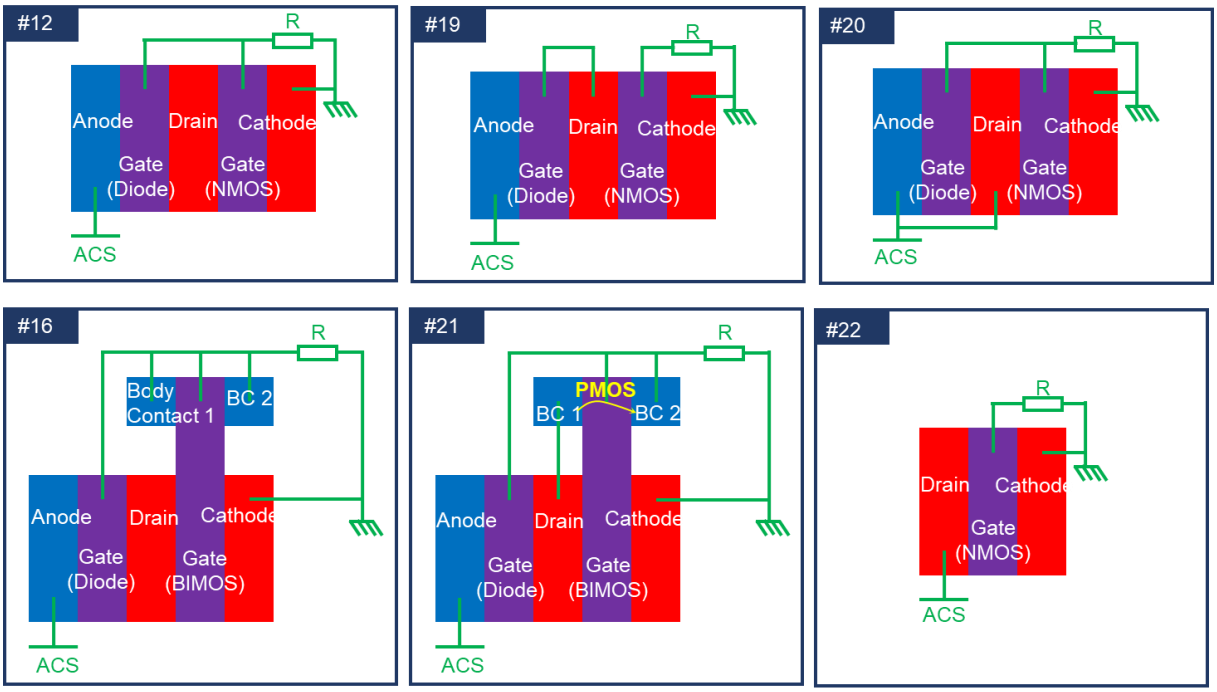


Figure 64: Topology of structures 12, 16 and 19 – 22.

Devices 1 to 17 all have a low leakage current around 2 nA at 1 V. However, the idea of further lowering the leakage is attractive, since the GDNMOS and GDBIMOS structures with high doping in the drain could potentially be used for low-power applications, provided that their trigger voltage is controlled by a trigger circuit situated – for example – on the back gate.

The goal of connecting the drain to the gate of the diode (device 19) is to enhance the barrier for the holes between the anode and the channel by increasing the voltage on the gate of the diode. Indeed, above 0.6 V between the anode and the drain, the diode starts conducting more current. When the diode is conductive, the voltage is raised on both the drain and the gate of the diode, thus making the diode less conductive. An equilibrium is reached in this loop with the self-biased diode gate. As expected, the leakage in the device 19 (70 pA) is reduced with respect to the devices 12 and 16 (2 nA), but its trigger voltage is more suitable for high voltage protection (Figure 65 and Figure 66).

Connecting the drain to the anode (device 20) is supposed to decrease the leakage current in the low-doped drain GDNMOS, but not in the high-doped drain GDNMOS. However, it is still interesting to investigate this connectivity on a high-doped drain structure, because it gives new options in terms of ESD design window. In fact, with high doping in the drain region, we no longer observe a SCR-like behavior. The device operates just like a diode in series with a NMOS. The high-doped drain GDNMOS triggers before a regular MOSFET only, but has a higher holding voltage and a higher leakage current. It triggers before the MOSFET because it benefits from the parasitic capacitance of the diode gate in addition to the one of the NMOS gate. Its holding voltage is higher than in MOSFET due to the diode gate which induces a barrier opposing the carriers to flow in the device. In this new configuration (device 20), the trigger voltage in the TLP curve is lowered with respect to devices 12 and 22, since both parasitic capacitances C_{AGD} and C_{DGM} are active and most of the surge can flow through the easiest path (*i.e.*, in the NMOS part of the device).

The holding voltage is low and the leakage is the same in comparison with the device 12 (Figure 65 and Figure 66).

We expect the same behavior for GDBIMOS structures by connecting the drain like in devices 19 or 20. The device 21 is a GDBIMOS where the body contact is connected to the drain. The trigger voltage is reduced and the leakage current is increased at V_{DD} = 1 V with respect to the device 16. As soon as the diode is conducting, the current flows in the first body contact of the BIMOS through the connection of the drain. The parasitic PMOS transistor constituted by the two body contacts of the BIMOS and its channel is turned ON very quickly, and therefore the gate of the BIMOS starts to be biased. This turns the whole BIMOS ON, and then the whole GDBIMOS becomes conductive. In the classical configuration (device 16), the GDBIMOS is turned ON dynamically mainly thanks to its parasitic capacitances, whereas in static mode, it is activated thanks to impact ionization and band-to-band tunneling. The parasitic PMOS is much faster than the RC circuit built by the parasitic capacitances, the impact ionization and band-to-band tunneling processes. That is why the device 21 is activated earlier, both in TLP and in DC measurements. Its leakage current at V_{DD} = 1 V is quite high - 30 nA – (if we want to use the protection for a V_{DD} = 1 V technology) but if the protection is designed for low-power applications (with $V_{DD} \leq 0.8$ V) the leakage becomes acceptable ($I_{leak} < 2$ nA). In fact, the DC curves of the devices 16 and 21 start to diverge after 0.6 V, which is the voltage for which the diode starts to conduct, and then the BIMOS is activated by a different mean in each device (Figure 65 and Figure 66).

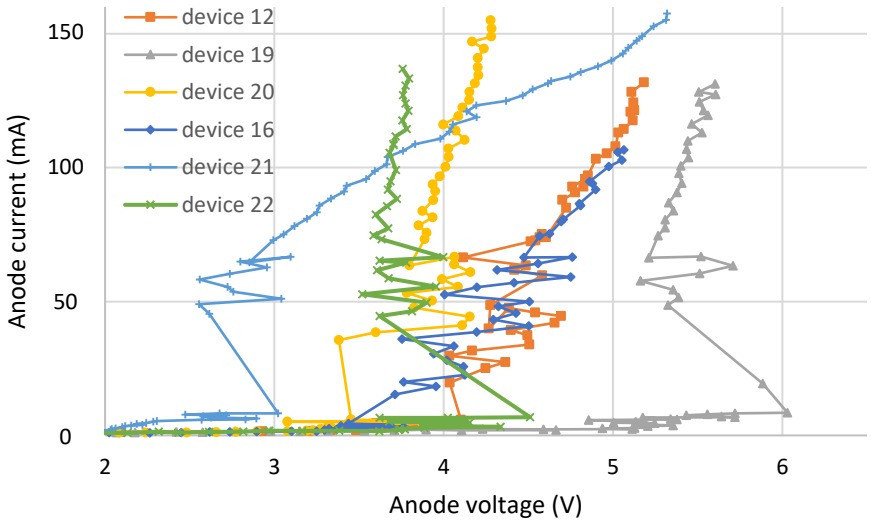


Figure 65: TLP measurements of GDNMOS (devices 12, 19 and 20), GDBIMOS (devices 16 and 21) and NMOS (device 22), that have different drain connectivities. Note that the TLP responses of devices 12 and 16 are not exactly the same as in Figure 53 and Figure 57, due to wafer-to-wafer variability.

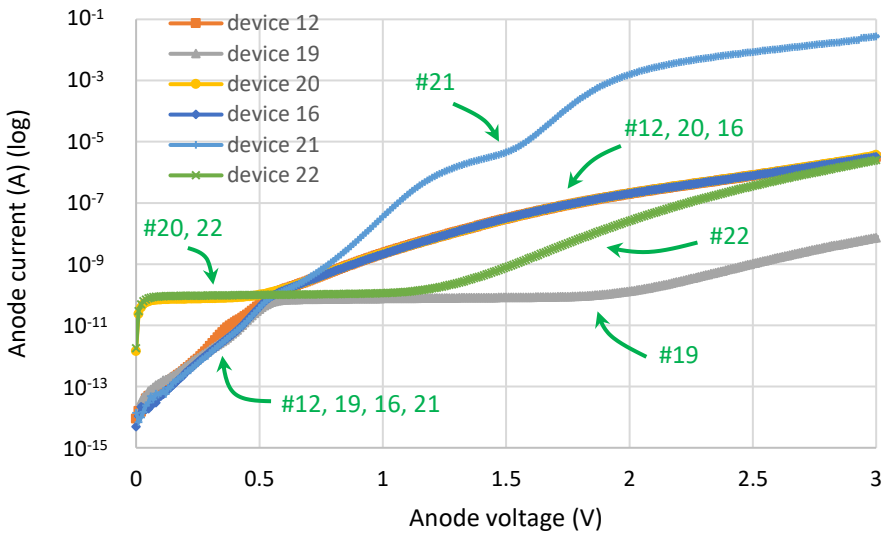


Figure 66: DC measurements: AVS of GDNMOS (devices 12, 19 and 20), GDBIMOS (devices 16 and 21) and NMOS (device 22) with different drain connectivities.

2. GDxMOS as a low-voltage protection

Theoretically the GDNMOS device could be considered also as a SCR since it is a P/N/P/N structure (Figure 67). Nevertheless, the too high doping level in the drain (N^+) of the NMOS prevents the SCR-like behavior to be observed. The operating mechanism of GDxMOS devices that feature a N^+-doped common drain is not relying on the parasitic bipolar transistors in the thin-film. Indeed, a too high doping level in the drain of the NMOS degrades the emitter efficiency of the PNP parasitic bipolar transistor composed of the diode and the channel of the NMOS. In order to obtain a SCR behavior, the emitter efficiency of each bipolar transistor has to be enhanced, either by decreasing the length of the base or by lowering the acceptor doping in the base, according to [109]:

$$\gamma_E = \frac{D_{nB} \cdot L_{pE} \cdot n_{iB}^2 \cdot N_{DE}}{D_{nB} \cdot L_{pE} \cdot n_{iB}^2 \cdot N_{DE} + D_{pE} \cdot W_B \cdot n_{iE}^2 \cdot N_{AB}}$$

The common-base current gain of a NPN bipolar transistor increases with the emitter efficiency γ_E. D represents the diffusion coefficient; L is the diffusion length of carriers (n for electrons and p for holes) and n_i is the intrinsic carrier density. N_D corresponds to the donor concentration and N_A to the acceptor concentration. W is the width. All those values are for the base (index B), emitter (E) or collector (C). In this study, a modification of the original GDNMOS structure is proposed in order to benefit from the SCR-like behavior.

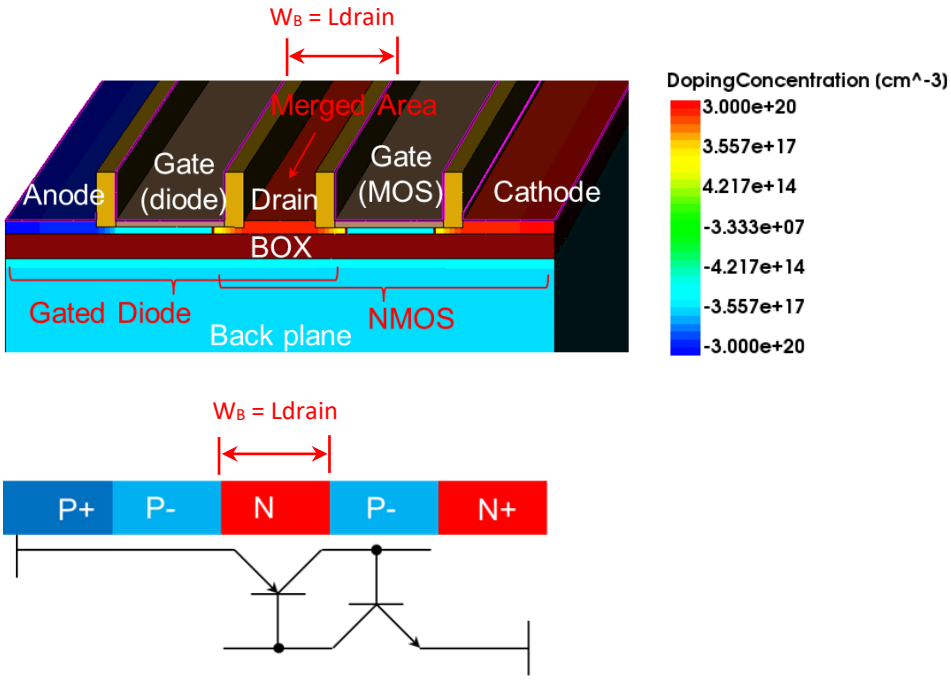


Figure 67: TCAD view (top) and corresponding thin-film schematic (bottom) of a GDNMOS structure.

a. Low-doped drain GDNMOS

The GDNMOS was implemented for protecting 'high' voltage chips designed in 28 nm FD-SOI technology with a 4 V breakdown voltage. As a consequence, this GDNMOS could not protect the 1 V operating core 28 nm UTBB FD-SOI devices efficiently. Indeed, their breakdown voltage is about 3 V [48]. It is due to the high doping level in the drain of the NMOS. We propose to investigate the operation of a GDNMOS version with lower doping level (LDD doping instead of N^+ doping) for the drain of the NMOS. This modification does not increase the cost of the device since the number of steps required for the process flow is decreased with SD implantation omitted.

We perform 3D TCAD simulations considering the classical semiconductor equations and taking into account the thermal effects. The initial condition for temperature is 300 K. The surge is an ACS stress. The simulations are plotted until the temperature of 800 K is reached. All the structures are 10 µm wide for current comparison. The structures are designed with minimum dimensions according to the design rules.

The solutions with N^+ doping and solely LDD doping in the drain of the NMOS are compared in terms of I-V curves (Figure 68), and the effect of changing the drain length is shown in Figure 69. Both the holding voltage and the trigger voltage are lowered for LDD doping (with respect to N^+ doping) and when the length of the drain is reduced, as expected.

Note that the studies of drain doping and drain length (Figure 68 and Figure 69) are done on GDNMOS devices that have both of their gates grounded. It has been shown in the "GDxMOS as a high voltage protection" part of the manuscript, that GDNMOS devices with N^+ doping in the drain and with both grounded gates are experiencing an early break, since the diode gate is not protected by an external resistor. So why studying the LDD doped GDNMOS with grounded gates (shouldn't they also have a robustness issue)? In fact, the N^+ doped GDNMOS with grounded gates are failing around 4 V of anode voltage (because of the difference between the anode and the diode gate voltages). With LDD doping, we can expect the devices to trigger earlier than 4 V. If the anode voltage never reaches 4 V, then the difference between the anode and diode gate voltage will be lower and it is possible that the LDD devices do not have an early break issue. This has to be confirmed by measurements.

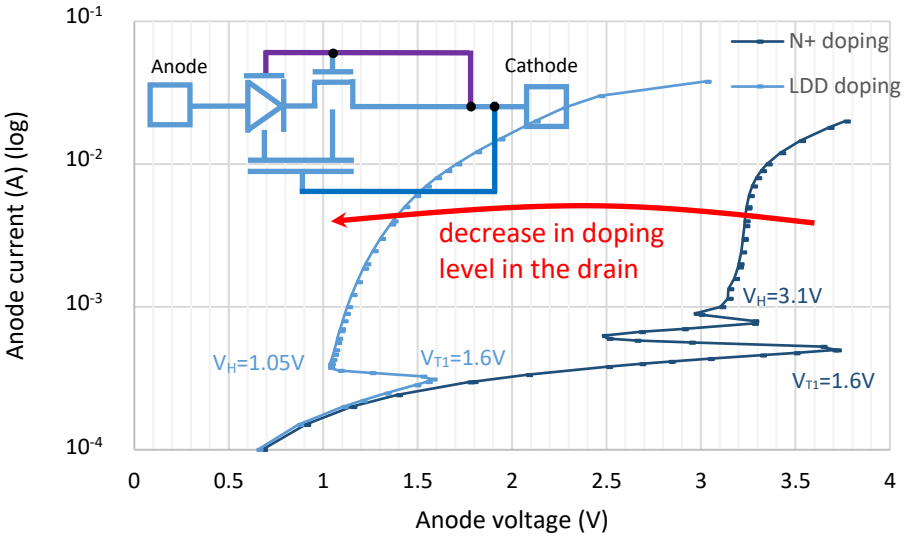


Figure 68: Electro-thermal ACS I-V TCAD simulation of GDNMOS devices with N^+ or N-LDD drain. Both gates are grounded.

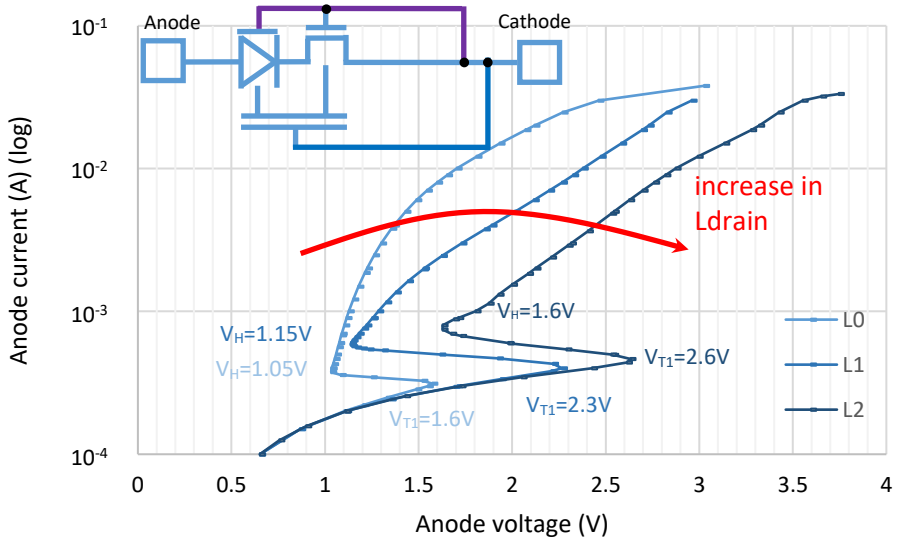


Figure 69: Electro-thermal ACS I-V TCAD simulation of GDNMOS devices with different drain lengths (L_0 = 114 nm, L_1 = 250 nm, L_2 = 350 nm). Both gates are grounded.

To further decrease the V_{T1} of the structure, a resistor can be plugged to the gates.

In the case where the gate of the NMOS is connected to the resistor, during a transient stress the structure is triggered on thanks to the parasitic capacitance between the drain and the gate of the NMOS. The anode voltage raises the drain voltage which increases the gate of the NMOS voltage through parasitic capacitance. With a small voltage on the gate, electrons are favored to flow in the NMOS, thus to turn on the first bipolar transistor of the SCR. As a consequence of the SCR becoming active, the current is spread along the width of the structure. Hence the value of the resistor is important for helping the voltage on the gate of the NMOS to be controlled by the parasitic capacitor more easily at first (Figure 70).

In the case where the gate of the diode is plugged to the resistor, I_{T1} decreases with an increase in the resistance value because the structure is more resistive before triggering. It is due to the parasitic capacitance between the anode and the gate of the diode. This capacitance allows the voltage on the gate of the diode to increase slightly due to capacitive coupling: in the channel of the diode the electrostatic doping tends toward N doping instead of P^-, and this does not help the carriers to flow from anode to channel of the diode. Therefore, the resistance of the structure increases. V_{T1} and V_H are also increased for the same reason; for small values of resistor, the changes are minor (Figure 71).

If the gates of the diode and NMOS are tied together and connected to a resistor, V_{T1} and I_{T1} decrease with an increase in external resistor value (Figure 72). Indeed, the parasitic capacitances of both gates are playing a role, and they help each other: an increase in voltage for one of the gates means that the other gate voltage increases also. As a consequence, the GDNMOS with both gates connected to a resistor benefits from the decrease in I_{T1} (the device is more resistive before triggering because of the biased gate of the diode) and from the decrease in V_{T1} (due to the NMOS gate being biased; the SCR is then activated thanks to the NPN bipolar transistor).

To summarize: if a resistor is connected to the NMOS gate (compared to both gates being grounded), V_{T1} decreases (Figure 70 and Figure 73). If in addition to connect the NMOS gate to a resistor, the gate of the diode is also connected to this resistor, the GDNMOS becomes more resistive before triggering, which induces a strong decrease in I_{T1}; V_{T1} is also decreased further because the parasitic capacitances of the gated diode also help the NMOS gate to be biased.

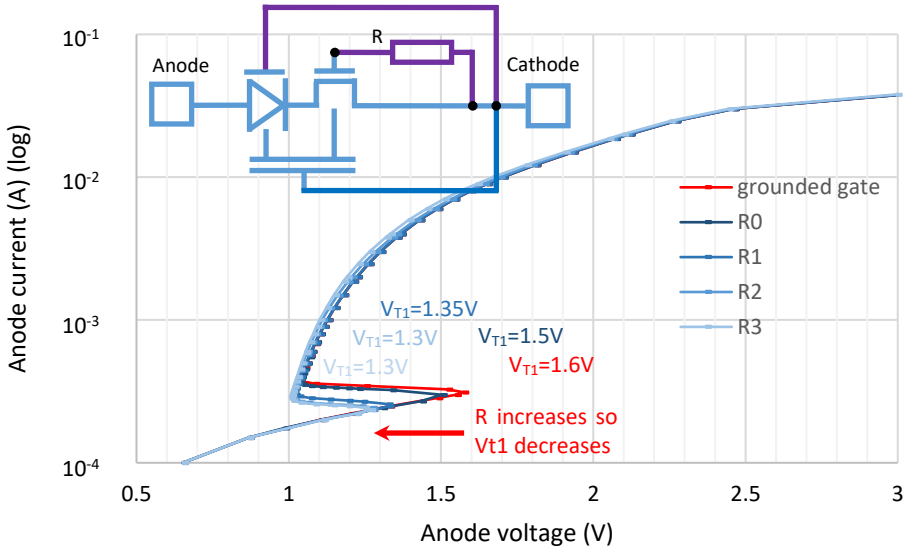


Figure 70: Electro-thermal ACS I-V TCAD simulation of GDNMOS devices with different values of resistor ($R_0 < R_1 < R_2 < R_3$) plugged to the gate of the NMOS. The diode gate is grounded.

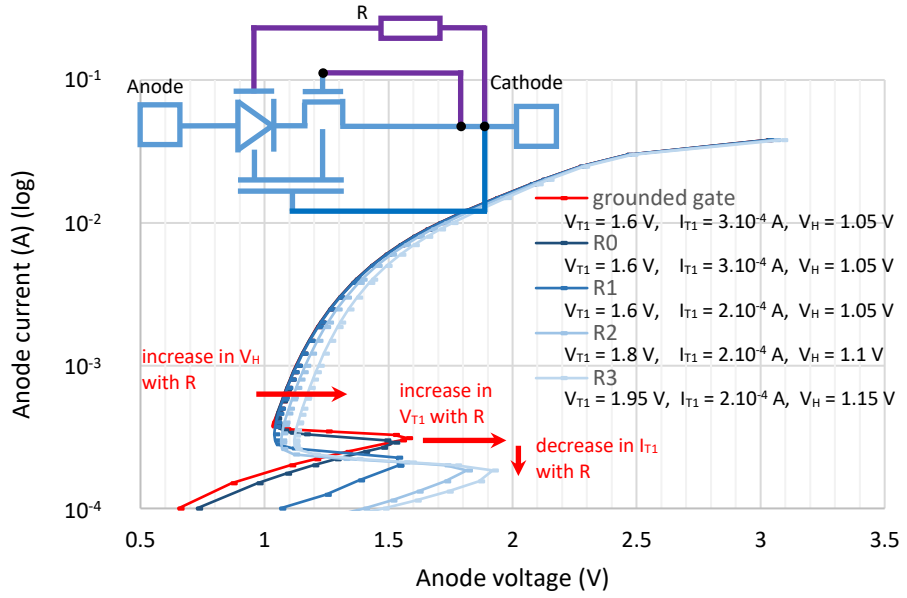


Figure 71: Electro-thermal ACS I-V TCAD simulation of GDNMOS devices with different values of resistor ($R_0 < R_1 < R_2 < R_3$) plugged to the gate of the diode. The NMOS gate is grounded.

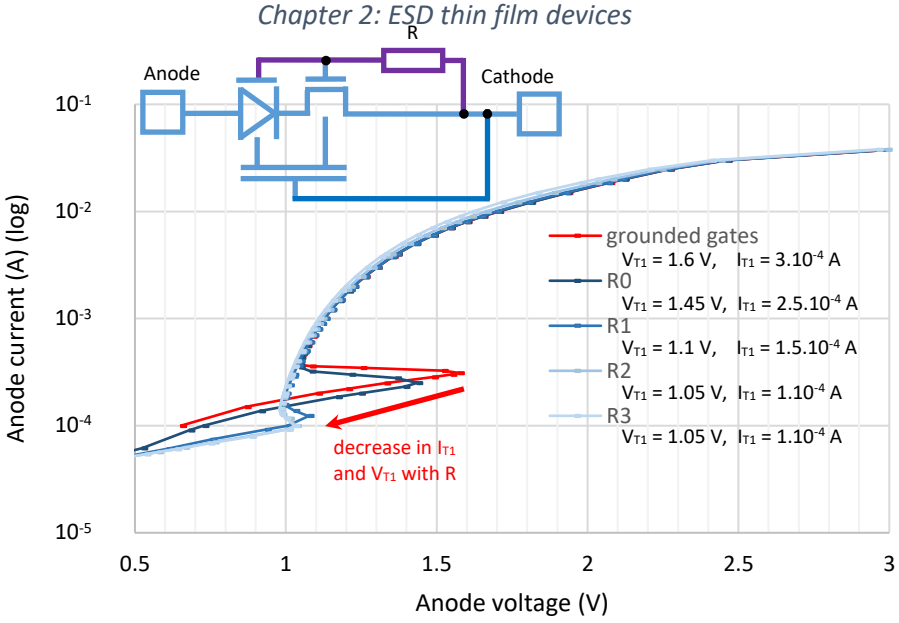


Figure 72: Electro-thermal ACS I-V TCAD simulation of GDNMOS devices with different values of resistor ($R_0 < R_1 < R_2 < R_3$) plugged to the gate of the diode and of the NMOS.

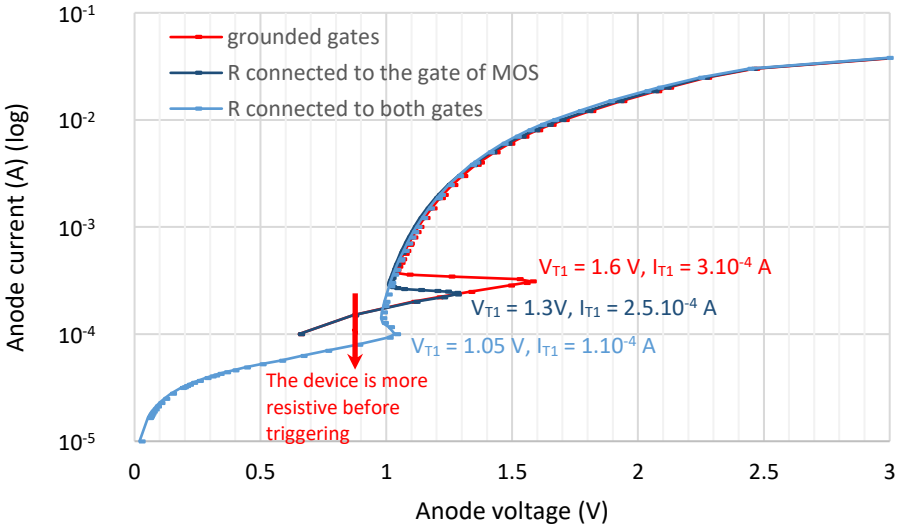


Figure 73: Electro-thermal ACS I-V TCAD simulation of GDNMOS devices with different connections of the gates: GDNMOS with grounded gates, GDNMOS with the gate of the diode grounded and the gate of the NMOS connected to the resistor R_2, and GDNMOS with both gates tied together and connected to the resistor R_2.

b. Low-doped drain GDBIMOS

As a reminder a BIMOS is a N-type MOS transistor on which a P^+ body contact is placed in order to have access to the channel of the NMOS, which is supposedly the base of a NPN parasitic Lateral Bipolar Junction Transistor (LBJT). For ESD applications this contact and the gate are both plugged to the same resistor.

The principle of combining the GDNMOS with a BIMOS is used to improve R_{ON} of the GDNMOS thanks to the BIMOS. Indeed, besides having the advantage of gates connected to a resistor, thus a reduced V_{T1}, the contacts of the BIMOS attract holes once it is turned on, therefore a voltage is applied on the gate all along the ESD and not only when the device triggers. The layout is very compact since the NMOS of the GDNMOS is easily transformed into a BIMOS (Figure 45).

3D TCAD simulation results show that the GDBIMOS designed with a low drain doping is effective to protect devices in a reduced ESD window with respect to the GDNMOS (Figure 74). In this comparison, both gates are connected to the resistor of the BIMOS. The higher the value of the resistor, the lower V_{T1}, I_{T1} (like in the GDNMOS) and R_{ON} (Figure 75). Those results have to be confirmed by measurements. Indeed, in section about the GDBIMOS with a highly doped drain, multi-triggering was preventing us to observe experimentally the body contact effect. However, this multi-triggering was only occurring at high voltage, so we can expect to observe the body contact effect with the LDD doping on the drain.

Also, in the case of LDD doping in the drain, the GDNMOS no longer operates as a gated diode in series with a NMOS, but as a SCR. Therefore, there is no more competition between R_{ON} of the diode part and R_{ON} of the NMOS part of the device; R_{ON} will all be due to the SCR. In this case, it will be interesting to compare the measured R_{ON} of a GDNMOS and GDBIMOS devices that have LDD doping.

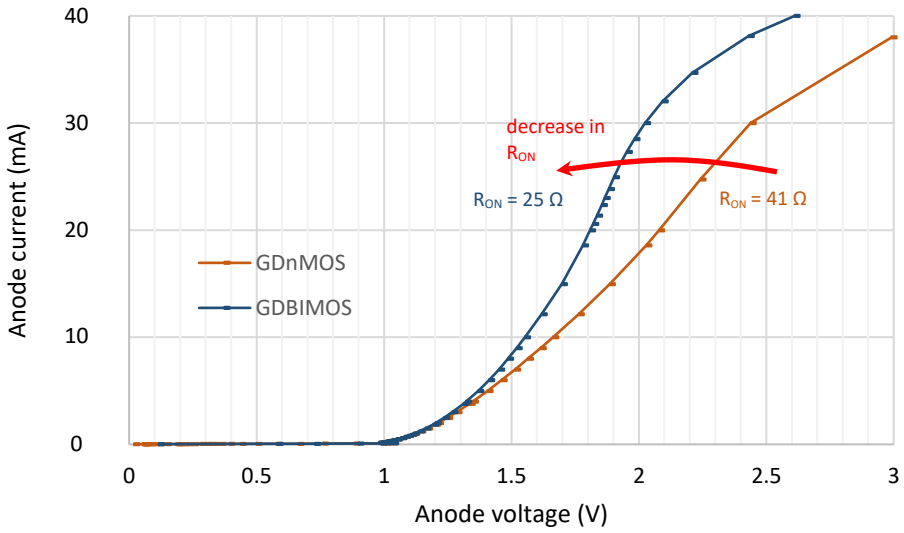


Figure 74: Electro-thermal ACS I-V TCAD characteristics of a GDNMOS and a GDBIMOS. Both devices have their gates connected to the resistor R_2 and N-LDD doping.

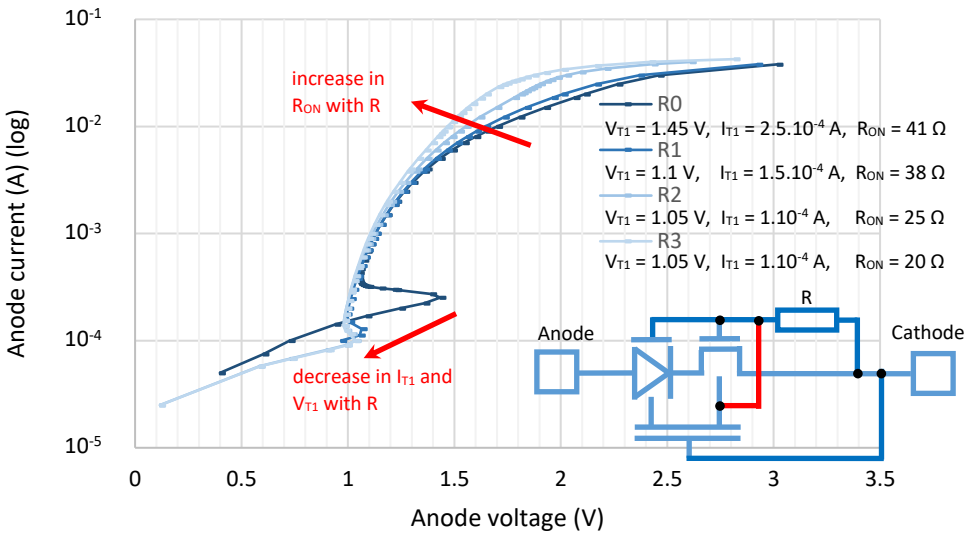


Figure 75: Electro-thermal ACS I-V TCAD simulation of a GDBIMOS with both gates connected to a resistor R of different values ($R_0 < R_1 < R_2 < R_3$).

Considering how the structure turns on, plugging the common back gate to the BIMOS gate makes no major difference. Indeed, the oxide of the back-plane is too thick to compete with the BIMOS and the SCR in this voltage range.

When the structure is about to trigger (at I=1.10^{-4} on Figure 76), the hot spot is situated at the junction between the body contacts and the channel for the GDBIMOS, and for the GDNMOS it is situated at the junction between the drain and the channel of the NMOS.

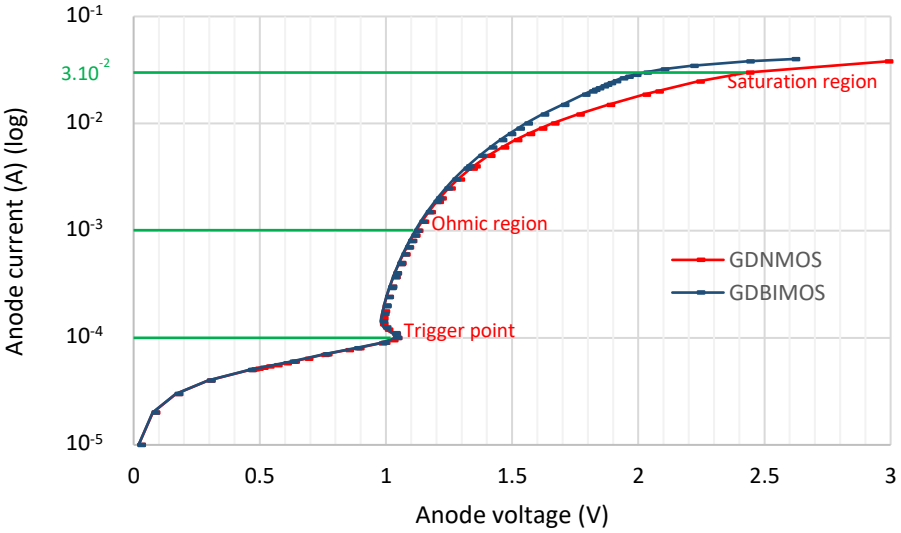


Figure 76: Electro-thermal ACS I-V TCAD simulation of a GDNMOS and a GDBIMOS with both gates connected to the resistor R_2. Those are the same curves than the ones in the Figure 74, except that the current is displayed with a logarithmic scale.

When the structure has triggered (at I = 1.10^{-3} A on Figure 76), the temperature profile is the same in the GDBIMOS as in the GDNMOS. The hot spot is situated at the junction between the drain and the channel of the NMOS or BIMOS (Figure 77).

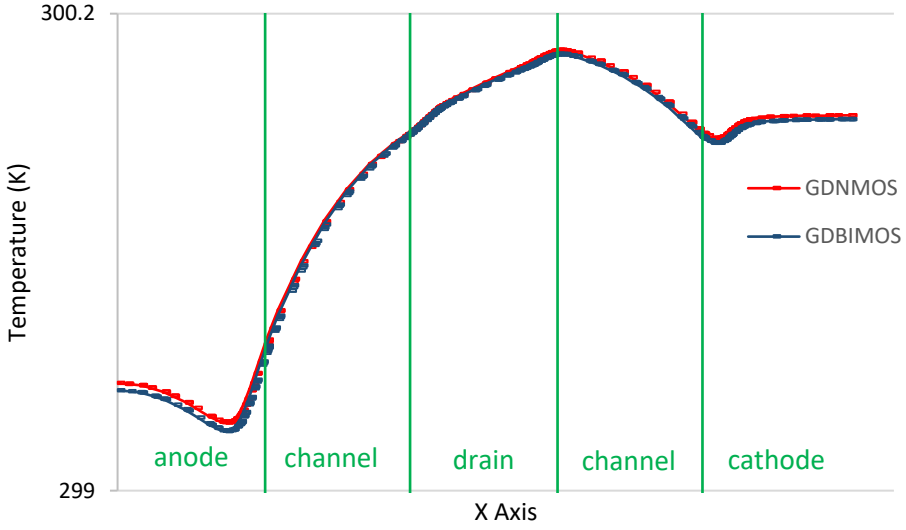


Figure 77: Temperature extraction at 1 ns during ACS stress in thermodynamic simulation in GDNMOS and GDBIMOS devices with both gates connected to the resistor R_2.

At I = 3.10^{-2} A on Figure 76, the hot spot is situated at the interface between the drain and the channel of the NMOS for the GDNMOS, and for the GDBIMOS it is situated in the middle of the channel of the BIMOS (Figure 78). The thermal dissipation is less effective in FD-SOI than in bulk. However, it is shown in the study that it is possible to reduce the temperature in the device for a same current. Thanks to the addition of a BIMOS in the device, the gates of the GDBIMOS are active when the current flows massively through the device, therefore the device is less resistive (Figure 74) and the hot spot of the structure has a lower temperature (Figure 78). Hence, the GDBIMOS is more robust in temperature than the GDNMOS.

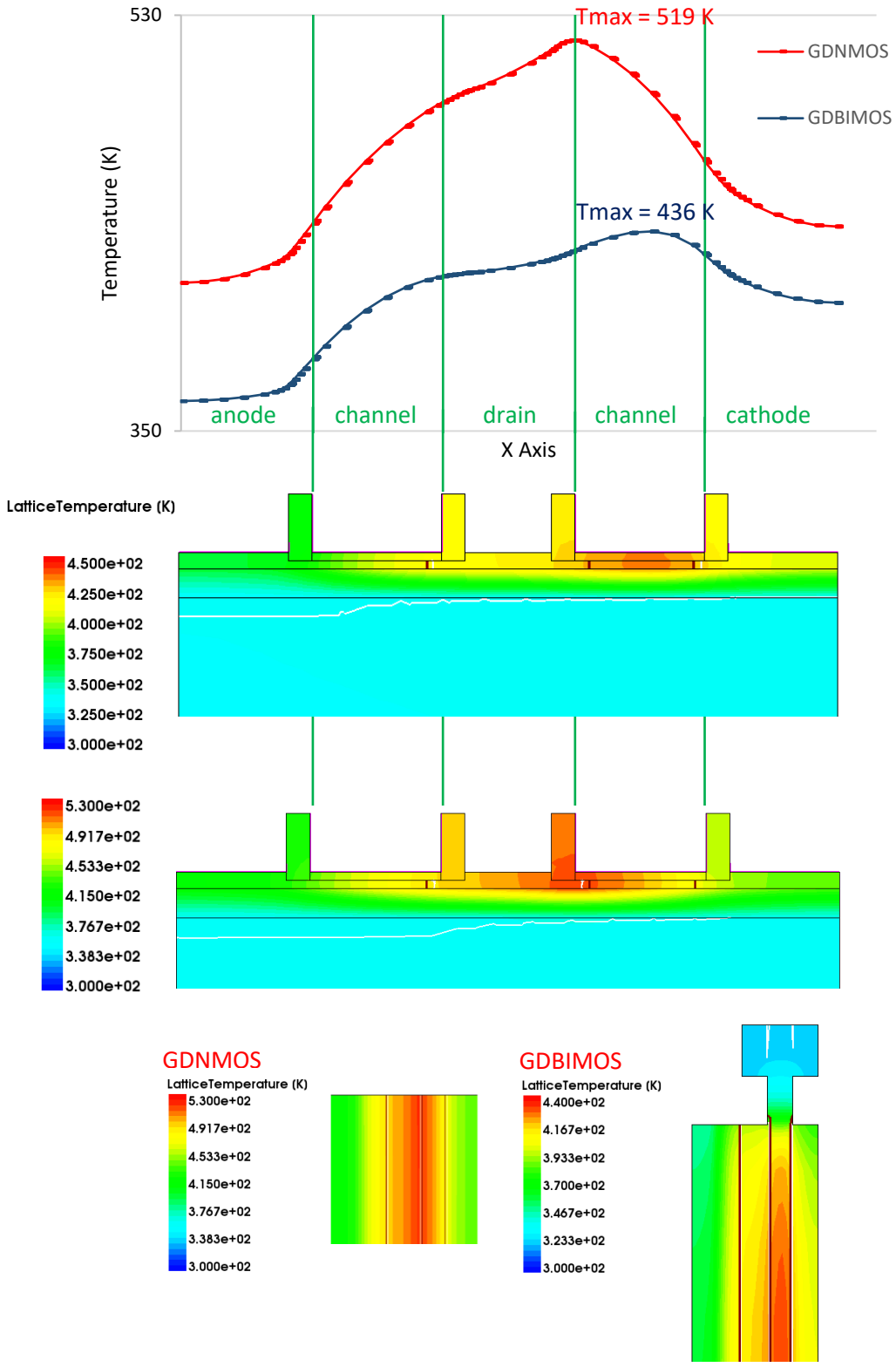


Figure 78: Temperature extraction at 30 ns during ACS stress in thermodynamic simulation in GDNMOS and GDBIMOS devices with both gates connected to the resistor R_2.

Having all the gates plugged to a resistor is also effective and advantageous in case of reverse ACS (negative ESD surge). Indeed, a grounded gate device is not robust against negative ESDs. The resistor provides additional protection to the device in case of overvoltage on the gates.

According to Figure 79, a reverse diode is requested to be compliant with the ESD window in case of reverse ACS, since it will clamp the voltage over 1 V.

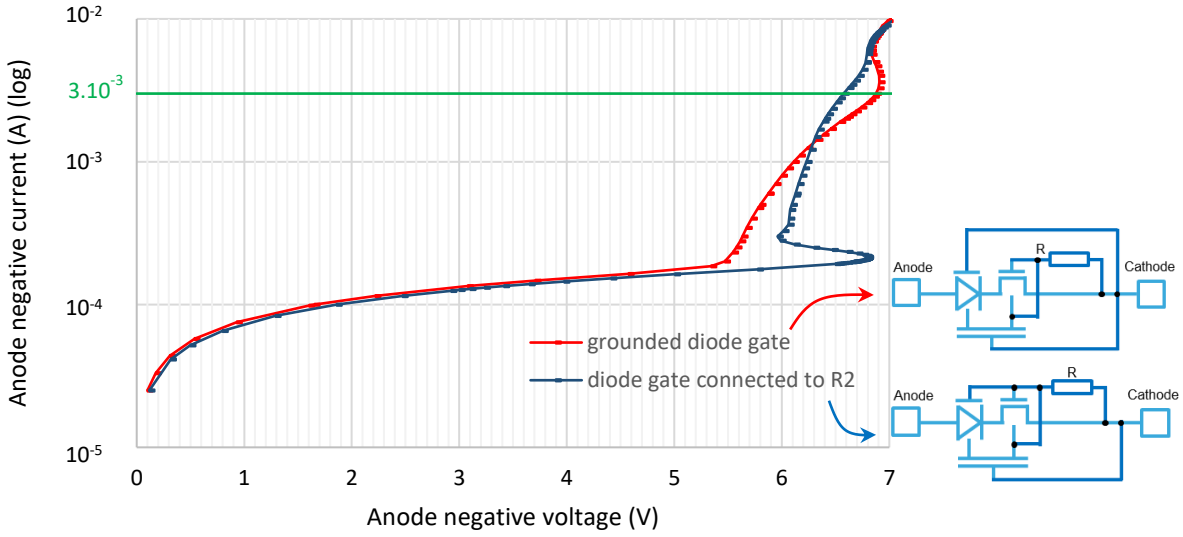


Figure 79: Electro-thermal reverse ACS I-V TCAD simulation of a GDBIMOS with both gates plugged to the resistor R_2, and with a GDBIMOS with only the BIMOS gate plugged to the resistor R_2 and the diode gate grounded.

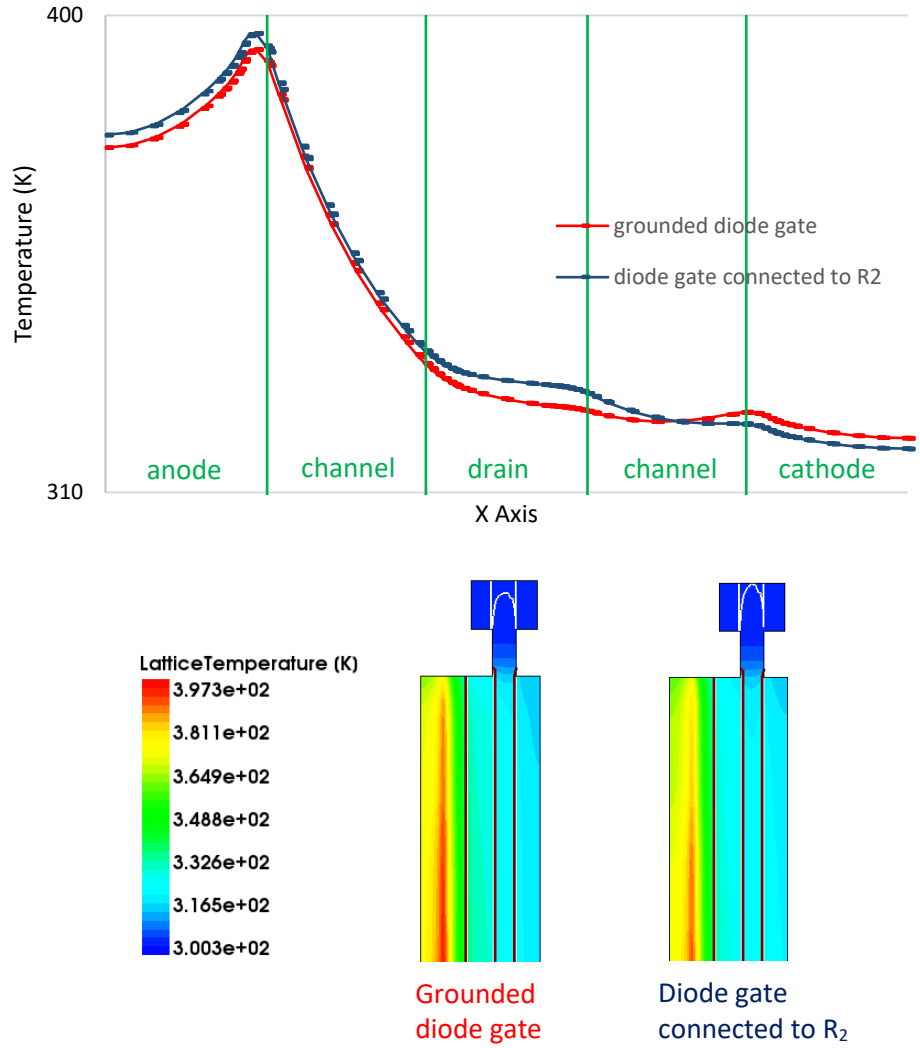


Figure 80: Temperature extraction during ACS stress in thermodynamic simulation for an anode negative current of I = 3.10^{-3} A (see Figure 79) in a GDBIMOS with both gates plugged to the resistor R_2, and in a GDBIMOS with only the BIMOS gate plugged to the resistor R_2 and the diode gate grounded.

For current leakage concern, both GDNMOS and GDBIMOS devices have huge leakage at 1 V if there is N-LDD doping in the drain instead of N^+ doping. However, at 0.6 V the structures feature a low leakage (8.10^{-10} A/µm) (Figure 81). Therefore, it is a good protection for low-voltage applications (≤ 0.6 V).

To be compliant with a 1 V application in terms of leakage, a solution could be to use two protection devices in serial mode or to change the terminals' connection (for example, plugging a resistor between the drain and the anode or between the drain and the gate of the diode), at the cost of impacting ACS characteristics.

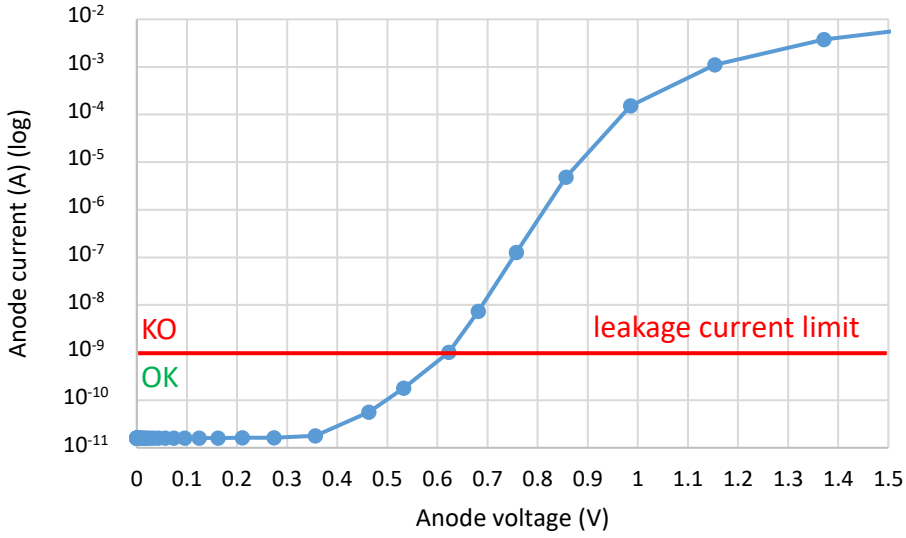


Figure 81: AVS I-V TCAD simulation of a GDBIMOS with both gates plugged to the resistor R_2 at room temperature.

3. Silicide management in the GDxMOS

In the previous section, we first presented the GDxMOS as a high voltage protection (4 V) with N^+ drain doping, then we showed simulations of the GDxMOS as a very low-voltage protection (0.6 V) with N-LDD doping. The goal of this section is to explore the GDxMOS device as a low-voltage protection (1 V) with N-LDD doping and a connected drain, thanks to the silicide management.

Having N-LDD doping in the drain is realized by putting a mask that prevents N^+ doping to be added (NOSD mask), but another mask also has to be added, the one that prevents the salicidation of the drain (SBLK mask). Normally ESD protections are in the bulk, therefore it makes no difference to add silicide or not. Indeed, the conduction of current is in the volume and silicide is only on the surface. But in the thin-film this is not the case anymore, because the silicide takes almost all the thickness of the thin-film. Silicide acts as if the silicon was highly doped, so removing the silicide is capital for getting a low-voltage SCR.

In the end of the previous section, it was explained that the GDxMOS with N-LDD doping (and no silicide) had a high leakage current, therefore being useful for very low-voltage ESD protection (0.6 V). A possible solution to be compliant with a low-voltage technology (1 V) would be to connect also its drain. However, a metallic connection on a region with no silicide is problematic and leads to a very poor yield in the factory. Indeed, there is a risk that the tungsten CA contact is transpiercing all the silicon, and leads to Schottky contact behavior instead of ohmic behavior. A possibility would be to increase the thickness of the raised epitaxy of silicon so that the contact is not transpiercing, but it implies to change the process (high cost, not guaranteed yield...). A better solution is a “partial silicide” (along the width of the device), *i.e.* the idea is to put a small part of silicide somewhere in the width of the drain so that the conduction is mainly within the part of the drain without silicide, and the drain is contacted on the silicide part (Figure 82).

The partial silicide can also be used to shift the V_{T1} and leakage current of the structure (by being wider or smaller), in order to obtain a set of protections for any ESD design window.

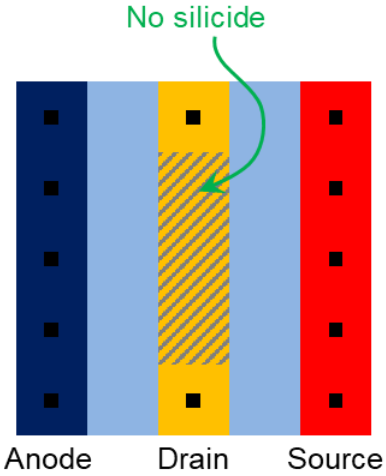


Figure 82: Top view of a GDNMOS device with partial silicide.

a. Silicide removal

As it can be seen in Figure 83, the trigger voltage of the GDxMOS is substantially lowered in the case of N-LDD doping and no silicide on the drain. In fact, if the drain doping is lowered without removing the silicide in the drain region, carriers are recombined due to the presence of the silicide (Figure 84). This happens because of the thin thickness of the thin-film silicon: only a small hole current is able to flow to the channel of the NMOS (Figure 85). Therefore, it is not possible to obtain a SCR, which basically relies on the two bipolar transistors: the PNP transistor where holes are flowing from the anode to the channel of the NMOS, and the NPN transistor where electrons are flowing from the source to the drain. In case of recombination in the drain, the PNP transistor is impacted (anode-drain-channel of the NMOS).

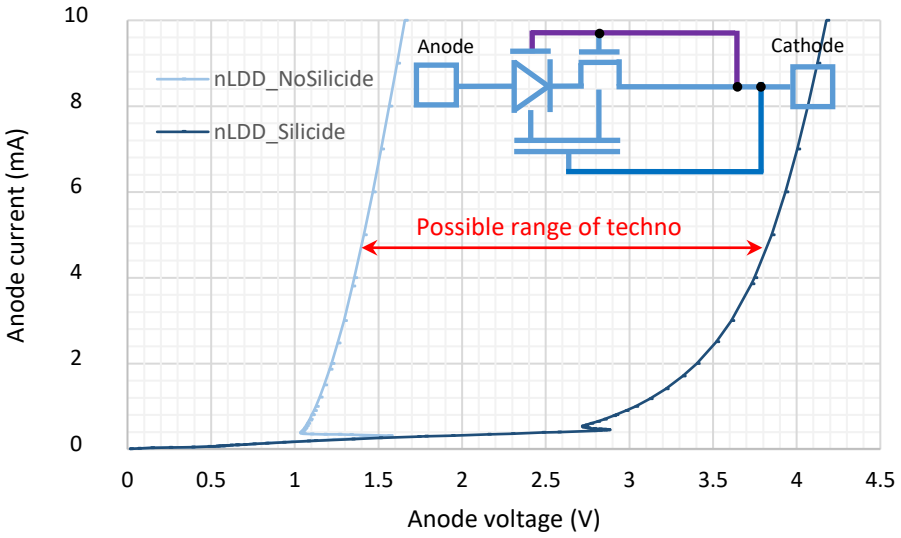


Figure 83: I-V TCAD ACS simulation of a GDNMOS with grounded gates.

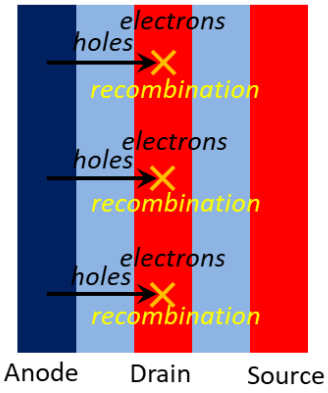


Figure 84: Schematic explaining the recombination effect in the drain. Holes that come from the anode are recombined with electrons in the drain region. Because of this, less holes are able to arrive in the channel of the NMOS region (this hole current turns ON the NPN bipolar transistor drain-channel-source).

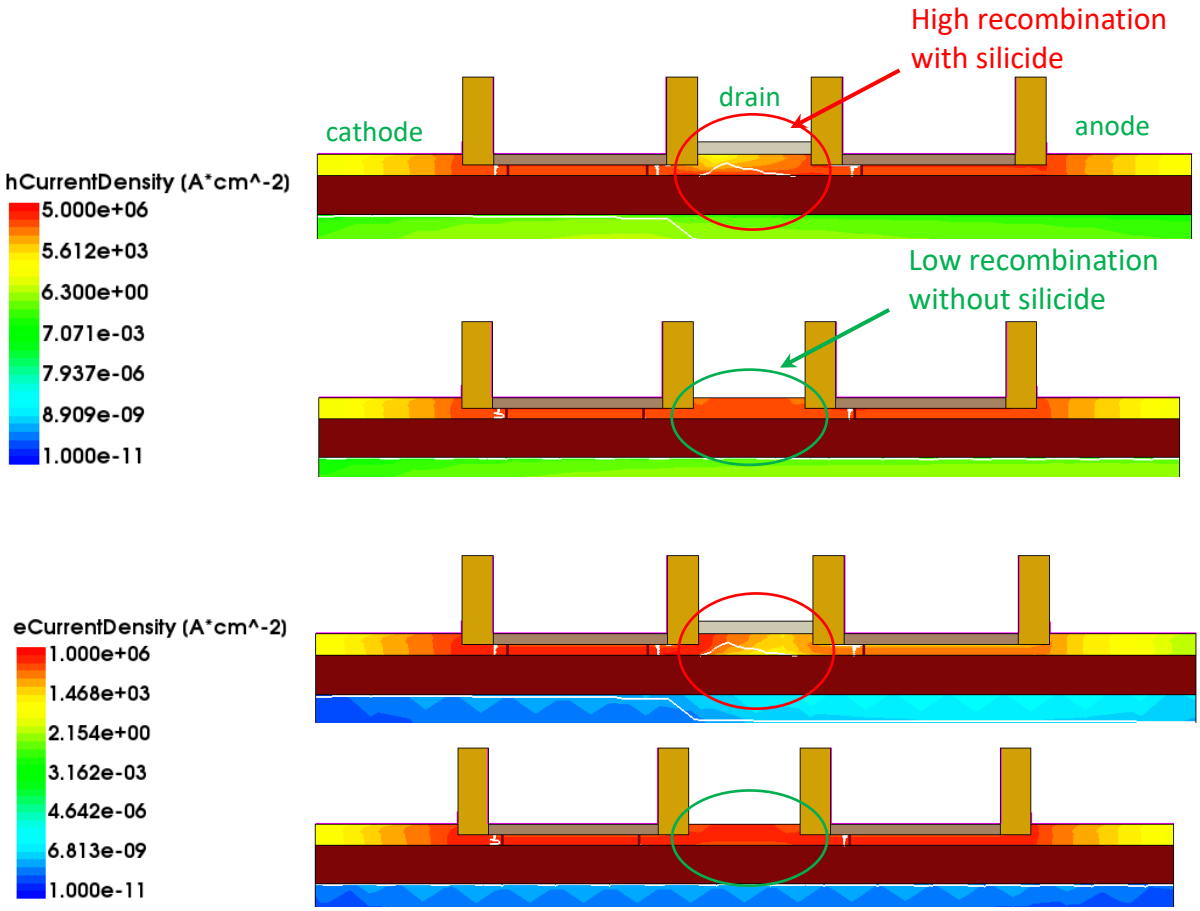
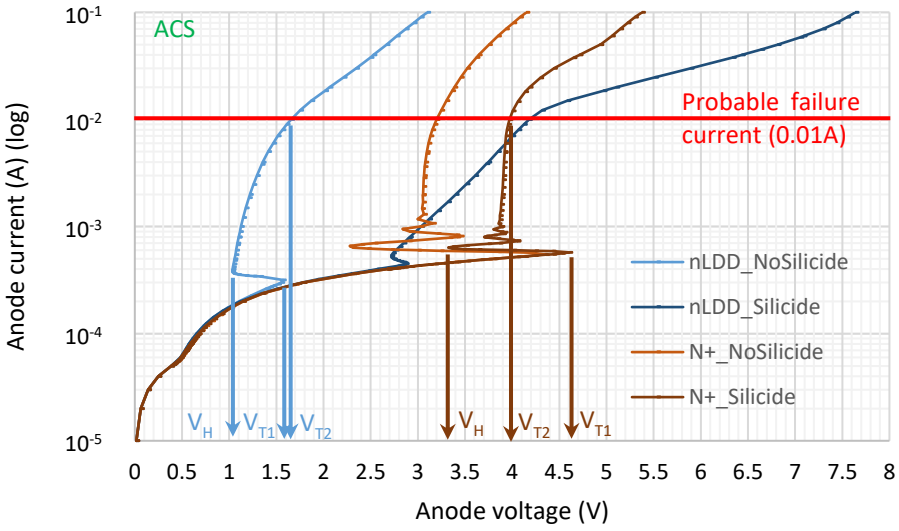
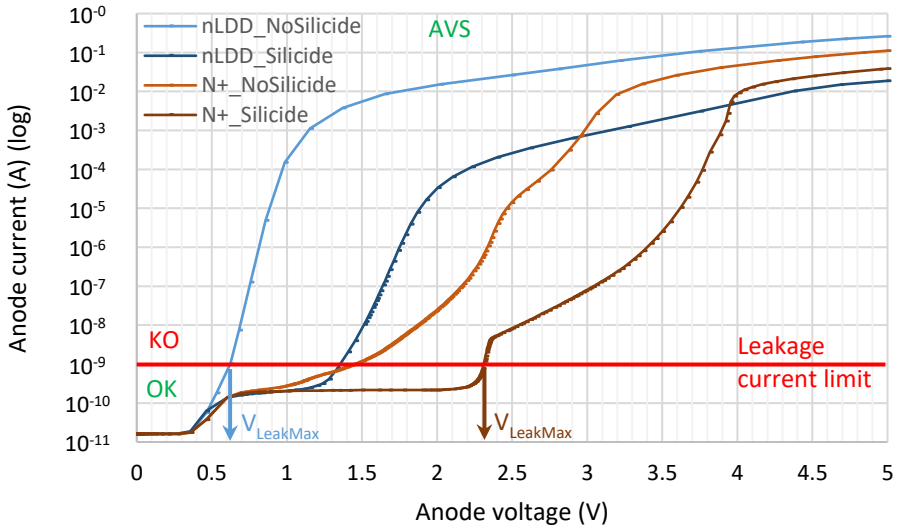


Figure 85: Hole current density (top) and electron current density (bottom) in the N-LDD doped drain GDNMOS with grounded gates, extracted from the TCAD ACS simulation in Figure 83. The N-LDD GDNMOS with silicide on the drain (silicide is the grey layer) is compared to the N-LDD GDNMOS without silicide. With silicide, electron and hole current density is lowered in the drain region, due to recombination.

As a result, four types of ESD protections are available (Figure 86): the GDxMOS with N-LDD or N^+ doping, and with or without silicide on its drain. When looking at the ESD design windows allowed by each possibility, one can conclude that the classical GDxMOS (the one with N^+ doping and silicide) is a high-voltage protection. No additional mask is needed for drawing its layout. To create a low-voltage protection, it is needed to use the SBLK (silicide block) mask OR the NOSD mask (for having only N-LDD doping). To obtain a very low-voltage protection, both the SBLK and the NOSD mask (no silicide and N-LDD doping) are used.





	V_{T1} (V)	V_H (V)	V_{T2} (V)	$V_{LeakMax}$	V_{DD_Max}	V_{BD_Max}
N-LDD, no silicide	1.6 V	1.0 V	1.7 V	0.6 V	0.6 V	1.7 V
N-LDD, silicide	2.9 V	2.7 V	4.2 V	1.3 V	1.3 V	4.2 V
N^+, no silicide	4.3 V	2.3 V	3.2 V	1.4 V	1.4 V	4.3 V
N^+, silicide	4.6 V	3.4 V	4.0 V	2.3 V	2.3 V	4.6 V

Design window

Figure 86: Top graph: I-V TCAD ACS simulation of a GDNMOS with grounded gates. Bottom graph: I-V AVS TCAD simulation. Table: extracted parameters from the graph. The failure voltage V_{T2} of the ESD device is taken on the ACS graph at I = 0.01 A, and the maximum voltage $V_{LeakMax}$ allowed in order to restrict the leakage current is taken on the AVS graph at I = 10^{-9} A. The maximum V_{DD} allowed (V_{DD_Max}) is the smallest parameter between V_{T1}, V_H and $V_{LeakMax}$. The minimum breakdown voltage (V_{BD_Max}) of the components to be protected is the highest parameter between V_{T1} and V_{T2}. The interval between these V_{DD_Max} and V_{BD_Max} represents the design window allowed by each ESD protection. This ESD design window is not taking into account the 10% of margin.

Measurements are expected to exhibit an even bigger difference of trigger voltage between the GDxMOS devices with or without silicide. Indeed, the TCAD simulations were performed with a layer of silicide NiSi above the layer of silicon. In reality there is a gradient of silicide which can be thought as a gradient of doping from N^- (near the BOX) to N^+ (close to the top of the structure, where there is silicide) (Figure 87). Due to this gradient, there is band curvature in the energy band diagram, which induces an internal potential $\Phi_0(x)$ (Figure 88). Therefore, an electric field is established in order to equilibrate the displacement of the charges at the thermodynamic equilibrium. As a result, silicide attracts holes, which is very detrimental for a SCR.

A small remark: it would really be interesting to investigate the GDBIMOS device with LDD drain doping further (by simulations as well as measurements) and to compare it to the GDNMOS device. Indeed, in the GDNMOS device, holes are coming from the anode and do not recombine in the drain (because of LDD doping and no silicide in the drain), so they reach the NMOS channel (which is also the base of the PNP bipolar transistor), hence the two bipolar transistors of the SCR become conductive. The question is, in the case of the GDBIMOS device, would the holes arriving in the channel of the BIMOS also be sucked in the body contacts of the BIMOS (additionally to turning the SCR ON)? This would increase the BIMOS gate voltage and explain the improvement of R_{ON} of the GDBIMOS device (with respect to a GDNMOS).

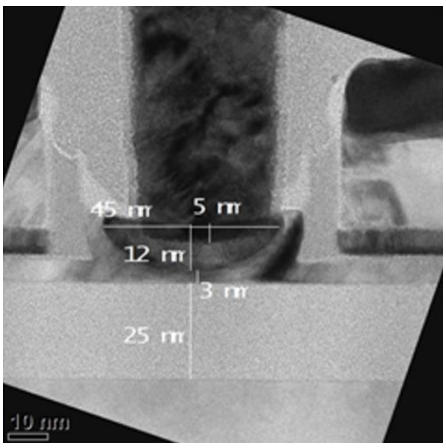


Figure 87: Post NiSi step TEM (transmission electron microscopy). Note that only 3 nm of silicon without silicide is remaining.

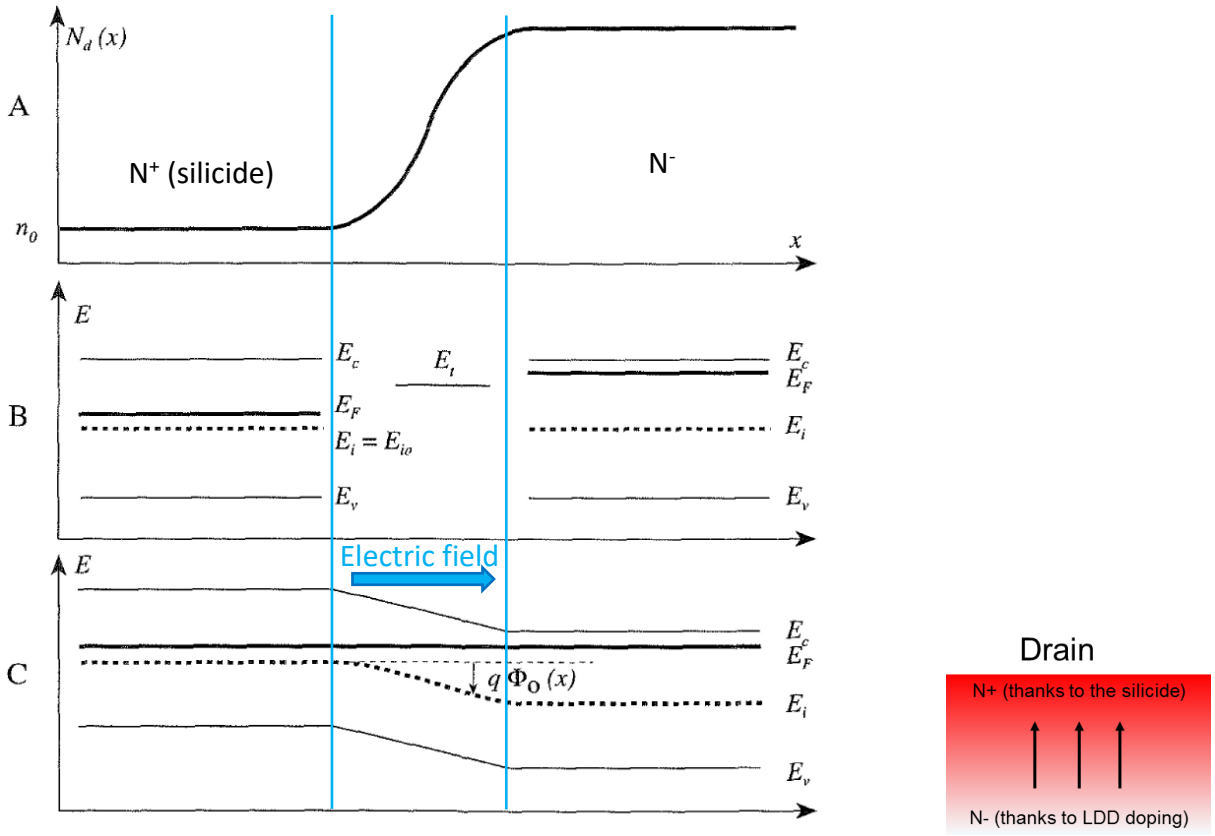


Figure 88: Left: In case of inhomogeneous doping (adapted from [110]). A: inhomogeneous doping profile. B: Fermi levels in the fragmented samples. C: band diagram of the complete structure. Right: Schematic of the drain of a GDxMOS.

b. Partial silicide

The idea of this section is to create a set of protections thanks to partial silicide along the width of the GDxMOS. There will be silicide on one side of the width and no silicide on the other; therefore, different conductive parts will be activated in the structure (Figure 89).

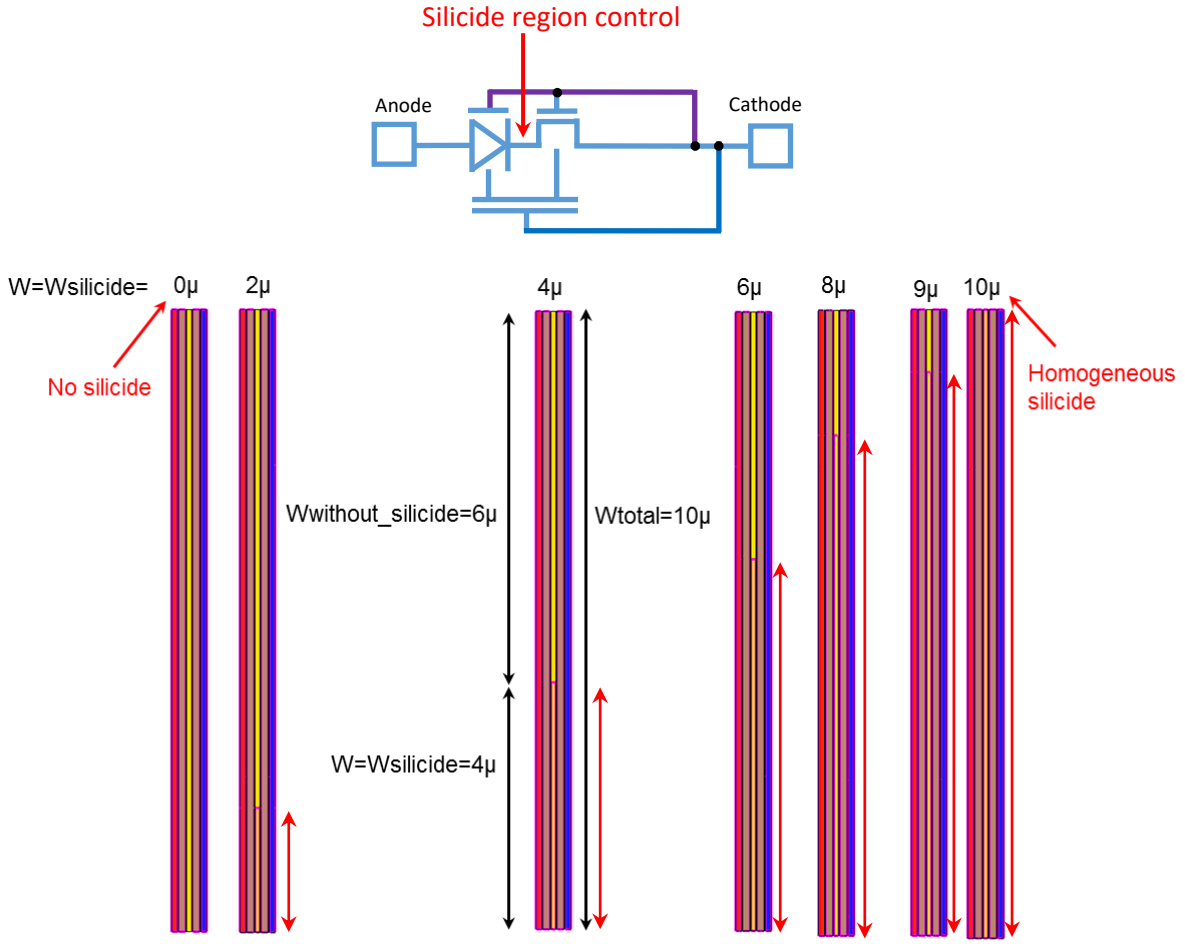
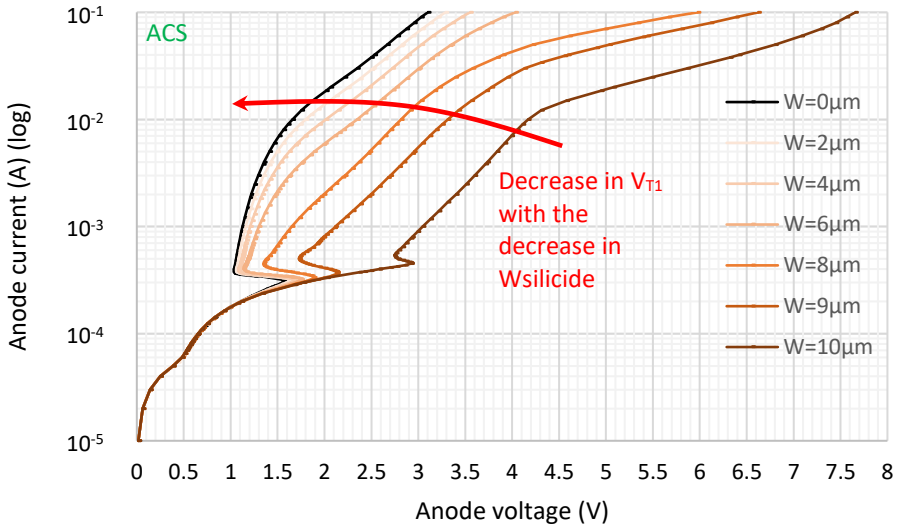
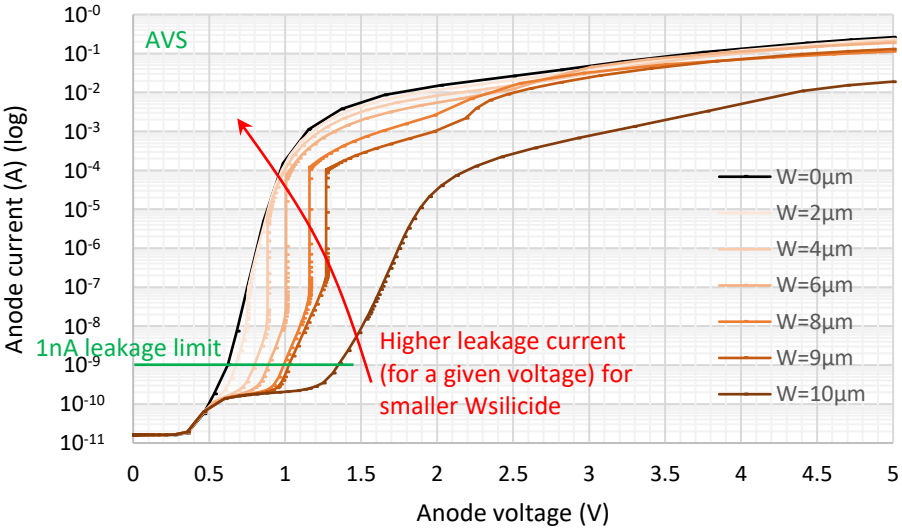


Figure 89: Schematic and top views of the studied GDNMOS structures. W (the width with silicide) varies from 0 to 10 µm. The total width of the finger is always 10 µm. (When W = 0 µm it means that the structure does not have silicide at all on its drain, and when W = 10 µm the drain is fully salicided.)

As seen in Figure 90, the wider the silicide, the higher the trigger voltage and the lower the leakage current. The parasitic bipolar transistor NPN activates because a certain quantity of holes is present *in* the NMOS channel. So, the more silicide there is, the less hole flux (from the anode to the channel of the NMOS) and therefore the activation of the bipolar occurs later (when there are more holes).

As a conclusion, partial silicide allows to tune the trigger voltage and leakage current according to the application, with no extra cost on silicon area and on trigger circuit design.





	V_{T1} (V)	V_H (V)	V_{T2} (V)	$V_{LeakMax}$ (V)
no silicide (W = 0 µm)	1.6	1.0	1.7	0.6
W = 2 µm	1.6	1.1	1.8	0.7
W = 4 µm	1.7	1.1	2.0	0.8
W = 6 µm	1.7	1.1	2.3	0.9
W = 8 µm	1.9	1.2	2.7	0.9
W = 9 µm	2.0	1.4	3.1	1.0
full silicide (W = 10 µm)	2.9	2.7	4.2	1.3

Very low voltage protection

Low-voltage protection

Design window

Figure 90: I-V TCAD simulation of a GDNMOS with grounded gates and N-LDD doping in the drain, for different silicide covers of the drain. Top: ACS. Bottom: AVS. Table: extracted values.

Note that the effect of silicide over the width is not linear: significant change (> 10%) on the ACS curves (with respect to the situation when there is no silicide at all) occurs from Wsilicide/Wtotal > 0.5 (Figure 91). This means that up to 5 µm of silicide can be deposited on the drain (with a finger of 10 µm) without impacting the performance of the protection. On the one hand, it means that the idea of adding silicide on some width of the drain in order to be able to put a tungsten contact on it (to connect the drain) is valid and will not prevent the protection to have a low trigger voltage, thanks to the part of the width without silicide and N-LDD doping. On the other hand, designers have to pay attention to keep a sufficiently high percentage of salicided width if they want to change the ACS and AVS characteristics of the protection.

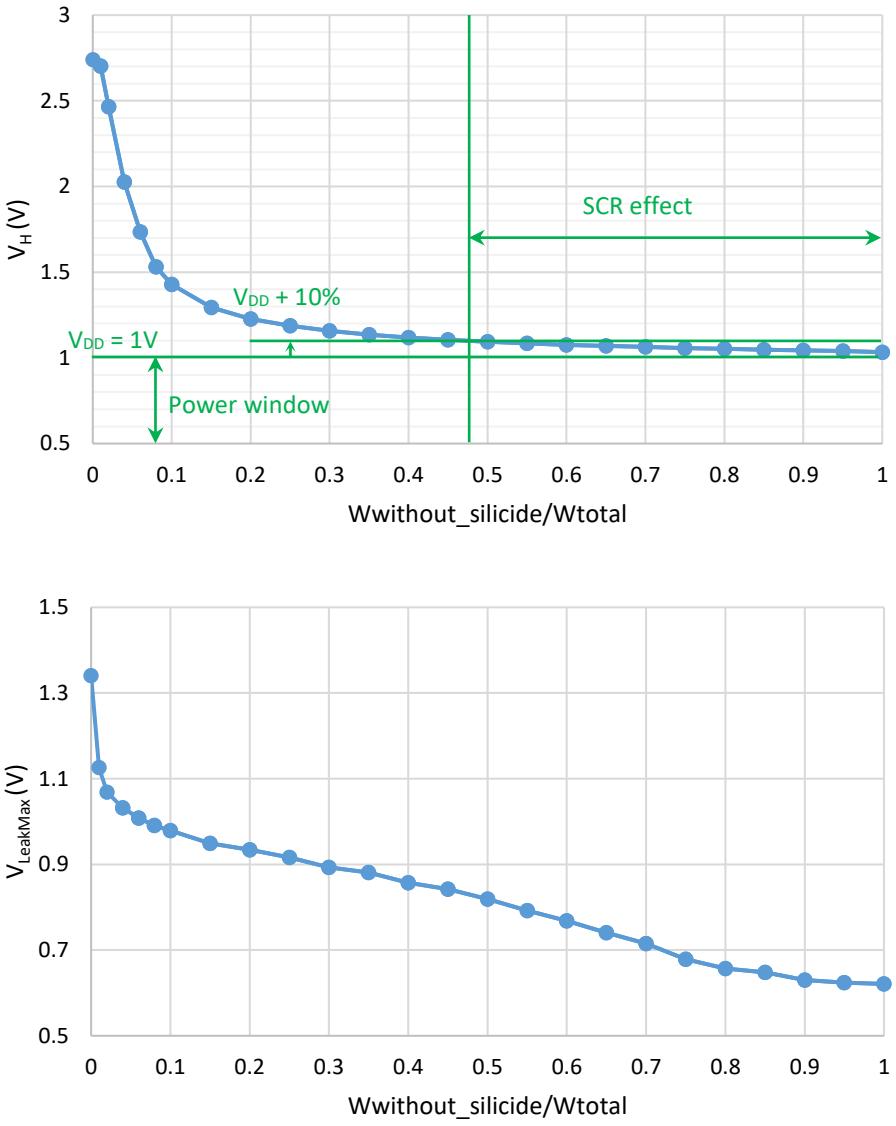


Figure 91: Holding voltage V_H (extracted from the ACS curve) and maximum voltage allowed as a V_{DD} in terms of leakage current for the protection $V_{LeakMax}$ (it is the voltage at 1 nA on the AVS curve) versus the proportion of width without silicide and the total width of the finger. Note that Wwithout_silicide = Wtotal – Wsilicide. When there is few silicide, the SCR effect is present in the device and the holding voltage V_H is close to V_{DD} = 1 V; as a result, the device is immune to latch up if there is a sufficient width of silicide on the drain.

Remark also that the dynamic ON resistance R_{ON} of the protection is modified with the silicide cover of the drain (Figure 90). This is because the part that has silicide will less participate in current conduction than the part without silicide.

The same study with partial silicide can be done with N^+ doping in the drain region (Figure 92). In this case, the wider the silicide, the lower the holding voltage, but the trigger voltage only slightly decreases. In fact, N^+ doping is already preventing holes to reach the channel of the NMOS, this is why changing the silicide width does not affect a lot the trigger voltage of the device.

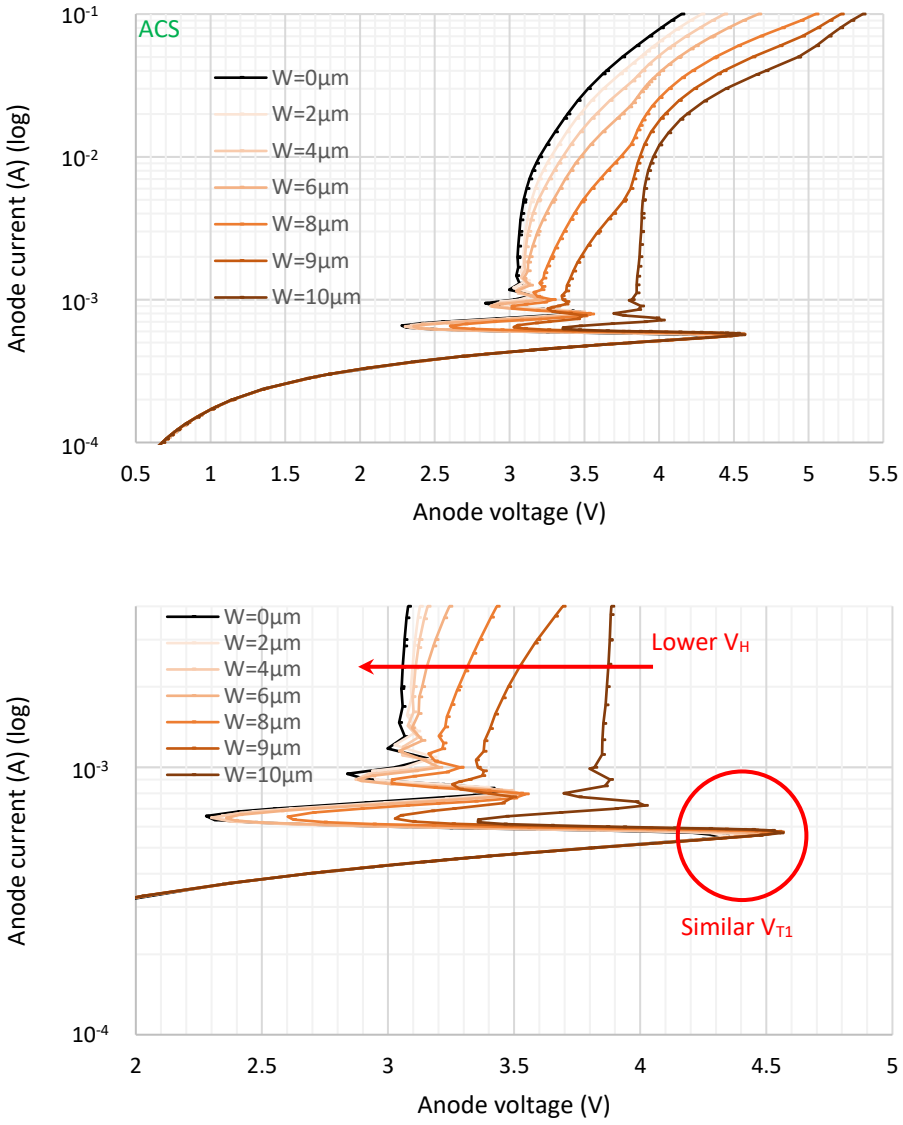


Figure 92: I-V TCAD ACS simulation of a GDNMOS with grounded gates and N^+ doping in the drain, for different silicide covers of the drain. Bottom: zoom on the curves.

c. Partial silicide and drain connected to the diode gate

GDxMOS with N-LDD doping and no silicide has a high leakage current, therefore it can be used as a very low-voltage protection (0.6 V). A possible solution to be able to utilize it as a low-voltage protection (1 V) is to connect its drain. Partial silicide could be employed so that the contacts CA are put on the silicide part of the drain, and the main conduction occurs in a region without silicide in order to benefit from the “low doping”.

Two connectivities of the drain are studied: when it is connected to the gate of the diode and when it is plugged to the anode. This section deals with the first type (Figure 93).

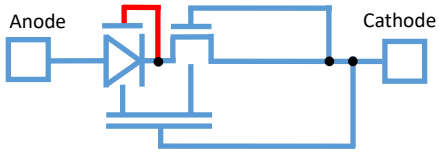


Figure 93: Schematic of the GDNMOS with the drain connected to the gate of the diode.

The leakage current problem of the N-LDD GDNMOS with no salicided drain comes from the hole current. As soon as holes are able to flow through the device, they are not stopped because they do not sufficiently recombine.

Of course, the higher recombination, the higher current also, since holes that arrive in the drain recombine with electrons that come from the source. However, this current is very low in comparison with the current that flows in the SCR when both its bipolar transistors are ON (and this occurs when a sufficient number of holes arrives in the channel of the NMOS).

By connecting the drain on the gate of the diode, the barrier for the holes between the anode and the channel is bigger, because the voltage on the gate of the diode increases as soon as carriers are flowing through the diode (self-biasing of the diode gate). As expected, the ACS and AVS characteristics of the N-LDD GDNMOS with floating drain and with the drain connected to the diode gate are shifted with respect to each other (Figure 94).

Figure 95 shows the ACS and AVS behavior of the N-LDD GDNMOS with the drain connected to the gate of the diode, for different silicide cover of the drain. It emphasizes that it is still possible to adjust the silicide width in order to obtain different trigger voltages, leakage currents, etc.

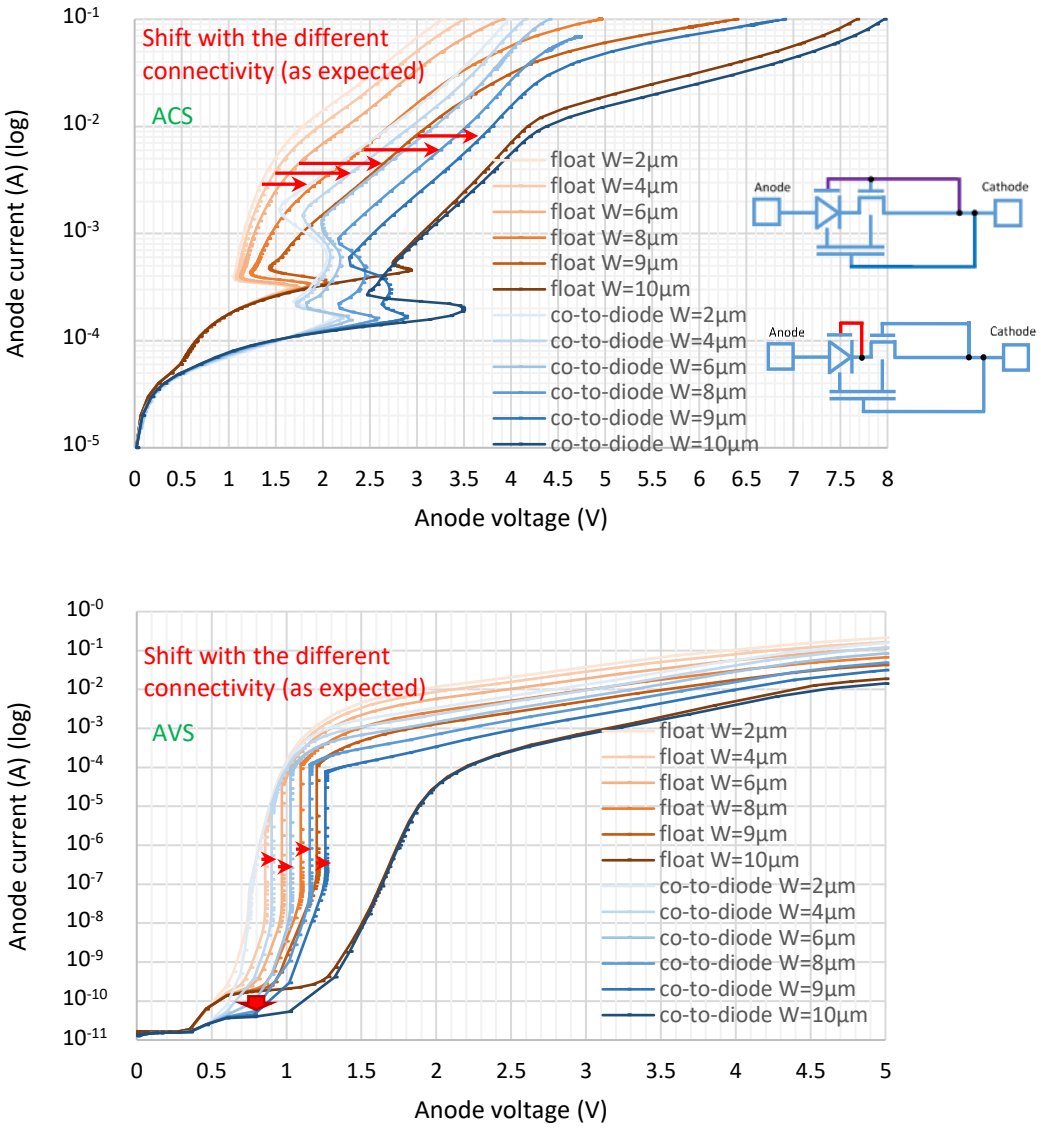


Figure 94: I-V TCAD simulation of a N-LDD GDNMOS with a grounded NMOS gate, for different silicide covers of the drain. ''float'' (the drain is left floating and the diode gate is grounded) is compared to ''co-to-diode'' (the drain is connected to the diode gate). Top: ACS. Bottom: AVS.

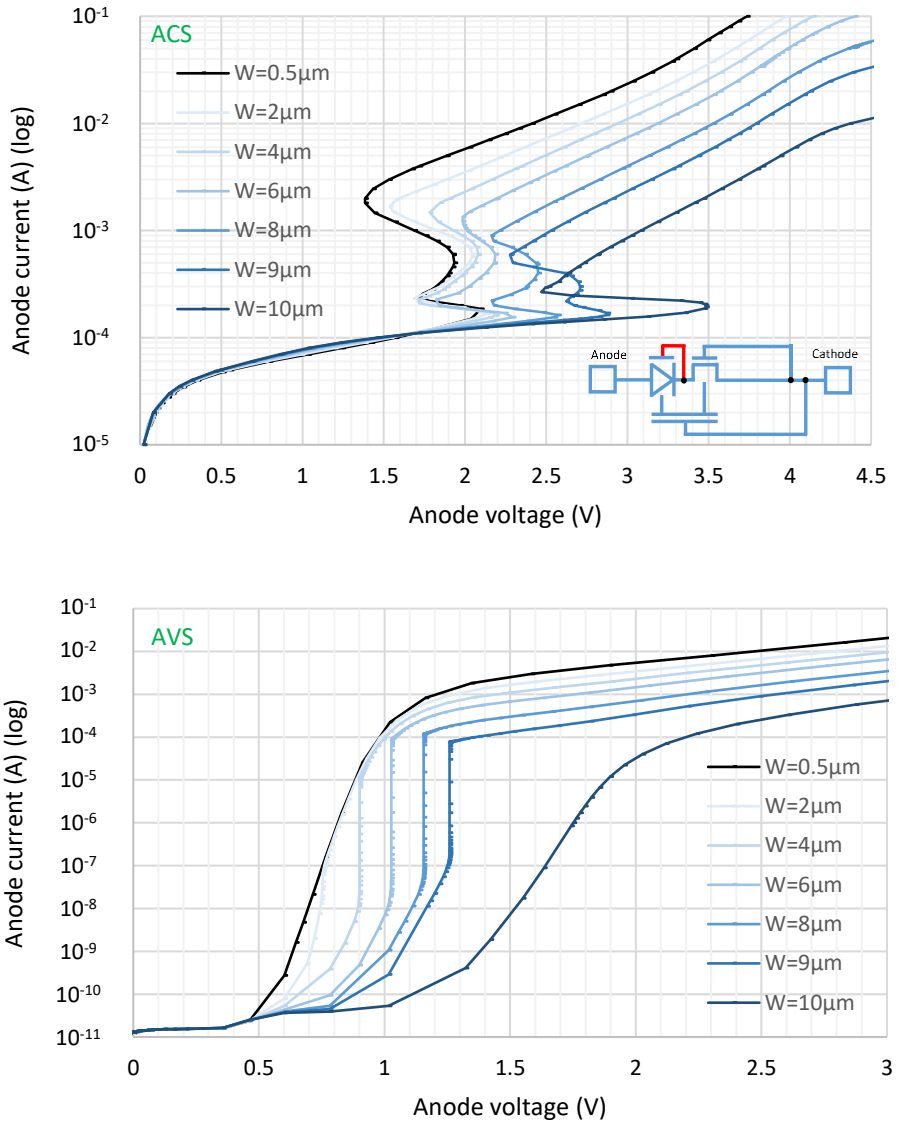


Figure 95: I-V TCAD simulation of a N-LDD GDNMOS with the drain connected to the diode gate and a grounded NMOS gate, for different silicide covers of the drain. Top: ACS. Bottom: AVS.

Note that a double-snap-back is observed. The first snap-back is due to the current starting to flow in the part that has no electrode. The second snap-back is due to the other side (with the electrode) starting to conduct current also (Figure 96).

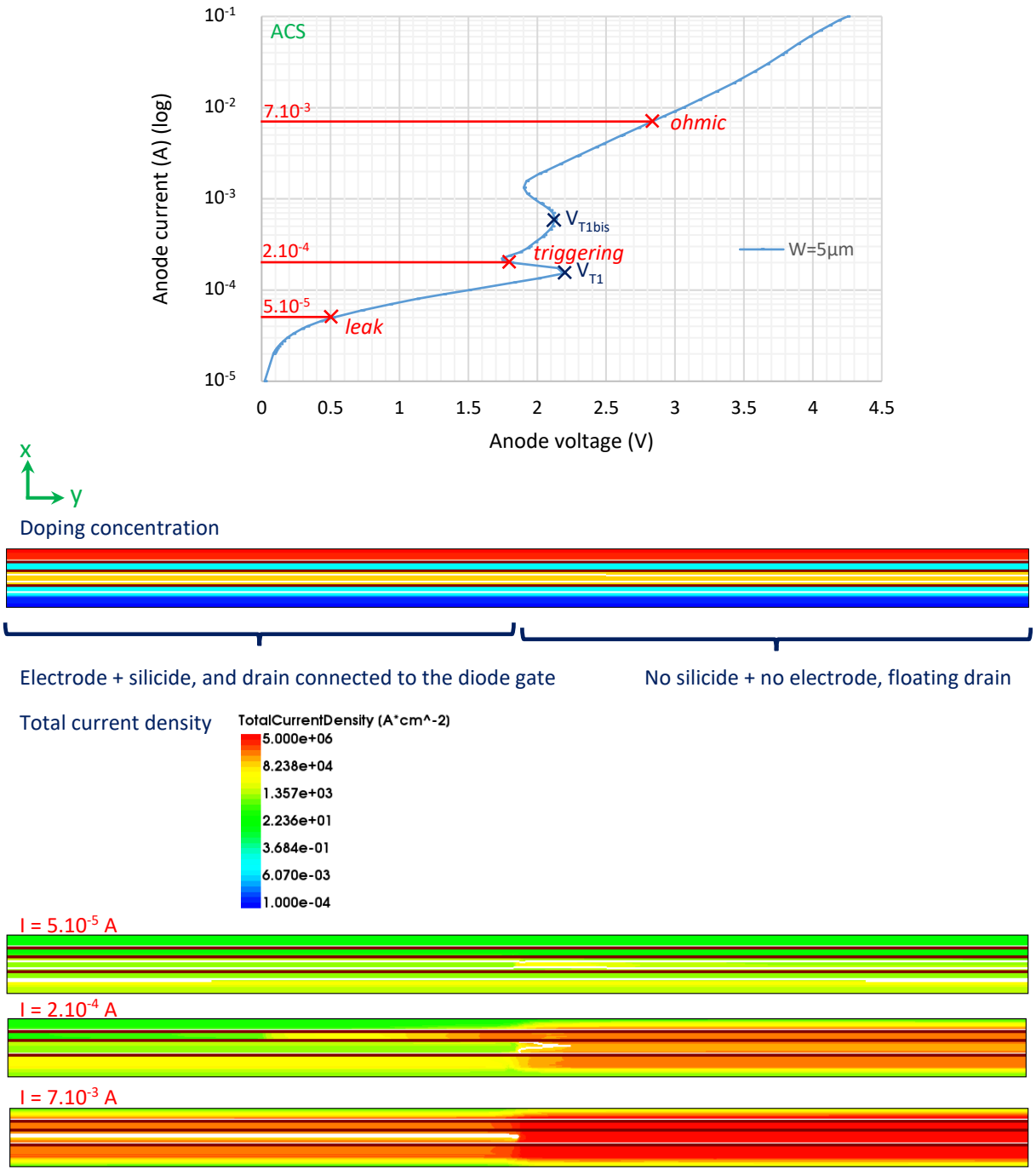
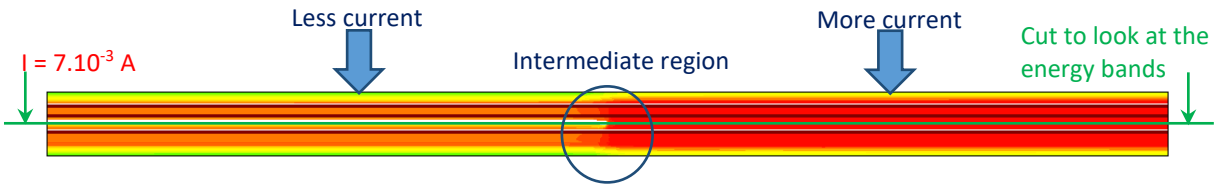
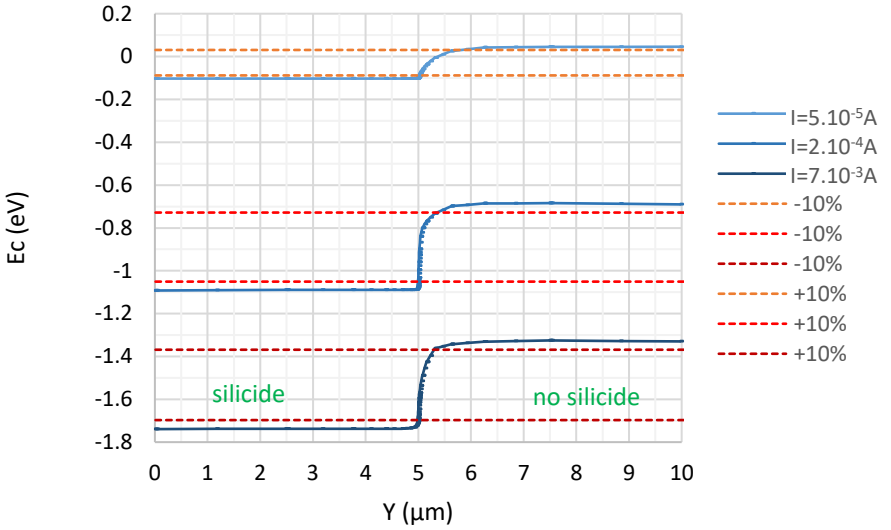


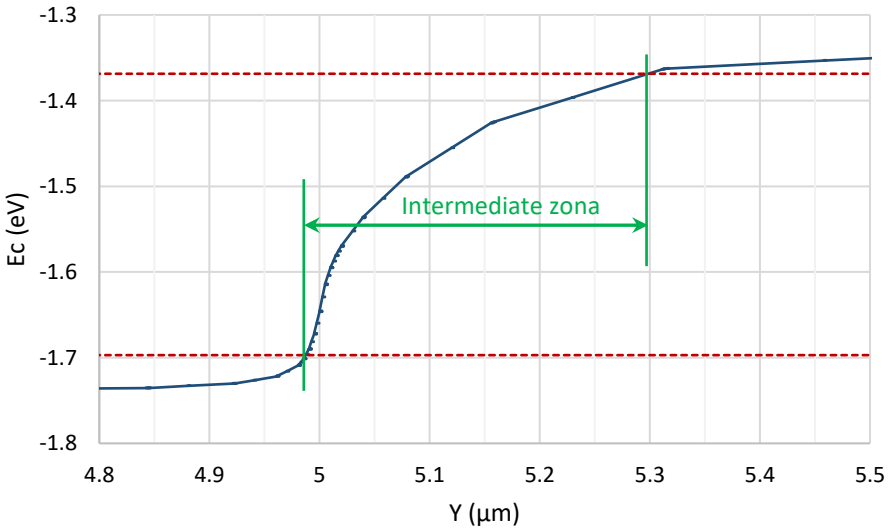
Figure 96: Top: I-V TCAD ACS simulation of a N-LDD GDNMOS with the drain (with a cover of silicide of 5 µm out of 10 µm of finger) connected to the diode gate and a grounded NMOS gate. Bottom: extracted top views from the structure.

Note that there is always less current flowing through the part that has the electrode than in the other one, even for a high current. There is an intermediate region where both the influence of the electrode and the one of the region without electrode are effective (the finger is progressively conductive along the width: it is less conductive close to the salicided region covered by the electrode and more conductive far from the electrode) (Figure 97). This is because the electrode acts like a gate, and it has an influence on the potential in the

drain until 1 µm after it is cut, thanks to the resistivity of the silicon without silicide. The effect of the electrode on the part that has no silicide can be seen on the energy bands.







	I = 5.10^{-5} A	I = 2.10^{-4} A	I = 7.10^{-3} A
Size of intermediate zona (µm)	0.80	0.36	0.31

Figure 97: Top: top view from Figure 96. Middle: conduction band in the drain versus the width of the structure for different currents. Bottom: zoom on the conduction band in the drain along the width of the structure at I = 7.10^{-3} A. Table: intermediate zona size extracted from the graphs.

d. Partial silicide and drain connected to the anode

The goal of plugging the drain of the GDxMOS to its anode (Figure 98) is to cope with the leakage current problem of the N-LDD GDxMOS that has a floating drain. The idea is to decrease the current flowing in the diode (in order to decrease the number of holes that will create the leakage) by providing another path for the current, thanks to the NMOS and the system “diode + NMOS = SCR” being in parallel.

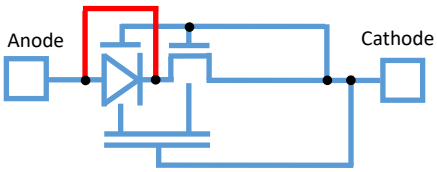


Figure 98: Schematic of the GDNMOS with the drain connected to the anode.

Simulation results are shown in Figure 99, Figure 100, Figure 101 and Figure 102. The ACS behavior of the GDNMOS with the drain connected to the anode is improved at very high current with respect to the GDNMOS with a floating drain (Figure 99 and Figure 101), except for W = 10 µm which is a special case. This improvement is due to the second snap-back (at V = 3.5 V). However, this needs to be verified by measurements. Indeed, the failure current may be sufficiently low in order never being able to experience the second snap-back. Another possible behavior, is that the failure current may be improved thanks to this second snap-back, because another part of the width of the structure will activate and thus more current will be able to flow inside the device. As expected, the drain being connected to the anode reduces the leakage current with respect to the floating drain (on the AVS curves Figure 102).

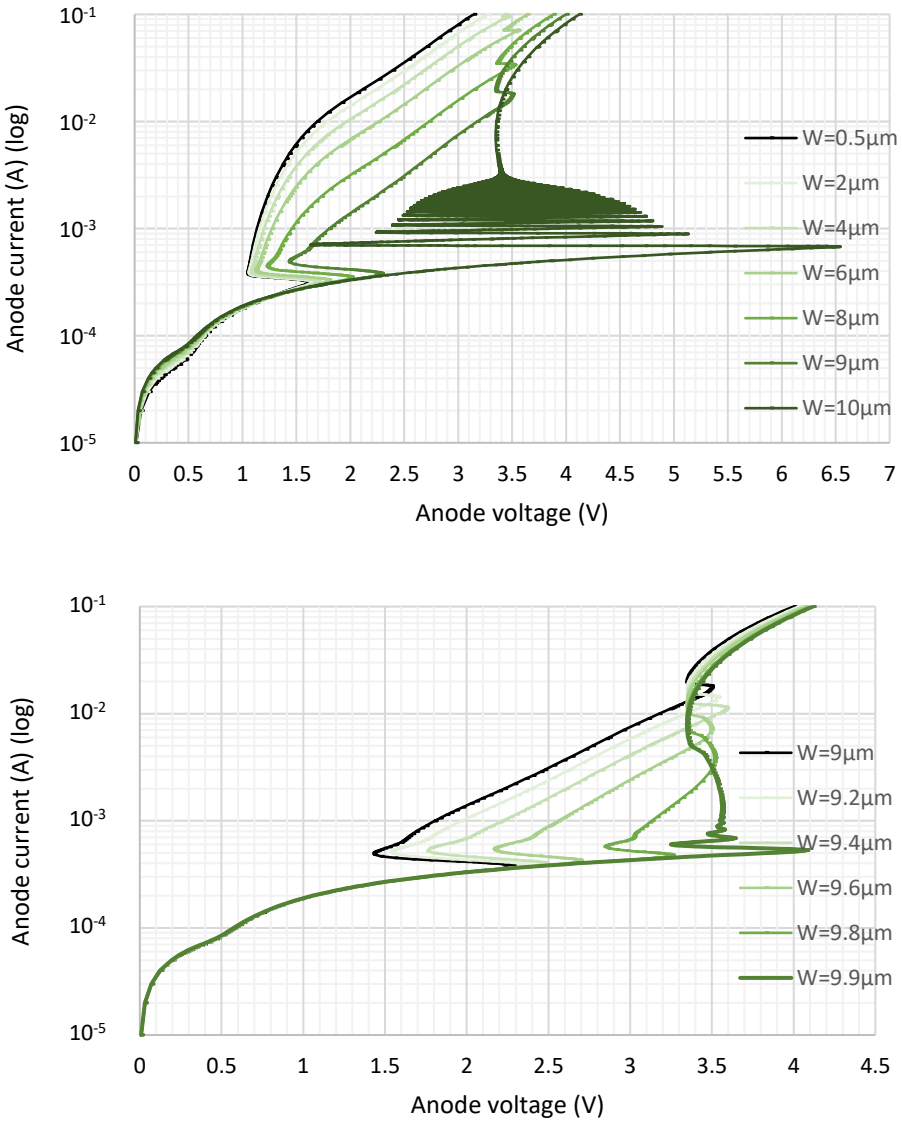


Figure 99: I-V TCAD ACS simulation of a N-LDD GDNMOS with the drain connected to the anode and grounded gates, for different silicide covers of the drain.

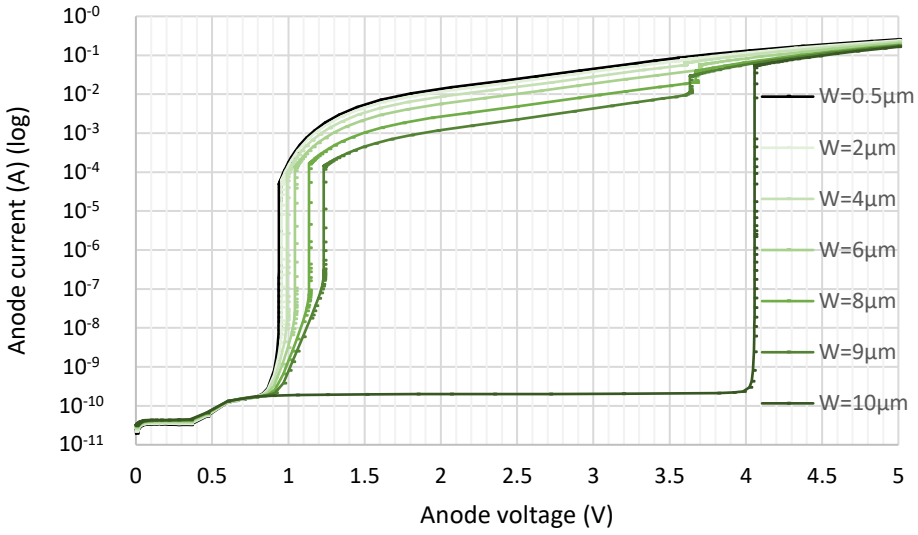


Figure 100: I-V TCAD AVS simulation of a N-LDD GDNMOS with the drain connected to the anode and grounded gates, for different silicide covers of the drain.

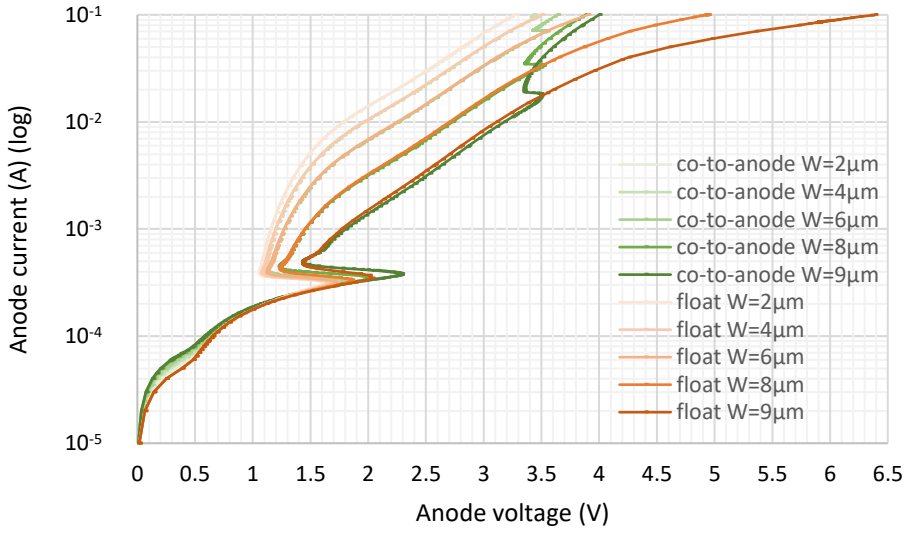


Figure 101: I-V TCAD ACS simulation of a N-LDD GDNMOS with grounded gates. ‘‘float’’ (the drain is left floating) is compared to ‘‘co-to-anode’’ (the drain is connected to the anode), for different silicide covers of the drain.

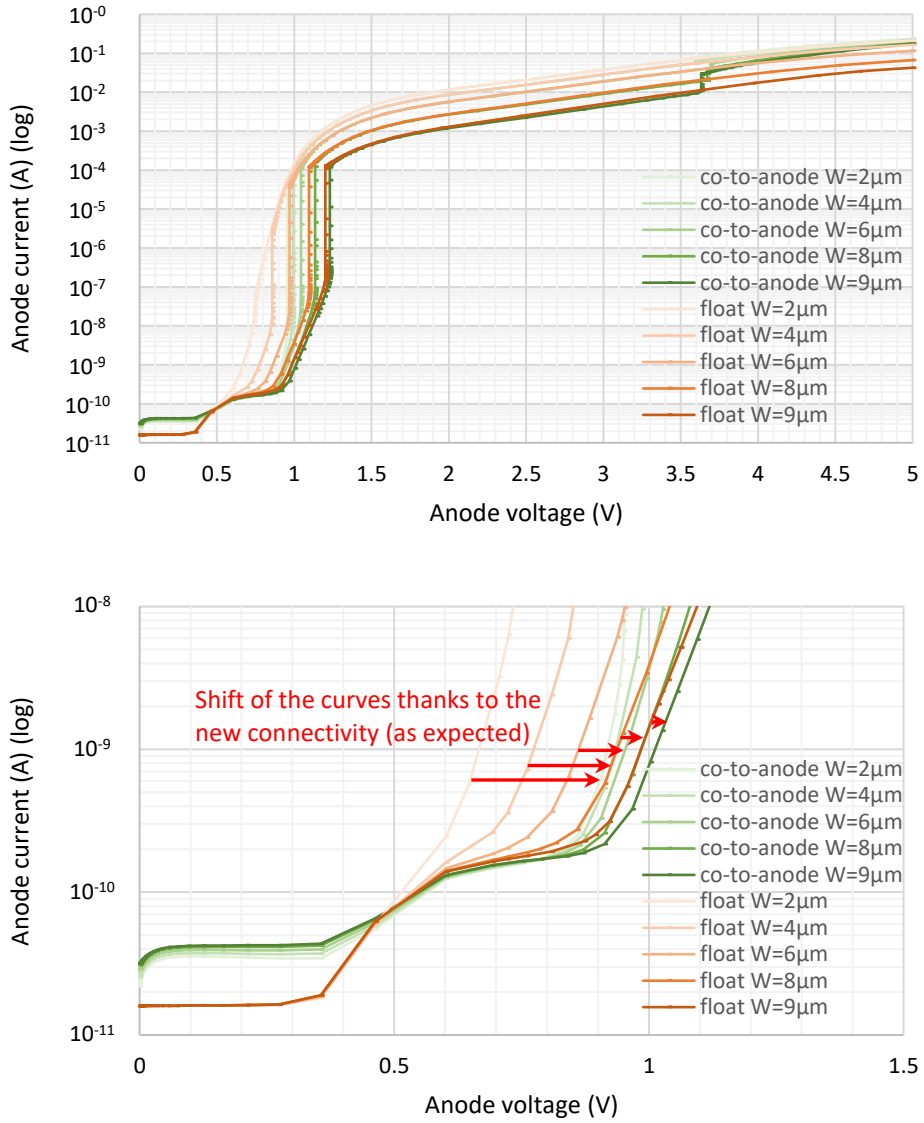


Figure 102: I-V TCAD AVS simulation of a N-LDD GDNMOS with grounded gates. ''float'' (the drain is left floating) is compared to ''co-to-anode'' (the drain is connected to the anode), for different silicide covers of the drain. Bottom: zoom.

Figure 103 shows that the first snap-back is due to the SCR part of the device (the one with no silicide), and the second snap-back is due to the NMOS part (the one with silicide and the connectivity between the drain and the anode).

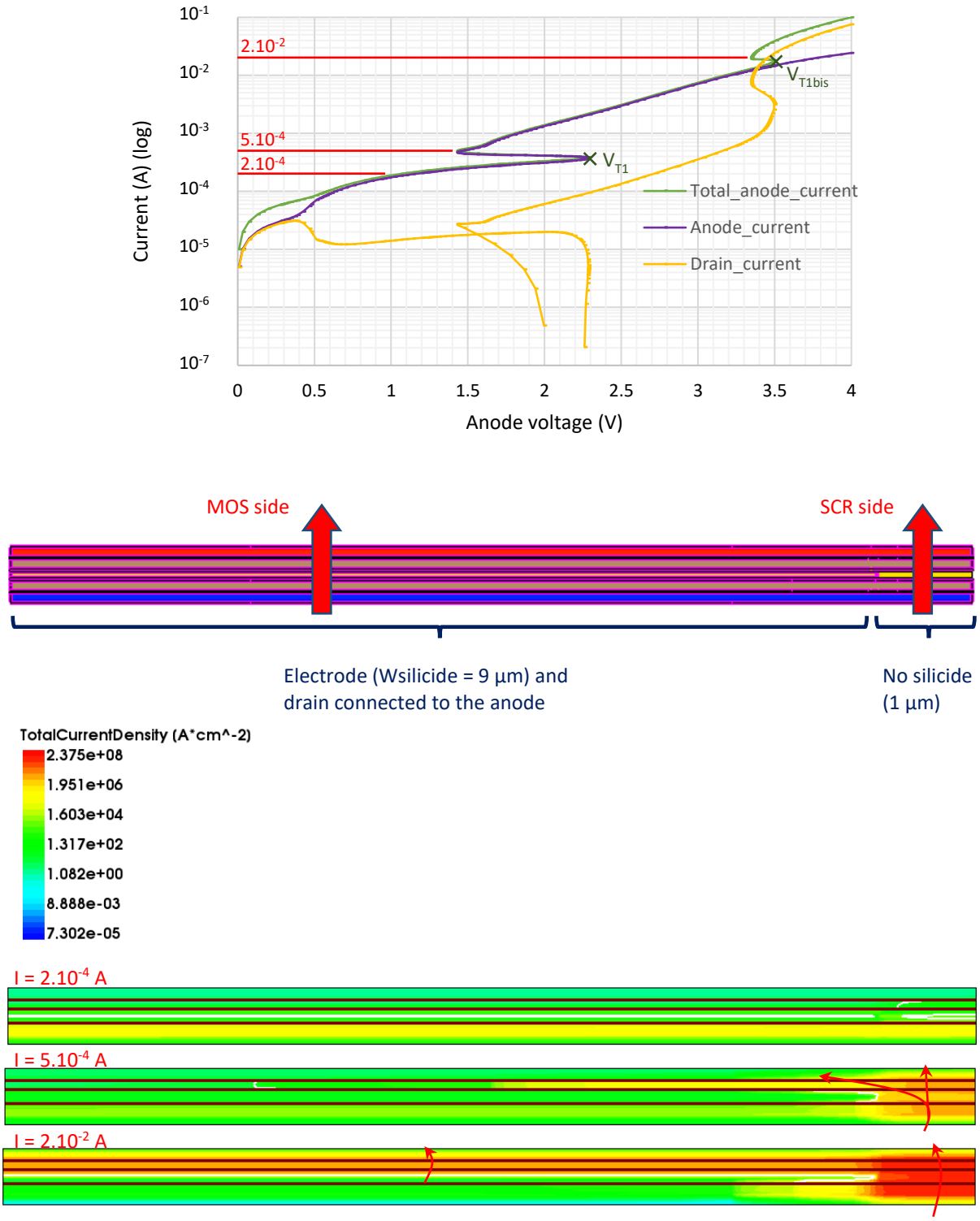


Figure 103: Top: I-V TCAD ACS simulation of a N-LDD GDNMOS with grounded gates and the drain connected to the anode on a salicided part of the width of the drain. Wsilicide = 9 µm. The contribution of the anode and the drain are detailed in the total current flowing through the device. Bottom: extracted views from the structure.

From the Figure 99, it can be seen that the wider the Wsilicide, the lower the second snap-back current I_{T1bis}. This is because the "SCR" part is smaller in width so it cannot conduct as much current as with a higher width (Figure 104). At some point, where a high current is already flowing through the SCR and the SCR starts to be saturated, it becomes easier to flow through the "NMOS" part (in terms of resistivity for carriers). Also, the wider the Wsilicide, the higher R_{ON} between I_{T1} and I_{T1bis}, because the width of the "SCR" part is lower. Those results have to be confirmed by measurement, since a too small "SCR" part in the device could eventually cause robustness issues.

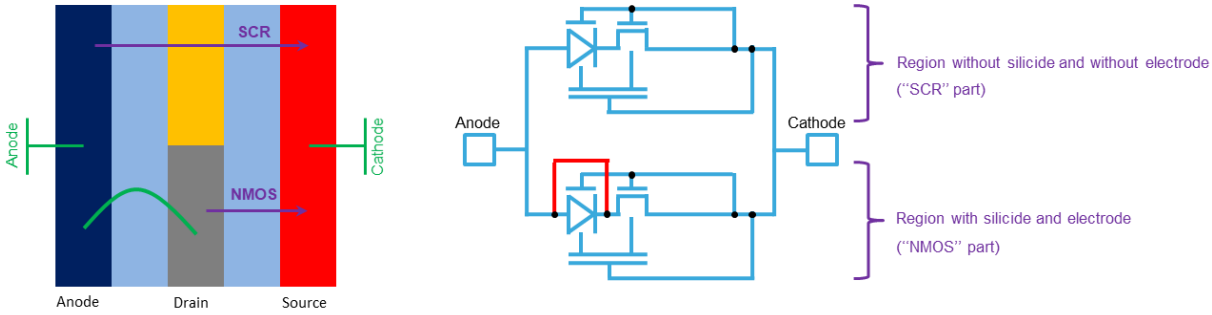


Figure 104: N-LDD GDNMOS with grounded gates and the drain connected to the anode on a salicided part of the width of the drain. "SCR" and "NMOS" parts of the device conduction are emphasized in a top view schematic of the GDNMOS (left) and in an electrical schematic (right).

When Wsilicide = 10 µm (Figure 105), there is no "SCR"-like conduction, because the SCR is shorted (connection between anode and drain), so there is only NMOS conduction. It can also be seen as a SCR on which the basis N of the bipolar transistor PNP is plugged to a high potential, thus blocking the positive feedback loop (between the two NPN and PNP bipolar transistors) that activates the SCR. This is why, when the drain is connected to the anode with silicide all over the width of the GDNMOS, the device has the same V_{T1} and R_{ON} as a NMOS transistor. Its current before the triggering is as high as a GDNMOS that has a floating drain; indeed, it corresponds to the leakage of the holes of the P^+ region.

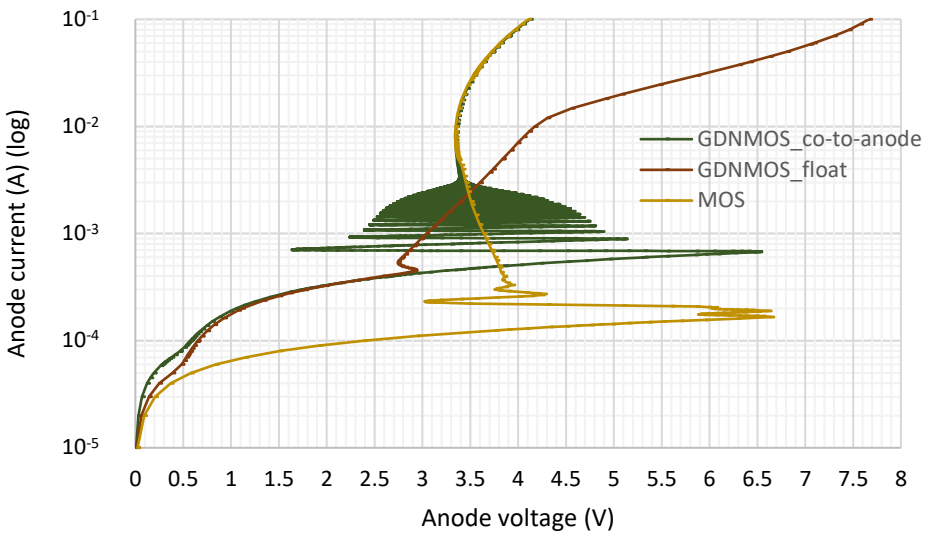


Figure 105: I-V TCAD ACS simulation of a N-LDD fully salicided (W = 10 µm) GDNMOS with grounded gates and with the drain connected to the anode or left floating. They are compared with a NMOS with a grounded gate and a N-LDD drain.

e. Fragmented partial silicide

In this section, let us take a finger of 10 µm of N-LDD GDNMOS and compare its behavior when it has 5 µm of silicide in one piece or fragmented into multiple pieces.

Wtot_silicide is the total width of silicide; it is always 5 µm in this experiment. Wsi is the width of one silicide piece (it is the width of one electrode in the TCAD simulation). Wnosi is the width without silicide between two electrodes. Nelec is the number of electrodes (of salicided part of the drain). Wsi = Wnosi. Therefore Wtot_silicide = Nelec · Wsi and Wfinger = 2 · Wtot_silicide (Figure 106).

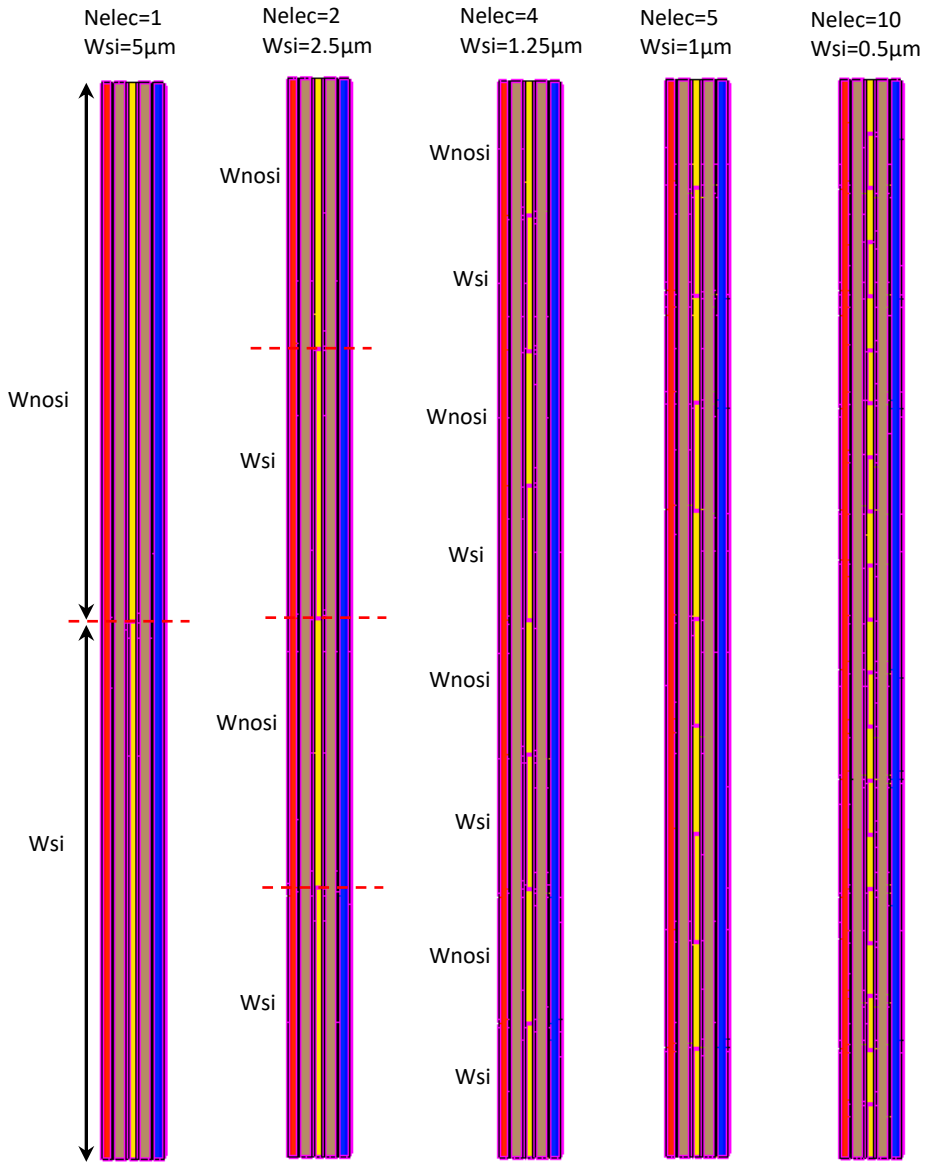


Figure 106: Top views of the studied N-LDD GDNMOS structures (both gates are grounded; the salicided regions control is in the drain). Electrodes (in pink) are placed on the salicided regions and there is no electrode where there is no silicide. The width of the fingers is always the same. The total width of silicide is always the same, but the salicided regions are cut in more and more pieces from one device to another.

If the electrodes are left floating, a higher number of electrodes leads to a higher V_{T1} (on the ACS), a higher R_{ON} (on the ACS) and an improved leakage current (on the AVS) (Figure 107), as if the width of silicide was slightly increasing (like in Figure 90).

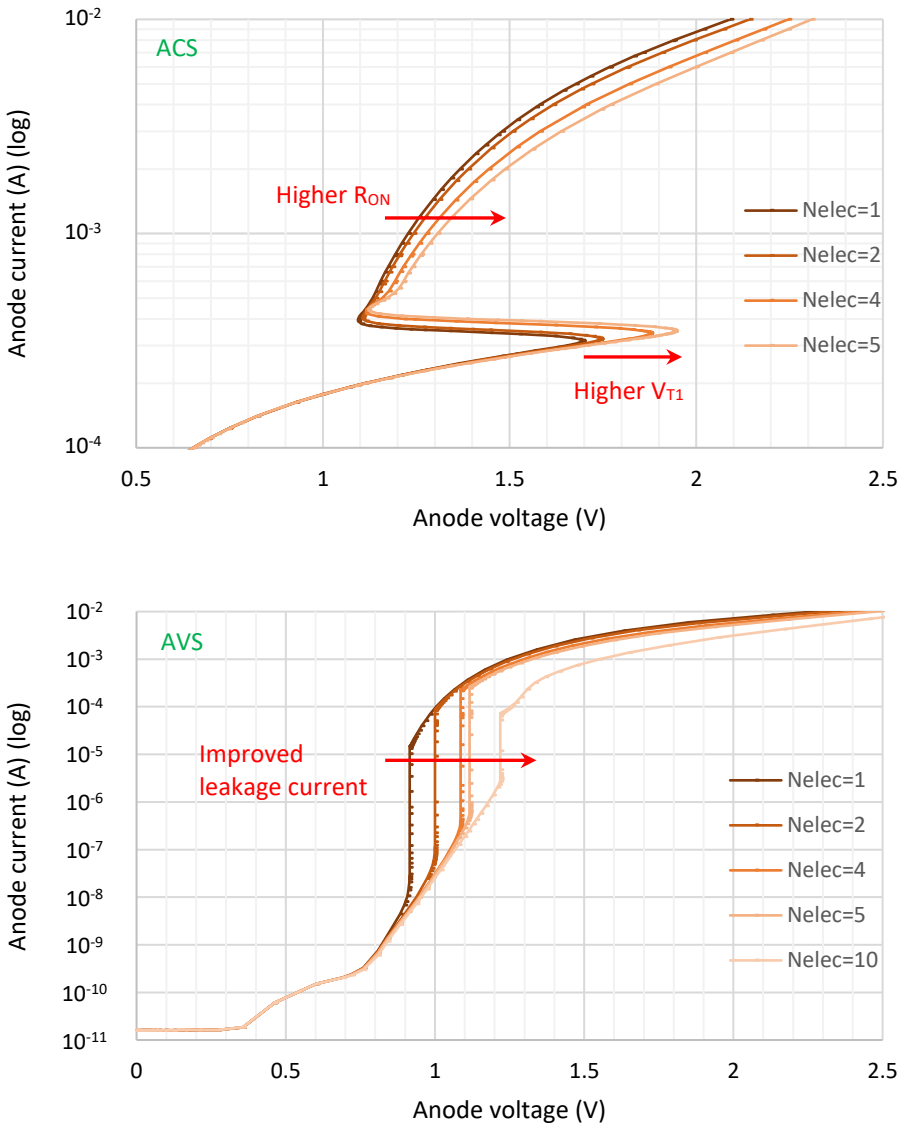


Figure 107: I-V TCAD simulation of a GDNMOS with grounded gates, N-LDD doping in the drain, the same total silicide cover but different number of electrodes on the drain. Electrodes are left floating. Top: ACS. Bottom: AVS.

Similarly, if the electrodes are connected to the gate of the diode, a higher number of electrodes leads to a higher V_{T1} and an improved leakage current (Figure 108), as if the width of silicide was increasing (like in Figure 95).

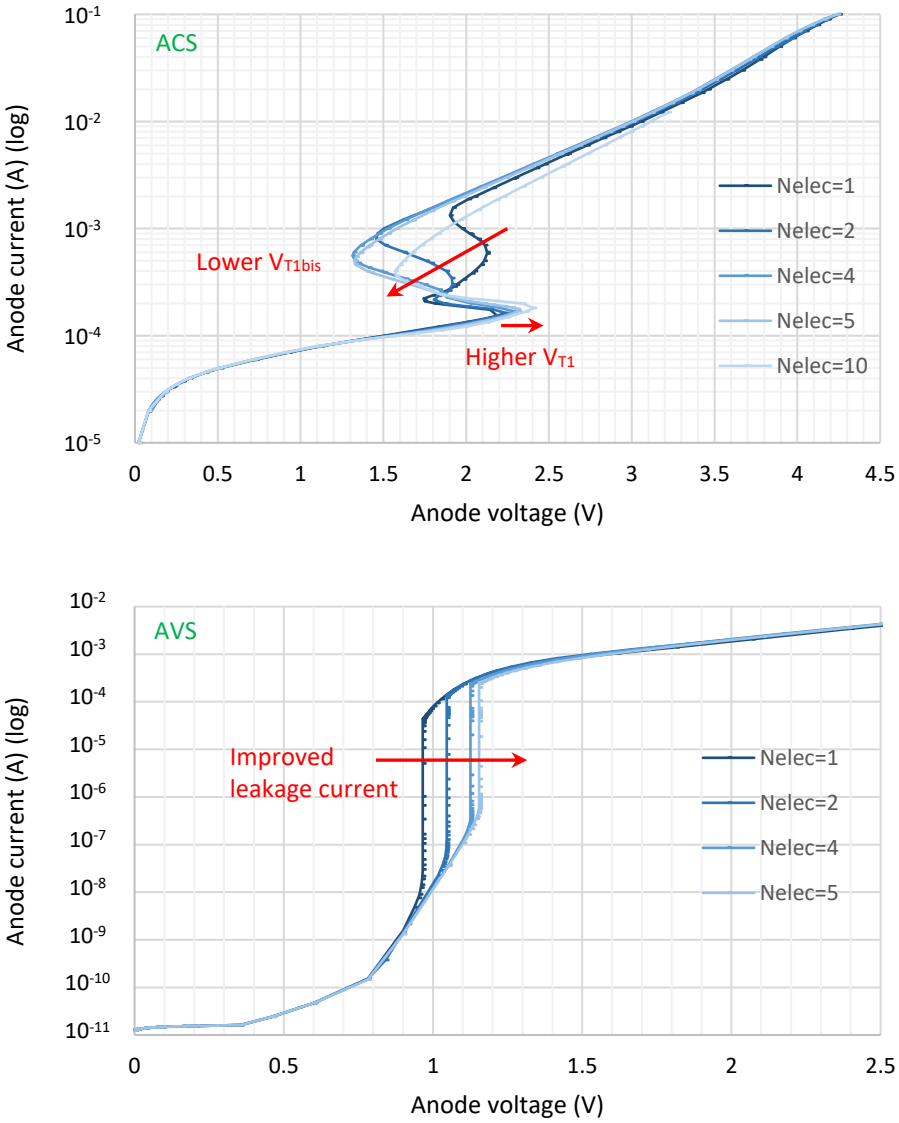


Figure 108: I-V TCAD simulation of a GDNMOS with grounded NMOS gate, N-LDD doping in the drain, the same total silicide cover but different number of electrodes on the drain. Electrodes are connected to the gate of the diode. Top: ACS. Bottom: AVS.

The explanation is that each electrode's influence on the potential in the drain extends until 1 µm after it is cut. The effect of the electrode on the part that has no silicide can be seen on the energy bands (Figure 109). If the number of electrode increases, the number of "borders" between electrode and no-electrode regions increases. If the width of the region with no electrode is too small, this region "sees" the influence of the two electrodes that are surrounding it. Therefore, even if the total width of electrode is always 5 µm, the device acts as if the total width of electrodes is more than 5 µm "electrostatically speaking".

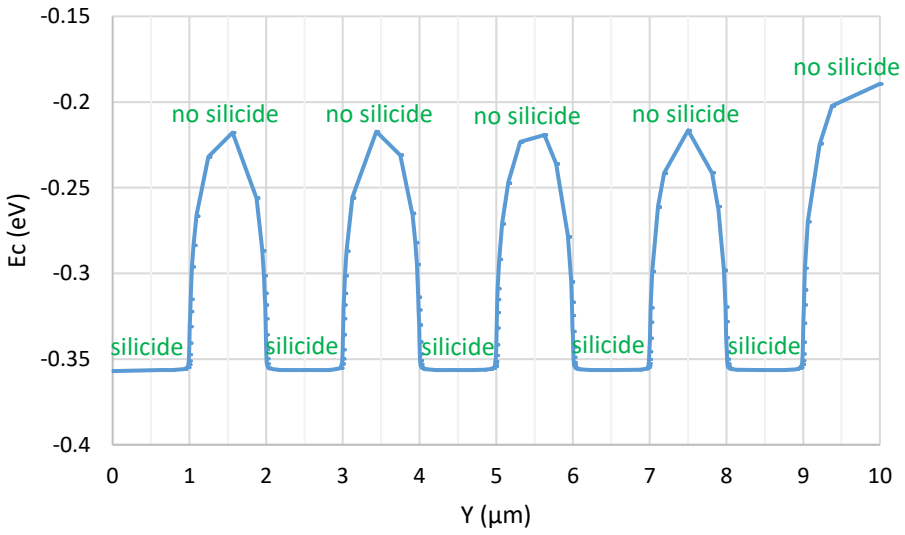


Figure 109: Conduction band in the drain along the width of the structure at I = 5.10^{-4} A for 5 electrodes being connected to the gate of the diode. Note that the region between 9 and 10 µm only sees the influence of one electrode, and that is why the energy band is different from the other regions without electrodes.

Note that when the electrodes are connected to the gate of the diode, a higher number of electrodes also leads to a lower V_{T1bis} (Figure 108). To explain this, an additional study would be needed. All we know is that the first snap-back is due to the SCR parts of the device with a floating drain (no electrode, so no connection between the drain and the gate of the diode, and no silicide, so less recombination of holes). The second snap-back is due to the other parts of the device triggering (with an electrode, silicide, and the connection between the drain and the diode gate). A higher number of electrodes means a wider total surface where there is no electrode i.e. no silicide, but where the presence of the nearby electrode still has an influence, with the diode gate being biased.

As a conclusion for this study, it is always possible to tune the number of electrodes, their size and their placement on the drain, in order to shift the ACS and AVS curves, which will make the device easier to fit some ESD design windows.

To summarize the simulations about silicide, low and very low-voltage solutions are available with no extra cost: the high-voltage GDxMOS layouts can be kept, and only the NOSD and SBLK masks have to be added. The structure stays compact and there is no or minor silicon area impact.

To conclude this chapter, GDNMOS and GDBIMOS were investigated through 3D electro-thermal TCAD simulations and measurements. Both devices show improvements in ESD performance (robustness, tunable to fit the ESD design window, and low leakage in the proper window).

For the robustness, it is advised to connect the gate of the diode to a resistor. A higher resistor is better. To improve further the robustness, the gate of the NMOS should also be connected to this resistor.

In order to fit into the desired ESD design window, connecting the gate of the NMOS to a resistor helps decreasing the V_{T1}. Connecting the back gate to this resistor helps to further decrease V_{T1}. A BIMOS can be merged into the GDNMOS to decrease the R_{ON}. Decreasing the doping level in the merged region is mandatory to get a low-voltage protection. Silicide removal is necessary to prevent recombination in the low doped drain. Partial silicide removal is useful in order to put contacts on the drain. Partial silicide can also be useful to shift the V_{T1} and leakage current of the device.

All the solutions are designed with standard process steps with no over cost. The protections are flexible for being able to cope with dedicated applications.

Chapter 3: BIMOS matrices

"Science means constantly walking a tightrope between blind faith and curiosity;
between expertise and creativity;
between bias and openness;
between experience and epiphany;
between ambition and passion;
and between arrogance and conviction – in short,
between an old today and a new tomorrow."
Heinrich Rohrer.

I. BIMOS dot topology

1. 1D BIMOS dot

Different topologies of BIMOS devices are available in thin-film 28 nm FD-SOI (Figure 110): (i) the so-called "classical" BIMOS [111], where the gate is a straight line, but the active is cut, in order to separate better the body contact from the source; and (ii) the T-gate BIMOS [112] [113] [114].

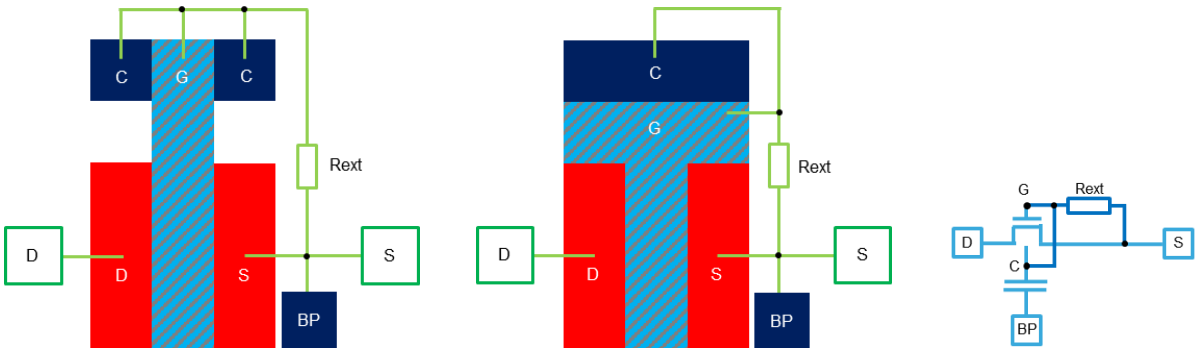


Figure 110: Left: classical BIMOS topology (external body contact with active bridge). Middle: T-gate BIMOS topology (body contact with polysilicon gate mask). The hatched region is used for the gate. The colors correspond to the doping layers (light blue is Pint in the channel, dark blue is P^+ and red is N^+). D, S, C, G, BP and Rext respectively stand for Drain, Source, Body Contact, Gate, Back-plane and external Resistor. Note that the T-gate BIMOS comprises additional parasitic diodes (between the body contact and the drain and between the body contact and the source) with respect to the classical BIMOS. Right: electrical schematic of a BIMOS, corresponding to both classical and T-gate BIMOS topologies.

The classical BIMOS has a long gate length (108 nm) (Figure 111) and its matrix topology is not optimized. The T-gate BIMOS has a small gate length (48 nm) and is difficult to port in matrix (except if we consider a matrix of NMOS with a peripheral ring of body contact, like the matrix which will be presented in the Chapter 3, II section). This is why a new design is proposed - the BIMOS dot - which offers a small gate length and is efficiently portable in a matrix, in addition of showing improved ESD performances.

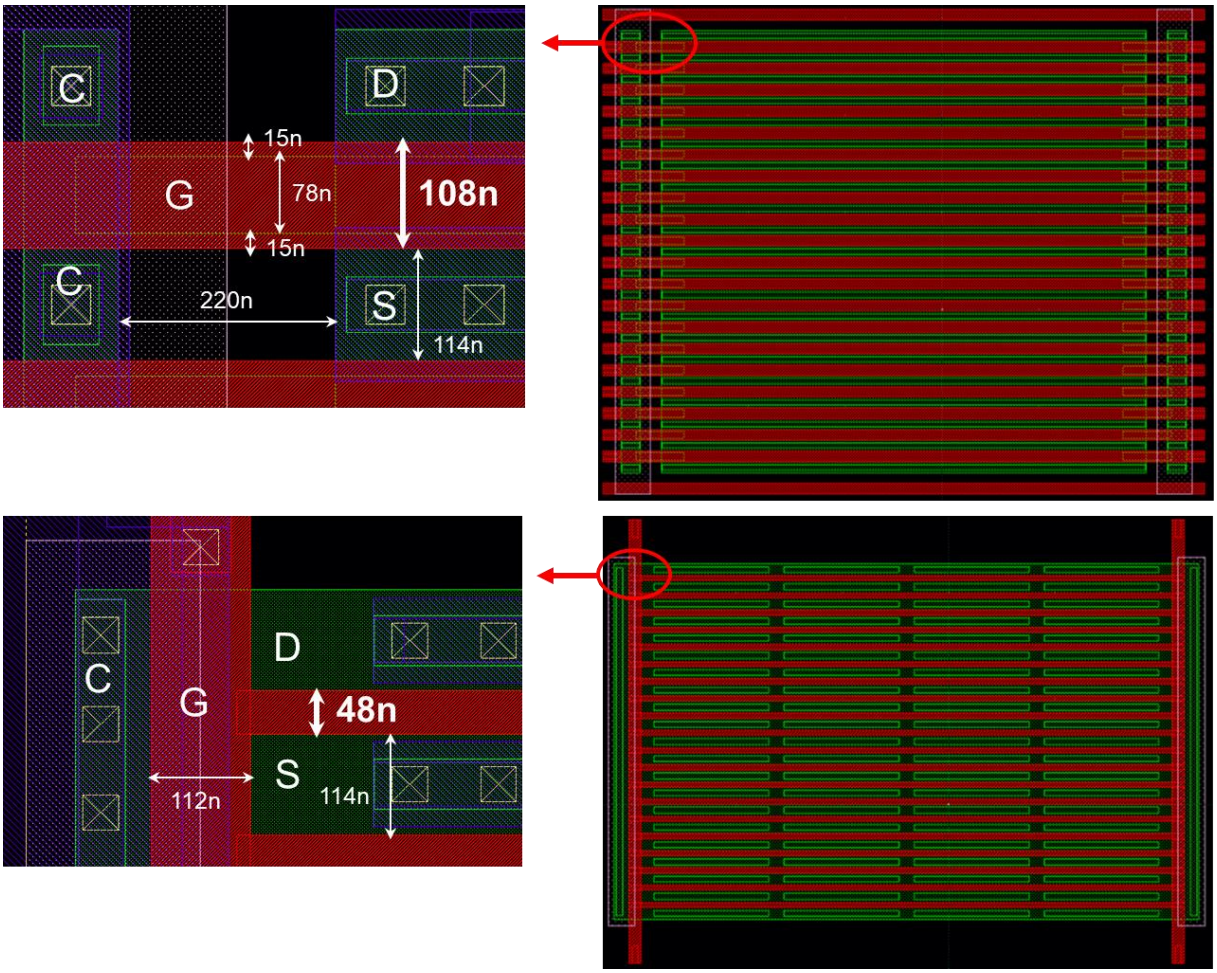


Figure 111: Top: Classical BIMOS. Bottom: T-gate BIMOS. Right: multi-finger layout (20 fingers of 5 µm). Left: zoom and dimensions. Note that the classical BIMOS is very constrained in terms of dimensions because of the bridge of active (to link the body contacts and the source and drain) that must be covered by the gate. D, S, C and G respectively stand for Drain, Source, Body Contact and Gate.

The BIMOS dot device consists in placing the body contact in the middle of the gate (Figure 112). An improvement is to reduce the gate length far from the body contact, so that the NMOS part of the BIMOS benefits from a better conduction of current. The minimal gate length of the BIMOS dot with a thick gate is 338 nm, and 48 nm for a BIMOS dot with a thin gate (Figure 113).

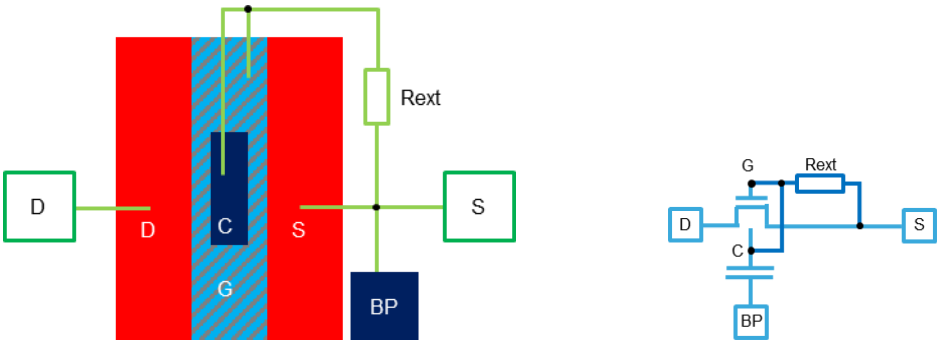


Figure 112: Left: BIMOS dot topology. Right: Reminder of the BIMOS schematic.

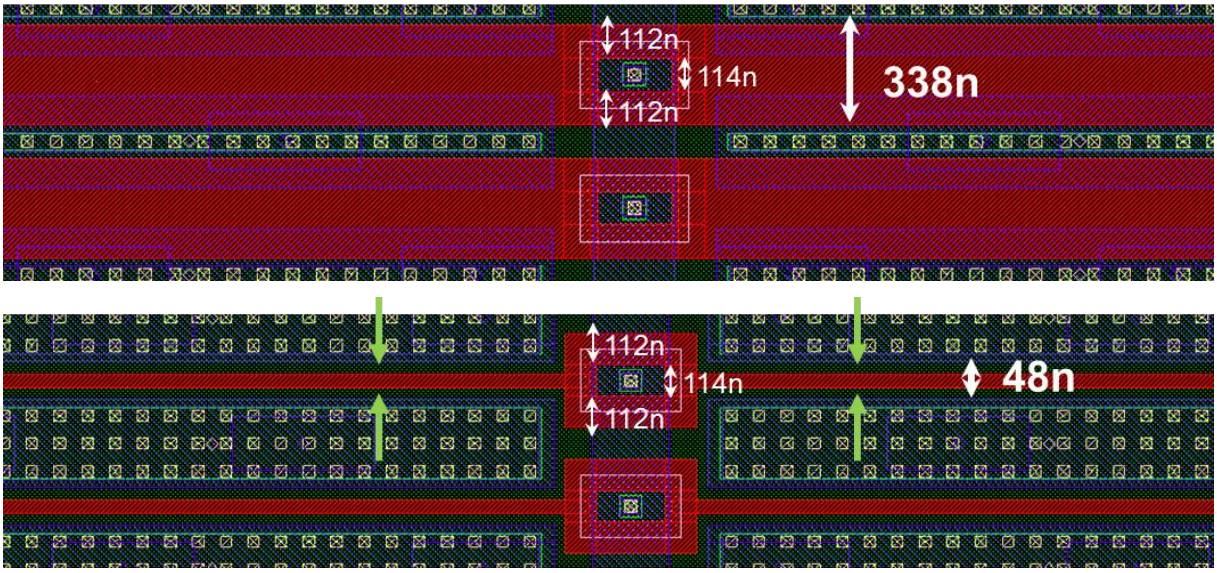


Figure 113: Top: Thick gate BIMOS dot topology. Bottom: thin gate BIMOS dot topology. Two fingers are displayed.

Three devices were compared (Figure 114): the classical BIMOS, and BIMOS dots with a thick gate and with one or two body contacts of 500 nm (each). The gate length of all the devices is 300 nm, for comparison purpose. The length of one finger is 5 µm for each device (dot comprised), for a total width of 100 µm (20 fingers). An external polysilicon resistor was connecting the gates and body contacts. The back gate was grounded.

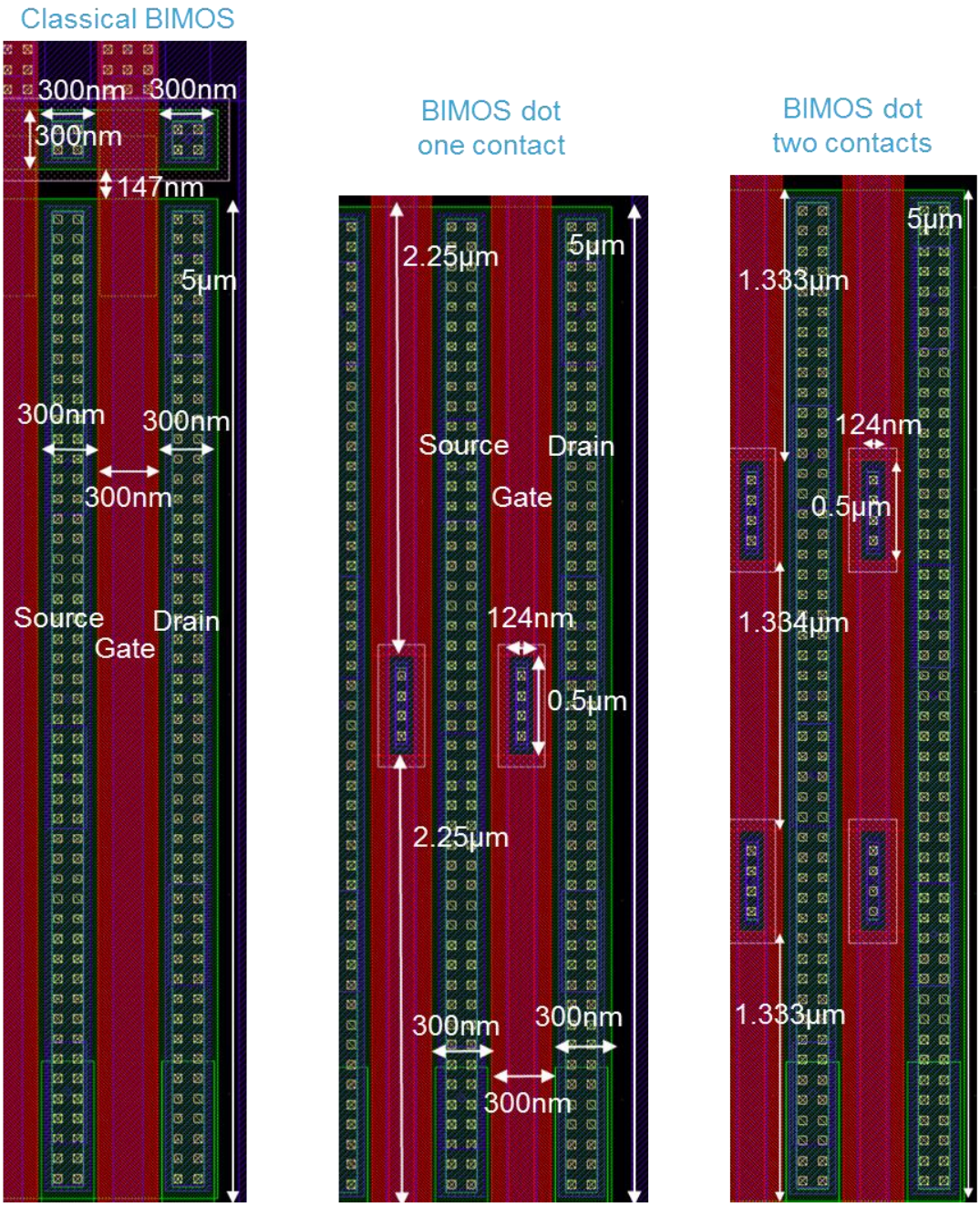


Figure 114: Layout of the three measured devices.

In the TLP measurements, the trigger voltage is similar (~ 4.2 V) but in the VF-TLP measurements, the trigger voltage decreases with the number of dots (~ 3.7 V for the classical BIMOS, ~ 3.5 V for the BIMOS dot with one body contact and ~ 3.4 V for the BIMOS dot with two body contacts) (Figure 115). This suggests that the BIMOS dot topologies are useful for CDM protection.

The trigger voltage decreases between the TLP and VF-TLP measurement because the rise time is shorter in VF-TLP (300 ps compared to 10 ns). For a given voltage pulse, a shorter rise time will take the gate node to a higher voltage through the parasitic capacitance between the drain and the gate (since the RC trigger circuit with the capacitor connected to the node with the signal and the grounded resistor is a high pass filter: when the entry signal

is rapid, the capacitor allows to increase the gate voltage very quickly, and the carriers do not have time to be evacuated through the resistor). As a consequence, the trigger voltage (linked to the threshold voltage of the gate of the NMOS) is reached for a lower value of drain voltage.

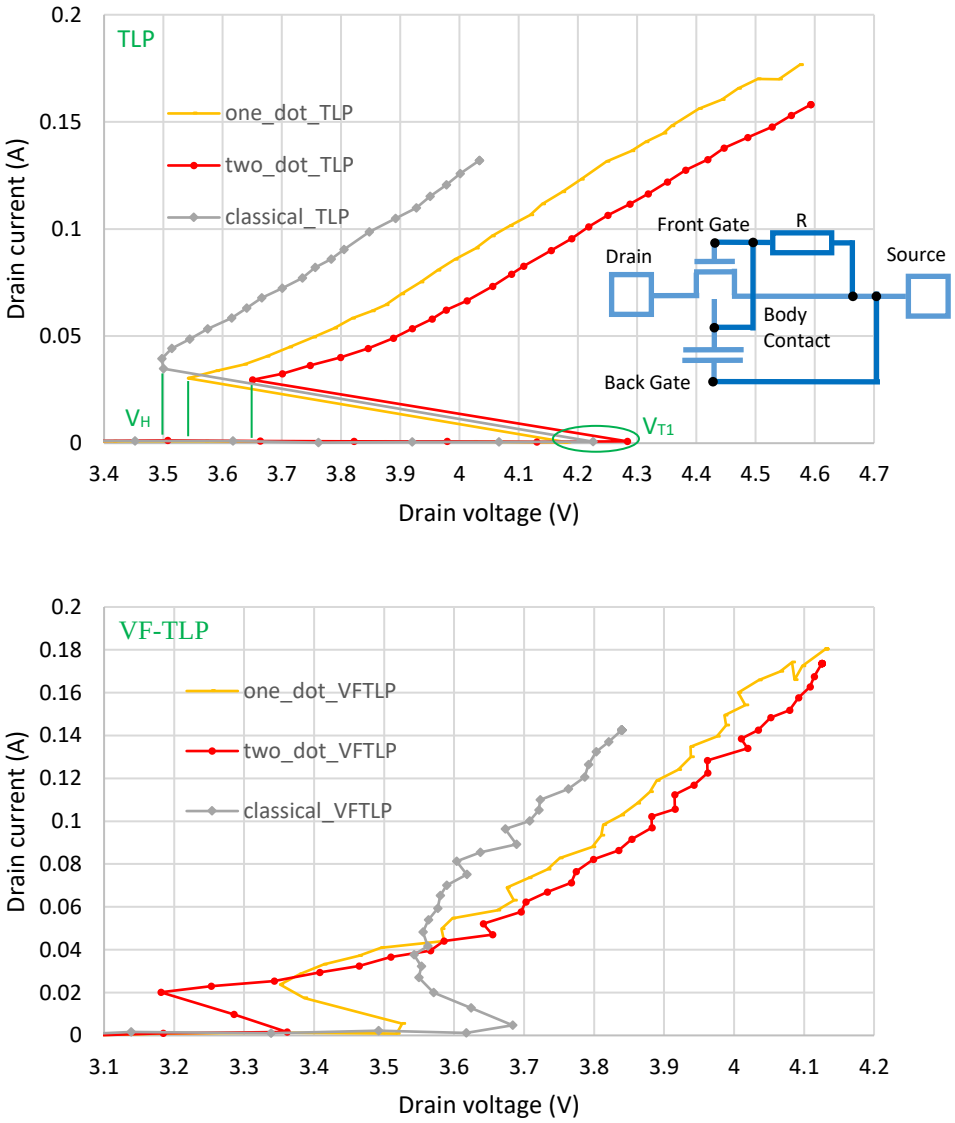


Figure 115: Comparison of the I-V characteristics of the classical BIMOS, the BIMOS dot with one body contact and the BIMOS dot with two body contacts. Top: TLP (100 ns duration and 10 ns rise time). Bottom: VF-TLP (5 ns duration and 300 ps rise time).

The difference of trigger voltage between the topologies could be due to different phenomena: (i) the difference of conduction currents in the devices, (ii) the difference in parasitic capacitance and resistance between the topologies, and (iii) the difference of parasitic diode.

TCAD ACS simulations have been performed to catch the different conduction currents in the BIMOS dot devices. Figure 116 shows the Drain current versus drain voltage

of the structures. The ACS behavior is a bit different than in the TLP curves (comparison between Figure 115 and Figure 116), indeed the models used in the simulations were not calibrated. Nevertheless, TCAD simulations can still provide valuable insights about the devices. Figure 117 shows the drain voltage and gate voltage versus time. Note that the drain voltage versus time curve in Figure 117 is analogous to the drain current versus drain voltage curve in Figure 116 since the ACS is a current ramp in 100 ns. When the drain voltage is increasing rapidly, the parasitic capacitance C_{DG} between the drain and the gate increases the gate voltage. This is what happens before the first snap-back of the device (slightly after 1.10^{-10} s of simulation): the voltage increase in the drain node provokes a voltage increase in the gate node. The change in voltage versus time is still rapid until 2.10^{-10} s of simulation ($R = dU/dI \propto dU/dt$ and R ~ 30 kΩ before the snap-back. As a matter of comparison, R_{ON} = 18 Ω). This is where the parasitic effects play a role. After this first snap-back, the parasitic capacitances do not play a significant role anymore.

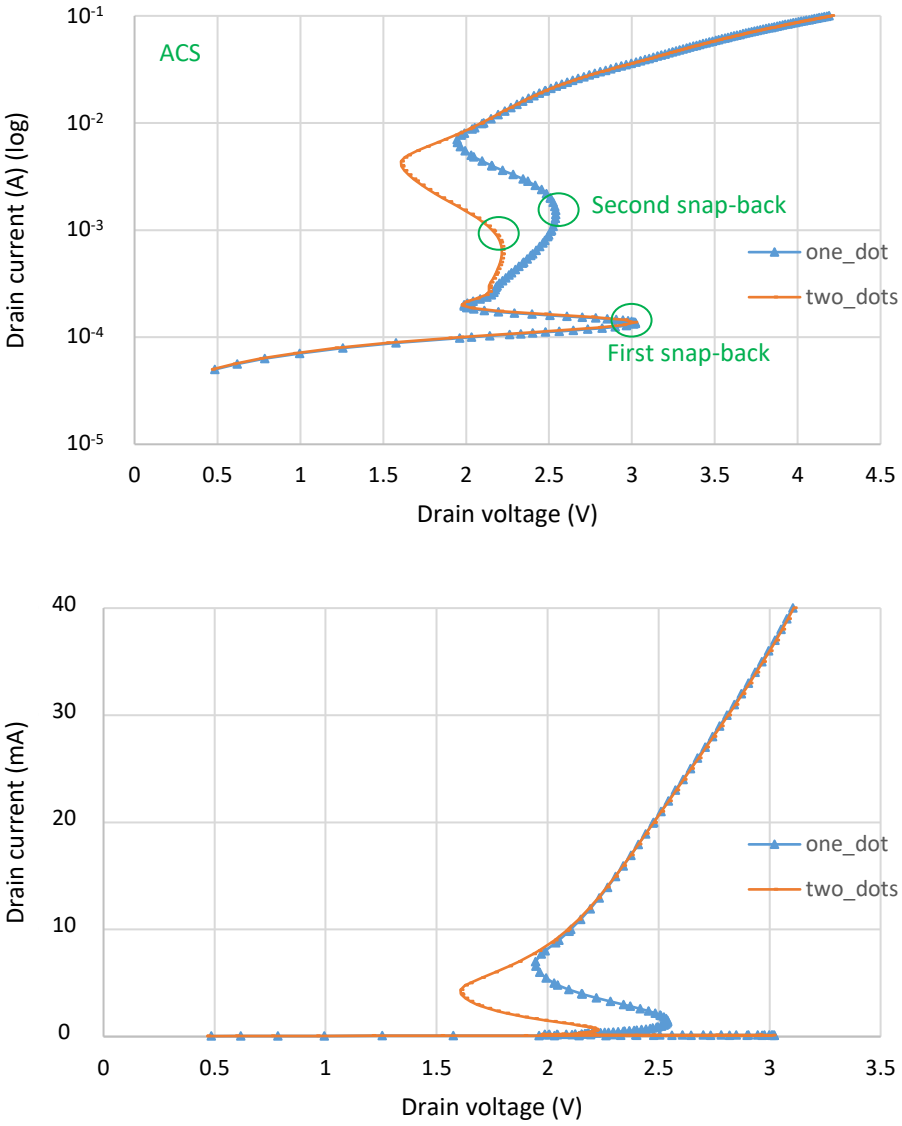


Figure 116: TCAD ACS simulation of the BIMOS with one and two dots. Drain current versus drain voltage. Top: logarithmic scale; Bottom: linear scale.

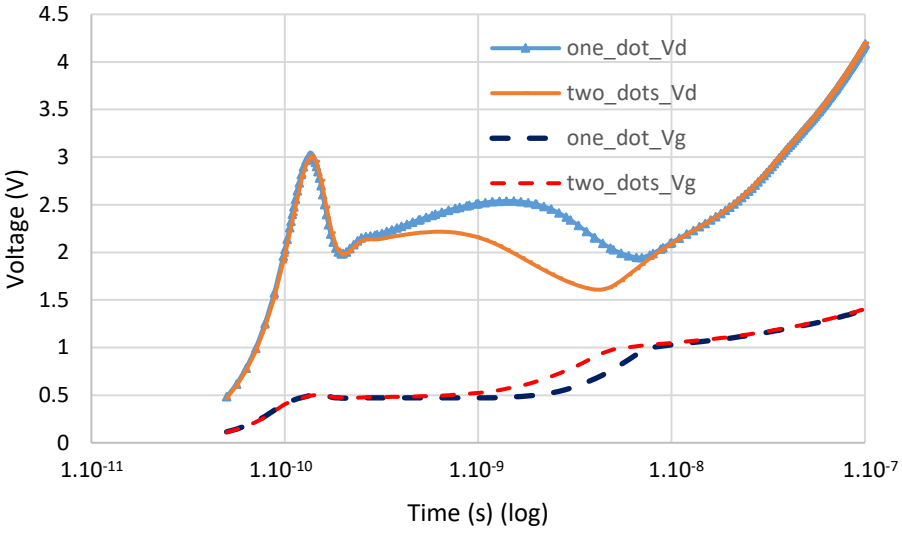


Figure 117: TCAD ACS simulation of the BIMOS with one and two dots. Drain voltage and gate voltage versus time.

Figure 118 shows different currents in the BIMOS devices versus time. Note that currents have been normalized to plot the curves in a logarithmic current scale. In reality, all source currents are negative, which means that electrons are coming from the source and flowing toward the rest of the device, and holes are coming from the device and flowing toward the source. All contacts' electron currents are positive which means that electrons are coming from the device and flowing toward the body contacts. All contacts' hole currents are positive until T = 2.10^{-10} s which means that holes are flowing from the body contacts to the rest of the device, and then the contacts' hole currents are negative which means that the holes are going from the device toward the body contacts.

It can be seen in Figure 118 that the first snap-back is due to source current conduction. It is coherent with the fact that the gate voltage was raised until the first snap-back of the device (Figure 117). A voltage on the gate activates subthreshold MOS conduction inside the device. With an increased drain current and drain voltage, impact ionization in the device near the drain increases, thus providing electrons and holes. After this first snap-back, holes that are created thanks to impact ionization are flowing through the body contact. Holes tend to go where the voltage is the lowest and/or where the doping is favorable to them, that is why they are attracted by the source and by the body contacts. It can be seen however that more holes are flowing toward the source than toward the body contacts - there is at least one decade between the hole current of the source and the one of the body contact: this means that about 10% of holes available from impact ionization contribute to the hole current in the body contact, and 90% are wasted in the source. This body contact current is very important because it helps raising the gate voltage. More and more holes are created thanks to impact ionization, thus slightly raising the gate voltage, until the second snap-back, which corresponds to the threshold of the NMOS (for a given hole current in the body contact). Then electrons start to flow through the body contact, thus raising even more the voltage of the body contact thanks to the external resistor, and the whole NMOS conduction is active.

The difference between devices with one dot and two dots lies in the fact that more holes can be caught by the body contacts (between the two snap-backs of the device), because the dot (which attracts holes) is closer if there are two dots in the structure than only one.

This conduction current theory explains well why the BIMOS with two dots triggers before the BIMOS with one dot (second snap-back in Figure 116). However, it can be seen in Figure 117 that the gate voltage of the BIMOS with one and with two dots are superposed before the first snap-back (and the I_D-V_D curves are superposed before the first snap-back). This corresponds to the moment where parasitic capacitive coupling is playing a role, meaning that *a priori*, there is no difference of parasitic capacitance between the BIMOS with one and two dots. Yet measured difference in trigger voltage increases from TLP to VF-TLP (Figure 115) between the BIMOS with one dot and the BIMOS with two dots, and a higher value of capacitor between the drain and the gate (in the case of the BIMOS with two dots, with respect to the BIMOS with one dot) would explain this increasing difference in trigger voltage. Indeed, a higher value of C_{DG} capacitor induces a higher voltage on the gate for a given drain voltage. Thus, the trigger voltage of the BIMOS can be reached for a lower drain voltage in the case of a higher parasitic capacitance. This difference in gate voltage (for a given drain voltage) is more important in case of rapid rise time. This would explain why a significant difference is observed between one or two dots in the VF-TLP curve and not in the TLP one.

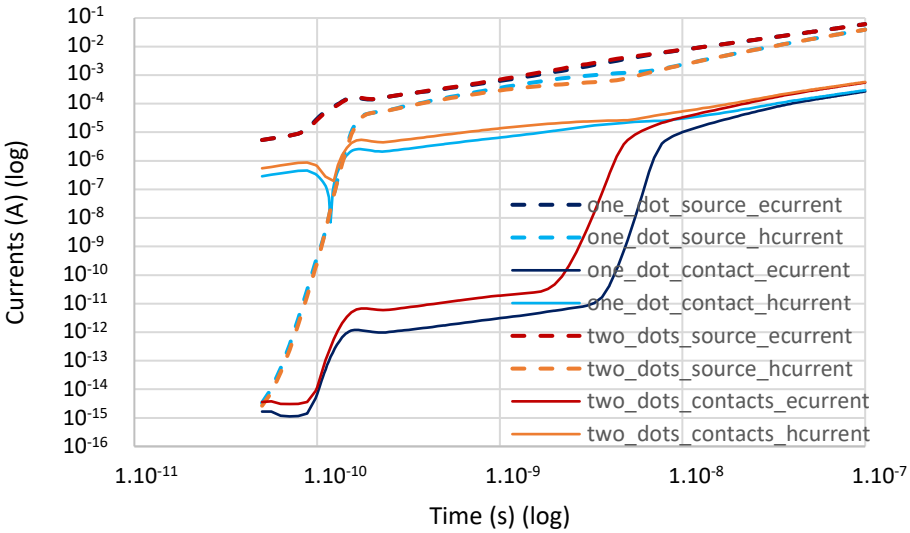


Figure 118: TCAD ACS simulation. Currents in the BIMOS with one dot (blue curves) and the BIMOS with two dots (red curves) versus time. Dashed lines correspond to currents in the source and normal lines to currents in the body contact (for the BIMOS with two dots, currents flowing in the contact 1 and 2 have been summed). Dark colors correspond to electron currents and light colors to hole currents.

Why would the BIMOS with two dots exhibit a larger parasitic capacitance than the BIMOS with one dot? A review of different parasitics present in the devices is carried out.

In our hypothesis, the value of the parasitic capacitance between the drain and the gate is higher in the case of two dots, thanks to the increase in parasitic capacitance between the body contact and the drain through the N^+/P junction (parasitic diode), since the space charge zone acts like a capacitance. The surface of this junction (drain to body contact) is twice bigger in the case of two dots, therefore the junction capacitor is expected to be approximately two times higher. The classical BIMOS does not benefit from this additional direct capacitor between the drain and the body contact (by topology), which is why its trigger voltage is even higher than the one of the BIMOS with one dot in VF-TLP.

There is also a parasitic capacitor due to the drain-to-channel junction. The channel is linked to the body contact because they are physically touching each other and they are built with the same type of carriers (P). This parasitic capacitor would play a role in both cases (one or two dots). The channel being Pint doped (intrinsic), it is highly resistive, so the path drain – channel – body contact comprises a junction capacitor followed by a high resistance (Figure 119). A lower channel resistor (in the case of two dots) allows the device to trigger more uniformly, *i.e.* the whole finger is involved at a given time, while a higher channel resistor can induce voltage differences and time delays in the channel.

Figure 120 shows the influence of the rise time on the three devices (for a TLP duration of 100 ns). The BIMOS with two dots is more sensitive to the rise time than the BIMOS with one dot, which in turn is more sensitive than the classical BIMOS (sensitive means that their trigger voltage and holding voltage change significantly with the change in rise time). The shorter the rise time, the lower the holding voltage. The shift of apparent holding voltage with the pulse risetime in Figure 120 is just a consequence of the shift of trigger voltage with the rise time and the finite load line of the pulser. This is coherent with the capacitor explanation for the differences between the devices in TLP and/or VF-TLP in Figure 115, based on the difference of rise time between TLP and VF-TLP measurements.

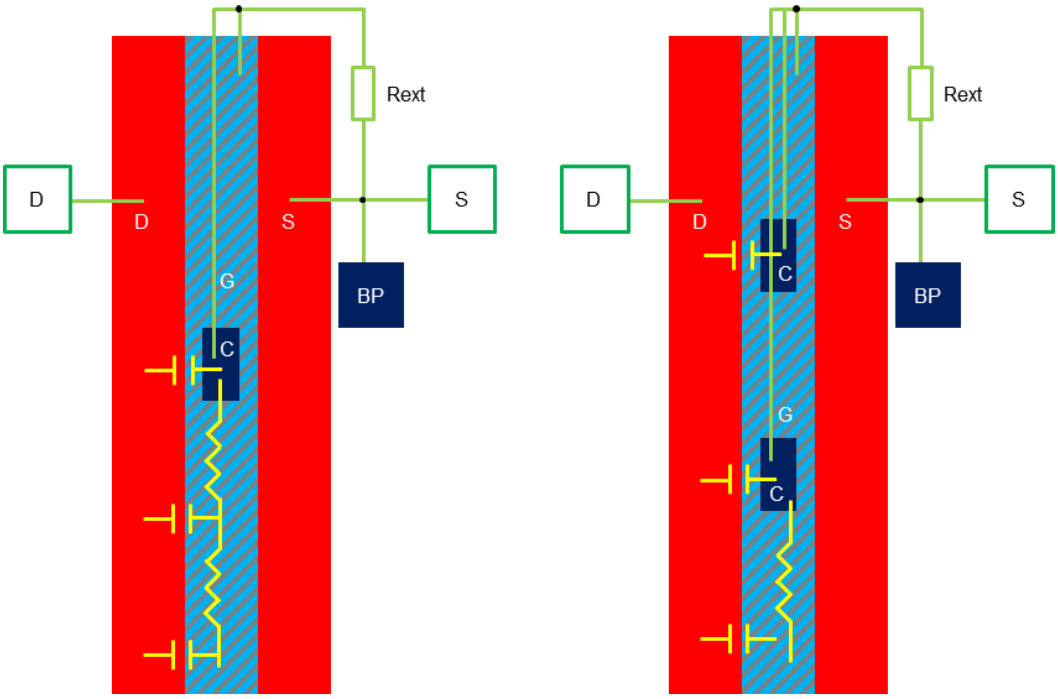


Figure 119: BIMOS with one (left) or two (right) dots with some of their junction parasitic capacitances and some of the channel parasitic resistors (in yellow).

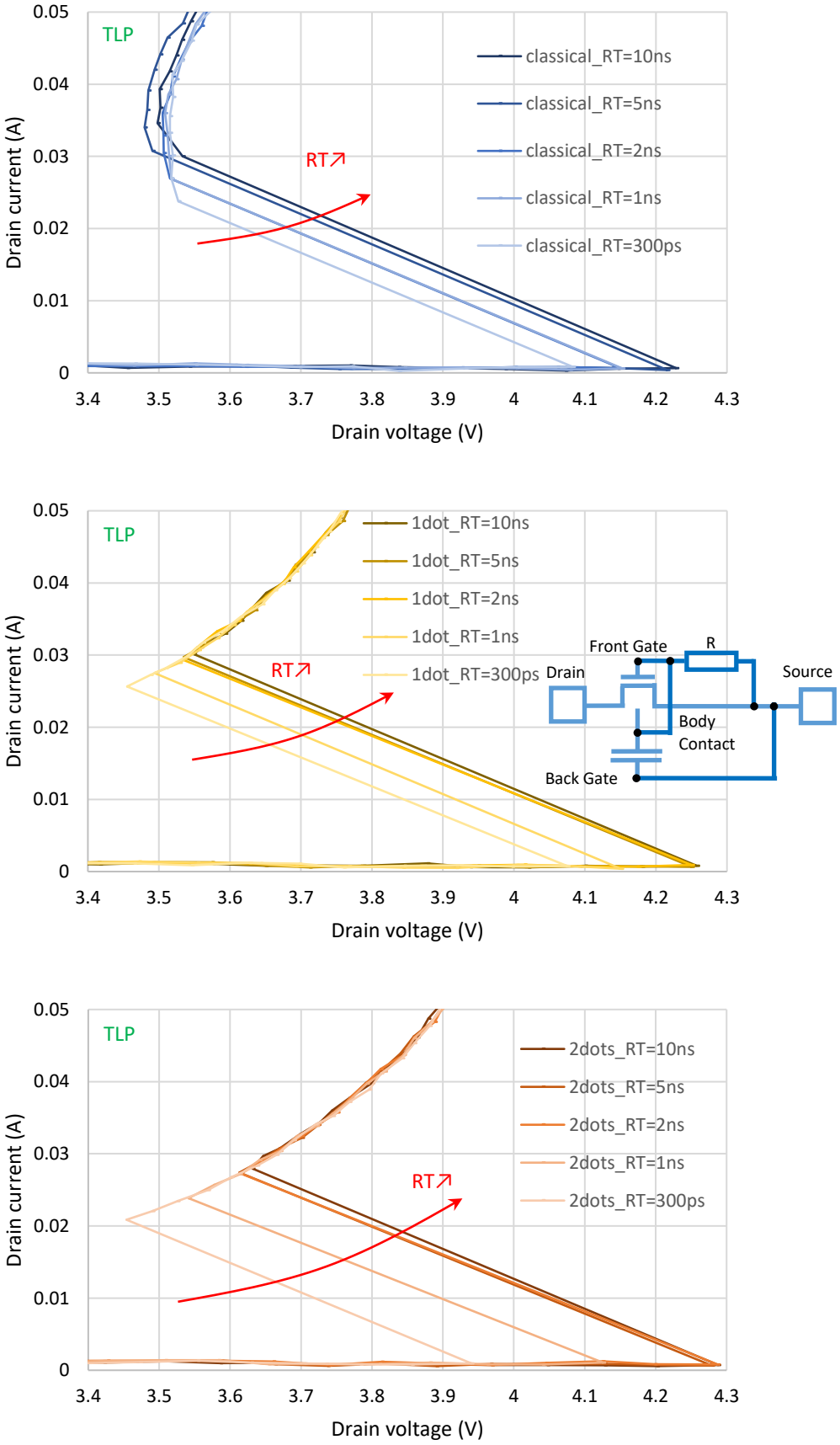


Figure 120: TLP measurement (100 ns duration) for different rise times. Top: Classical BIMOS. Middle: BIMOS with one dot. Bottom: BIMOS with two dots.

The last element to be considered (in the eventual modification of the trigger voltage of the devices due to their topology difference) is the parasitic diode $P^+/P/N^+$ that is situated between the body contact and the source. This diode is forward biased; therefore, it induces a hole current from the body contact to the source. This hole current reduces the total hole current that goes from the device to the body contact and that contributes to the rising of the body contact voltage. Therefore, this parasitic diode (wider in the case of the BIMOS dot) is counterbalancing the increase in gate voltage. The gate voltage still increases with time, but less than if there was no parasitic diode.

The classical BIMOS is probably failing for a significantly lower I_{T2} and V_{T2} than the BIMOS dot structure (Figure 115) because its topology is less robust given that the silicon is bended for building the bridge.

The 28 nm FD-SOI UTBB technology allows us to use the back-plane to tune the threshold voltage of the front gate, so it is interesting to see the effects of this feature on the ESD devices. Back-plane biasing was investigated on the BIMOS with two dots in Figure 121. The trigger voltage is reduced for a positive back-plane biasing, which is explained by the back-plane acting like a back gate. The same tendencies (voltage shift in the I_D-V_D TLP characterization with the back-plane biasing) are observed on the classical BIMOS and on the BIMOS with one dot devices.

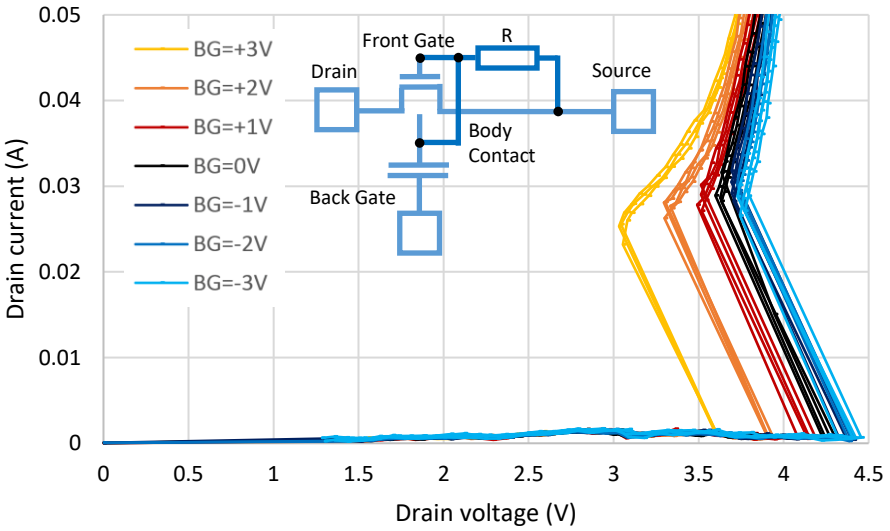


Figure 121: TLP measurements (100 ns duration and 10 ns rise time) of the BIMOS with two dots for different constant back-plane biasing.

DC characterizations were performed to investigate the leakage current in the devices. The drain voltage was swept and the resulting drain current was measured. Measurements were stopped before the triggering of the devices (Figure 122). No significant difference is observed between the devices. The reason is that the parasitic capacitors do not play a role anymore in quasi-static operation. The trigger voltage would however probably change from one device to another, because of the differences in conduction currents.

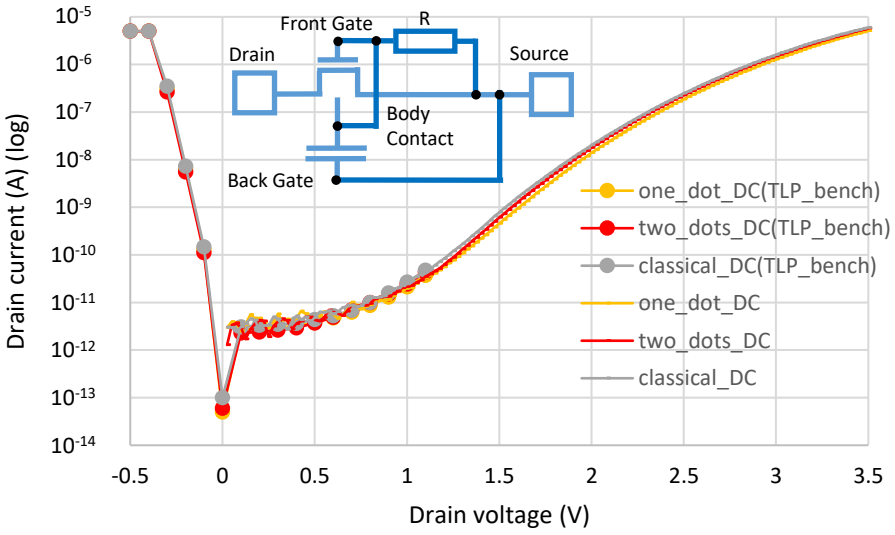


Figure 122: DC measurements: a voltage sweep is performed on the drain and the resulting drain current is measured. DC measurements have also been done with the TLP bench.

DC characterizations were also performed on the same devices except that their gate (connected to the body contact) was left floating instead of being connected to a resistor. This time the trigger voltage of the devices happens in the 3.5 V window of acquisition of the measurements (Figure 123). The floating gate acts as if an infinite resistor was plugged between the gate and the source. With the increase in voltage on the drain, the leakage current in the channel increases. Holes are attracted in the body contact and cannot be evacuated through the external resistor. A voltage is established on the front gate and the device triggers significantly earlier than if there was an external resistor between the gate-body contact and the source. The BIMOS with two dots triggers before the BIMOS with one dot, which triggers before the classical BIMOS. Indeed, the access to the body contact is more difficult for carriers in the case of the classical BIMOS, because it is situated at the end of the finger and the channel is resistive. Once devices have triggered the channel is much less resistive and the position of the body contact along the finger is not important anymore. The classical BIMOS can drive more current when it is ON because its conduction width is 5 µm (to be compared with 4.5 and 4 µm for the BIMOS with one dot and two dots respectively).

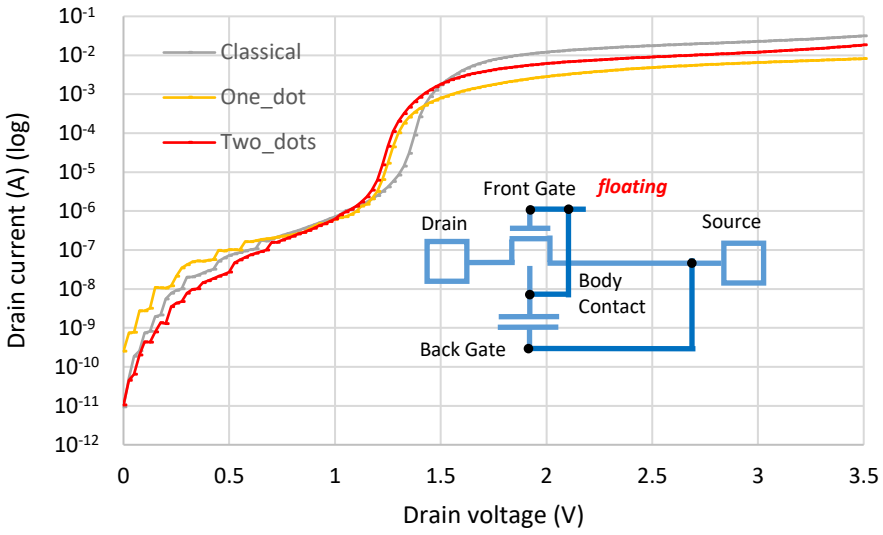


Figure 123: DC measurements: a voltage sweep is performed on the drain and the resulting drain current is measured. The front gate of the devices is connected to the body contacts and this node is left floating.

To conclude this experiment, TLP measurements comparing a BIMOS dot and the classical BIMOS show that the BIMOS dot has the same trigger voltage V_{T1}, a smaller holding current I_H and a higher failure current I_{T2} (Table 2). The latter is very important, because it means that the device has a better robustness and current propagation. VF-TLP shows that the BIMOS dot has a smaller trigger voltage V_{T1} (particularly in the case of two dots, probably because it has more capacitance) and a higher failure current I_{T2}. This study has to be pursued regarding fabrication variability. DC characterizations show a leakage current of less than 10^{-10} A at 1 V with a resistor connected to the gate and the body contact, which is compliant with the protection of GO1 transistors. The leakage increases up to 10^{-6} A at 1 V if the gate and body contact are tied together but left floating - which is not compatible with a V_{DD} = 1 V technology and shows the importance of having an external resistor.

		V_{T1} (V)	I_{T1} (mA)	V_H (V)	I_H (mA)	V_{T2} (V)	I_{T2} (mA)	R_{ON} (Ω)
TLP	Classical	4.23	0.73	3.50	39	4.03	132	5.9
	1 dot	4.18	0.36	3.54	30	4.57	177	6.5
	2 dots	4.28	0.92	3.65	30	4.59	158	6.8
VF-TLP	Classical	3.68	4.75	3.55	27	3.84	143	3.4
	1 dot	3.53	5.47	3.35	24	4.13	180	3.9
	2 dots	3.36	1.54	3.18	20	4.13	174	3.6

		V_{T1}	I_{T1}	V_H	I_H	V_{T2}	I_{T2}	R_{ON}
TLP	1 dot	-1%	-51%	+1%	-23%	+13%	+34%	+10%
	2 dots	+1%	+26%	+4%	-25%	+14%	+20%	+15%
VF-TLP	1 dot	-4%	+15%	-6%	-12%	+8%	+27%	+15%
	2 dots	-9%	-68%	-10%	-25%	+7%	+22%	+4%

Smaller V_{T1}

Higher I_{T2}

Table 2: Top: Values extracted from the Figure 115. Bottom: comparison with the classical BIMOS. The most important is the smaller trigger voltage in VF-TLP of the BIMOS dot with respect to the classical BIMOS, and its higher failure current in TLP and in VF-TLP. Note that the higher R_{ON} is due to the silicon area taken by the body contact. Future study could compare devices with the same width of conduction.

2. Matrix of BIMOS dot

Let us compare four optimized devices in 28 nm FD-SOI: the classical BIMOS, the T-gate BIMOS, the BIMOS dot with a thick gate and the one with a thin gate. The minimal dimensions have been selected in order to be compliant with the design rules and to achieve the maximum performance of each device (Figure 124). The associated multi-finger layout is drawn in Figure 125: each device has 20 fingers for a total width of 100 µm. The area of the BIMOS dot is much larger than the total area of the classical and the T-gate BIMOS. The area concern can be mitigated thanks to the matrix topologies.

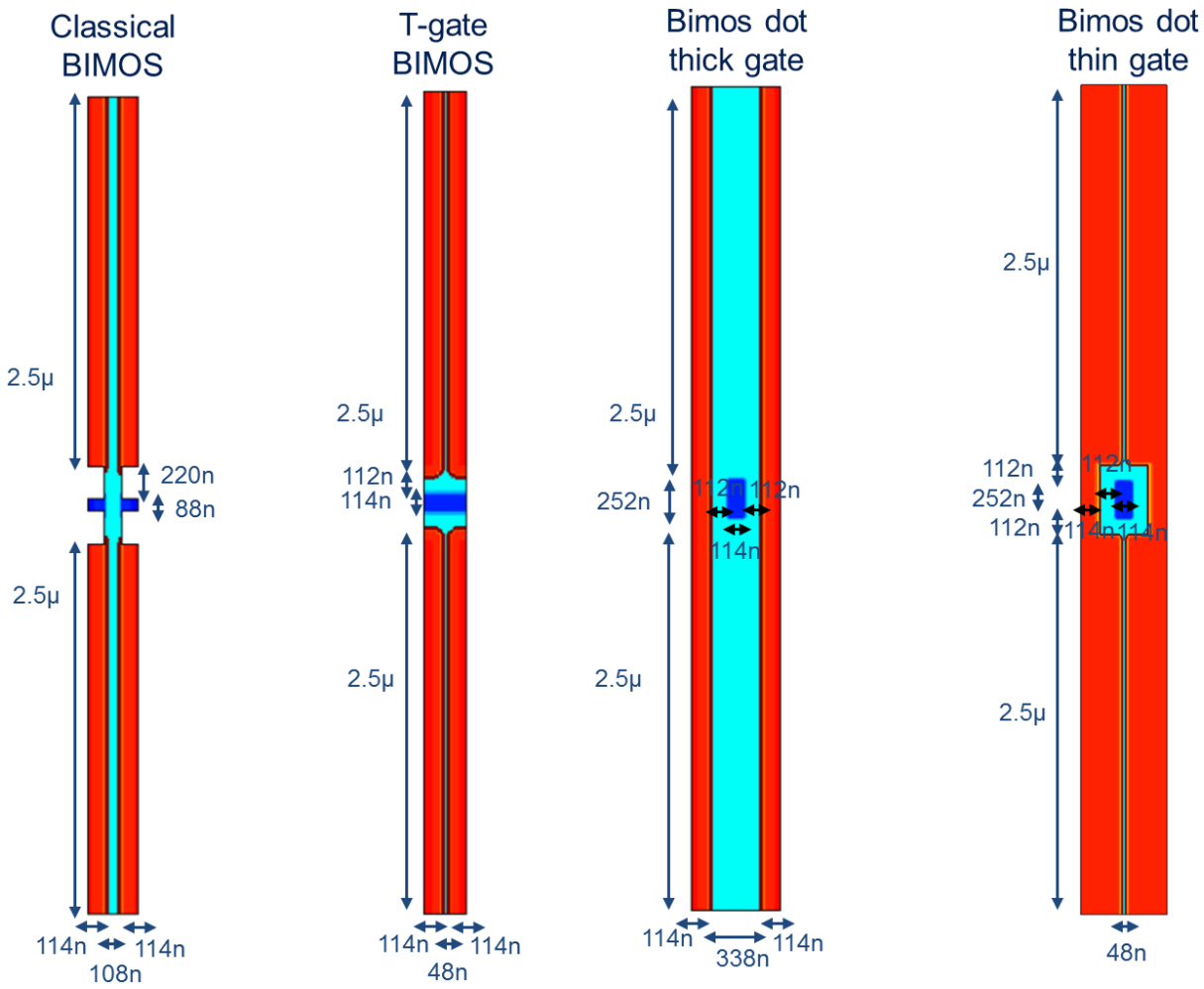


Figure 124: TCAD top view of the four compared devices with their dimensions.

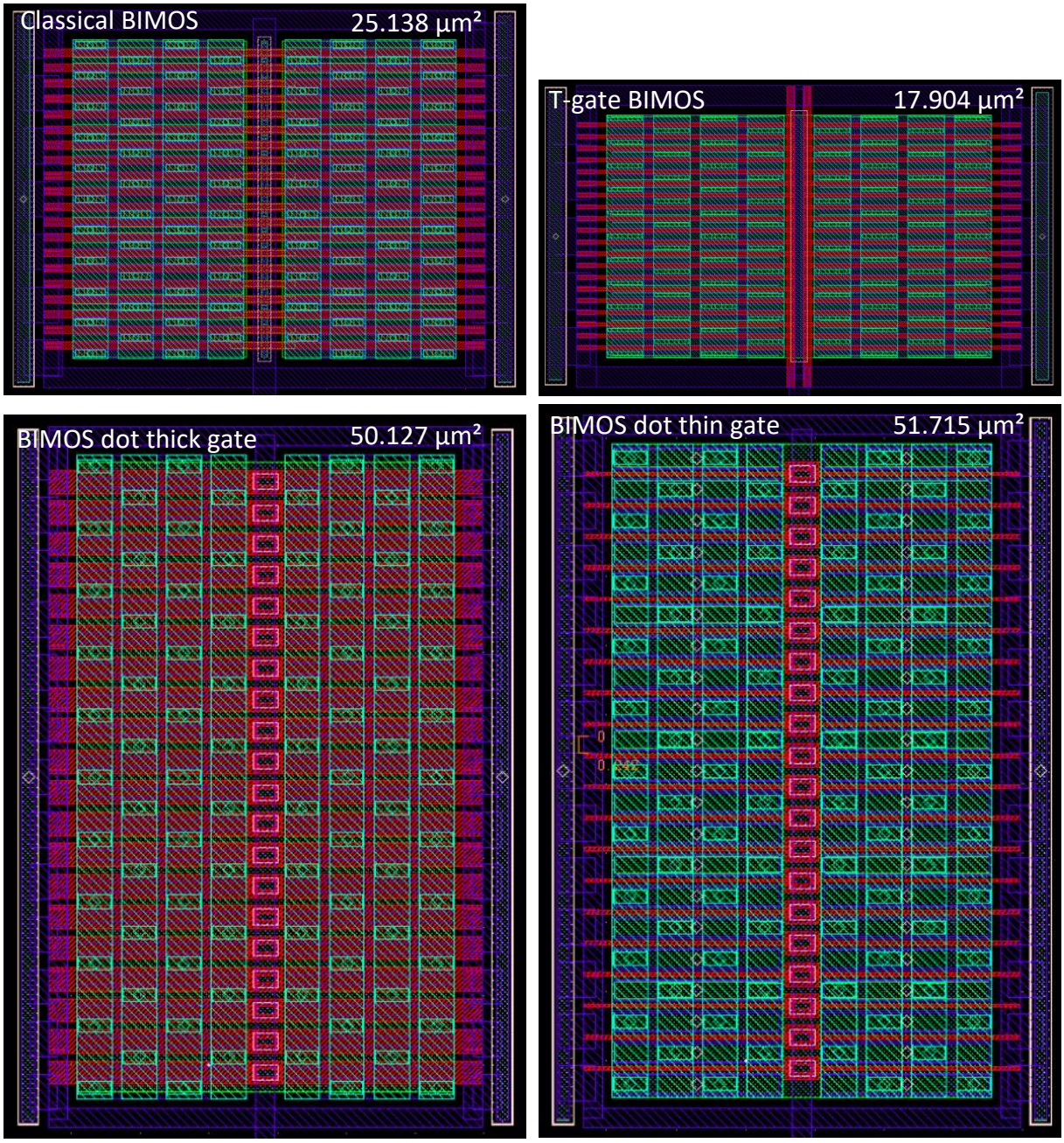


Figure 125: Layout of the four devices. Top left: classical BIMOS. Top right: T-gate BIMOS. Bottom left: BIMOS dot with a thick gate. Bottom right: BIMOS dot with a thin gate.

ACS simulations were performed (Figure 126). The high trigger voltage of the classical BIMOS is attributed to its relatively long gate length (108 nm). Indeed, the leakage current that flows through the device before the triggering cannot be high if the gate is long, and less time is given for carriers to flow in this high resistive channel. The T-gate BIMOS and the thin BIMOS dot have the same gate length (48 nm), so it is possible to compare their performance. The thin BIMOS dot has the lowest trigger voltage in ACS, hence it seems to be the most suitable protection for GO1 transistors of the UTBB FD-SOI technology. However, if the comparison was made with the same area, the T-gate BIMOS would benefit from more fingers than the thin BIMOS dot, therefore its failure current I_{T2} would probably be higher. Nevertheless, when ported into a matrix device, the thin BIMOS dot would recover its advantage.

Measurements show that the difference between the T-gate device and the BIMOS dot with a thin gate is negligible (Figure 127). The parasitic capacitances of the Back End Of Line (BEOL) metallic connections are probably masking the capacitive improvement between the dot and the T-gate topology. Apart from this, the tendencies are confirmed: devices with a longer gate have a higher trigger voltage.

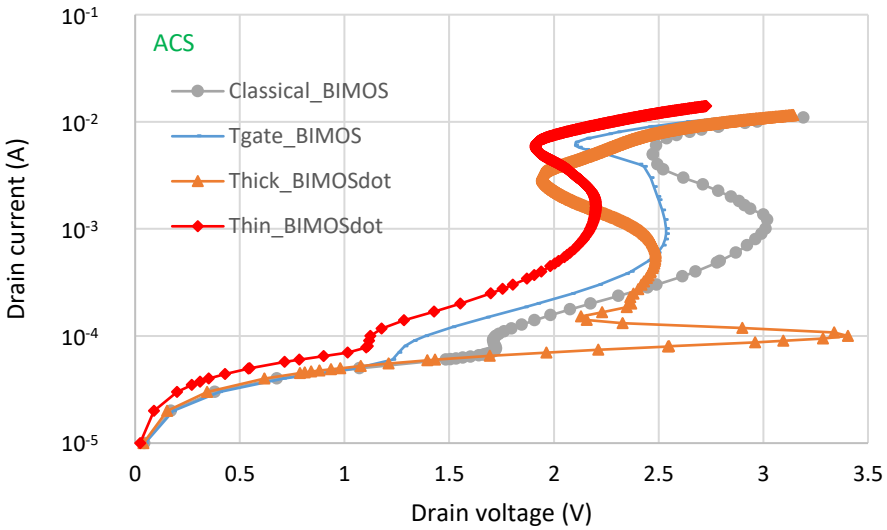


Figure 126: TCAD ACS simulation. Comparison of three devices with minimal dimensions: the classical BIMOS, the T-gate BIMOS and the BIMOS dot with a thin gate.

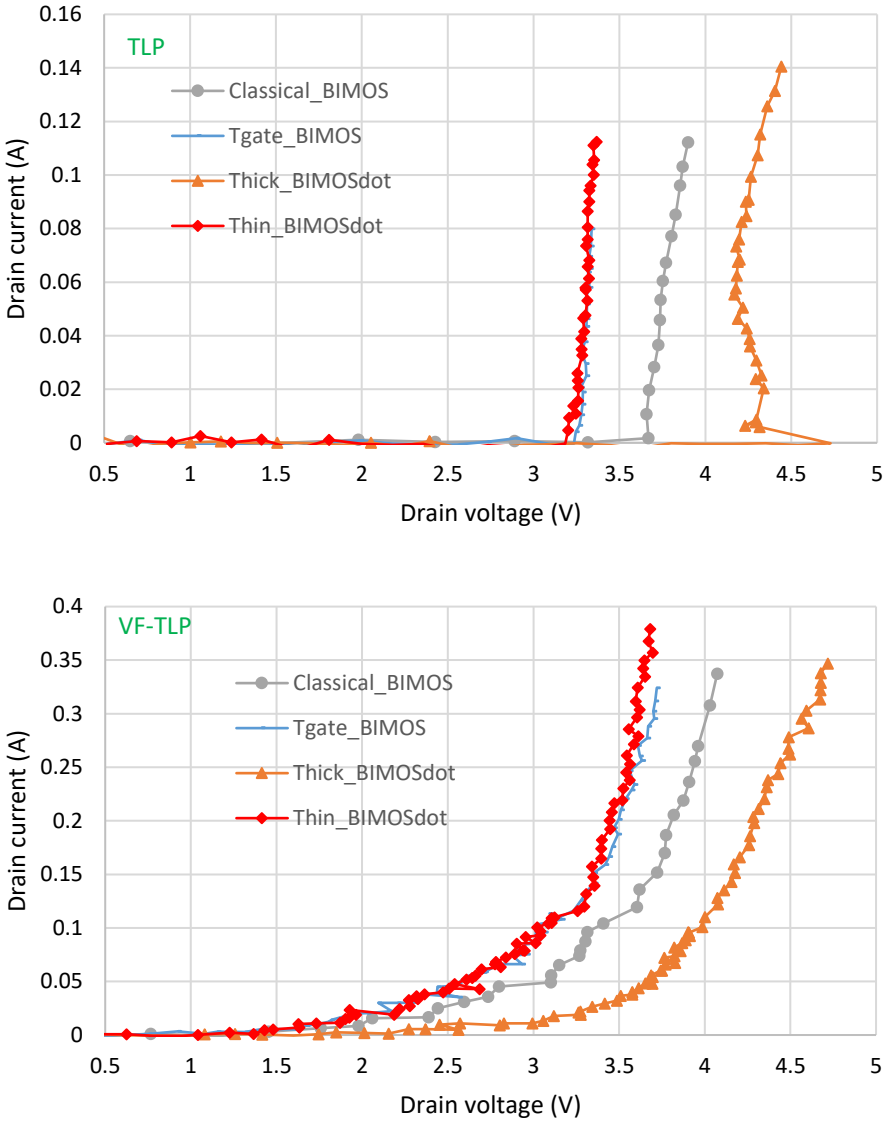


Figure 127: TLP and VF-TLP measurements on the four test devices: the classical BIMOS, the T-gate BIMOS, the BIMOS dot with a thick gate and the one with a thin gate.

Let us consider the matrix version of the classical BIMOS and the BIMOS dot with a thin gate (the T-gate BIMOS is not portable in matrix) (Figure 128). The bridge of active silicon of the classical BIMOS leads to design constraints that impose a minimal gate length of 108 nm and a minimal drain/source to body contact distance of 147 nm. Therefore, the total length wasted by the presence of a body contact inside the matrix is Wbc = 672 nm in the case of the classical BIMOS. In the BIMOS dot there are less design constraints, therefore the minimal dimensions of the body contact part of the device is Wbc = 414 nm.

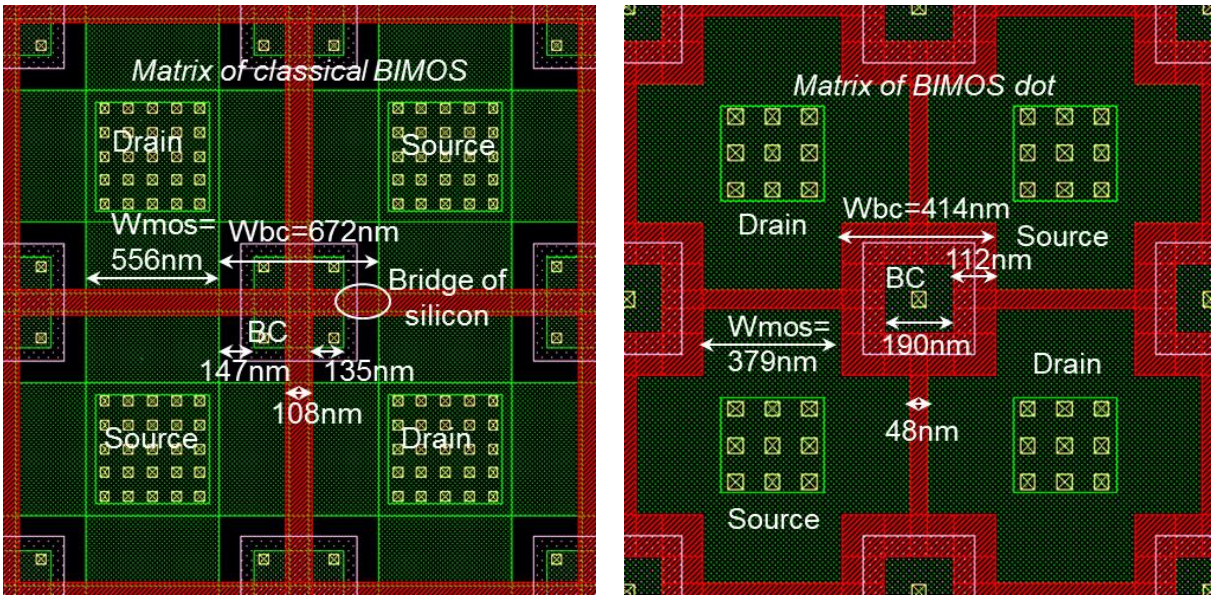


Figure 128: Layout view of the classical BIMOS matrix (left) and the BIMOS dot matrix (right) with their dimensions. Green color stands for the RX layer (active thin-film). Red stands for the PC layer, which is polysilicon. When PC and RX layers are superposed, there is a gate stack. STI is in black. Yellow crosses correspond to CA contacts between RX and M1 (the first metal layer). The white layer with small dots is the BP layer. When this layer is superposed to RX, it means that the doping is P^+. If there is no layer BP, the doping in the active is N^+. Note that in the case of the BIMOS dot, active is everywhere, thus there is no STI/active constraint; in the classical BIMOS the bridge of active leads to severe design constraints (STI/active constraints and active/gate constraints).

The area of each matrix has been optimized (the smallest possible) with respect to the size of its body contact (Wbc). As seen in the previous paragraph, Wbc is reduced at the minimal dimensions according to the design rules. Let us define Wmos as the width where the conduction can occur in one of the small NMOS (one drain or one source). nfx and nfy are the number of fingers in the x or y direction. Wcond is the total width of the conductive channel (taking into account all the small NMOS of the matrix). We decide Wcond to be 100 µm, in order to expect a failure current I_{T2} of the order of 0.1 A in TLP and 0.4 A in VF-TLP.

$$Wcond = [nfx \cdot (nfy + 1) + nfy \cdot (nfx + 1)] \cdot Wmos$$

and

$$Area = [Wmos \cdot (nfx + 1) + Wbc \cdot nfx] \cdot [Wmos \cdot (nfy + 1) + Wbc \cdot nfy]$$

If we choose nfx=nfy=nf, then:

$$Wcond = 2 \cdot Wmos \cdot nf \cdot (nf + 1)$$

So reversely:

$$Wmos = \frac{Wcond}{2 \cdot nf \cdot (nf + 1)}$$

Also:

$$\begin{aligned} Area &= [Wmos \cdot (nf+1) + Wbc \cdot nf]^2 \\ &= \left[\frac{Wcond}{2 \cdot nf \cdot (nf+1)} \cdot (nf+1) + Wbc \cdot nf\right]^2 \\ &= \left[\frac{Wcond}{2 \cdot nf} + Wbc \cdot nf\right]^2 \end{aligned}$$

So, we have a function $y=(a/x+b.x)^2$ (with y=area, x=nf, and a and b are constants). The minimum of this function occurs for $x=\sqrt{(a/b)}$. Hence the minimal area is when:

$$nf_{minArea} = \sqrt{\frac{Wcond}{2 \cdot Wbc}}$$

The classical BIMOS matrix has Wbc = 672 nm, so the optimized number of fingers is nf = 9. From this it is possible to calculate Wmos; we get Wmos = 556 nm with nf = 9. The total area is equal to $(11.6)^2$ µm² (it is the width of the device in each dimension). Wcond can be recalculated as 100.1 µm. The same procedure can be applied for the matrix of BIMOS dot. The obtained dimensions are summarized in Table 3. If we take the area of the classical BIMOS as a reference, the BIMOS dot matrix is 39% less silicon surface consuming.

	Classical BIMOS matrix	BIMOS dot matrix
Wbc (nm)	672	414
nf	9	11
Wmos (nm)	556	379
Wcond (µm)	100.1	100.1
Area (µm²)	134.6	82.8

Table 3: Summary of the dimensions and parameters of the BIMOS matrices.

Measurements (Figure 129) show that the BIMOS dot matrix has better performances (lower trigger voltage) than the classical BIMOS matrix, due to its shorter gate length (48 nm instead of 108 nm). In addition, the gate all-around of the BIMOS dot matrix allows a better integration than the silicon bridge of the classical BIMOS matrix: it is more robust, more design rule compliant, and the silicon area of the whole matrix is reduced (which gives margin to increase the number of fingers and the failure current I_{T2} thanks to a higher total conduction width).

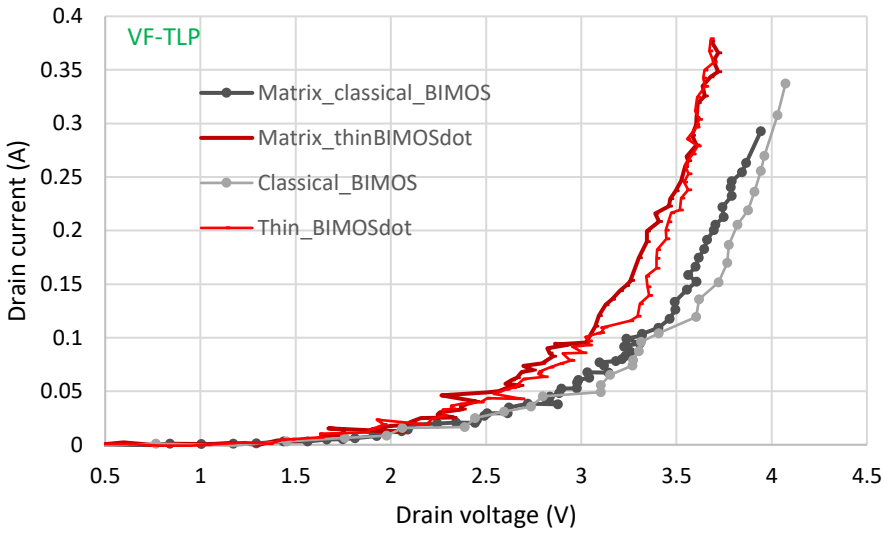


Figure 129: VF-TLP measurements of the classical BIMOS and the thin BIMOS dot, in matrix version or not.

II. Comparison of different BIMOS devices

In the previous section, the BIMOS dot topology has shown its advantage when ported into a matrix. It is partly due to the silicon area that was saved at each body contact of the matrix. But does a BIMOS matrix really need a body contact next to each small NMOS (one drain and one source)? This section will answer to this question by comparing various thin-film BIMOS topologies. Among them, a 2D matrix of BIMOS that only has one peripheral body contact will be proposed. Those topologies were fabricated with the 28 nm thin-film with high-k dielectrics and metal gate FD-SOI CMOS technology.

1. Devices description

The following devices were measured (Table 4 and Figure 130): devices 1 – 6 are 2D-matrices of NMOS transistors in the x and y direction. Those NMOS matrices are surrounded by a P^+ region connected to the gate, in order to form a BIMOS. Devices 1 - 4 are designed with the minimal dimensions compatible with the design rules; there is only one tungsten contact (to link the active and the first metal layer) per source or drain. The number of fingers in each direction, nf, was calculated knowing the width of each drain or source Wmos and the theoretical maximal conduction in the thin-film, that is slightly above 0.1 mA/µm. We selected nf = 16 fingers, so that the theoretical total current flowing through the device could reach 0.1 A, which results in a total silicon area taken by the device of 18 µm² ("small" area). Devices 5 and 6 have larger width Wmos in order to enable two tungsten contacts per source/drain in each direction. They have 13 fingers in each direction, so that the total current could also be 0.1 A. The total silicon area is 25 µm² ("large" area). Devices 7 – 12 are conventional 1D BIMOS. They have all been designed with the smallest dimensions (in terms of gate length, length of source/drain Wmos, etc.) according to their topology. Their number of fingers has been adapted so that the silicon area is similar to that of the matrices, for comparison purpose. Devices 7 and 8 have an H-gate shape, devices 9 and 10 a π-gate shape, and devices 11 and 12 are the classical BIMOS, where the gate is a straight line, but the active is cut, in order to better separate the body contact from the source.

Devices 7, 9 and 11 take the same silicon area as devices 1 – 4; devices 8, 10 and 12 have the same area as devices 5 and 6. Devices 1 – 10 feature a gate length of 48 nm, but devices 11 and 12 have longer gates (108 nm), because of the physical process constraints due to their geometry. The values of the external polysilicon resistor that is plugged to the gate and the body contacts of the BIMOS are given in Table 4, with R1 < R2 < R3. Device 4 has its body contact connected to the gate, but the node is left floating, so it is as if an infinite resistor was plugged to the node. In all variants, the sources and drains are fully covered with silicide.

The aim of this design of experiment is to compare different topologies of BIMOS in order to optimize the ESD performance of devices. We chose to compare devices with same area. Indeed, the area concern is very important since ESD devices take substantial space, in particular when they are placed in the IO pads and designed for standing HBM pulses. Silicon footprint is very expensive. Also, the larger the total width of the structure, the higher the current that it can handle before failing. Some topologies may take less silicon area for a given number of fingers than other topologies, thanks to the minimum design rules being respected. Therefore, if there is a given area assigned to ESD protections in the IO pad, more fingers could be designed for this "smaller" topology, thus enabling a higher total width; the device could be more robust to high currents. So, the final question to be answered is: for a given area (small or large), which topology (matrix, H-gate, π-gate or classical) is the best as an ESD protection for the 28 nm FD-SOI technology? The aim of comparing structures that have large or small area is to show whether the small 2D matrix - that only has one tungsten contact per source/drain - could be used as an ESD protection. Indeed, in the small area devices, the dimensions have been pushed to the limit of the technology. Due to its topology, the 2D matrix robustness could potentially be even more impacted by the minimum design rules, since each small NMOS square has to be perfectly processed for the device to be operational. The goal of the area comparison is thus to assess the intrinsic robustness of the smallest topology toward ESD. It shows the limit of the structure and therefore the limit of the design.

TLP, VF-TLP and DC measurements were performed at room temperature on several dies. The stress was applied on the drain, while the source and the back gate were grounded.

<table>
<tr><th>Device</th><th>Structure</th><th>Area</th><th>Gate length</th><th>Resistor</th></tr>
<tr><td>1</td><td rowspan="6">Matrix</td><td rowspan="4">Small</td><td rowspan="10">48 nm</td><td>R_1</td></tr>
<tr><td>2</td><td>R_2</td></tr>
<tr><td>3</td><td>R_3</td></tr>
<tr><td>4</td><td>Infinite</td></tr>
<tr><td>5</td><td rowspan="2">Large</td><td>R_2</td></tr>
<tr><td>6</td><td>R_3</td></tr>
<tr><td>7</td><td rowspan="2">H-gate</td><td>Small</td><td rowspan="8">R_2</td></tr>
<tr><td>8</td><td>Large</td></tr>
<tr><td>9</td><td rowspan="2">π-gate</td><td>Small</td></tr>
<tr><td>10</td><td>Large</td></tr>
<tr><td>11</td><td rowspan="2">Classical</td><td>Small</td><td rowspan="2">108 nm</td></tr>
<tr><td>12</td><td>Large</td></tr>
</table>

Table 4: Different thin-film BIMOS device topologies fabricated in 28 FD-SOI.

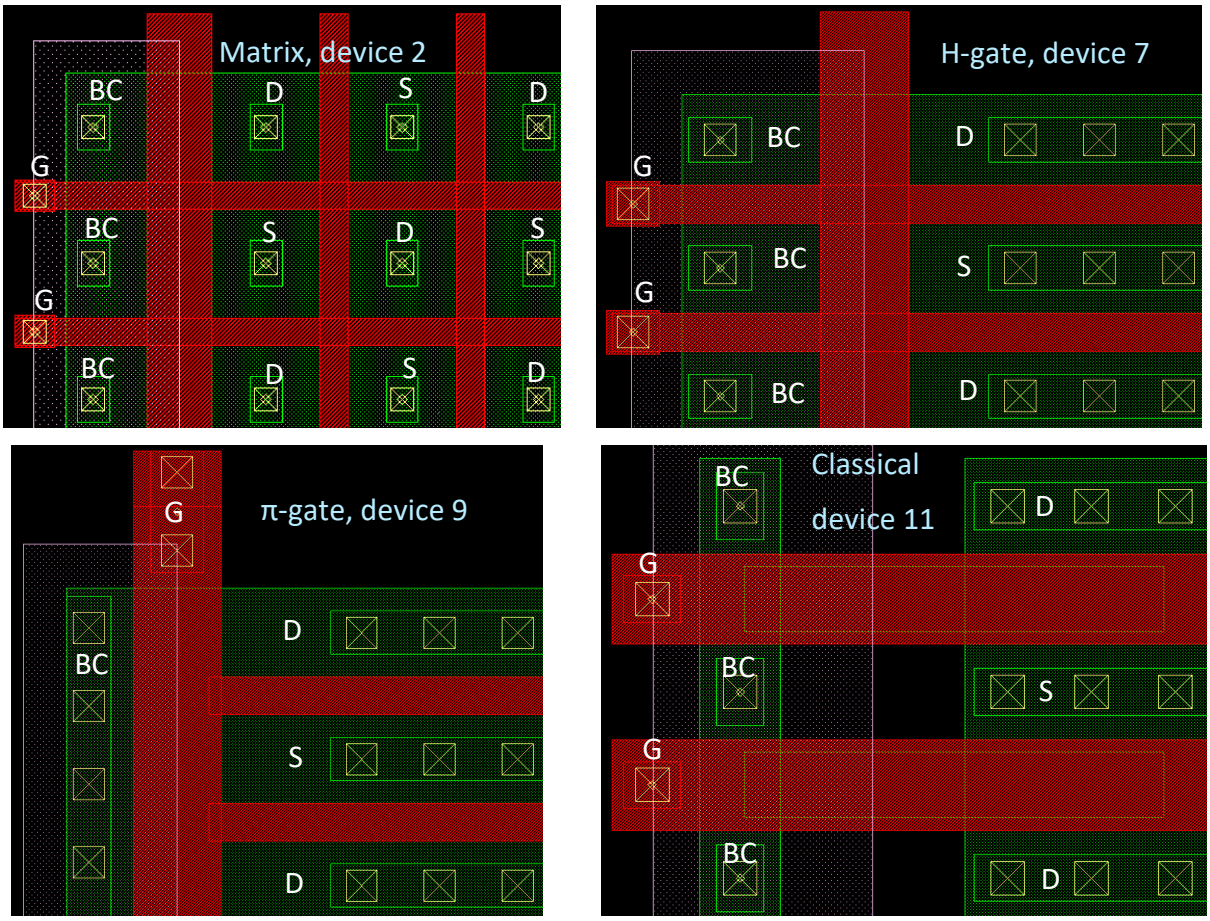


Figure 130: Four layout views for devices 2, 7, 9 and 11. D: Drain, S: Source, G: Gate, BC: Body Contact.

2. Results and discussion

Devices 1, 2 and 3 show the effect of the resistor value connected to the gate of the BIMOS (Figure 131 and Figure 132). In terms of silicon footprint, it would be better to totally remove the resistor, or to reduce its silicon area which lowers its value. The resistor is helping to control the trigger voltage of the device: the device activates sooner with a higher resistor value, since it allows a higher bias both on the base of the BIMOS and on the gate in a shorter time. This time constant τ is dependent on the resistor value R and the parasitic capacitor between the drain and the gate C_{DG}. C_{DG} is constituted by the capacitance of the spacer that lies between the drain and the gate, the capacitance of the thin oxide layers that separate the gate from the channel (its contribution is important in the overlap region of the drain under the gate), and the capacitance of the diode junction between the drain and the channel (since the body contact P^+ is touching the channel and is connected to the gate). C_{DG} being 4.10^{-14} F approximately, the order of magnitude of τ is 0.4 ns for the resistor R_1, 2 ns for R_2 and 4 ns for R_3. Therefore, for the TLP curve a difference of trigger voltage is still observed between R_2 and R_3, because the time constant of the device is shorter than that of the TLP measurement (100 ns), whereas these curves are superposed for the VF-TLP measurement (1 ns pulse) (Figure 131).

Since thin oxide transistors of the 28 nm FD-SOI technology cannot withstand high voltages, efforts have to be made to associate them with ESD protections that trigger as early as possible. It is recommended from our data to plug a resistor at least as high as R_2.

The quasi-static behavior of devices 1 – 4 (Figure 132) shows that the higher the resistor, the lower the trigger voltage. This rule holds whatever the value of the resistor. This is because the time constant of the experiment is too long to show such transient behavior. Since the gate node of the device 4 is left floating, the device triggers very early in terms of voltage. It starts to have a consequent leakage (the criterion of maximum allowed leakage current being 10 nA) at the voltage bias of $V_{LeakMax}$ = 1.1 V. If the targeted technology has a V_{DD} of 1 V, this means that there is a narrow margin for the device not to be too leaky in the normal operational range. As a consequence, it is advised to plug a resistor to the gate, even if this resistor removal would reduce the silicon footprint of the ESD devices.

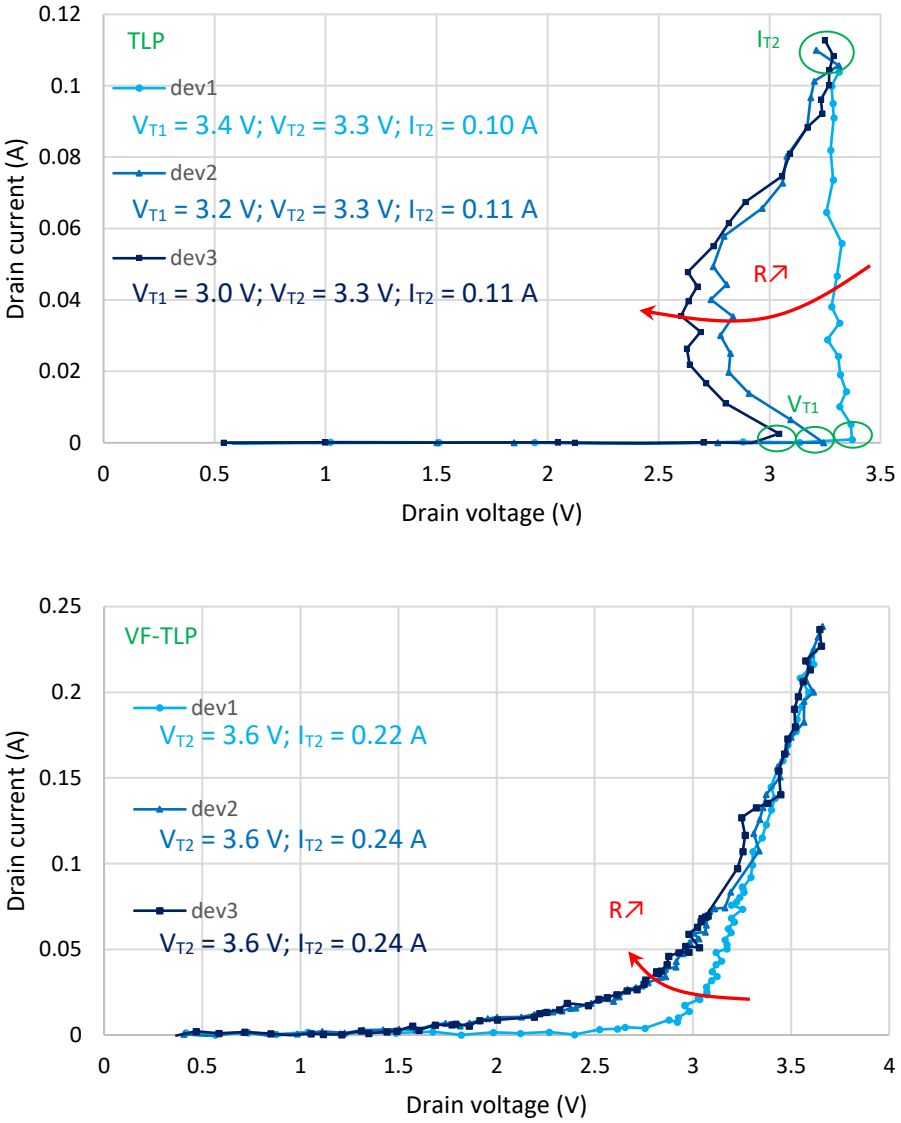


Figure 131: Measurement of devices 1 - 3. Top: TLP. Bottom: VF-TLP.

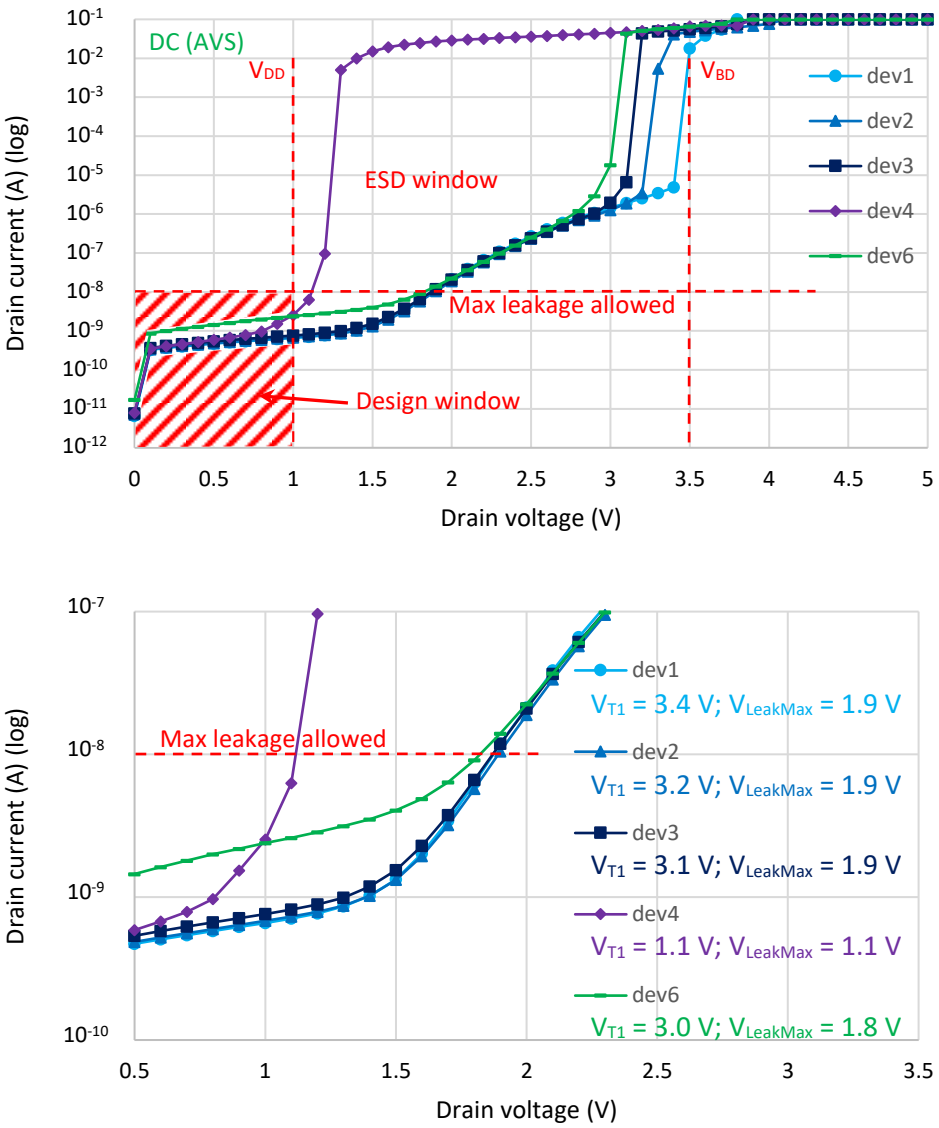


Figure 132: DC measurement of devices 1 – 4 and 6. Bottom: zoom.

When comparing the different device topologies (matrix, H-gate, π-gate and classical) via TLP measurements (Figure 133), it is noted that the classical BIMOS curve is very separated from the others, that are quite superposed. The trigger voltage of the classical BIMOS and its dynamic ON resistance are higher than the ones of the other topologies, which is problematic if the target is to protect thin oxide transistors. The classical BIMOS is however much less leaky than the other topologies (Figure 134). These significant differences are attributed to the gate length that is much longer in the classical BIMOS. The leakage study is not really meant for comparing the different topologies with respect to each other, but to verify if each topology (designed with the minimal dimensions) is reaching the leakage target for a given technology. The ESD specification is done with the most aggressive dimensions in order to determine their intrinsic performance. The leakage in the matrix, H-gate and π-gate topologies can only be relaxed if their gate length is increased. The failure current I_{T2} (Figure 133) of the classical BIMOS is also lower with respect to the other topologies, due to the fact that less fingers could be included in the structure (at same area). This BIMOS is still interesting because in case a very low leakage was needed (low power applications), the gate length should be increased (whatever the topology), and then all the topologies could have similar gate length, making the classical BIMOS competitive again. The eventual benefit of the classical BIMOS topology is that there is much less conduction via the parasitic diode between the body contact and the source (where the holes are attracted by the low-voltage of the N^+ doped source). This has to be confirmed in a new study where all gate lengths are similar.

When the matrix features a “large” area, its behavior is the same as for the H-gate and the π-gate BIMOS. However, for a “small” area, its dynamic ON resistance is affected (Figure 133). This could be explained by (i) the rounding of the gates, that takes a too significant portion of the width of each NMOS for the small dimension matrix; or (ii) the high access resistance due to the unique tungsten contact per MOS. The same reasons can explain why the leakage current at 1 V of the matrix (device 2) is less than the one of the H-gate and the π-gate BIMOS (devices 7 and 9) (Figure 134). It is worthwhile noticing that even if the body contact is very far from each MOS, the matrix BIMOS is still working like the other BIMOS topologies. This result is quite important because the 2D matrix is the first step toward designing a 3D matrix. The simplification of the layout - by placing the body contacts far from the NMOS matrix – would probably be required to imagine 3D matrix of BIMOS, where, for example, a NMOS gate could control a channel in the thin-film of the FD-SOI transistor as well as a channel in a new layer of silicon that would be fabricated on top of the gate (Figure 135).

There is not much difference between the H-gate and the π-gate topology, which shows that designers could choose any of them.

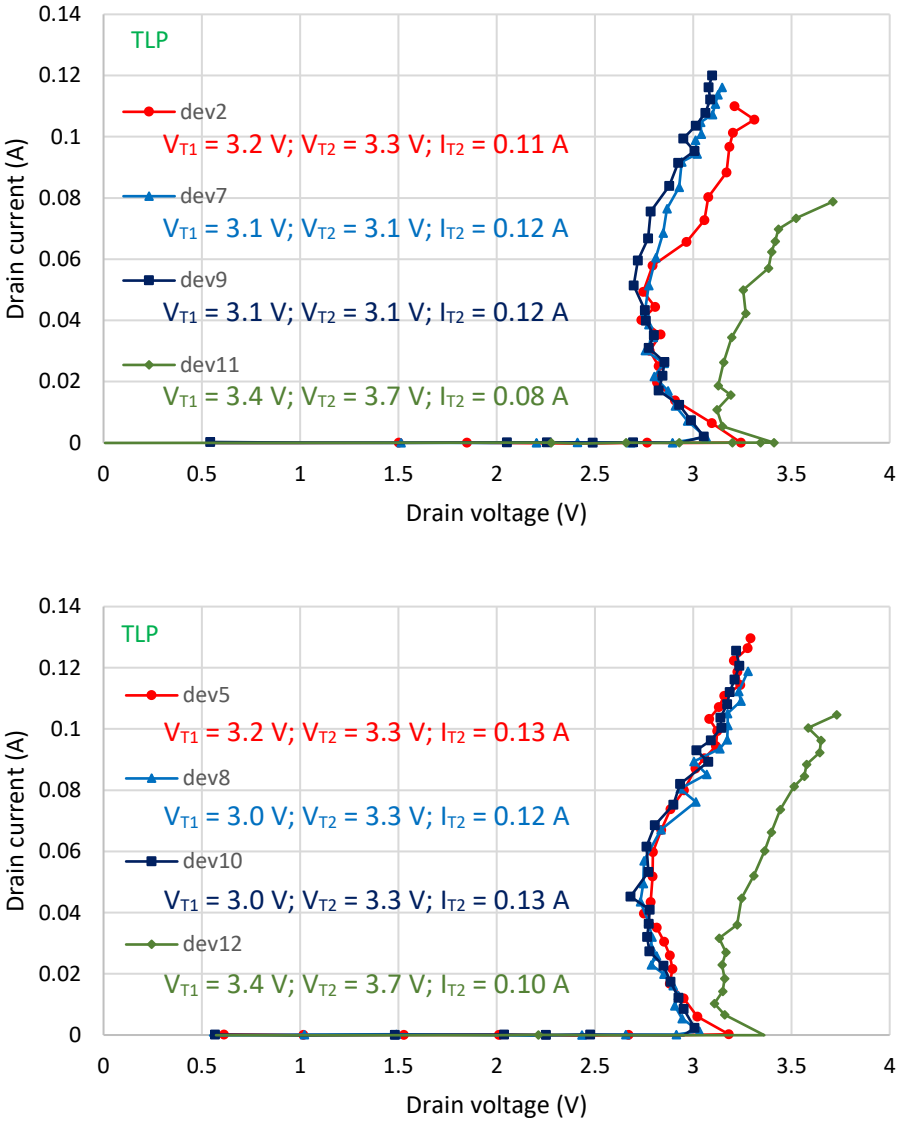


Figure 133: TLP measurement. Comparison of the matrix (devices 2 and 5), the H-gate (devices 7 and 8), the π-gate (devices 9 and 10) and the classical (devices 11 and 12) BIMOS topologies. Top: small area. Bottom: large area.

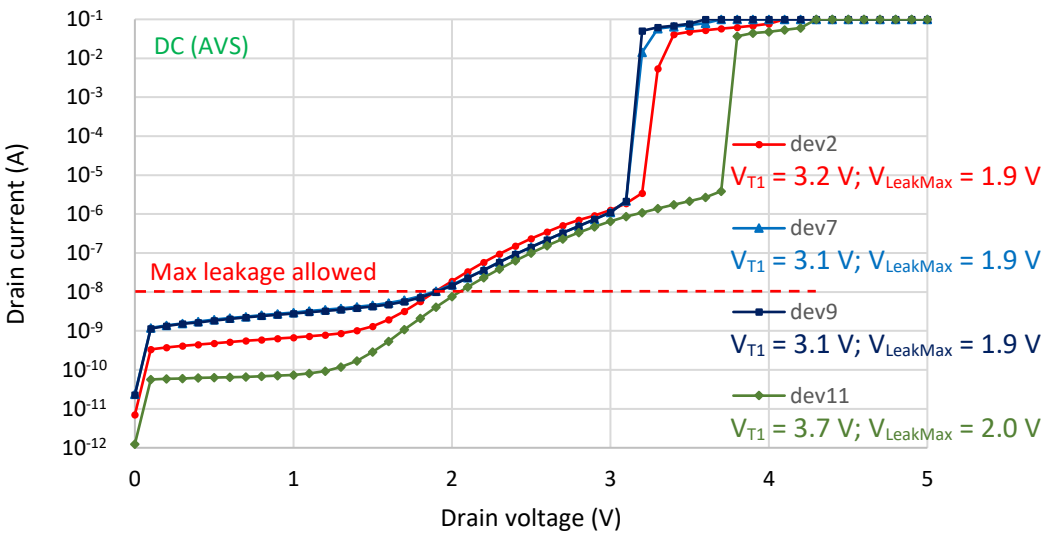


Figure 134: DC measurement of devices 2 (matrix), 7 (H-gate), 9 (π-gate) and 11 (classical).

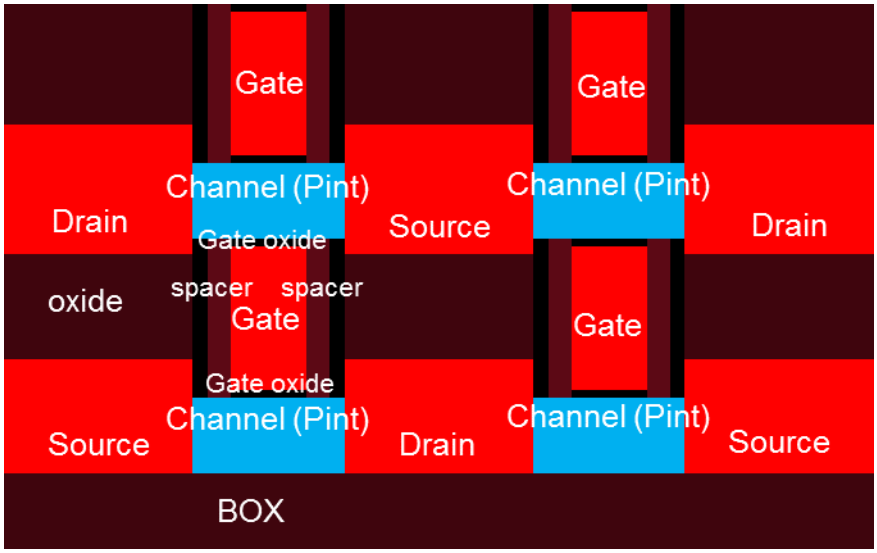


Figure 135: Example of a polymorphic homothetic 3D [115] ESD NMOS protection with thin layer deposition on top of the gates to add the upper layers. The gates are therefore controlling the bottom thin-film as well as the top thin-film silicon layer. Oxides are displayed in brown, N^+ doping in red and Pint doping in blue.

The same trends can be observed with the VF-TLP measurements (Figure 136): the classical BIMOS has a very different behavior compared to other topologies due to its gate length and finger number. The dynamic ON resistance of the matrix is affected primarily in the "small" area device; the H-gate and π-gate topologies give similar results. All devices are failing for a voltage of $V_{T2} \approx 3.7$ V approximately.

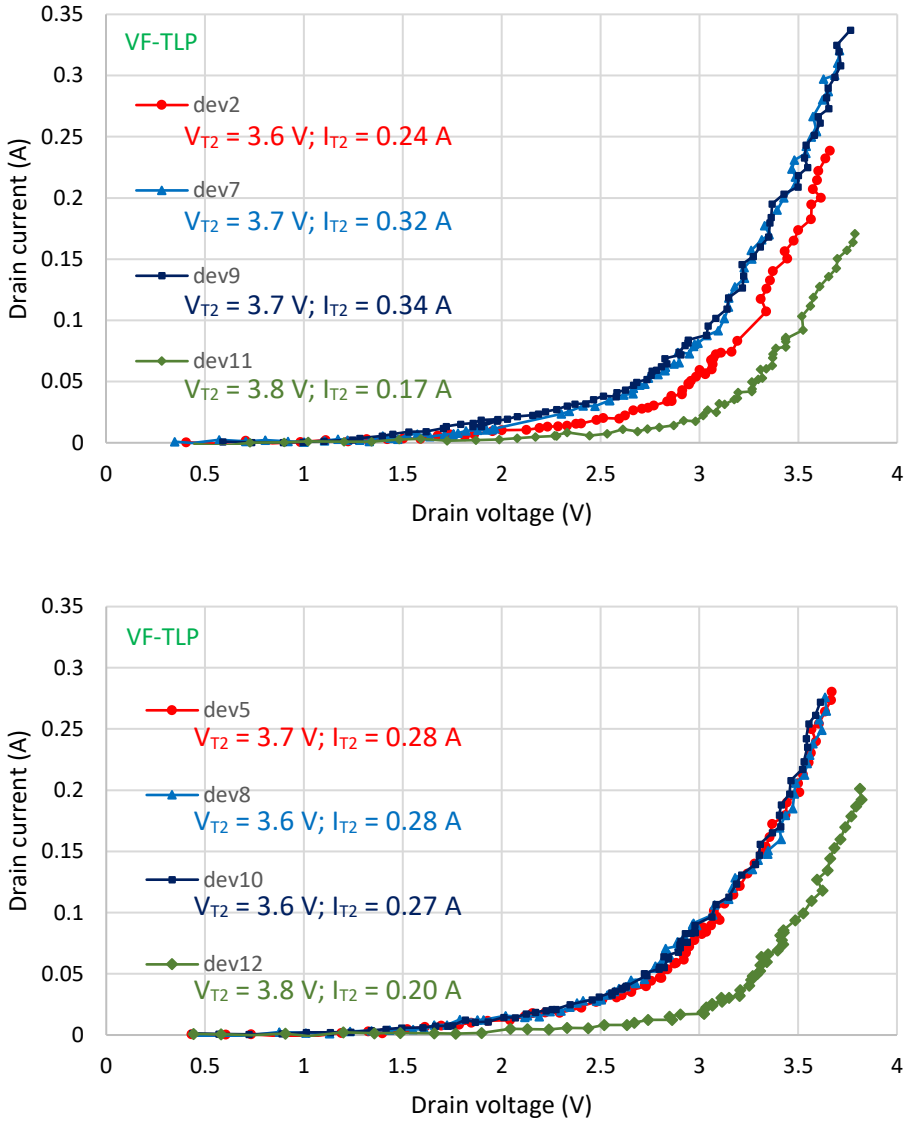


Figure 136: VF-TLP measurement. Comparison of the matrix (devices 2 and 5), the H-gate (devices 7 and 8), the π-gate (devices 9 and 10) and the classical (devices 11 and 12) BIMOS topologies. Top: small area. Bottom: large area.

Note also that in all our TLP measurements, the condition I_{T2} = 1 mA/µm is respected. The devices robustness is increased in VF-TLP (higher failure current I_{T2} and voltage V_{T2}), because the transferred energy is reduced thanks to the shorter time.

Another interesting result is that no multi-triggering is observed, neither in the TLP nor in the VF-TLP measurements. This shows that even if the devices could be designed without silicide on the drain, this is actually not needed because they are stable enough.

As it can be seen in Figure 137, where some waveforms of the device 2 for the VF-TLP measurements are shown, there is no over-voltage that may destroy the thin-oxide transistors to be protected.

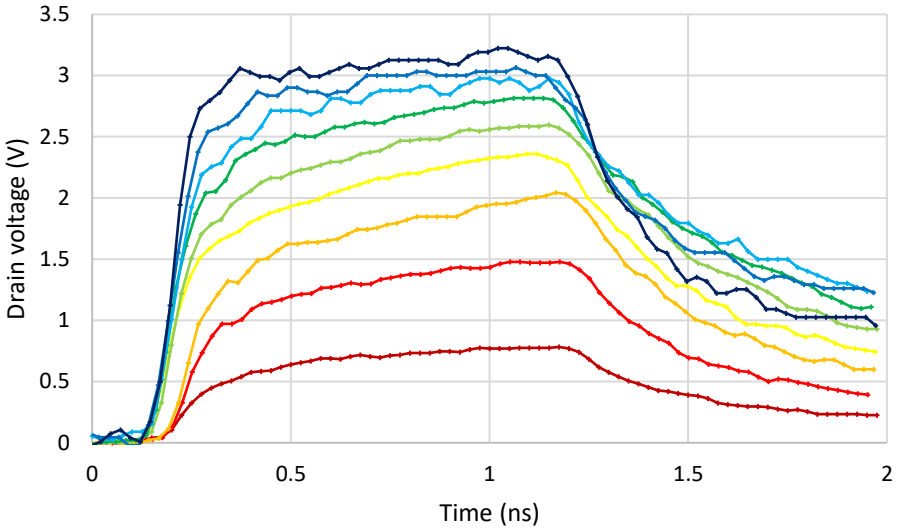


Figure 137: VF-TLP waveforms in device 2.

As a conclusion for this experiment, different thin-film BIMOS topologies were tested for ESD protection application. The devices are all process compliant, robust, reconfigurable and latch up free. The I-V characteristics of the matrix, the H-gate and π-gate BIMOS show that these devices can fit into similar ESD design windows. The so-called classical BIMOS is still interesting provided that a long gate (> 100 nm) is needed for securing a small leakage current in specific applications.

These results are promising, since the 2D matrix can provide some tracks for designing a 3D matrix. Here it was proven that the 2D matrix device was as performant as classical 1D devices. Also, even if the body contact of the 2D matrix was situated at the periphery of the device only - and not at each NMOS node like in the BIMOS dot topology -, it was still well behaving. Additional studies would be required to determine how far from the center of the matrix the body contact can be placed and still influence the conductivity in the whole matrix, without observing multi-finger effects like multi-triggering. The suppression of all those additional small body contacts in the matrices allows to gain substantial silicon area (so much, that the BIMOS matrices start to become competitive against the 1D devices in terms of silicon footprint, as seen in this section).

Chapter 4: 3D ESD protections in FD-SOI

"You can't use up creativity.
The more you use, the more you have."
Maya Angelou.

I. FD-SOI silicon continuity with bulk

Merging two ESD devices - or merging the protection device with its trigger circuit - allows to gain silicon area, to reduce some parasitics and to get new paths of conduction inside the merged device. Bulk ESD protections and bulk trigger circuits already exist in merged versions [75] [116]. An example of thin-film ESD protection merged with a thin-film trigger circuit is the GDBIMOS device studied in the chapter 2. It is advantageous to place the protection in bulk silicon since it would benefit from a volumic conduction current and therefore exhibit a higher failure current I_{T2} and a higher level of ESD protection, while the trigger circuit can be situated in the thin-film because it does not need to undergo the whole ESD surge. Bulk protections and thin-film trigger circuits only exist in merged version when considering the displacement currents (capacitive coupling), but not with the conduction currents. For example, [117] and [118] propose to use the Hybrid FD-SOI process to stack active components on top of each other. The thin silicon film over the BOX is used for the upper devices and lower devices are situated under the BOX (Figure 138). Therefore, this method creates a device stacking thanks to the BOX (which separates the two devices) and allows the area to be reduced. It constitutes an example of 3D integration. The goal of this chapter is to explore ESD devices with 3D conduction current. Therefore, the idea of the Figure 138 will be modified for getting conduction current additionally to displacement currents. The aim of this section is to obtain a bulk protection that would be fully merged with a thin-film trigger circuit (Figure 139).

With the silicon merge, new topologies and devices are allowed, with alternative conduction path, direct potential control, thermal homogeneity… The leverage of a brick of process will be used to address 3D with a 2D technology without over-cost, since it is part of one of the core processes of STMicroelectronics. The right merging solution will be investigated through TCAD simulation solely, since the new ideas presented in this chapter arose at the end of the time allocated for the PhD. TCAD is a powerful tool that allowed us here to explore new solutions and to provide a proof of concept. Silicon is under process and measurements will follow in a second time; they will be available for a new generation of PhD students.

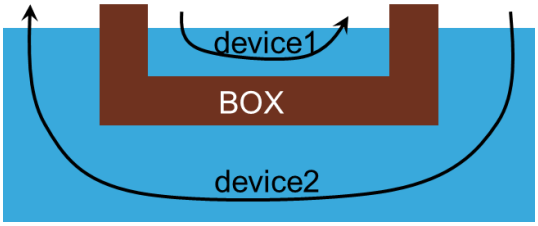


Figure 138: Device over device principle using the BOX as a separator. Note that there is no silicon merge between the two devices.

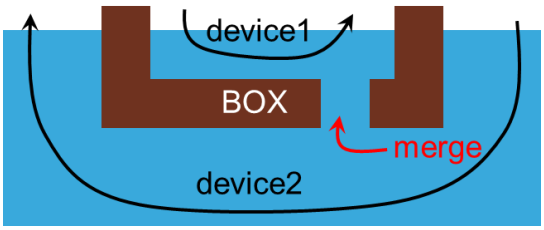


Figure 139: Proposal in this study: fully merged devices.

Merging the top device (above the BOX) and the bottom device (under the BOX) is possible with the NOSO process of STMicroelectronics M28 route [119], which comprises two masks: NOSOI and SSTI. Figure 140 illustrates this process, that allows bulk and thin-film devices to be realized at the same height of the wafer. Figure 141 is an example of realization, where Dual Isolation by Trenches and Oxidation (DITO) is used to separate devices (instead of STI) [120]. A hard mask is deposited and lithography is performed with the NOSOI mask. Then the thin-film is oxidized in the NOSOI region (since it was left unprotected). A resist is deposited and lithography is performed with the SSTI mask. The oxide is etched in the SSTI region (left unprotected). The hard mask and the resist are removed after having grown silicon via epitaxy, in order to equalize all the silicon levels.

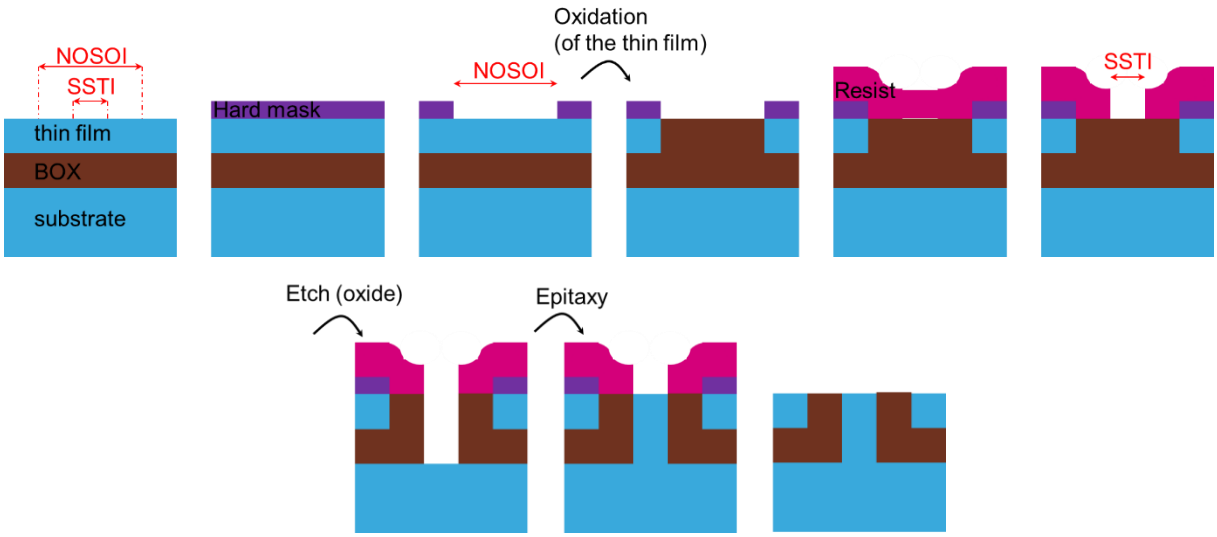


Figure 140: NOSO process described in [119].

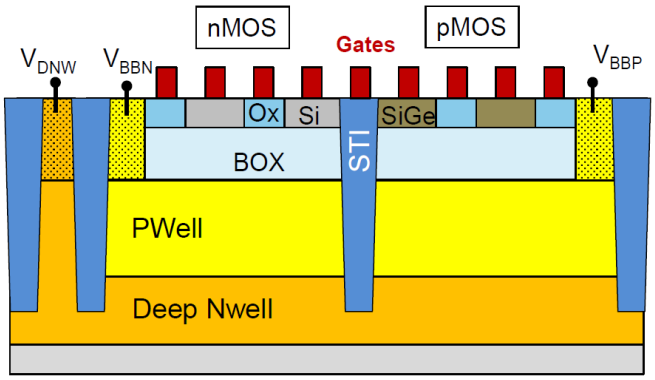


Figure 141: An example of realization of the NOSO process, where a Dual Isolation by Trenches and Oxidation (DITO) is used to separate devices [120].

If the SSTI mask is larger than the NOSOI mask, then silicon continuity is obtained between the substrate and the thin-film. A small bump of silicon is observed above the silicon film at the BOX junction (Figure 142). Because of this bump, minimal distances between layers have to be observed. For example, at least 55 nm between the border of Hybrid layer (which will create the NOSOI mask) and the PC layer (for the gate stack) is required when drawing layouts.

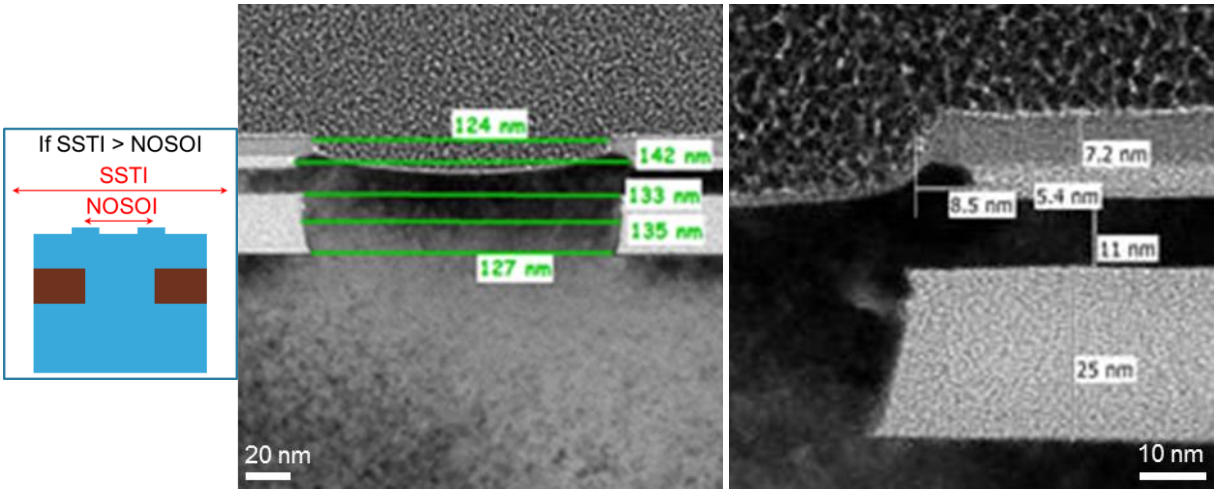


Figure 142: Post epitaxy TEM views. The NOSO process is valid for achieving silicon continuity between the substrate and the thin-film.

In this section, the NOSO process is used to build a fully merged device with a silicon communication between the thin-film and the substrate, therefore obtaining 3D conduction. First a BIMOS merged SCR and then a BIMOS with the external resistor embedded in the substrate will be presented.

II. 3D BIMOS merged SCR with silicon continuity

The idea is to merge two devices: one in bulk (SCR with volumic conduction) and one on the thin-film (BIMOS as a trigger circuit).

1. BIMOS merged SCR using P-doped trigger

An example of possible BIMOS merged SCR is given in Figure 143. The merge has been done in the Body Contact of the BIMOS (which is also the P-doped trigger Gp of the SCR) and in the cathode of the SCR (Source of the BIMOS). The trigger Gp has been chosen to control the activation of the SCR, because the NPN bipolar transistor is faster to activate than the PNP. The principle is the following: as soon as the BIMOS enters in conduction, the system Gate - Body Contact increases in potential. Therefore, Gp goes from a low-voltage (SCR blocked) to a high voltage (SCR opened).

3D TCAD simulations at room temperature were done to evaluate the performance of the device (electro-thermal simulations would be too long to process). The TCAD view of the device is shown in Figure 144, and layout views along with the dimensions are shown in Figure 145. For calculation time saving, the Niso doping was not simulated. Simulation results are displayed in Figure 146.

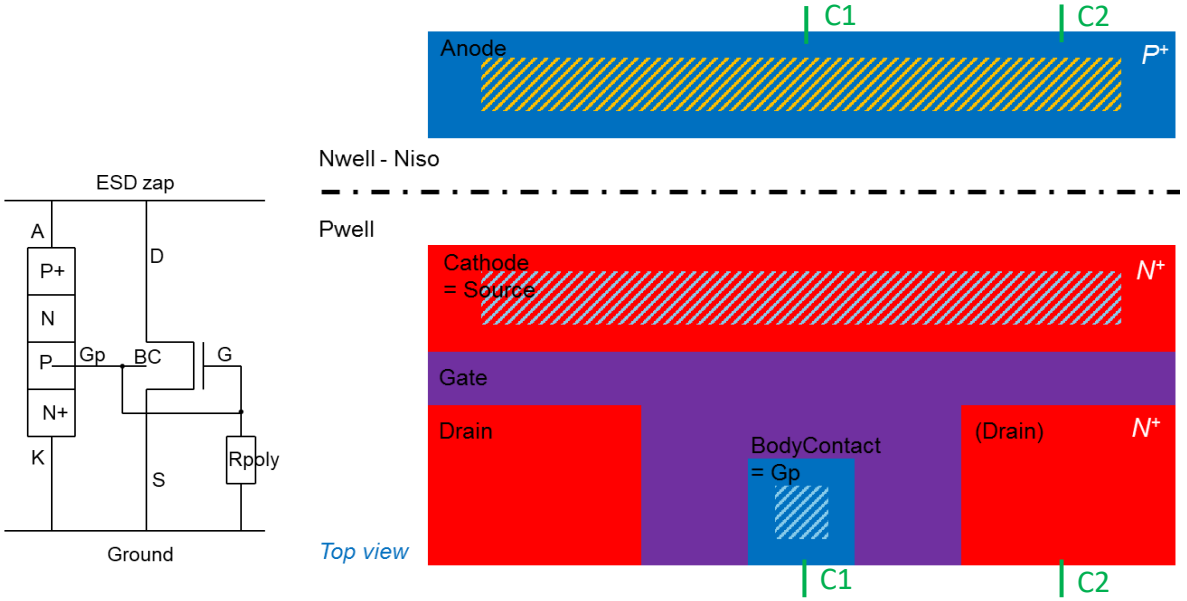


Figure 143: Right: Top view of the 3D BIMOS merged SCR with the use of the trigger Gp. Red color is for N^+ doping, blue for P^+ doping, and violet for the gate and the gate stack. Hatched regions correspond to the merged regions; under the BOX the doping layers are Nwell (under the anode) and Pwell (under the rest of the device). C1 cross section provides a correspondence with Figure 144. C2 corresponds to the cross sections seen in Figure 148. Left: corresponding schematic.

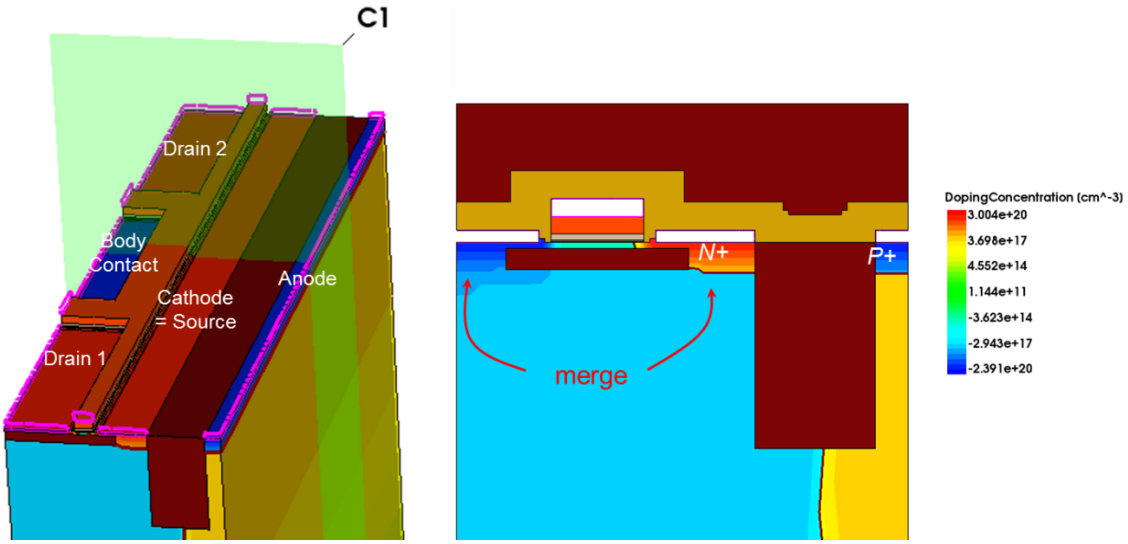


Figure 144: Left: TCAD view of the device. Right: Cross section (C1).

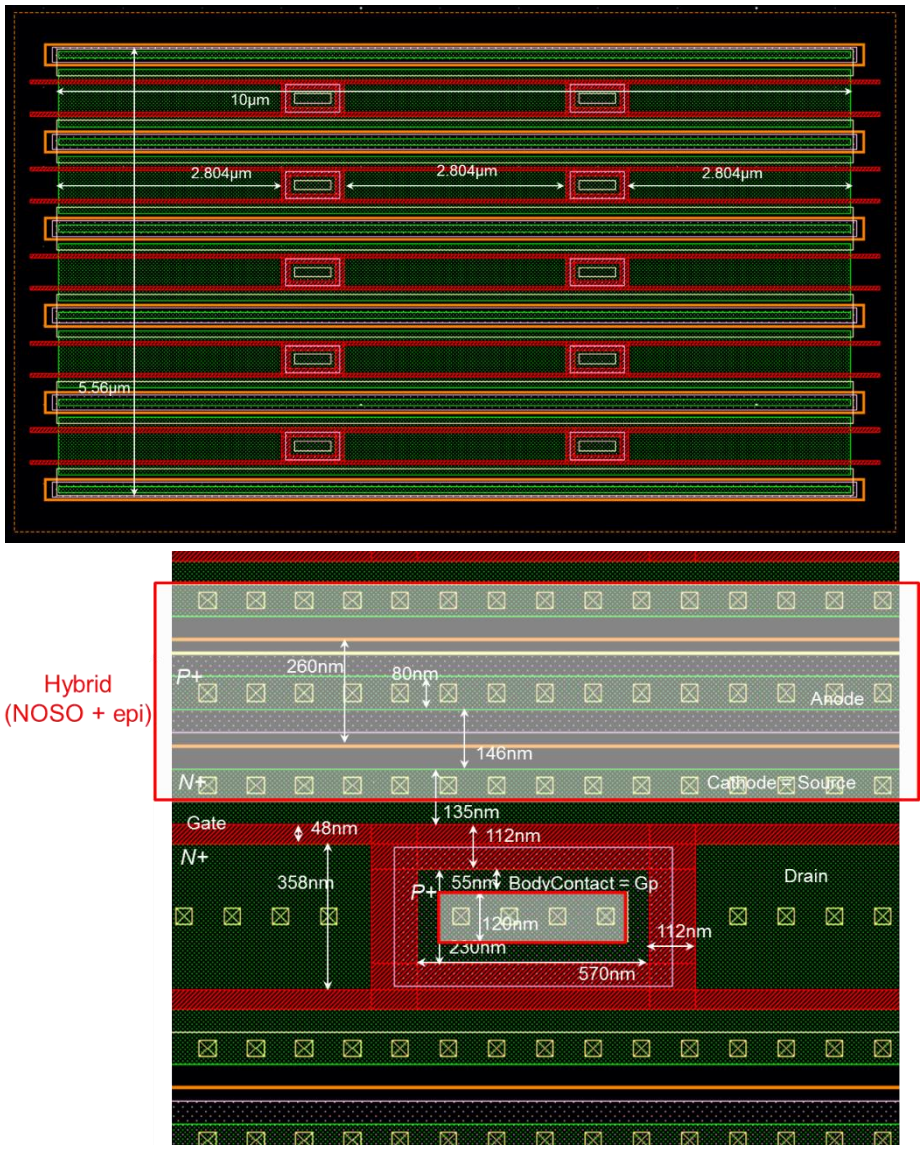
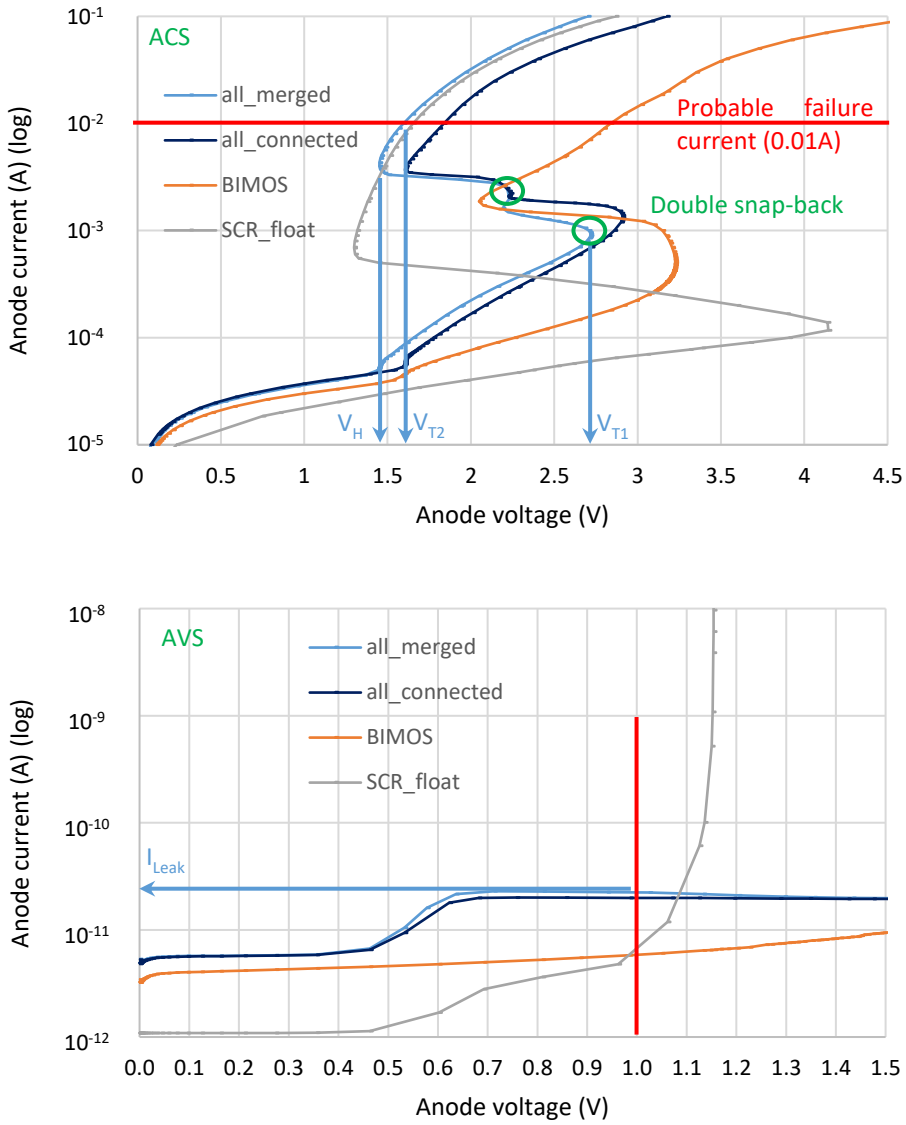


Figure 145: Top: Layout view of the device with its dimensions. Bottom: zoom.



	V_{T1} (V)	V_H (V)	V_{T2} (V)	I_{Leak} (A)
SCR (float)	4.1	1.3	1.7	7.10^{-12} with LU
BIMOS	3.2	2.0	2.8	6.10^{-12} without LU
BIMOS+SCR connected	2.9	1.6	1.8	2.10^{-11} without LU
BIMOS+SCR merged	2.7	1.5	1.6	2.10^{-11} without LU

Figure 146: TCAD simulation of a simple BIMOS (drain current versus drain voltage), a simple SCR with floating triggers, the SCR and BIMOS connected together electrically such as to respect the schematic in Figure 143, and the SCR merged BIMOS. Top graph: ACS. Bottom graph: AVS simulation. Table: extracted parameters from the graph. The failure voltage V_{T2} is taken on the ACS graph at I = 0.01 A. The leakage current is taken on the AVS graph at V = 1 V. We specified that the SCR is potentially undergoing Latch Up (LU) issues because it is triggering right after the 1 V limit on the AVS.

The table in Figure 146 shows the extracted parameters (the trigger voltage - or first trigger voltage - V_{T1}, the holding voltage V_H, the expected failure voltage V_{T2} and the leakage current I_{Leak}) from the simulation graph. It allows to compare a simple BIMOS device, a SCR with floating trigger, the SCR and BIMOS connected together electrically such as to respect the schematic in Figure 143, and the SCR merged BIMOS.

The leakage current in all the solutions is acceptable (Figure 146). Special care is eventually needed to verify by measurements that the SCR is not subject to Latch Up. The SCR has a very good holding voltage and dynamic ON resistance (if the goal is to protect GO1 transistors of the 28 nm FD-SOI technology), but it has a too high trigger voltage. Therefore, it needs a trigger circuit. The BIMOS could be the trigger circuit of the SCR, because its trigger voltage is smaller than the one of the SCR. The BIMOS connected to the SCR is therefore a good solution, because it has the advantage of a lower trigger voltage thanks to the BIMOS and a good ON conduction thanks to the SCR. The conductions of both the SCR and the BIMOS are improved further because they help each other to raise the required voltages. Merging the SCR and the BIMOS is the best solution. The trigger voltage is decreased because the carriers of one device help the other to be triggered. Also, the body contact of the BIMOS acts like a positively biased back gate for the BIMOS, since the BOX is opened. The positive bias on this back gate helps to reduce the threshold voltage of the BIMOS. In terms of silicon footprint, the merged device takes 16% less silicon area than the connected device (Table 5).

	X (µm)	Y (µm)	Area (µm^2)
SCR	10	4.26	42.6
BIMOS	10	2.374	23.74
Connected	10	4.26+2.374	66.34
Merged	10	5.56	55.6

Improvement area 16%

Table 5: Total silicon dimensions and area of the four compared devices: SCR, BIMOS, SCR + BIMOS connected, and SCR + BIMOS merged.

Figure 147 and Figure 148 explain the double snap-back shape of the “All_merged” curve seen in Figure 146, by showing the contribution of the SCR and the BIMOS currents. The trigger voltage V_{T1} of the merged device corresponds to the triggering of the BIMOS. With the BIMOS being active, the voltage on the body contact is raised faster (Figure 150), therefore the SCR triggers. The device reaches its holding voltage V_H when the SCR and the BIMOS are conductive.

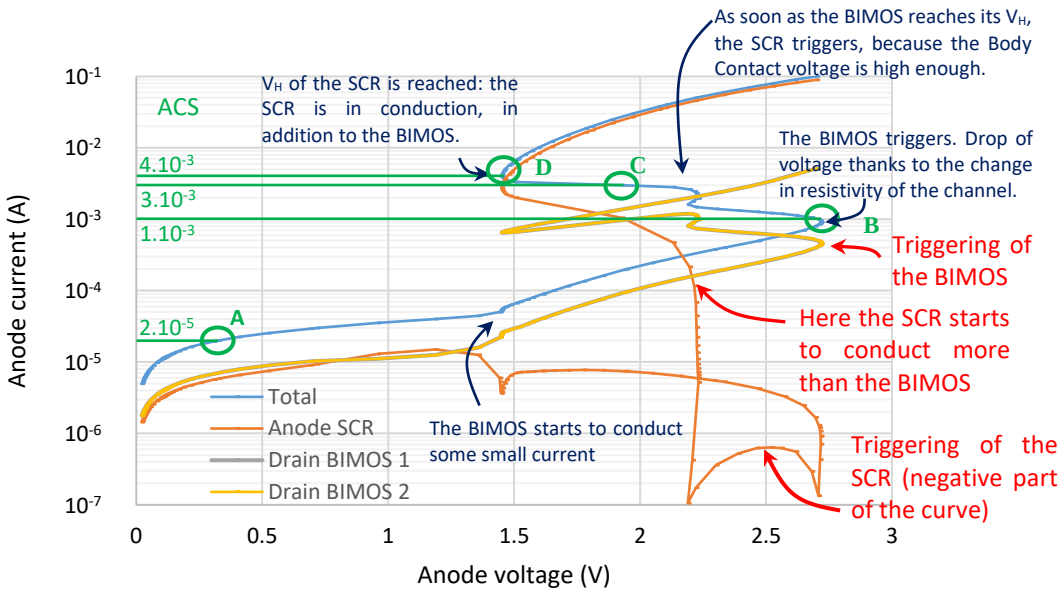


Figure 147: TCAD ACS simulation of the merged device. All the contributions to the total current (which corresponds to the Anode current of the "all_merged" curve in Figure 146) are shown. The anode current of the "all_merged" curve in Figure 146 corresponds to the addition of the anode current of the SCR part of the merged device with the first and the second drain currents (of the BIMOS part).

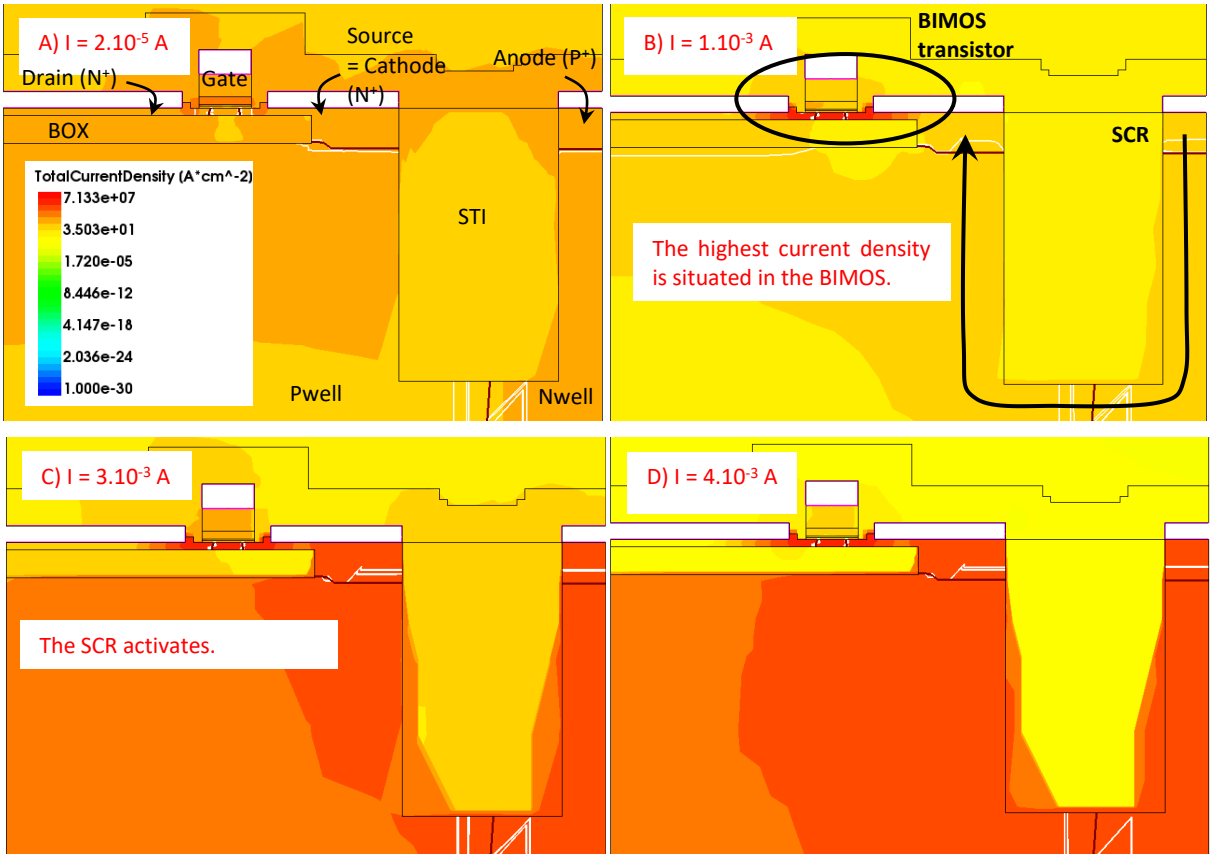


Figure 148: Extracted current density in cross section C2 (see Figure 143) of the merged device from the curve in Figure 147 (points A, B, C and D).

Note that the triggering voltage of the SCR alone occurs when Gp is around 0.89 V (Figure 149), therefore the SCR in the merged device starts to drive significant current when Gp (the Body Contact) is also around 0.8 V (Figure 150). This shows that the second snap-back (Figure 146) of the merged device (which is due to the SCR triggering) is indeed a consequence of the body contact (which is also Gp) being at a sufficiently high voltage.

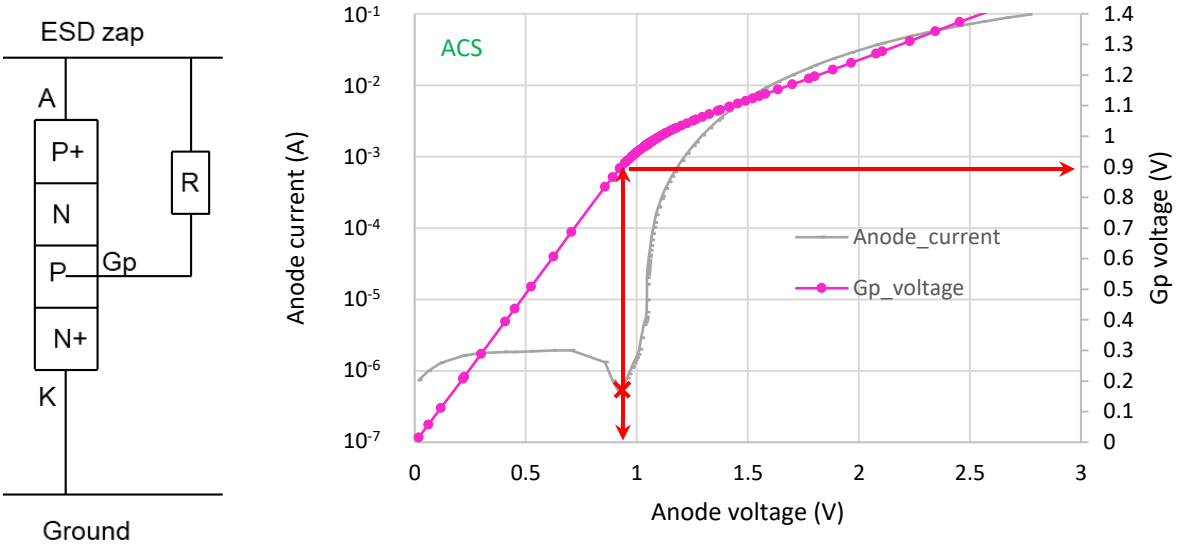


Figure 149: Left: schematic of the studied SCR; a resistor has been plugged between Gp and the anode (subject to the surge). Right: TCAD ACS simulation of the SCR shown left. The triggering of the SCR occurs when the body contact is around 0.89 V.

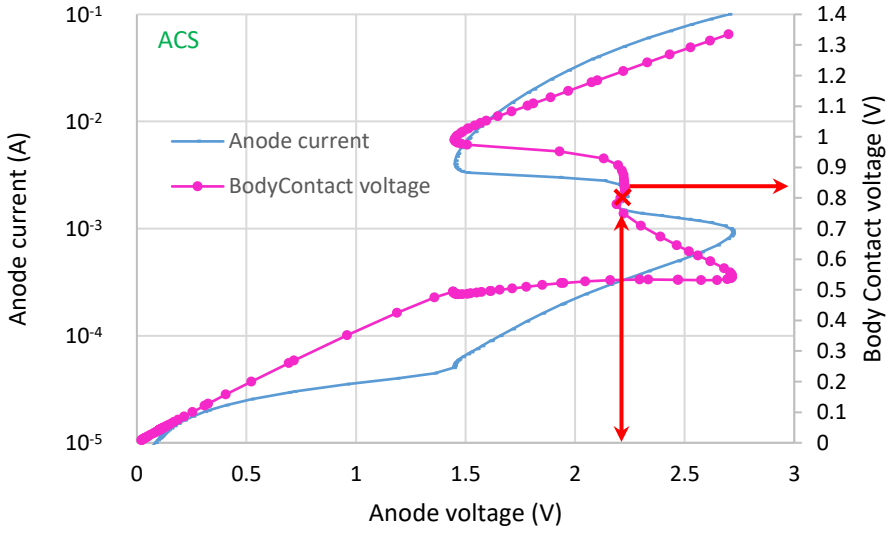


Figure 150: TCAD ACS simulation of the merged device. Total anode current versus anode voltage (same curve as in Figure 146 and Figure 147), and Body Contact voltage versus Anode voltage. The triggering of the SCR occurs when the body contact is around 0.8 V.

The same schematic and working principle than the ones of the previous device can be used with a different topology (layout). For example, in Figure 151, the SCR is longer in size (the anode to cathode path is longer than the one in Figure 145), the BIMOS topology is different (π-shape of the gate) and the BIMOS is really on top of the SCR. This device has not been simulated since the high number of BIMOS fingers to be taken into account leads to too many nodes in the meshing, which would cause a too long simulation time.

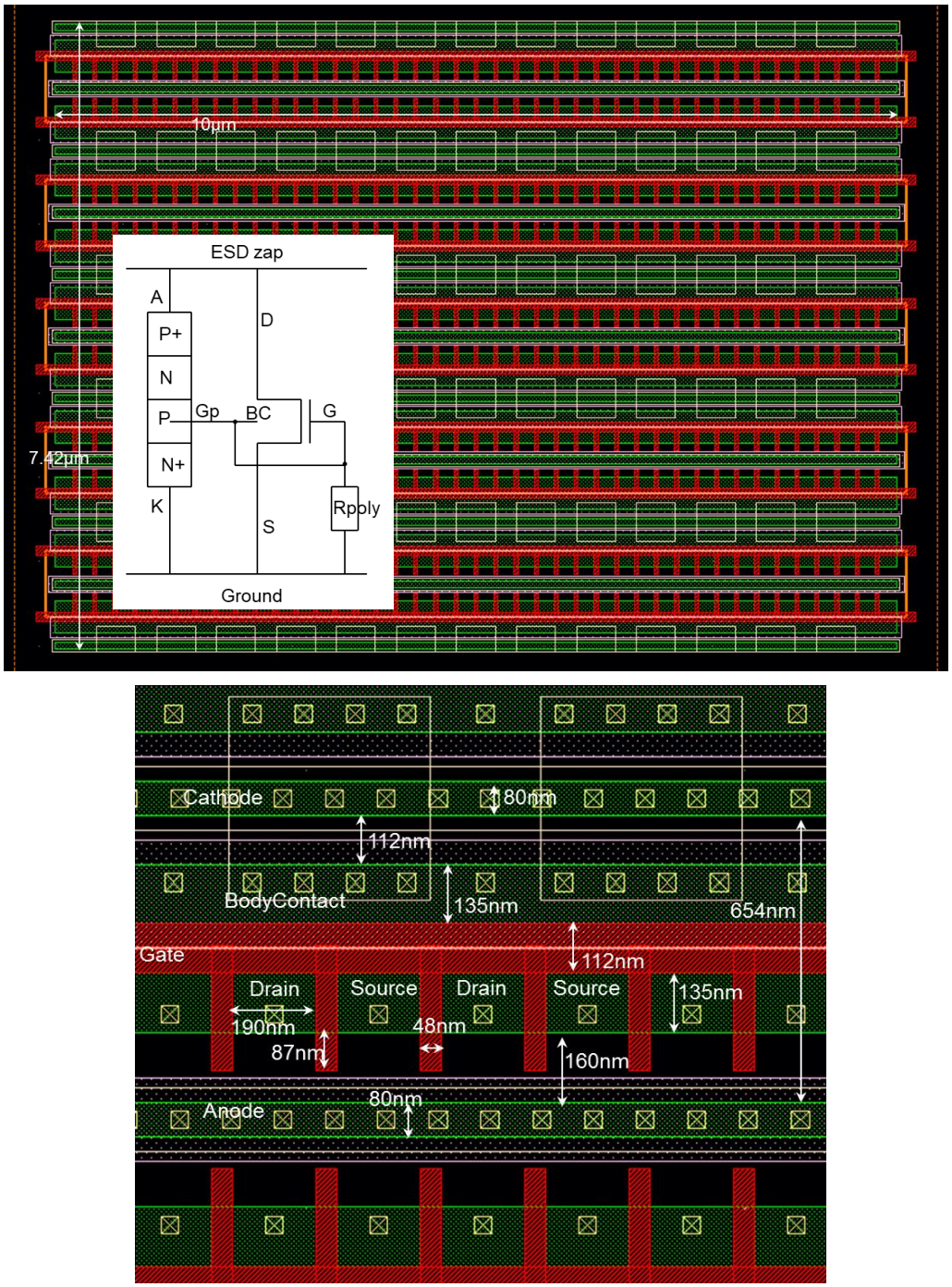


Figure 151: Layout views of another topology with its dimensions. Inset: schematic (same as in Figure 143). Bottom: zoom.

2. BIMOS merged SCR using N-doped trigger

The merge can also be done in the Gn trigger of the SCR (instead of the Gp), and it also leads to a 3D BIMOS merged SCR, even if the schematic and working principle is a bit different. Figure 152 and Figure 153 give an example of topology where the merge has been done in the drain of the BIMOS (which is also the trigger Gn of the SCR) and in the cathode of the SCR (source of the BIMOS). In this device, the SCR is activated thanks to its N-doped trigger. As soon as the PN diode (with the P-doping in the anode and the N-doping in the trigger Gn) lets a sufficient amount of current flowing (with an increase in voltage on the anode), the voltage increases on the drain of the BIMOS. When the BIMOS enters in conduction (with a sufficiently high drain voltage), it goes from a high resistive state to a low resistive state. Therefore, Gn is put to a low-voltage node (because the NMOS is like a small resistor). When Gn is grounded, the SCR is opened.

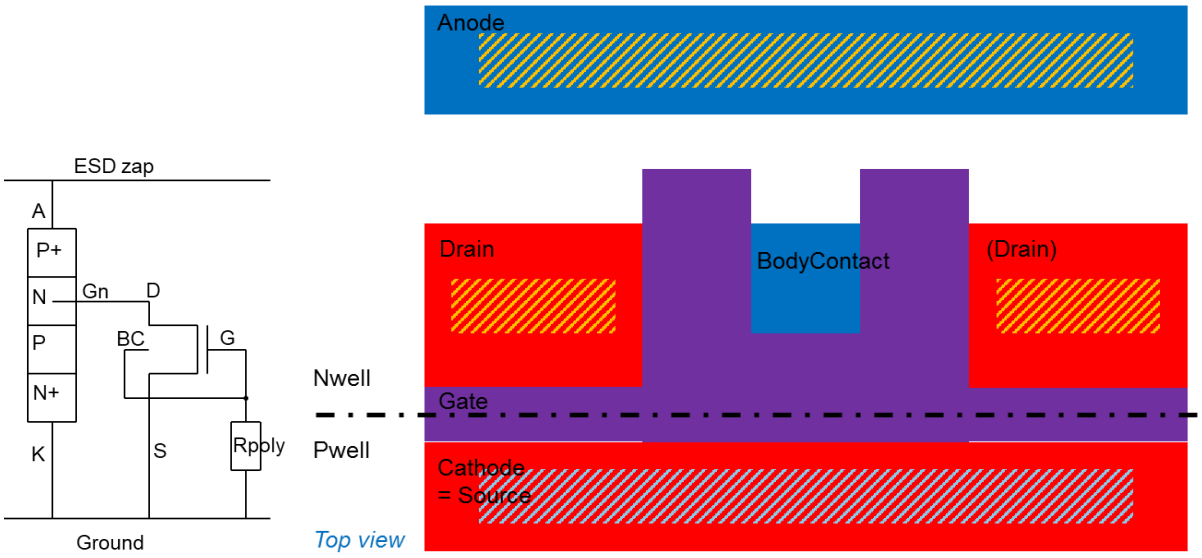


Figure 152: Right: Top view of the 3D BIMOS merged SCR with the use of the trigger Gn. Red color is for N^+ doping, blue for P^+ doping, and violet for the gate and the gate stack. Hatched regions correspond to the merged regions; under the BOX the doping layers are Pwell (under the cathode) and Nwell (under the rest of the device). Left: corresponding schematic.

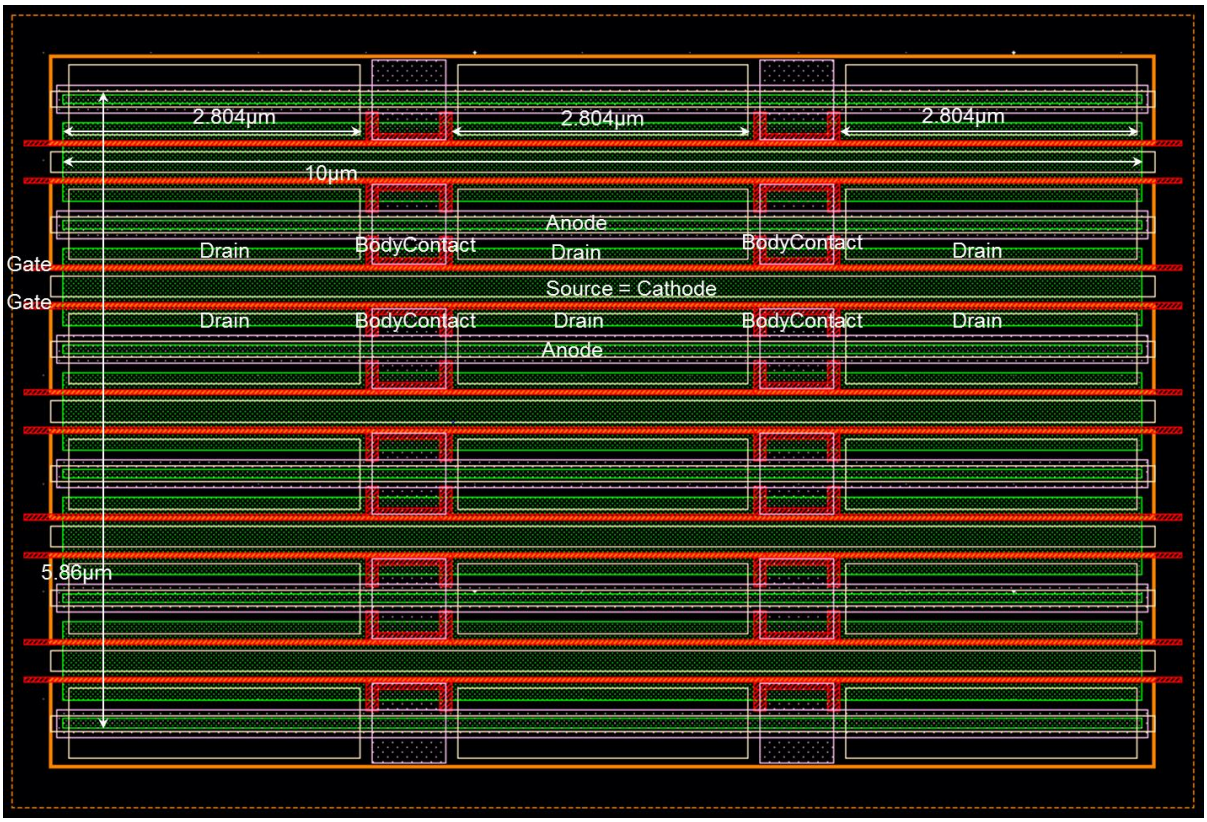
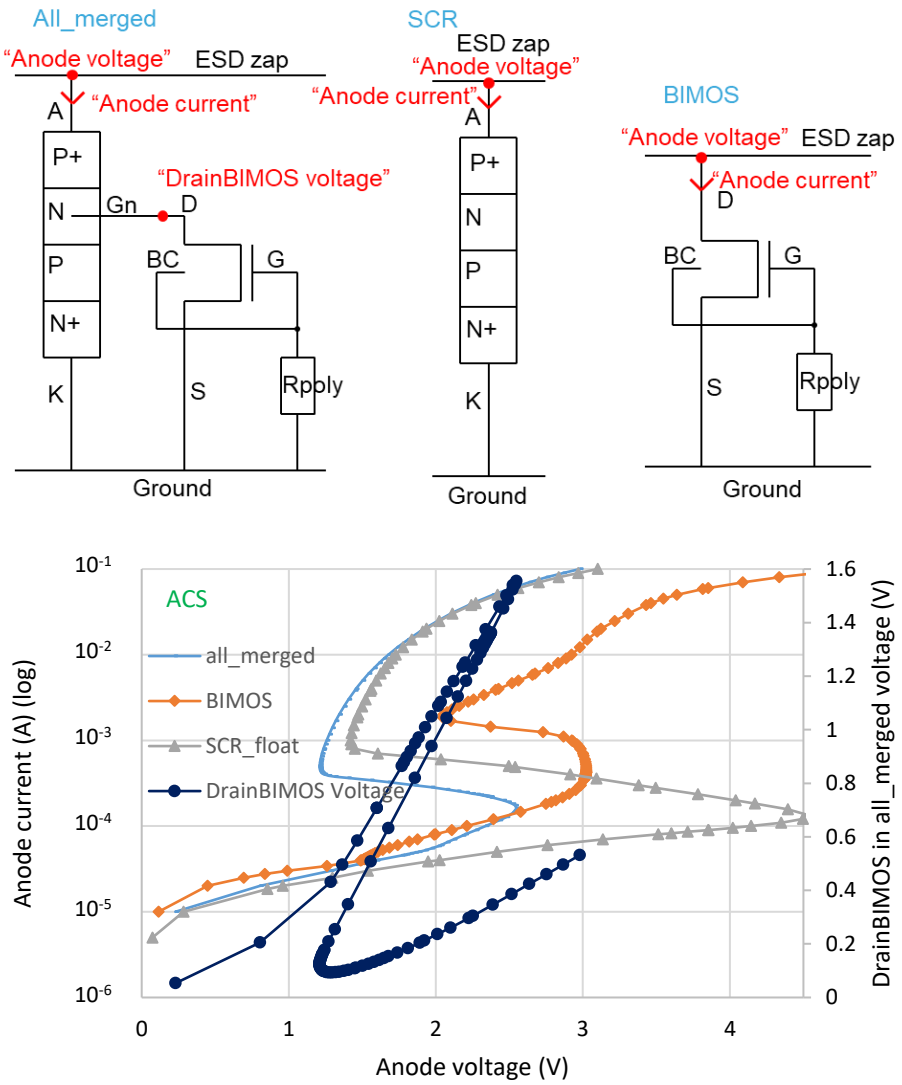


Figure 153: Layout view of the device corresponding to Figure 152 with some dimensions. The device has a total area of 10 µm x 5.86 µm for a total conduction width (of the SCR) of 100 µm. The BIMOS has an additional conduction width of 2.804 x 3 µm and a gate length of 48 nm.

Figure 154 shows the behavior of the BIMOS Gn-merged SCR, and compares it with a simple BIMOS and SCR that have the same dimensions and topology. On the “anode current” versus “anode voltage” curve of the simple BIMOS, it can be seen the BIMOS’ conduction starts around 1.5 V (even if it triggers only at 3 V). On the “drainBIMOS voltage” versus “anode voltage” curve of the merged device, it can be seen that the voltage of the drain of the BIMOS drops drastically to 0 after reaching 1.5 V. This happens at Vanode = 2.5 V. This difference of voltage (2.5 V versus 1.5 V) corresponds to the P^+/N diode between the anode and the drain. The drop of voltage on Gn activates the SCR.



	V_{T1} (V)	V_H (V)	V_{T2} (V)
SCR (float)	4.5	1.4	1.7
BIMOS	3.0	2.1	2.9
BIMOS+SCR merged	2.5	1.2	1.7

Figure 154: TCAD simulation of a simple BIMOS (drain current versus drain voltage), a simple SCR with floating triggers Gn and Gp, and the merged structure. Top: associated schematics. Bottom graph: Comparison of the "anode current" versus "anode voltage" for the three devices. The "DrainBIMOS voltage" versus the "anode voltage" of the merged device is also plotted (right scale). Table: extracted parameters from the curves. The failure voltage V_{T2} of the ESD device is taken on the ACS graph at I = 0.01 A.

To summarize, the BIMOS is a good candidate as a trigger circuit for a SCR. BIMOS merged SCR benefits from the reduced trigger voltage - thanks to the BIMOS - and from a good conduction current - thanks to the SCR. With the use of the NOSO process it is possible to merge devices that are on the thin-film and under the BOX. Multiple topologies and schematics for the 3D BIMOS merged SCR can be used. The advantages of the proposed devices are the following: (i) devices are compact, so thanks to the merging there is less connection resistivity, and other problems due to the distance between the thin-film and the bulk structures are avoided; (ii) silicon area is saved; (iii) conduction current is allowed between the thin-film structure and the bulk structure, as well as direct potential control, thermal homogeneity, and so on; (iv) improvement of the ESD solutions (they can fit new ESD windows). As a conclusion, using the NOSO process to build merged devices-on-devices enables 3D conduction of current, which enhances their performance. A new set of structures can be imagined in this direction.

III. In-situ coupled bias resistance

1. In-situ coupled bias resistance in thin silicon film

The classical approach for using a BIMOS device as an ESD protection consists in plugging the body contact to the gate, and link it to the ground via an external polysilicon resistor. This allows a positive ESD surge to flow from the drain to the source of the BIMOS. For the negative ESD surge, an external diode is plugged between the source and the drain of the BIMOS.

A new way for using the BIMOS device would be to replace the external resistor by an embedded resistor in the channel (Figure 155). The gate and the first body contact would be connected together, and the second body contact would be grounded. This configuration allows a parasitic resistor between the first and the second body contact (in the channel). It also allows a parasitic diode between the second body contact (P^+ doped region) and the drain (N^+ doped). Instead of three devices (the BIMOS, and the external diode and resistance), only one compact device is required. The intention here is to reduce the silicon area. However, those parasitic elements cannot be used as easily as the external ones. The well-functioning of such a BIMOS device with a resistor distributed in the channel as an ESD protection has to be assessed experimentally.

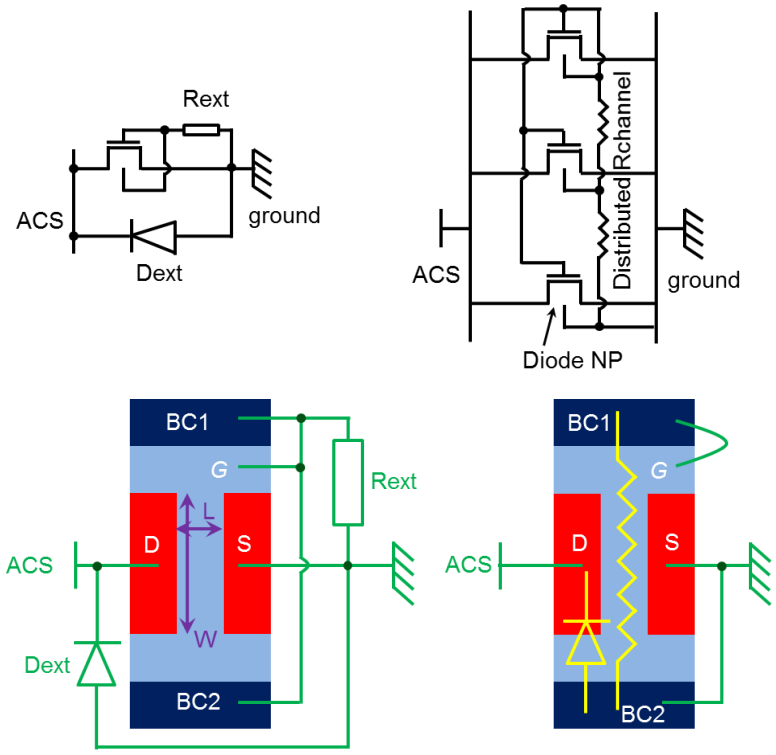


Figure 155: Top: schematics. Bottom: Top view of the T-gate BIMOS topology with two body contacts, with the external connectivity in green and the internal elements in yellow. Left: External resistor (classical circuit). Right: Embedded resistor (new circuit).

The resistance of the Pint doped channel is very high (few MΩ). It is difficult to calculate the exact value of the embedded resistor seen by the BIMOS device (for comparing it to the external resistor) because it is distributed over the channel. Indeed, it is possible to describe the BIMOS with an embedded resistor as multiple NMOS devices with a small width that represent sections of the BIMOS. For example, in Figure 155 (right), three NMOS devices have been used to describe schematically the BIMOS. The top NMOS consists of some drain-channel-source junctions with the channel placed directly next to the body contact that is connected to the gate. The whole resistor along the width of the channel is helping this part of the device to trigger early. The middle NMOS consists of drain-channel-source junctions, with the channel connected to some resistor that has its other terminal connected to the body contact (and gate) node, and the channel is also connected to some other resistor that is grounded. The bottom NMOS has its channel next to the grounded body contact. Also, the value of the embedded resistor changes when there is conduction in the channel.

The very high resistance of the Pint doped channel is very good in ACS performance, because it reduces the trigger voltage. However, it tends to activate the device too soon in AVS (the trigger voltage of the device in AVS would be below 1 V) so the leakage would be too high. Therefore, to decrease the value of the resistor, we can adjust few parameters that can be found in the formula for calculating the value of a simple resistor: $R = \rho \cdot \frac{W}{L \cdot T_{Si}}$ (with ρ the electrical resistivity of the material, W the width of the MOS, L the length of the NMOS and Tsi the thickness of the thin-film in the channel), with the approximation that the resistor is not distributed. The parameters that can be changed by a designer to get a different value of embedded resistor are therefore:

- The width W of the MOS, which has to be the smallest possible to reduce the value of the embedded resistor. Figure 156 shows the effect of the width on the electrical behavior of the BIMOS with an embedded resistor. Since the width has to be very small for staying in the ESD window, the device would be good for being used as a CDM protection or a trigger circuit, but not as an HBM protection. Few fingers can be added for increasing again the dynamic ON resistance of the BIMOS (that will be decreased by the reduction in W of the MOS). Also, those additional fingers will act like parallel resistors, so they reduce the total value of the resistor connected to the gate node.
- The length L of the NMOS has to be the longest possible, for reducing the value of the embedded resistor, but this tends to degrade the ACS (the device trigger voltage will be higher).
- A negative voltage can be applied on the back gate in order to increase the threshold voltage of the AVS characteristic (Figure 157). This negative voltage turns electrostatically the channel into a P-doped region, so the electrical resistivity of the material is changed, which results in a smaller value of embedded resistor.

3D TCAD simulations have been performed at room temperature in order to find empirically a dimension (length L and width W), so that a given T-gate BIMOS exhibits a similar electrical behavior with an embedded or an external resistor of the order of tens of kΩ. In this experience, the back gate is grounded. Figure 158 shows that a T-gate BIMOS with 114 nm width (minimal dimension for implementing metallic contacts on the source and drain) can have a similar behavior whether a typical external resistor is connected to its body contacts or whether it has the connectivity that allows only the embedded resistor to play a role in the triggering, if its length is 300 nm.

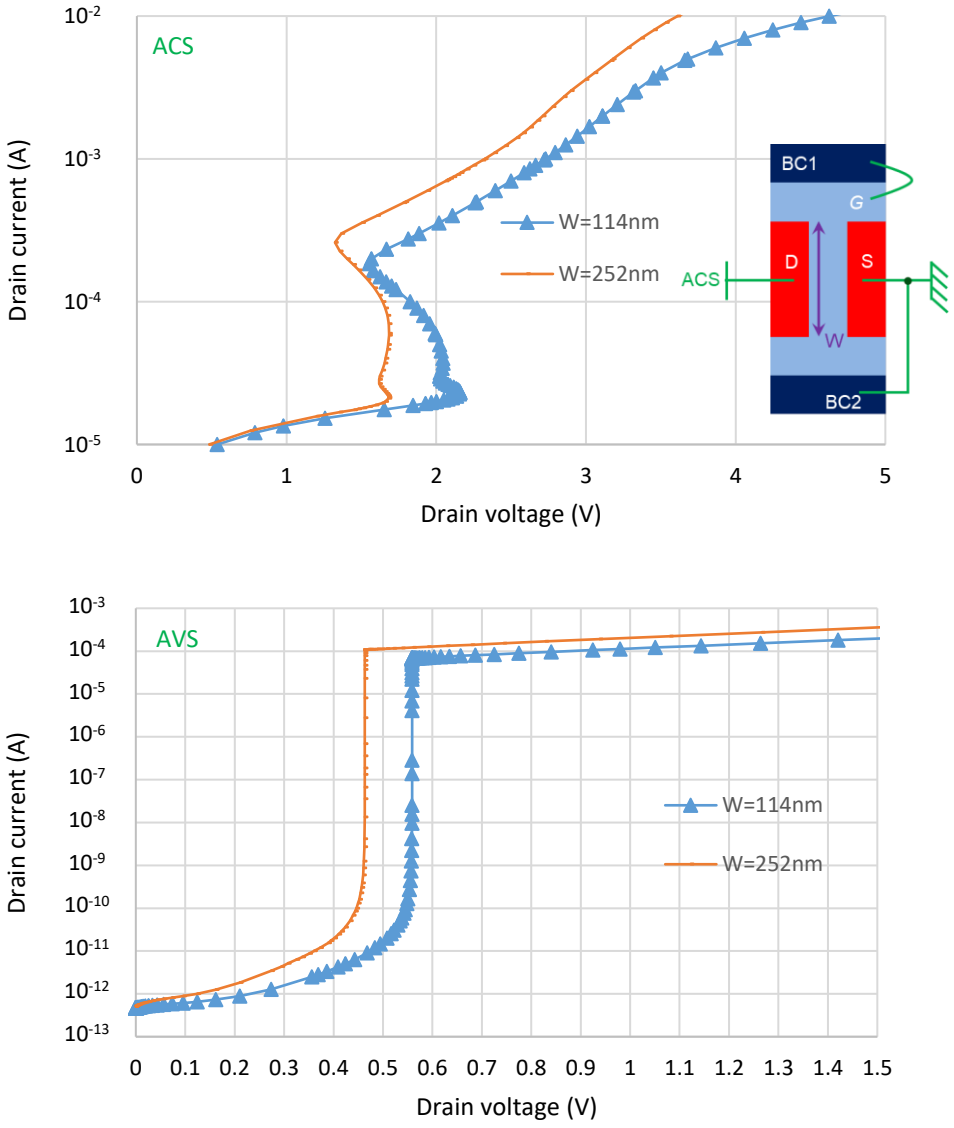


Figure 156: TCAD simulations of a BIMOS with a length of 48 nm. Comparison between the BIMOS with a width of 114 nm (minimal dimension) and a width of 252 nm. Top: ACS. Bottom: AVS.

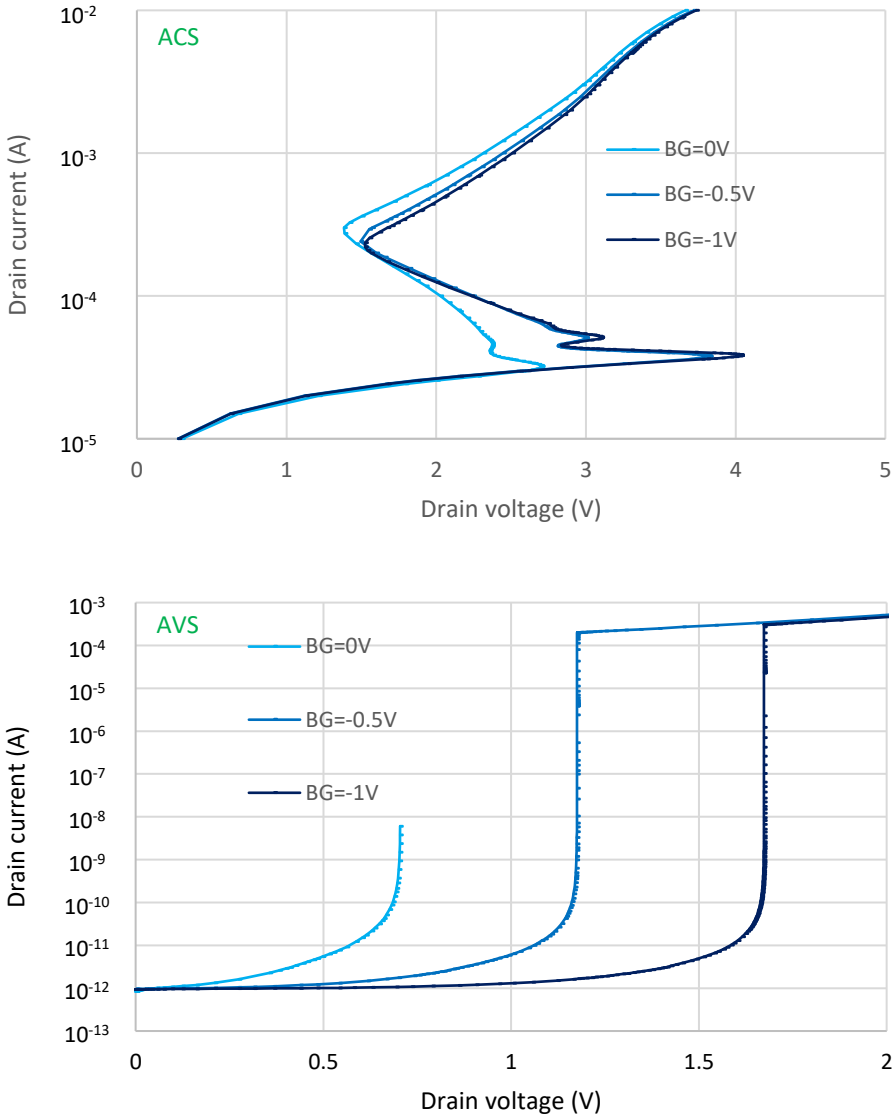


Figure 157: TCAD simulations of a BIMOS with a length of 48 nm, a width of 114 nm (minimal dimensions) and two fingers. Comparison between the BIMOS with a back gate voltage of 0 V, -0.5 V and -1 V. Top: ACS. Bottom: AVS (the AVS curve with a grounded back gate is limited to 0.7 V of drain voltage because of lack of convergence of the simulation). Note that the negative voltage on the back gate shifts significantly the AVS curves, which is interesting for applications.

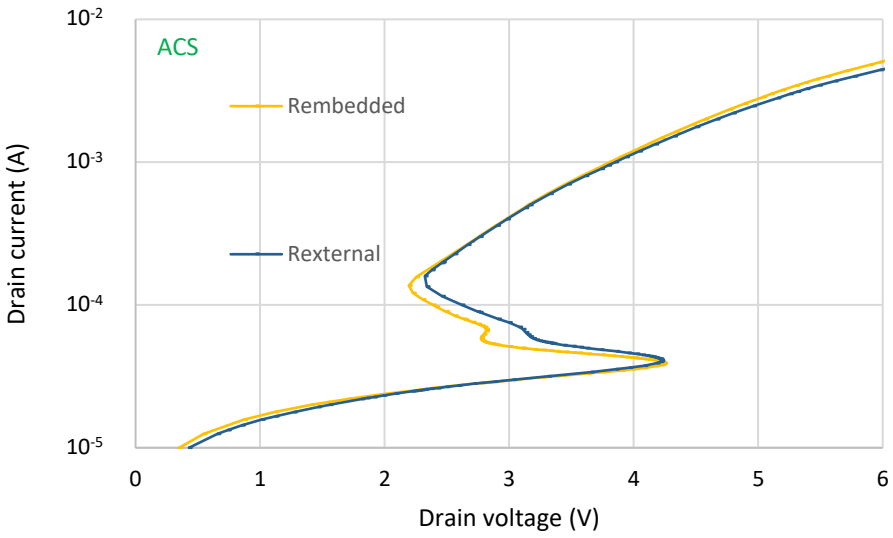


Figure 158: TCAD ACS of BIMOS with a width of 114 nm and a length of 300 nm: comparison between the BIMOS with an external resistor and the one with an embedded resistor (see the connectivities in Figure 155).

2. In-situ coupled bias resistance in hybrid bulk

Our idea is to distribute the resistor over the BIMOS device itself, and to use the P^+ body contact of the BIMOS to form a reverse diode for the negative ESD. In the previous section, this goal was achieved by using the channel as a resistor. In this section, the Pwell will be used as a resistor by opening the BOX (thanks to the NOSO process) inside the P^+ body contacts of the BIMOS (Figure 159). An opening of the BOX is performed in each body contact so that they can communicate together via the Pwell that is under the BOX (Figure 160). This Pwell resistor is additional to the channel resistor (in parallel). Efforts do not have to be made for having only the Pwell resistor without the channel resistor. Indeed, the channel resistor is significantly higher, which leads to the whole current flowing through the Pwell resistor instead. Since the Pwell resistor is situated on the back gate, the voltage on the first body contact (the one connected to the gate) will also help the BIMOS to trigger: a higher voltage on the back gate is shifting the energy bands in the channel, making it more conductive.

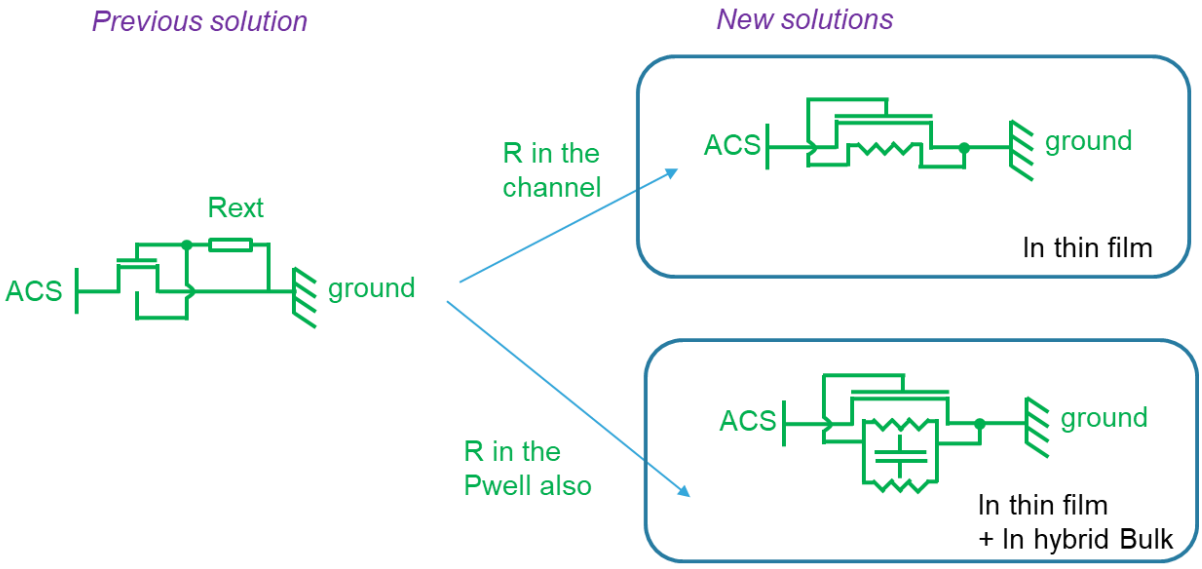


Figure 159: The external resistor of the BIMOS can be inserted in the device to become an embedded resistor. The resistor can be distributed over the channel, or it can also be embedded in the back gate of the BIMOS (by opening the BOX).

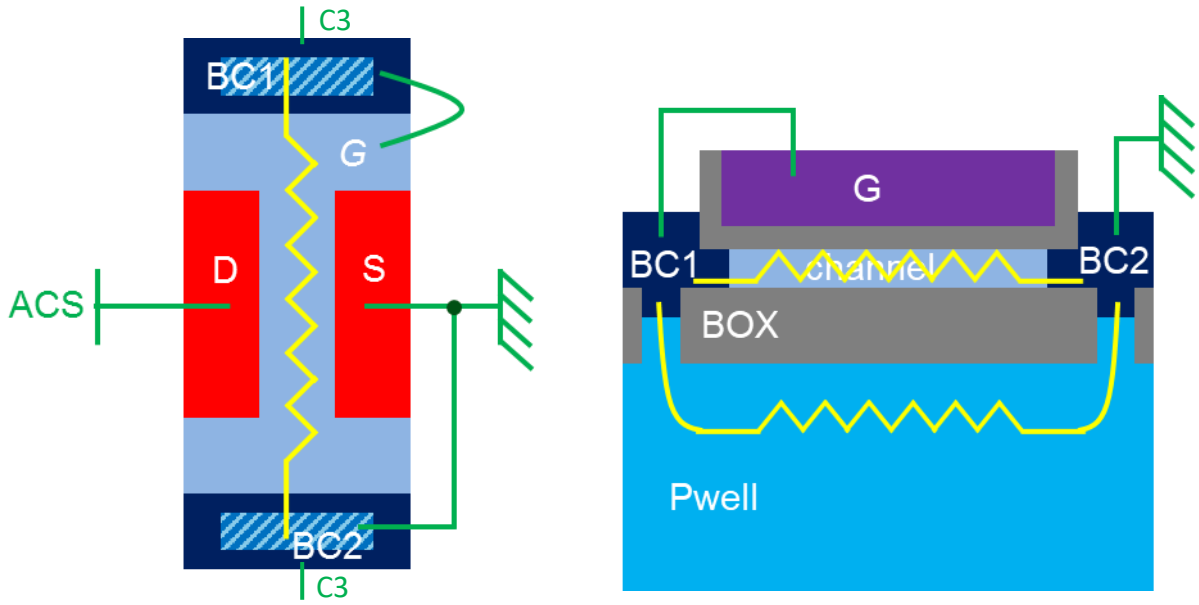


Figure 160: Left: Top view of a BIMOS with an embedded resistor in its Pwell. The hatched regions correspond to openings in the BOX. Right: Cross section C3.

Care is needed to define the dimensions of the BIMOS. Since the Pwell doping is higher than in the channel, the embedded resistor value tends to be too low for being used instead of an external resistor that would be around tens of kΩ. A minimal number of fingers has to be used (ideally one finger), and the STI (limiting the current flow in the embedded resistor) next to the drain region should be the closest as possible to the STI next to the source (this implies minimal dimensions for the drain, gate and source length), so that the resistor is not too wide. The length of the resistor should also be maximized (high width W of BIMOS transistor). A CDM protection could typically have one finger and a small gate length, but its width is of the order of magnitude of 1 µm. The simulated electrical response of BIMOS with different widths, 500 nm (Figure 161) and 2 µm (Figure 162), are shown. On those two graphs the device with an embedded resistor in the Pwell is compared to devices that have different values R_1, R_2, R_3 of external resistors plugged to their gate (with $R_1 < R_2 < R_3$). Due to its width, the 500 nm wide BIMOS with an embedded resistor has a similar behavior as the BIMOS with an external resistor R_1. The 2 µm wide BIMOS with an embedded resistor has a closer electrical behavior to that of the BIMOS with an external resistor R_2. As a matter of comparison, R_3 is the value of resistor that would typically be used in a classical BIMOS circuit with an external resistor. It is possible to decrease further the trigger voltage of the BIMOS embedded by increasing its width, but it would not be reasonable in terms of silicon area (for using it as a CDM protection). In conclusion, the 2 µm wide BIMOS with an embedded resistor could be an acceptable trade-off, provided that measurements follow the same trends.

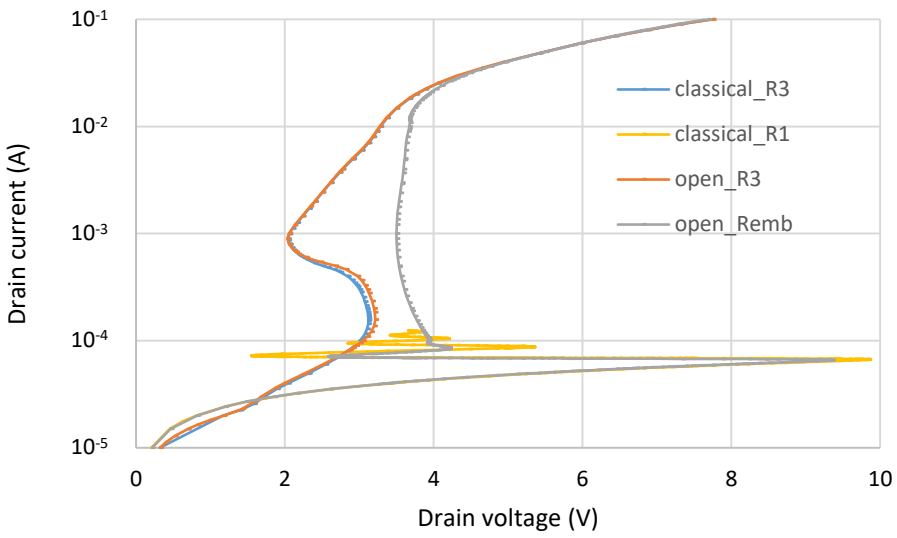


Figure 161: TCAD ACS simulations of BIMOS devices with a length of 48 nm (minimal dimensions) and a width of 500 nm. "classical" in the legend is opposed to "open": it is to compare a BIMOS without BOX opening and a BIMOS where there are openings in the two body contact regions (like in Figure 160). "Remb" means that the electrical circuit allows an embedded resistor in the Pwell (like in Figure 160), i.e. *that the first body contact is plugged to the gate and the second one is grounded. "R3" or "R1" corresponds to the value of the external resistor that is plugged between the gate + body contacts and the ground.*

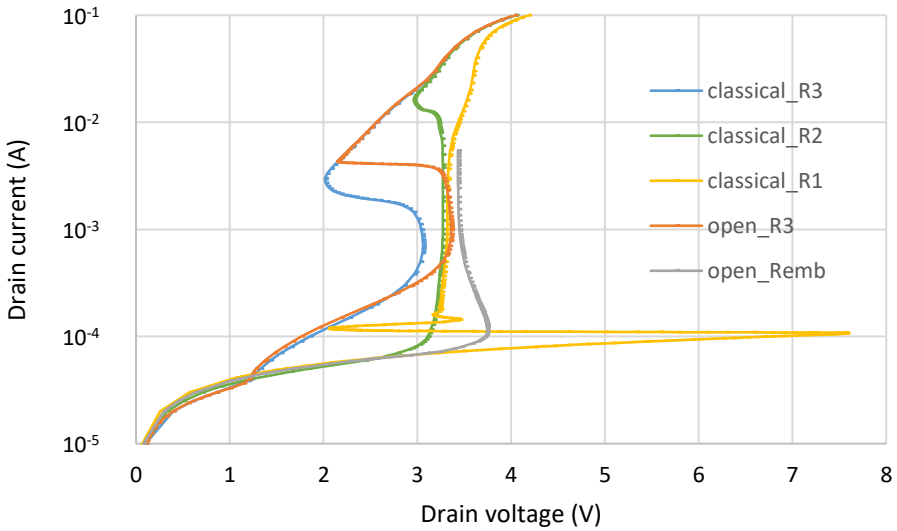


Figure 162: TCAD ACS simulations of BIMOS devices with a length of 48 nm and a width of 2 µm. "classical" (without BOX opening) in the legend is opposed to "open" (openings in the two body contact regions). "Remb" means that the electrical circuit allows an embedded resistor in the Pwell. "R3", "R2" or "R1" corresponds to the value of the external resistor. "open_Remb" curve is restricted to 5 mA because the simulation was stopped for saving computation time.

Note also from Figure 161 and Figure 162 (when comparing “classical_R3” and “open_R3”) that opening the BOX and keeping the electrical circuit with the external resistor (both body contacts and gate linked to the ground via the external resistor) does affect the ACS curve with respect to not opening the BOX. With the opening, the device triggers later. This is contradicting the fact that the back gate (which is now directly connected to the front gate) helps to reduce the trigger voltage. An explanation can be that the voltage on the body contact node is raised slower since it is connected to a big volume (the substrate); therefore, it needs an increased number of carriers for raising its voltage. This hypothesis has to be verified in an extended study.

According to quasi-static simulations, both the embedded version and the one with an external resistor have a low leakage current (around 10^{-11} A) when the drain voltage is 1 V.

In the reverse ACS simulation, the surge (ramp in current) is applied to the node called "ground" and the node called "drain" is grounded. The trigger voltage of the embedded BIMOS is lower than the one of the classical BIMOS in reverse ACS (Figure 163), thanks to the parasitic diode between the grounded body contact and the drain. There is no risk of latch up in a reverse ACS since the operating circuit is not supposed to be biased reversely.

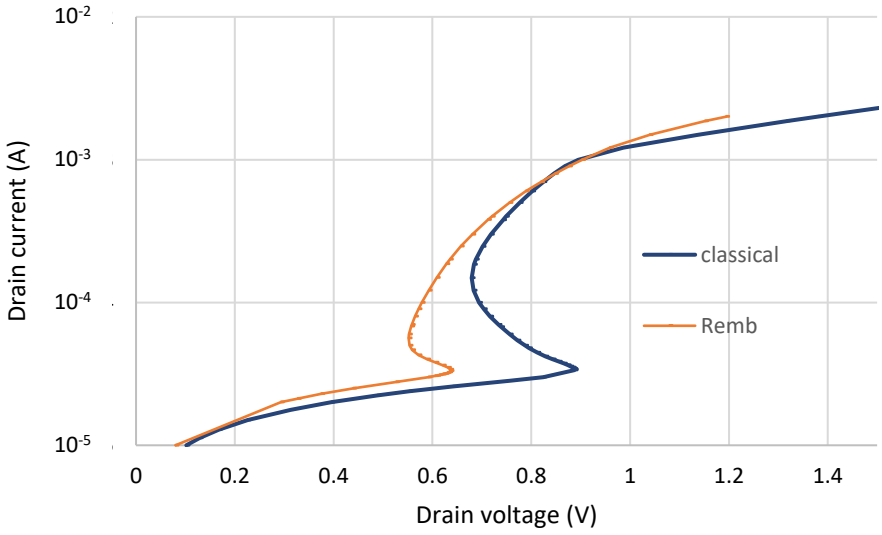


Figure 163: TCAD reverse ACS simulations of a 48 nm gate length and 2 µm gate width BIMOS. "classical" (without BOX opening, and with an external resistor of R_3 plugged between the gate + body contacts and the "ground") in the legend is opposed to "Remb" (with openings in the two body contact regions, and the first body contact is plugged to the gate while the second one is grounded).

Measurements need to be done in 28 nm FD-SOI technology in order to find the real trigger voltage of the embedded BIMOS, and to assess if this device can be used as a CDM protection. The 2 µm BIMOS seems to be a good candidate for future investigations.

Other embedded devices could be imagined (with openings in the BOX and embedded resistor). For example, an embedded resistor can be added in the BIMOS merged SCR from the previous section Figure 164 and Figure 165). Studies have to be performed in order to verify that the device presented in Figure 164 and Figure 165 is triggering uniformly. Indeed, in the part of the device near the second body contact (the one that is grounded), this body contact could prevent the given SCR section to trigger early (because it can act as a grounded trigger Gp).

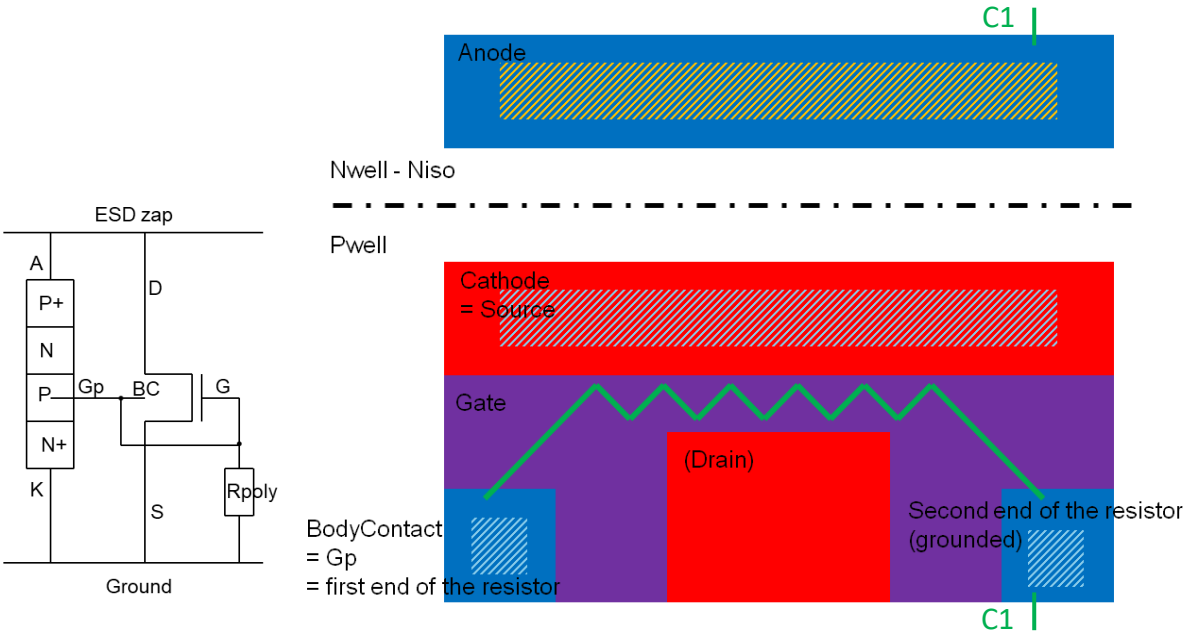


Figure 164: Left: schematic of the BIMOS merged SCR. Right: Top view of the 3D BIMOS merged SCR with an embedded resistor. It is the same BIMOS merged SCR as in Figure 143 except that the connectivity in the body contacts is changed. Red color is for N^+ doping, blue for P^+ doping, and violet for the gate and the gate stack. Hatched regions correspond to the merged regions; under the BOX the doping layers are Nwell (under the anode) and Pwell (under the rest of the device). The embedded resistor is displayed in green. C1 corresponds to the cross-section seen in Figure 165.

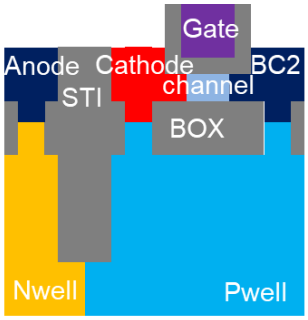


Figure 165: Cross-section C1 (mentioned in Figure 164).

To summarize, with the coupled bias resistance in the channel or in the Pwell it is possible to get by simulation only one merged device - ultra-compact (area reduction also thanks to metal connection relaxation) - that is ESD efficient. Measurements have to be done to verify that the device can be used as a CDM protection (including VF-TLP measurements). The idea is possible to generalize for other nodes and for other gate oxide thicknesses. Other structures that have an external resistor can be studied for placing the resistor in the channel or the Pwell.

To conclude this chapter, new variants of ESD devices have been introduced. Their advantage relies on the idea of 3D conduction, of silicon merging (between top and bottom silicon around the BOX) and device merging. It has been shown that it is possible to create 3D ESD protection devices even with a 2D technology (28 nm FD-SOI technology). The proof of concept has been exposed thanks to TCAD simulations. In the future, silicon characterizations need to be done for proving the feasibility and providing silicon data for assessing the devices performance more precisely.

General conclusions

"Dreaming, after all, is a form of planning."
Gloria Steinem.

The context of this manuscript is to investigate existing ESD protection devices and improve them, in line with the concept of 3D technology, such as for example homothetic and polymorphic 3D devices without metal interconnections.

It was decided to start with the study of FD-SOI ESD devices in the thin-film. There is already a need for improving ESD devices that are in the thin-film because the conduction in the device is not volumic and the level of protection that can be achieved is reduced. There is still room for improvement, in order to broaden the possibilities of fitting into different ESD design windows. The thesis includes a boost solution to reduce the trigger voltage. It consists in adding a finger connected to the drain in a NMOS or a BIMOS device, so that a parasitic capacitance that helps the device to activate sooner is added. This technique could also be used for other devices, for instance with a GDNMOS, which is a gated diode merged with a NMOS transistor. Speaking of the GDNMOS device, it has been intensively studied in the thesis, leading to improvements in the understanding of the trigger mechanism. In fact, the GDNMOS, NMOS and BIMOS devices built in the thin-film are mainly activated through their parasitic capacitances. This phenomenon is taking over the other phenomena such as band-to-band tunneling or impact ionization. Those mechanisms still play a role, but are not as dominant as in the case of devices built in bulk.

The aim of the GDNMOS study was to make the device operate as if it was an SCR. It was underlined that the drain was too heavily doped for this. Lowering the doping, along with removing the silicide present on the drain, was effectively helping to decrease the trigger voltage of the GDNMOS, leading to a similar electrical behavior as an SCR. Yet the leakage current in the lightly doped GDNMOS was unacceptably high. A new idea was proposed to cope with this leakage current. It consists in partially removing the silicide along the width of the drain. This solution allows to shift the electrical parameters, thus enabling to get a whole set of devices that are usable for different supply voltages, hence for different technologies. While studying GDNMOS devices, the importance of plugging external resistances to the gates was underlined. As a matter of fact, not only the resistor helps shifting the electrical parameters, but also it is vital for a robustness purpose. A brand-new thin-film device was also conceived: the GDBIMOS, which is a GDNMOS merged with a BIMOS structure. Theoretically, the BIMOS could improve the dynamic ON resistance of the GDNMOS. This new device also enables additional connectivities. We could imagine for example to connect one body contact to the gate of the diode and the other one somewhere else, just to build a parasitic PMOS transistor in the GDBIMOS device. Some connectivities were investigated in this manuscript, yet there is still room for finding interesting possibilities of connection, since this device has the advantage of being reconfigurable.

Connectivity analyses, addition of new fingers and merging of different devices together, are not the only leverages to improve the performance of devices and understand their physics better. Topology investigation is also possible. With the idea in mind of getting closer to a 3D architecture, 2D devices were studied. By 2D we refer to 2D conduction of current; we thereby typically think of matrices. The easiest ones to be fabricated are NMOS or BIMOS matrices, since they have less terminals to handle than a GDNMOS for example. The BIMOS dot topology is proposed and leads to a BIMOS dot matrix that is more efficient than a classical BIMOS matrix with a body contact at each corner of source or drain. Different 1D BIMOS topologies were also compared to a matrix of NMOS with an external ring of body contact. Results showed that squares of sources or drains should be as small as possible in order to beat 1D devices in terms of electrical performances, although process constraints do not allow yet to reduce sufficiently the dimensions. This result should be kept in mind, since it is probably only a question of time before process constraints are relaxed. One day, we may even think of a 3D matrix of BIMOS.

Merging different devices together also went inspirational for thinking of a new type of structure with 3D conduction. The idea was to dig a hole in the BOX in the middle of the thin-film device, in order to merge it with another device present in the substrate. Thereupon a 3D BIMOS merged SCR was presented. It presents the advantage of using the surface conduction for the trigger circuit (which is the BIMOS) and the conduction in the volume for the SCR, leading to a high robustness and level of ESD protection. The same technique of opening the BOX can also be used to no longer require an external resistor in a BIMOS device.

All the devices studied during the thesis have been designed and simulated. A great care has been taken so that the fabrication was feasible, even in the case of 3D structures. Process compliance was an important criterion since I had the wish of immediate application. Some devices were fabricated and measured. Others, like the 3D BIMOS merged SCR or the BIMOS with embedded resistor still need to be fabricated. The next person that will handle a similar topic still has room for more understanding and innovations. The verification of the simulation results with silicon measurements has to be pursued. From an industrial point of view, ensuring a good yield and reducing the shift of electrical parameters due to process variability is essential for commercialization.

This work paves the way of ESD protection devices that do not exist yet in the industry. Several innovative devices were proposed, including 3D designs.

Appendix 1: TCAD setup

A set of physical device equations that describes the carriers' distribution and conduction mechanisms have to be specified for the TCAD computation. The system of coupled equations to be solved in our TCAD simulations at each meshed node is the following:

- **Poisson equation**

The Poisson equation is used to compute the electrostatic potential φ for a given charge distribution ρ. Knowing the electric field ε:

$$\vec{\varepsilon} = -\vec{\nabla}(\varphi)$$

and Gauss's law:

$$\vec{\nabla} \cdot \vec{\varepsilon} = \frac{\rho}{\epsilon}$$

the Poisson equation is:

$$\Delta\varphi = -\frac{\rho}{\epsilon}$$

where ϵ is the local material dielectric constant and ρ is the positive net charge density:

$$\rho = q \cdot (p - n + N_D^+ - N_A^-)$$

with n and p respectively the electrons' and holes' densities; q is the elementary electronic charge; and N_D^+ and N_A^- are respectively the concentration of ionized donors (n-doped semiconductor) and acceptors (p-doped). The charge density contribution from traps and fixed charges has been neglected in this equation.

- **Continuity equations**

The charge conservation principle allows to obtain an equation describing the time and space evolution of the charges' concentrations. The time variation of the number of electrons n and holes p depends on the currents $\vec{J}_n$ and $\vec{J}_p$ due to the spatial movement of carriers, and on their generation (G_n and G_p) and recombination (R_n and R_p) rates.

For electrons:

$$\frac{\partial n}{\partial t} = \frac{1}{q} \cdot \vec{\nabla} \cdot \vec{J}_n + (G_n - R_n)$$

For holes:

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \cdot \vec{\nabla} \cdot \vec{J}_p + (G_p - R_p)$$

- **Transport equations**

The total current of carriers is the addition of a conduction current due to an electric field, and a diffusion current due to a gradient of carriers (drift-diffusion model).

Drift current corresponds to carriers' motion induced by an electric field ε. The microscopic Ohm's low states that the current density $\vec{J}_{drift} = \sigma \cdot \vec{\varepsilon}$ where σ is the electrical conductivity due to the free electrons in the conduction band and to the free holes in the valence band. σ depends on the elementary electronic charge q, the mobility of the carriers and their density.

Diffusion is the natural tendency of a gas to make the particle's concentration spatially uniform, and the diffusion intensity is proportional to the concentration gradient. Since particles in a semiconductor carry a charge, a current density is associated to their diffusion.

The total drift-diffusion current for each type of carrier is therefore:

$$\vec{J}_n = q \cdot \mu_n \cdot n \cdot \vec{\varepsilon} + q \cdot D_n \cdot \vec{\nabla}(n)$$
$$\vec{J}_p = q \cdot \mu_p \cdot p \cdot \vec{\varepsilon} + q \cdot D_p \cdot \vec{\nabla}(p)$$

with μ_n and μ_p the mobility of carriers, and D_n and D_p the diffusion coefficients of carriers. They are linked by the Einstein relation if the approximation is made that thermal equilibrium (absence of any energy exchange with the environment) holds:

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{k_B \cdot T}{q} = V_T$$

with V_T the electrical equivalent of the temperature; k_B is the Boltzmann constant and T is the temperature.

- **Contact equations**

All contacts (electrodes) on semiconductors are ohmic, subject to charge neutrality and equilibrium. When all contacts of a device are biased to the same voltage, the device is in equilibrium, and the electrons' and holes' densities are described by a constant Fermi potential. For an electrode placed on a semiconductor region, the charge is computed from Gauss's law, and the space charge in the doping well is taken into account.

For contacts on insulators, the electrostatic potential is:

$$\varphi = \varphi_F - \varphi_{MS}$$

with φ_F the Fermi potential at the contact (it is equal to the applied voltage because the contact is not resistive), and φ_{MS} is the work function difference between the metal and an intrinsic reference semiconductor. In our simulations the value of φ_{MS} has been chosen thanks to the calibration of the TCAD curves with a reference NMOS device. For an electrode placed on an insulator region, the charge is computed from Gauss's law.

- **Circuit equations**

The connectivity of the circuit is considered. Some external resistances, current and voltage sources, or other additional external devices can be added (lumped elements).

- **Heat equations**

For electro-thermal simulations, the self-heating of the device is taken into account by adding the lattice temperature equations to the previous set of equations:

$$\frac{\partial}{\partial t}(C_L \cdot T) - \vec{\nabla} \cdot \left(k \cdot \vec{\nabla}(T)\right) = -\vec{\nabla} \cdot [(P_n \cdot T + \phi_n) \cdot \vec{J}_n + \left(P_p \cdot T + \phi_p\right) \cdot \vec{J}_p]$$

with C_L the lattice heat capacity, T the lattice temperature, k the thermal conductivity, P_n and P_p the electron and hole thermoelectric power respectively (it accounts for the Seebeck effect), ϕ_n and ϕ_p the electron and hole quasi-Fermi potential respectively.

A thermal electrode (thermode) is defined in order to apply a temperature boundary condition. For example, room temperature is applied on an electrode situated under the bulk region situated below the BOX of the device.

Physical models are selected to be included in the numerical resolution of the previous equations. In our simulations, the default model parameters were left unchanged. The following models were used:

- **Boltzmann statistics**

Electrons' and holes' densities were computed from the electron and hole quasi-Fermi potentials using the Maxwell-Boltzmann statistics.

In order to estimate the density of free carriers, the energy distribution function is computed. It is the product of the number of available states by the probability for each state to be occupied. In thermal equilibrium, the occupation probability for available states is given by the Fermi-Dirac distribution. Boltzmann statistics is a simplification of the Fermi-Dirac statistics for non-interacting particles. If the Boltzmann statistics holds (for a non-degenerate semiconductor, *i.e.* when the semiconductor is not highly doped), the following formula can be derived (in equilibrium):

$$n = N_C \cdot e^{-\frac{E_C - E_F}{k_B \cdot T}}$$

$$p = N_V \cdot e^{-\frac{E_F - E_V}{k_B \cdot T}}$$

where E_F is the Fermi level: it is the thermodynamic work required to add one electron to the material; it corresponds to the energy level where the occupation probability is ½. E_C and E_V are respectively the lowest energy of the conduction band and the highest energy of the valence band. N_C and N_V are the effective band edge densities of states.

- **Semiconductor band structure**

A silicon band gap narrowing model was included to determine the intrinsic carriers' concentration.

For an un-doped semiconductor, we have n = p = n_i; n_i being the intrinsic density. If we consider a non-degenerate semi-conductor in thermal equilibrium, the mass action law is:

$$n_i{}^2 = n \cdot p = N_C \cdot N_V \cdot e^{-\frac{E_G}{k_B \cdot T}}$$

with $E_G = E_C - E_V$ (the gap energy).

The effective band gap results from the band gap reduced by band gap narrowing. This band gap narrowing occurs when the semi-conductor is degenerate (with high doping or with high current injections).

In fact, donor impurities create energy levels in the band gap near the conduction band, and acceptor impurities create energy levels near the valence band. If the doping level is important, the density of states of these dopants is increasing, and their effect significantly affects the effective band gap.

By default, band gap narrowing is active in all TCAD simulations. Different band gap models are available, and the Slotboom model has been chosen [60]. This model is based on measurements in n-p-n and p-n-p transistors with different base doping concentrations.

In devices that contain different materials, the electron affinity (the difference between the lowest energy in the conduction band E_C and the vacuum level) is also important. Along with the band gap, it determines the alignment of conduction and valence bands at material interfaces. The affinity is affected by band gap narrowing.

- **Generation-recombination models**

In our TCAD simulations, the SRH model (Shockley–Read–Hall recombination with doping-dependent lifetime) was used along with Auger recombination model, avalanche van Overstraeten-de Man model (electron–hole pair generation by impact ionization) and non-local path band-to-band tunneling model.

SRH recombination, also called trap-assisted recombination, is an indirect mechanism: transitions of carriers are assisted by recombination centers (inside the band gap). Those defect energy levels (traps) are due to lattice defects. The net recombination rate ($U = R - G$) depends on the electron and hole lifetime. In the SRH model the electron and hole lifetimes depend on the doping, the field and the temperature. This process is significant in silicon material.

The Auger recombination is a direct mechanism of band-to-band transition: carriers are directly exchanged between the conduction and the valence band. When an electron and a hole are recombined, it releases an energy that is transferred to a third particle. In case the third particle is an electron, it is transferred at a higher energy level into the conduction band; if it is a hole, it is pushed deeper into the valence band. Auger recombination rate is important at high carrier densities.

Impact ionization happens when the electric field exceeds a threshold field. Free carriers are accelerated by the electric field and collide with the lattice atoms, thus creating electron-hole pairs. Those electron-hole pairs are accelerated in turn. If the space charge regions are wider than the mean free path between two ionizing impacts (wide space charge regions are obtained thanks to light doping in the junction wells or when the voltage across the junction is high), those electron-hole pairs also collide, thus creating new pairs. This chain reaction of charge multiplication is called avalanche and typically causes the junctions breakdown.

The band-to-band tunneling model implemented in the recombination section of TCAD addresses the non-local generation of electrons and holes caused by band-to-band tunneling. Band-to-band tunneling occurs when the energy bands are sufficiently bended by the electric field (due to an applied potential or to a sharp doping difference); then electrons in the valence band - at a certain location A - can reach the conduction band in a different location B thanks to direct or phonon-assisted band-to-band tunneling, without the assistance of traps. For direct tunneling, an electron tunnels without the absorption or emission of a phonon. In the indirect tunneling process, the electron changes its momentum because of the absorption or emission of a phonon. Phonon-assisted tunneling process is dominant in indirect semiconductors such as silicon. As a consequence, holes are generated at location A and electrons are generated at location B, and the net hole and electron recombination rates at each location can be calculated considering the non-local path band-to-band tunneling.

- **Mobility models**

Free carriers in the material gain momentum from electric fields and lose momentum through scattering with perturbations to the spatial periodicity of the lattice potential. Phonon scattering, doping-dependent mobility degradation, mobility degradation at interfaces (with transverse field dependence), and mobility degradation due to high field saturation were considered in our TCAD simulations. The different mobility contributions are combined by Matthiessen's rule.

Mobility due to Phonon Scattering is only dependent on the lattice temperature. It is degraded in doped materials because the carriers scatter with the charged impurity ions.

In the channel region of a MOSFET, the high transverse electric field (perpendicular to the semiconductor–insulator surface) forces carriers to interact strongly with the interface. Thus, the mobility degradation model at interfaces is very important to be implemented. Mobility is degraded at interfaces because the carriers are subject to scattering by acoustic surface phonons and surface roughness.

For small electric fields, the carrier drift velocity is proportional to the electric field, therefore the mobility of carriers is constant. For large electric fields, carriers gain enough energy to excite high-energy lattice vibrations (optical phonons), which slows them down. As a consequence, the drift velocity saturates for large electric fields.

Appendix 2: AVS behavior of the BIMOS

In the case of AVS, we observe the quasi-static behavior of the BIMOS device. An extensive work has been performed by Thomas Bedecarrats [103] to understand the BIMOS in DC. Here is a summary of some of his work: the I-V curve can be divided into four regions (Figure 43).

In the region 1, the increase in drain voltage polarizes the NMOS part of the BIMOS (drain-channel-source). The MOSFET polarization makes electrons flow from source to drain (channel current). Since the gate voltage is zero, this channel current $I_{Channel}$ is a diffusion current, and its intensity is driven by the voltage between the source and the gate. Since the gate voltage stays at zero (because no charges are going in the gate node), $I_{Channel}$ stays constant, even if the drain voltage increases.

In the region 2, the electrical field (situated at the frontier of the drain and the channel) that is built by the voltage difference between the drain and the gate – is becoming sufficiently high to generate electron-holes pairs by band to band tunneling. Electrons go to the drain (because it's the highest potential), and holes go to the body contact (because it is a P-doped low potential). It is the band to band current I_{BTBT}. When the holes reach the metallic contact of the Body Contact, they recombine with electrons of the contact. A positive charge appears in the Body Contact node, thus the Body Contact potential V_{BC} increases. In this region, V_{BC} is comprised between 0 and V_{TH} (the threshold voltage of the MOS). Because of the connectivity of the circuit, with the body contact connected to the gate and to an external resistor, an increase in body contact potential can also be seen as an increase in gate potential and an increase in the resistor voltage.

If the resistor voltage increases, a small current I_R flows through the resistor. This current evacuates the charges that are accumulated in this Body Contact node; therefore, the Body Contact potential decreases.

If the gate potential increases, the channel current increases by MOSFET effect. The voltage difference between the drain and the gate decreases, so the electrical field at the drain and channel frontier decreases, which reduces I_{BTBT}.

If the Body Contact potential increases, the gated diode built by the Body Contact (P-doped) and the source (N-doped) is positively biased. Therefore, a diode current I_{Diode} starts to flow (holes going from the body contact to the source and electrons from the source to the Body Contact). As long as the Body Contact voltage is not high enough, this current is low, and in the region 2 it is negligible in front of I_R and I_{BTBT}.

In the end, a stationary state is reached, where $I_R = I_{BTBT}$, and a charge is stored in the Body Contact node. The Body Contact voltage is a function of the drain voltage. The final drain current that is seen corresponds to $I_{Channel}$ and I_{BTBT}.

In the region 3, the BIMOS goes from OFF state to ON state. This happens when the V_{BC} reaches V_{TH} (V_{BC} increased thanks to I_{BTBT}). Because V_{BC} is also the gate voltage, an increase in V_{BC} increases also the MOS conduction, so $I_{Channel}$ increases. With an increase in $I_{Channel}$, an impact ionization current I_{II} (more electrons flowing in the channel will create

more impact ionization by ionizing the atoms of the silicon lattice) appears, and its contribution becomes much greater than I_{BTBT}. Like for I_{BTBT}, the generated (by impact ionization) holes flow toward the Body Contact, and the electrons go to the drain. With this new current I_{II}, the charge that is stored in the Body Contact node is much bigger, therefore V_{BC} increases. The current I_R that discharges the Body Contact node is limited by the Body Contact voltage, and I_R is not sufficient anymore to evacuate the charges brought by I_{II}. The increase in Body Contact voltage increases $I_{Channel}$ very much, and so on. The feedback loop responsible from this run-away phenomenon (where V_{BC} increases independently from the drain voltage) can therefore be described as such: V_{BC} increases => $I_{Channel}$ increases => I_{II} increases => V_{BC} increases …

In the region 4, V_{BC} is higher than V_{TH}. The current is mostly constituted by $I_{Channel}$. In this region, I_{Diode} is significant (and I_R negligible in front of I_{Diode}). A stationary state is reached, where I_{II} = I_{Diode}, and the body contact voltage is again a function of the drain voltage.

Appendix 3: Résumé étendu en français

"Quand google ne trouve pas quelque chose,
il demande à Chuck Norris."

This section offers an extended summary of the manuscript content in French.

Chapitre 1 : Introduction

I. Contexte et objectifs

Les circuits intégrés qui utilisent des transistors MOSFET (de l'anglais « Metal Oxide Semiconductor Field Effect Transistor ») ont grandement évolués lors de ces dernières décennies, grâce à la miniaturisation des composants élémentaires. Il est cependant très difficile de miniaturiser les transistors à l'extrême, et certaines limitations physiques de la technologie planaire sont tellement fortes qu'une façon de continuer la course à la miniaturisation est de changer l'architecture intrinsèque du MOSFET. L'une des voies d'amélioration est la technologie FD-SOI (pour « Fully Depleted Silicon-On-Insulator ») (par opposition à la technologie traditionnelle dans le substrat), qui consiste à placer les transistors dans un film mince de silicium, qui se situe au-dessus d'une couche isolante d'oxyde appelée BOX (pour « Buried Oxide »), elle-même située au-dessus de la couche principale de silicium qui constitue le wafer, appelée substrat. Une autre façon d'améliorer les performances des circuits est d'adresser l'intégration dans les trois dimensions (3D), afin d'augmenter la densité des composants par surface et de réduire les interconnexions métalliques. Pour un empilement 3D avec une précision de l'ordre d'un transistor, l'une des techniques les plus prometteuses est celle proposée par le projet Coolcube™, qui consiste à fabriquer directement des couches de transistors (avec leur premier niveau d'interconnexion) les unes sur les autres (pour en savoir plus, voir la Figure 31 et la référence [10]). D'autres implémentations 3D peuvent aussi être imaginées. Ce domaine est relativement nouveau et beaucoup d'opportunités et de solutions peuvent être envisagées. Toutes ces technologies, qu'elles soient planaires ou en 3D, sont extrêmement sensibles aux décharges électrostatiques (ESD pour « Electrostatic Discharge ») et doivent être protégées contre ce stress. En effet, les événements ESD peuvent survenir à n'importe quel moment de la vie du composant et impliquent de forts courants et tensions. Dans ce manuscrit, certaines solutions vont être présentées en termes de schéma, de topologies et de performances de protection.

Ainsi l'objectif principal de la thèse est de concevoir des composants de protection contre les ESD sur film mince de silicium en technologie 28 nm FD-SOI de chez STMicroelectronics (technologie silicium sur isolant « Silicon-On-Insulator » (SOI) entièrement déplété « Fully Depleted » (FD)). Cette technologie est caractérisée par un film de silicium, un oxyde enterré ultra minces (UTBB), et par une grille métallique avec oxyde à haute permittivité (« high-k »).

Notamment, le premier chapitre donne des éléments pour comprendre la technologie et l'importance de protéger les circuits contre les ESD. Les modèles principaux d'ESD sont donnés, ainsi que la définition de ce qu'on attend de la part d'une protection contre les décharges électrostatiques. Un état de l'art des protections ESD existantes est également fourni. Les simulations numériques et les mesures effectuées tout au long de cette thèse sont introduites.

Le chapitre 2 est dédié aux protections ESD en 1D dans le film mince, 1D au sens où le courant circule dans une direction privilégiée dans le composant. Classiquement les protections ESD ne sont pas réalisées dans le film mince : le BOX est ouvert et la protection est réalisée dans cette ouverture, afin qu'elle puisse bénéficier d'un courant volumique dans le substrat. Cependant il est intéressant de pouvoir réaliser des protections efficaces dans le film mince. Entre autres, l'une des raisons est de rendre possible la protection des couches du dessus de la technologie Coolcube™. Des protections ESD existantes dans le film mince sont étudiées et améliorées dans ce chapitre.

Dans le chapitre 3, les matrices, qui sont des protections avec une conduction de courant en 2D, sont particulièrement étudiées. Elles démontrent des performances intéressantes tant d'un point de vue topologique qu'électrique.

Enfin dans le chapitre 4, il est proposé d'investiguer des protections innovantes avec un courant de conduction en 3D mettant en jeu à la fois la couche de film mince et le substrat massif.

II. Présentation du MOSFET en technologie FD-SOI

La technologie FD-SOI présente de nombreux avantages : les effets de canal court sont réduits, certains effets parasites sont supprimés (de par la présence du BOX qui isole le composant actif des autres et des éléments parasites du substrat), et il y a moins de chemins pour les courants de fuite et de « latch up ». L'un des seuls éléments parasites restant est la capacité du BOX, mais celle-ci peut être considérée comme une opportunité de mieux contrôler le composant en utilisant le substrat comme une grille arrière. Le contrôle électrostatique de la grille avant est meilleur grâce aux jonctions qui sont moins profondes en technologie FD-SOI. Enfin, les composants dans le film mince ont des mécanismes de mise en conduction différents qui offrent des possibilités additionnelles d'implémentations.

Le MOSFET, lorsqu'il est réalisé dans le film mince, possède quatre terminaux : la grille (avant), le substrat (qui peut être considéré comme une grille arrière), la source et le drain (Figure 4). Il est principalement utilisé comme un interrupteur. En effet, avec une tension sur les grilles, le transistor peut laisser passer du courant entre la source et le drain

(ce qui correspond à un état passant), et si les grilles sont branchées à la masse, le courant ne circule pas (et le transistor est dit bloqué). Si le transistor est de type N (d'où le nom du composant NMOS), alors les porteurs dans le canal sont des électrons. Les différentes caractéristiques du NMOS en technologie 28 nm UTBB FD-SOI sont les suivantes : Il est possible de doper le substrat de type P (on l'appelle alors Pwell), ou de le doper de type N (Nwell). Le canal, qui se situe sous la grille, est laissé avec son dopage intrinsèque, c'est-à-dire qu'il est faiblement dopé de type P (dopage Pint). Son épaisseur (7 nm), et celle du BOX (25 nm) sont très fines. La source et le drain sont surélevés et très fortement dopés N (dopage N^+). Des extensions de dopage un peu plus faible (dopage N-LDD) sont ajoutées au niveau de la jonction entre le canal et la source ou le drain afin de limiter le champ électrique latéral dans le NMOS. La grille métallique est faite de poly-silicium et de nitrure de titanium (choisi pour son travail de sortie). En dessous de chaque contact métallique il y a du siliciure pour réduire la résistance d'accès. Les transistors sont isolés les uns des autres par des tranchées profondes d'oxyde que l'on nomme STI (« Shallow Trench Isolation »).

III. Les décharges électrostatiques

1. Définition des ESD et importance des protections

Une décharge électrostatique (ESD) est le courant soudain entre deux objets chargés. Les ESD peuvent arriver n'importe quand dans la vie d'un circuit (pendant sa fabrication ou après), et impliquent différents ordres de grandeur de courant et de tension. L'environnement de fabrication des circuits étant très contrôlé dans l'industrie, les circuits peuvent subir des ESD allant jusqu'à 4 kV, et plusieurs ampères en quelques nanosecondes, ce qui peut déjà s'avérer très destructif si le circuit n'a pas été protégé en ce sens. Une grande partie des retours de circuits défectueux sont attribués aux ESD, c'est pourquoi il existe un fort besoin de réaliser des composants de protection contre les ESD et de les implanter judicieusement dans les circuits.

2. Stress ESD standards

L'ESDA (Association ESD) a été créée pour établir des standards, et afin de partager les dernières améliorations entre les chercheurs et les industriels, en matière de recherches pour parer aux ESD. Dans ce cadre, trois modèles d'ESD ont été établis en tant que standards principaux utilisés en micro-électroniques :

- Le modèle HBM (modèle du corps humain chargé, de l'anglais « Human Body Model »), qui décrit une décharge provenant d'un humain chargé qui touche un composant électronique.
- Le modèle MM (modèle de la machine chargée, de l'anglais « Machine Model »), pour une décharge provenant d'une machine chargée.

- Le modèle CDM (modèle du composant chargé, de l'anglais « Charged Device Model »), qui est utilisé pour décrire un composant lui-même chargé, qui se décharge sur un autre objet.

Ces différentes configurations sont chacune représentées par un circuit équivalent RLC, ainsi une équation différentielle du second ordre peut décrire ces ESD.

Dans ce manuscrit, il sera principalement question de protéger des circuits contre des décharges HBM et CDM. L'intensité de la décharge HBM dépend de la tension de précharge, considérée comme valant entre V_{HBM} = 1 kV pour les applications RF (fréquence radio), et 4 kV pour les applications militaires. L'ESDA est constamment en train de faire évoluer ces exigences car il est de plus en plus difficile de fournir des protections ESD capables de supporter de telles tensions. On peut retenir qu'une décharge HBM peut typiquement délivrer un courant pic de 1 A (ce qui correspond à V_{HBM} = 2 kV), et d'une durée de 100 ns avec un temps de montée de 10 ns. La décharge CDM dure environ 1 ns avec 100 ps de temps de montée. Malgré sa courte durée, elle peut être très destructive car elle implique des courants de l'ordre de la dizaine d'ampères.

3. La fenêtre de conception ESD

Il existe deux stratégies complémentaires afin de protéger les circuits contre les décharges électrostatiques. La première consiste à empêcher les décharges de se produire, grâce à la maîtrise de l'environnement de fabrication du circuit (mettre à la masse toutes les surfaces pouvant toucher les composants, contrôler l'humidité dans l'air, etc.). Ce n'est malheureusement pas suffisant, d'où la seconde stratégie, qui a pour but de dévier la décharge dans le circuit grâce à des composants de protection dédiés, afin que la décharge n'affecte pas les composants du cœur du circuit.

Le rôle de la protection ESD est d'évacuer une quantité suffisamment importante de courant, tout en limitant la tension aux bornes de la zone à protéger, en cas d'événement ESD. La protection doit être transparente lorsque le circuit fonctionne normalement, ce qui implique que pour une tension de fonctionnement normale V_{DD}, le courant de fuite I_{Leak} de la protection doit être le plus faible possible, et le composant de protection ne doit pas être actif. La caractéristique électrique I-V idéale d'une protection (voir la Figure 14) correspond à un composant qui est normalement fermé, mais apte à s'ouvrir de façon abrupte, pour une tension donnée. Cette tension de déclenchement est appelée V_{T1}. Si la courbe présente un retournement, on appelle tension de maintien V_H la plus petite tension appliquée au composant lorsque celui-ci est ouvert. La résistance dynamique à l'état passant est appelée R_{ON}. I_{T2} et V_{T2} correspondent respectivement au courant et à la tension de casse du composant de protection. La protection ESD doit s'activer avant que les composants du circuit intégré ne cassent, pour une tension d'avalanche V_{BD}. Ainsi, il y a ce qu'on appelle la fenêtre de conception ESD (en anglais « ESD design window »), qui établit que la protection ESD doit être ouverte (c.a.d. avoir une faible impédance) pour un intervalle de tension $[(V_{DD}+10\%) – (V_{BD}-10\%)]$. Si V_H est trop proche de V_{DD}, il y a un risque de « Latch Up » : une fois que la protection est activée, elle reste ouverte même si la décharge est déjà passée et que le circuit est à nouveau en fonctionnement normal ; d'où la marge de 10%. La fenêtre ESD dépend de la technologie employée du circuit intégré. La protection ESD doit être

efficace et robuste : elle ne doit pas casser avant d'avoir suffisamment protégé le circuit intégré, donc son courant de casse I_{T2} doit être le plus haut possible. En fait, I_{T2} doit excéder la valeur de courant pic de la décharge HBM (qui dépend de la norme HBM à appliquer). Le composant de protection doit s'activer très vite, car les ESD se produisent en un temps très court (1 – 100 ns).

Non seulement le composant de protection doit être conforme à la fenêtre ESD lors de mesures quasi-statiques, mais il doit également l'être avec d'autres méthodes de mesures. En effet, certaines surtensions transitoires peuvent apparaître avant que la protection n'atteigne son comportement quasi-statique, et ces surtensions peuvent être à l'origine de l'échec de la protection du circuit considéré.

De nos jours, il est de plus en plus difficile de faire correspondre les protections ESD à la fenêtre ESD, car la différence entre V_{BD} et V_{DD} diminue. Une autre difficulté concerne la surface de silicium que prennent ces protections. En effet, afin d'obtenir un grand I_{T2}, les protections doivent être dessinées avec une grande largeur W, afin que la décharge puisse circuler à travers une large section de silicium. Les solutions proposées sont assujetties à la contrainte de minimiser leur taille effective (les protections ESD prennent une place significative sans même servir à la fonctionnalité du circuit), ce qui explique aussi pourquoi les normes HBM et CDM se font régulièrement mettre à jour.

4. Les stratégies de protection

Les stratégies de protection sont implémentées afin de mutualiser les composants de protection, pour ne pas avoir à en utiliser un trop grand nombre (ce qui nuirait à la surface du circuit).

Pour les décharges HBM et MM, le circuit primaire de protection est utilisé. Des composants de protection sont placés à chaque entrée ou sortie du circuit. Ces composants peuvent être unidirectionnels ou bidirectionnels (les protections bidirectionnelles prennent en général plus de surface), mais quelle que soit la stratégie, les protections doivent être placées de sorte à parer aux événements ESD qui ont une polarité positive ou négative, et ce pour n'importe quelle combinaison de plots d'entrée/sortie. Un circuit de déclenchement peut aider les protections à se déclencher dans la fenêtre ESD. Deux catégories de stratégies de protection peuvent être décrites : la stratégie locale, et la stratégie globale (avec protection centrale et distribuée). Pour avoir une description plus exhaustive, voir la section sur les stratégies de protections dans la partie en anglais du manuscrit.

Pour les décharges CDM, il est complexe de prévoir le chemin d'un ESD car la décharge provient du circuit lui-même. Les éléments de protection doivent être petits et localisés de façon stratégique dans le circuit ; ils sont prévus pour se déclencher en cas d'événements rapides. Il s'agit du circuit secondaire de protection, qui est complémentaire au premier.

IV. Outils de caractérisation

1. Mesures DC, TLP et VF-TLP

Des mesures quasi-statiques (DC) sont présentées dans ce manuscrit, et consistent à augmenter progressivement la tension sur le composant de protection tout en monitorant le courant qui en résulte. Le temps de mesure est suffisamment long pour être considéré comme infini (tous les phénomènes transitoires ont le temps de disparaître). Ces mesures fournissent des informations sur le courant de fuite I_{Leak} du composant pour chaque tension d'alimentation nominale possible V_{DD}.

Les mesures DC fournissent également des informations intéressantes à propos du fonctionnement du composant ESD. Cependant, afin de réaliser une caractérisation électrique plus fidèle au modèle HBM, une meilleure technique de mesure est le TLP (de l'anglais « Transmission Line Pulse »). Une impulsion carrée (par exemple d'une durée de 100 ns) est envoyée au DUT (de l'anglais « Device Under Test », il s'agit du composant testé) par une ligne coaxiale de transmission précédemment chargée. L'impulsion carrée de courant et de tension aux bornes du composant résultantes sont mesurées. Une moyenne est faite sur la valeur de courant et de tension. Typiquement, 30% de l'impulsion est sélectionnée pour faire cette moyenne, et cette fenêtre d'acquisition est située après les phénomènes transitoires, tels que l'« overshoot » (une surtension pouvant survenir au début de l'impulsion). Ces moyennes représentent un point de la courbe I-V du composant (voir la Figure 20). Après cette première mesure, le courant de fuite du composant est mesuré pour une tension donnée (par exemple pour $V = V_{DD}$) grâce à une mesure DC non-destructive. Si le composant n'est pas considéré comme endommagé (ce qui se voit au courant de fuite), alors l'intensité de l'impulsion carrée envoyée sur le DUT est augmentée, afin d'obtenir le point suivant de la courbe I-V. La courbe I-V entière est ainsi tracée. La mesure TLP est stoppée quand le composant est endommagé. Le dernier point I-V qui est mesuré avant la casse du composant est considéré comme étant I_{T2}-V_{T2} (le courant et la tension de casse du composant).

La mesure VF-TLP (de l'anglais « Very Fast TLP ») est similaire à la méthode TLP, sauf que les impulsions sont bien plus courtes et comparables à la durée d'une décharge CDM. Cette méthode permet d'avoir des informations supplémentaires sur le comportement du composant ESD et à propos de sa capacité à protéger le circuit contre des décharges très rapides telles que les décharges CDM. Cependant aucune corrélation ne peut être faite entre une décharge CDM et la mesure VF-TLP [59], étant donné que pour le VF-TLP la mesure se fait entre les deux bornes du composant, tandis que lors d'une décharge de type CDM, le circuit lui-même se décharge jusque dans une entrée ou une sortie.

Toutes nos mesures sont faites à température ambiante et sur plusieurs échantillons. Les impulsions des mesures TLP durent 100 ns avec un temps de montée de 10 ns. Chaque point des courbes TLP provient d'une moyenne effectuée sur les tensions et courants pour l'intervalle de temps entre 70 ns et 90 ns. La mesure DC pour vérifier que le composant n'est pas détruit est effectuée à 1 V (tension nominale). Les mesures TLP sont stoppées quand le courant de fuite à 1 V a augmenté de plus de 500 % par rapport au courant de fuite avant

que le composant n'ait subit d'impulsions. La durée de l'impulsion des mesures VF-TLP est de 1 ns avec un temps de montée de 100 ps. La mesure DC (pour vérifier que le composant n'est pas cassé) est également faite à 1 V.

2. Outil TCAD : les simulations ACS et AVS

L'outil TCAD (de l'anglais « Technology Computer Aided Design ») Synopsys Sentaurus™ est idéal pour simuler numériquement un composant électrique. Tout d'abord la structure physique doit être générée. Les différentes régions ainsi que les matériaux sont définis, les dopages sont placés ainsi que les électrodes (où les conditions électriques et thermiques vont être appliquées). Ensuite le composant est maillé : il est discrétisé en une grille de nœuds non-uniforme.

Nos composants sont maillés en 3D ; le but est d'être le plus conforme possible par rapport aux procédés de fabrication réels. Les simulations TCAD ont d'abord été calibrées grâce à un transistor NMOS de la technologie considérée. Sa courbe quasi-statique I_D-V_G (courant de drain, tension de grille) a été ajustée grâce aux dopages, au travail de sortie de la grille, et à d'autres paramètres, afin d'obtenir la même tension de seuil V_{TH} que dans les mesures. L'empilement de grille n'est pas simulé afin de réduire le temps de simulation ; seule une couche de SiO_2 (dioxyde de silicium) est créée et le travail de sortie de la grille est sélectionné. Le but de nos simulations n'est pas d'obtenir les paramètres ESD exacts tels que V_{T1} ou V_H, mais de mieux comprendre les phénomènes physiques impliqués dans le comportement électrique de la structure, et de comparer plusieurs composants sans avoir à attendre les résultats des mesures.

Un ensemble d'équations qui décrit la distribution des porteurs et les mécanismes de conduction sont spécifiés. Le système d'équations couplées à résoudre dans chaque nœud du maillage est le suivant :

- Equation de Poisson (pour le calcul du potentiel électrostatique pour une distribution de charge donnée)
- Equations de continuité (pour décrire la variation temporelle du nombre de porteurs, qui dépend de leur mouvement ainsi que des taux de génération et de recombinaison)
- Equations de transport (le courant total des porteurs correspond à l'addition d'un courant de conduction qui est dû à un champ électrique, et d'un courant de diffusion dû au gradient de porteurs)
- Equations de contact (pour les électrodes)
- Equations de circuit (connectivité)
- Equations classiques de la chaleur (pour les simulations électrothermiques)

Les modèles physiques sont sélectionnés et inclus dans la résolution numérique des équations précédentes. Dans nos simulations, les paramètres des modèles par défaut n'ont pas été modifiés. Les modèles suivants ont été utilisés :

- La statistique de Boltzmann à la place de la statistique de Fermi-Dirac (permettant d'estimer la densité de porteurs libres)

- La structure de bande des semi-conducteurs (avec le rétrécissement de la bande interdite en cas de semiconducteur dégénéré)
- Les modèles de génération-recombinaison (recombinaison Shockley-Read-Hall et Auger, ionisation par impact (modèle Avalanche), et génération non locale des porteurs causée par effet tunnel de bande à bande ont été pris en compte)
- Les modèles de mobilité (pour prendre en compte les perturbations de la mobilité liées aux phonons, au dopage, aux interfaces, et aux forts champs électriques)

Pour la simulation électrique de la courbe I-V des composants dans TCAD, la méthode TLP aurait été trop longue à reproduire, c'est pourquoi une autre technique a été utilisée : l'ACS (de l'anglais « Average Current Slope »). Il s'agit d'appliquer au composant une rampe en courant. Le temps de montée de l'ACS est de 100 ns, de sorte à imiter un test TLP pour une décharge HBM (Figure 21). La tension est initialisée à 0 V et après elle est libre de changer. Une courbe I-V est ainsi obtenue. Elle est dépendante de la pente du courant, qui est ajustée grâce au temps de montée et à l'amplitude du courant. Dans nos simulations la rampe en courant va de 0 à 0,1 A.

Il a été montré [61] que les simulations ACS avec une durée de 100 ns sont une bonne première approximation avant d'effectuer les mesures TLP. En effet, de par la courte durée des ACS, le flux de porteurs et les mécanismes de déclenchement des structures sont conservés. Un avantage de l'ACS est que le courant est augmenté progressivement et tous les points I-V de la courbe sont corrélés (contrairement au TLP). Par conséquent, certains effets sont pris en compte, tels que les surtensions observées dans les impulsions de tension du TLP. Ces surtensions ne font pas partie de la courbe TLP puisque les données I-V proviennent de la moyenne faite sur les données après stabilisation, ce qui conduit à des courbes I-V TLP qui sont parfaitement placées dans la fenêtre ESD et pourtant le composant peut se retrouver incapable de supporter les stresses ESD auxquels il était destiné. Pour un temps de montée de la durée d'une impulsion d'un TLP, l'analyse ACS est particulièrement conçue pour l'étude du déclenchement des composants ; cependant, tout comme pour n'importe quelle simulation, les déductions tirées de l'analyse d'un ACS doivent être évaluées prudemment, et uniquement utilisées pour l'indication de tendances générales. Elles doivent être vérifiées par des mesures.

Les simulations AVS (de l'anglais « Average Voltage Slope ») consistent à faire subir au composant une rampe en tension. Un temps de montée de 1 ms est choisi afin de simuler une mesure DC de balayage en tension. Il a été choisi de commencer à 0 V et terminer la rampe à 5 V. Le courant résultant (qui passe à travers le composant) versus la tension aux bornes du composant donne la courbe I-V quasi-statique. La tension maximale de l'AVS n'est pas importante car même si cela change la pente de la rampe en tension, l'échelle de temps est suffisamment longue pour que le comportement quasi-statique du composant soit atteint de toutes façons. Le but principal de la simulation AVS est de déterminer le courant de fuite que le composant aurait pour un V_{DD} donné et qui peut être choisi parmi toutes les valeurs de la rampe en tension.

V. Les composants de protection contre les ESD

La solution technologique de l'hybride permet d'enlever le BOX afin de placer des composants dans le substrat juste à côté de composants dans le film mince sur un wafer FD-SOI. Les composants ESD sont typiquement placés dans le substrat, car leur performance est significativement réduite dans le film mince (à cause de sa finesse extrême). Les principaux composants ESD qui sont couramment utilisés en technologie 28 nm FD-SOI sont les diodes, MOSSWI (interrupteurs MOS), GGNMOS (de l'anglais « Grounded Gate NMOS »), BIMOS (de l'anglais « Bipolar MOS ») et SCR (de l'anglais « Silicon Controlled Rectifier », et également appelé thyristor). Ces composants peuvent être adaptés pour être conçus dans le film mince avec la contrainte de la pleine désertion de ce même film mince de silicium. D'autres composants peuvent également servir de protection, tant que leur caractéristique électrique respecte bien la fenêtre ESD.

1. Diode de protection

Pour les applications ESD, la diode, qui est un composant unidirectionnel, est polarisée en inverse par rapport au circuit (elle est bloquée), et elle est polarisée en direct par rapport à la décharge ESD (pour pouvoir laisser passer beaucoup de courant). Il est possible de mettre plusieurs diodes en série pour modifier la tension de seuil de cette protection ESD, cependant cela augmente également la surface et les effets parasites. La diode à grille est un type de diode pour laquelle une grille est placée entre ses zones de dopage N^+ et P^+ (Figure 22).

2. Protections à base de NMOS

Le MOSSWI (de l'anglais « MOS switch ») est un interrupteur à transistor MOS. La grille d'un NMOS est mise à la masse quand il n'y a pas d'événement ESD, et une tension est appliquée sur la grille de sorte à laisser passer le courant entre drain et source en cas d'ESD. Il y a besoin d'un circuit appelé circuit de déclenchement (cela peut être un passe-haut RC par exemple, comme dans la Figure 23), afin de détecter l'ESD et de piloter la grille (en fonction de la fenêtre ESD). Une diode peut être ajoutée en parallèle pour assurer la bidirectionnalité du système de protection.

Une autre possibilité est d'utiliser le NMOS en tant que GGNMOS. La grille et la source sont connectées ensemble (Figure 24) de sorte à bloquer l'effet MOS, et c'est le transistor bipolaire NPN parasite de la structure (la source, le canal et le drain correspondent respectivement à l'émetteur, à la base et au collecteur) qui va jouer un rôle dans l'activation de la structure. Lors d'un ESD positif, des porteurs chauds sont générés par ionisation par impact au niveau de la jonction drain/canal. Le courant de trous crées va vers le substrat, ce qui augmente le potentiel entre la base et la source. Du fait de ce potentiel, des électrons

sont injectés de la source vers le drain, et ce faisant, activent tout le transistor bipolaire parasite. Lors d'un ESD négatif, la grille est polarisée (puisqu'elle est connectée à la source), ce qui ouvre le GGNMOS.

Le BIMOS (pour son effet BIpolaire et MOS) est un transistor NMOS auquel on rajoute une zone dopée P^+ et appelée « body contact » en anglais ; dans ce manuscrit on l'appellera contact de canal. Ce contact P^+ est placé de sorte à avoir accès à la base du bipolaire NPN parasite (qui est le canal du MOSFET). Pour les applications ESD, ce contact et la grille sont connectés ensemble à une même résistance externe de polarisation (Figure 25). Le BIMOS s'active à la fois en quasi-statique et en dynamique. Lors d'un ESD, l'ionisation par impact à la jonction drain/canal active le transistor bipolaire parasite. Comme la base est aussi branchée à la grille du transistor (grâce au contact de canal), la partie MOS est activée et une couche d'inversion est créé. Plus la valeur de résistance est haute, plus la tension d'activation est basse, puisque cela implique une valeur de tension plus haute sur la grille. Le BIMOS peut être activé de façon dynamique grâce à ces capacités parasites entre le drain et la grille et entre le drain et le contact de canal. Il peut également subir des ESD négatifs ; dans ce cas, la grille et le contact de canal sont polarisés grâce à la résistance externe, ce qui ouvre le BIMOS.

3. Protections à base de SCR

Le SCR, également appelé thyristor, est une structure dopée P/N/P/N. Cela peut être vu comme deux transistors bipolaires fusionnés, un transistor NPN et un transistor PNP, qui sont bouclés entre eux (Figure 26). Si la base des deux transistors bipolaires est flottante, le SCR peut s'activer mais sa tension de déclenchement est très haute. Afin de déclencher le SCR dans la bonne fenêtre ESD on peut utiliser un circuit de déclenchement qui va soit baisser la tension sur la base dopée N du transistor PNP, soit augmenter la tension sur la base dopée P du transistor NPN.

Le SCR est unidirectionnel mais il est possible de brancher deux SCR tête bêche afin d'obtenir ce que l'on appelle un triac (Figure 27), qui est bidirectionnel. La « beta-matrice » (Figure 30) est une architecture de protection ESD qui se comporte comme un réseau de triacs. Elle peut être considérée à la fois comme un composant de protection et comme une stratégie de protection globale.

Des efforts ont été faits pour concevoir des SCR dans le film mince, ce qui a conduit à des structures comme le Z^2-FET (de l'anglais « Zero subthreshold swing and Zero impact ionization FET ») et le GDNMOS (de l'anglais « Gated Diode NMOS »). Cependant, ces structures n'ont pas le même fonctionnement qu'un SCR. Le GDNMOS est une diode à grille (GD) qui partage sa cathode avec le drain d'un transistor (MOS). En théorie ce dispositif peut être considéré comme un SCR puisqu'il est constitué de dopages P/N/P/N (Figure 29). Le GDNMOS sera expliqué plus extensivement dans le chapitre 2.

Chapitre 2 : Protections ESD dans le film mince

Le but principal de ce chapitre est de mieux appréhender les protections ESD dans le film mince et en particulier de proposer des améliorations dans la mesure du possible.

I. Boost capacitif pour NMOS et BIMOS

Quand le BIMOS se situe dans le film mince, ses capacités parasites jouent un rôle important dans son mode de déclenchement dynamique. Notamment, avec la résistance externe de polarisation branchée à la grille et au contact de canal, les capacités parasites C_{DB} (entre le drain et le contact de canal ; il s'agit de la capacité due à la jonction N^+/Pint) et C_{DG} (entre le drain et la grille ; il s'agit de la capacité des séparateurs en parallèle avec la capacité de l'oxyde de grille au niveau du chevauchement du drain sous la grille), forment un circuit RC passe-haut (Figure 32 et Figure 33). Ce dernier permet l'augmentation de la tension sur la grille lorsque la tension sur le drain augmente rapidement. Ainsi, augmenter la valeur de la résistance externe de polarisation (Figure 41) ou des capacités parasites permet de diminuer la tension de déclenchement du BIMOS. Augmenter la valeur de résistance est dangereux car cela mène à une tension de déclenchement éventuellement trop basse en AVS (d'où un risque de « latch up »), étant donné que tous les courants de fuite dans le canal sont transformés en une tension non négligeable à travers cette résistance. D'où l'identification du besoin d'augmenter les valeurs des capacités parasites.

En ce sens, nous proposons d'utiliser un « boost » capacitif en réalisant une extension du silicium actif au niveau du drain et en rajoutant une grille avant supplémentaire. Quatre composants sont comparés (Figure 34) : un BIMOS avec un doigt (« BIMOS_1finger »), avec deux doigts (« BIMOS_2fingers »), et deux BIMOS avec un doigt plus un drain étendu (pour avoir le boost capacitif) et qui ont la même surface silicium que le BIMOS_2fingers. L'un d'entre eux a un drain étendu flottant et ses grilles connectées (« BIMOS_capaboost_float »), ce qui lui rajoute une capacité parasite par rapport au BIMOS_1finger, et l'autre (« BIMOS_capaboost ») a des grilles et son drain étendu connectés afin d'avoir deux capacités parasites en plus par rapport au BIMOS_1finger.

Des simulations électrothermiques ACS ont été effectuées sur ces quatre composants. Elles commencent avec une température initiale de 300 K et sont stoppées lorsque le point chaud du composant atteint 800 K. Ce point chaud se situe dans le canal, près du drain.

Pour un courant de drain donné, la température maximale dans le BIMOS_1finger est plus haute que dans le BIMOS_2fingers (Figure 37, Figure 38 et Figure 39), puisque le deuxième doigt aide à évacuer du courant, ce qui relaxe chaque doigt. Le BIMOS_capaboost n'a qu'un seul doigt de conduction, comme le BIMOS_1finger. Cependant la température maximale à l'intérieur de son doigt conducteur est plus faible que dans le BIMOS_1finger puisque le drain étendu permet d'étaler de façon plus uniforme la température dans le composant. De plus, grâce aux capacités parasites du BIMOS_capaboost, ce composant se

déclenche plus tôt ; en conséquence, à 10 mA, le BIMOS_capaboost a une tension de drain similaire à celle du BIMOS_2fingers, mais plus basse que celle du BIMOS_1finger. Pour un courant donné, la température est plus basse dans un dispositif qui a une tension plus basse ; ceci explique pourquoi la température maximale est similaire dans le BIMOS_2fingers et dans le BIMOS_capaboost à 10 mA. En termes de comparaison, la température maximale est plus élevée dans le BIMOS_capaboost que dans le BIMOS_2fingers à 1 mA, à cause du décalage en tension de drain. Une température maximale la plus faible possible dans le dispositif est bénéfique du point de vue de sa robustesse intrinsèque.

La Figure 40 permet de comparer les 4 composants avec un ACS identique. Deux retournements peuvent être observés. Le premier se produit pour une tension V_{T1} qui est la vraie tension de déclenchement, lorsqu'il commence à s'établir un courant significatif dans le composant. Le deuxième se produit à V_{T1bis} ; c'est ce retournement qui est observé lors de mesures TLP. En effet l'instrument de mesure TLP ne peut pas mesurer des courants inférieurs à 5 mA. Le boost capacitif est un phénomène qui peut être observé sur l'ACS près de V_{T1} et non près de V_{T1bis}, étant donné que les changements de capacité ont une importance pour des temps faibles ($\tau = RC$). Comme dans l'ACS la rampe de courant est réalisée en 100 ns, seuls les plus petits courants sont influencés par le changement de capacité, étant donné qu'ils sont mesurés lorsque le temps de simulation est encore court.

Si une courbe TLP était réalisée, elle ne montrerait pas de différence de tension de déclenchement entre les composants (car la courbe TLP ne peut montrer que V_{T1bis}). Cependant le composant aillant le plus petit V_{T1} supporterait sans doute mieux un ESD que les autres protections aillant une tension de déclenchement apparemment similaire sur les courbes TLP mais qui ont en réalité un V_{T1} plus haut. Ceci, pour deux raisons :

Premièrement, la surtension (« overshoot ») de la protection qui a le V_{T1} le plus grand serait plus élevée que pour les autres composants. En effet la surtension est un phénomène qui se produit lors de l'établissement de la tension, pendant le temps de montée de l'impulsion TLP, c'est-à-dire pour un temps très court. Il correspond donc à ce qui peut être observé pour des faibles courants dans l'ACS (lorsque le temps de simulation est court et que le composant n'a pas encore atteint son équilibre). La surtension ne se voit pas dans les courbes TLP, et pourtant une surtension trop intense peut endommager la protection. La seule façon de voir une surtension est de regarder les chronogrammes de tension en fonction du temps du TLP. Une autre façon d'observer des effets de temps courts sur le composant est de réaliser des caractérisations VF-TLP.

La deuxième raison est que la protection qui a le V_{T1} le plus faible peut être amenée plus facilement à se déclencher lorsqu'un événement ESD qui a une faible énergie se produit. Par faible énergie on considère ce qui pourrait arriver si le réseau de protection ESD endure un faible courant anormal. Par exemple, pendant l'ESD, la protection se déclenche et la majorité du courant est évacué. Lorsque l'ESD est terminé, la protection se referme. Un courant résiduel I_ε, qui n'a pas eu le temps de s'évacuer avant la fermeture de la protection, peut s'avérer problématique. En effet, même s'il est faible, il est transformé en une forte tension aux bornes de la protection, étant donné que la protection agit comme un élément hautement résistif. Cette forte tension peut détruire les composants à protéger. Ce phénomène de haute tension qui se produit lors de la fermeture de la protection peut être observé sur les chronogrammes de tension en fonction du temps de la courbe TLP, en regardant ce qui se passe après l'impulsion de tension, pour des temps supérieurs à 100 ns. D'autres phénomènes d'ESD à faible énergie peuvent également être cités (voir dans le

chapitre 2 en anglais). Une tension de déclenchement V_{T1} plus faible signifie que même si la protection n'est pas encore capable d'évacuer la totalité du courant, un faible courant commence à passer dans la protection et ce pour une tension plus faible que la protection qui a un V_{T1} plus grand. Du coup si un I_ε arrive sur la protection, elle va le laisser passer vu qu'elle sera déjà un peu ouverte, ce qui évite le problème de création d'une forte tension.

Grâce à ses capacités parasites supplémentaires, le BIMOS_capaboost se déclenche plus tôt que le BIMOS_1finger et que le BIMOS_2fingers : son V_{T1} est plus petit (Figure 40), comme attendu. Le BIMOS_2fingers et le BIMOS_capaboost_float ont un V_{T1} similaire puisqu'ils ont tous les deux le même nombre de capacités parasites, et ils se déclenchent avant le BIMOS_1finger. Le BIMOS_1finger, BIMOS_capaboost et BIMOS_capaboost_float ont un R_{ON} similaire (les courbes sont parallèles entre V_{T1} et V_{T1bis}), car ils n'ont qu'un doigt de conduction. La résistivité en conduction R_{ON} du BIMOS_2fingers est plus petite que les autres composants puisqu'il a un deuxième doigt de conduction (c'est pourquoi la pente entre V_{T1} et V_{T1bis} du BIMOS_2fingers est un peu plus abrupte que les autres).

L'approche originale du « boost » capacitif peut aussi être utilisée dans un MOS. Le doigt en plus rajoute des capacités parasites, et cela se voit à la décroissance de la valeur de V_{T1} sur la courbe ACS, tout comme pour le BIMOS. En revanche sur l'AVS il n'y a pas de différence entre le NMOS classique et celui qui bénéficie du boost capacitif (Figure 42), puisque l'AVS est long (1 ms), ce qui fait disparaître les effets capacitifs. La différence entre le BIMOS_1finger, BIMOS_2fingers et BIMOS_capaboost sur l'AVS est due à la connexion entre la grille et le contact de canal. Le courant de fuite (pour des faibles tensions) est plus élevé dans le BIMOS_2fingers et le BIMOS_capaboost que dans le BIMOS_1finger, parce que ces composants ont deux doigts qui permettent un courant de fuite entre l'anode et le contact de canal à travers la jonction, comme si leur largeur était deux fois plus grande. Ainsi, la tension de déclenchement du BIMOS_2fingers est la plus faible, car tous les courants de fuite qui passent dans le contact de canal sont transformés en une tension sur les grilles à travers la résistance externe de polarisation. Le BIMOS_capaboost a une tension de déclenchement plus élevée, puisque le courant passe moins facilement dans le doigt additionnel qui est branché au potentiel haut (connectivité de drain) que dans l'autre doigt. Le BIMOS_1finger a la tension de déclenchement la plus haute puisqu'il a moins de courant de fuite.

Pour conclure sur cette partie, la technique du boost capacitif est une bonne solution pour réduire la tension de déclenchement des composants de protection ESD, et donc pour augmenter leur efficacité. Elle peut être utilisée sur le BIMOS et sur d'autres composants (par exemple le MOS).

II. Le GDxMOS, protection ESD pour haute et basse tension

Le GDNMOS (de l'anglais « Gated Diode merged NMOS », ce qui signifie diode à grille fusionnée avec un NMOS) est un composant de protection contre les ESD qui comprend notamment une diode à grille fusionnée avec un transistor NMOS (Figure 29). L'anode et la cathode du composant correspondent respectivement à l'anode (dopée P^+) de la diode et à

la source (N^+) du MOS. La partie commune est la cathode de la diode, qui est également le drain du MOS. Le GDNMOS est conçu dans le film mince, et l'intérêt de ce composant est qu'en principe, un SCR (jonctions PNPN) peut être formé. Dans cette section l'étude du GDNMOS est approfondie par rapport à l'état de l'art, et un nouveau composant, le GDBIMOS (de l'anglais « Gated Diode merged BIMOS », ce qui signifie diode à grille fusionnée avec un BIMOS), est présenté et comparé au GDNMOS. Le GDBIMOS est très compact (en termes de surface de silicium) puisque le NMOS du GDNMOS est aisément transformable en un BIMOS (Figure 45). Nous avons nommé cette famille de composants GDxMOS, le « X » pouvant être remplacé par un « N » ou un « BI » pour obtenir un GDNMOS ou un GDBIMOS.

1. Le GDxMOS en tant que protection haute tension

Dans cette étude, le dopage du drain reste conventionnel. Les GDNMOS testés sont listés selon leur connectivité (Table 1).

Tous les composants avec la grille de la diode mise à la masse ont un problème de robustesse avec une casse prématurée (Figure 48). La raison est l'absence de résistance pour protéger la grille de la diode, qui est située juste à côté de l'anode. La résistance permet à la grille de la diode d'être polarisée à travers la capacité parasite C_{AGD} (capacité entre l'anode et la grille de la diode). Ainsi la différence entre V_A (tension d'anode) et V_{GD} (tension de la grille de la diode) est réduite pendant l'événement ESD. Sans la résistance, la différence de potentiel entre l'anode et la grille de la diode mise à la masse peut atteindre la tension de casse de l'oxyde qui les sépare. La valeur de la résistance doit être suffisamment grande pour éviter la casse prématurée. En effet, la tension de la grille peut monter plus haut avec une valeur de résistance plus grande, et ainsi la casse est retardée et un plus grand courant peut passer dans le composant, d'où une hausse de I_{T2} (courant de casse). Pour une même valeur de résistance, l'I_{T2} peut être encore augmenté si en plus de brancher la résistance à la grille de la diode, on la branche à la grille du MOS. En effet, en faisant cela on rajoute l'effet de la capacité parasite C_{DGM} (entre le drain et la grille du MOS). Les deux capacités parasites vont ainsi aider à la croissance de la tension sur les grilles, et ainsi contribuer à protéger la zone critique de la grille de la diode. La valeur de I_{T2} atteint un maximum de 0,12 A lorsque seule la grille de la diode est connectée à une résistance, même si la valeur de résistance augmente, tandis que si les deux grilles sont branchées à la résistance, la valeur de I_{T2} continue d'augmenter avec la valeur de la résistance (pour les mêmes valeurs de résistances utilisées). La raison est qu'avec seulement la grille de la diode branchée à la résistance, la caractéristique électrique I_A en fonction de V_A est la même quelle que soit la valeur de la résistance (Figure 51), tandis qu'avec les deux grilles branchées à la résistance, la tension de déclenchement V_{T1} baisse avec une résistance plus grande (Figure 53). En fait, la tension de casse V_{T2} des composants vaut environ 5 V (Figure 49), et pour un V_{T2} et une résistance à l'état passant R_{ON} fixés, I_{T2} augmente avec V_{T1} qui diminue (Figure 50). En réalité la résistance de conduction R_{ON} change d'un composant à l'autre à cause de la différence de valeur de résistance, mais la différence de R_{ON} engendrée n'est pas suffisante pour compenser le décalage de V_{T1}. Les composants qui ont leur grille de diode branchée à la masse n'atteignent pas le V_{T2} maximum possible à cause de la casse prématurée.

Lorsque la grille de la diode est branchée à une résistance et que la grille du NMOS est mise à la masse, V_{T1} ne change pas avec le changement de valeur de la résistance (Figure 51). En fait, quand la tension d'anode augmente, la capacité parasite C_{AGD} induit une augmentation de V_{GD} à travers la résistance qui est branchée à la grille de la diode. Il s'ensuit que plus la résistance est forte, plus V_{GD} est élevé. Un V_{GD} plus grand réduit la quantité de courant (pour une tension d'anode donnée) qui peut circuler à travers la diode avant son déclenchement. La tension de la grille module les barrières d'énergie le long du composant, ce qui empêche les trous provenant de l'anode d'être diffusés dans la cathode de la diode (Figure 52). Ainsi, avec une résistance plus grande, la diode est plus résistive avant le déclenchement. Lorsque les potentiels sont tels que les bandes d'énergie de l'anode coïncident avec celles du canal, il n'y a plus de barrière qui empêche l'injection de trous de l'anode vers le canal. A ce moment, l'anode commence à contrôler le niveau des bandes d'énergie dans le canal, au détriment de la grille. Lorsque la diode se déclenche, la tension de l'anode atteint la valeur pour laquelle toutes les bandes d'énergie (dans l'anode, le canal et la cathode) sont presque au même niveau, et les trous circulent dans toute la diode. Le R_{ON} de la diode ne dépend pas de la valeur de la résistance branchée à la grille car le nombre de porteurs circulant dans la diode est si important que la tension de grille ne peut plus agir sur le canal. Les trous qui circulent à travers la diode sont recombinés dans le drain du NMOS. La tension d'anode se retrouve presque intégralement sur le drain du NMOS puisque la diode est très conductrice. Puisque la grille du NMOS n'est pas branchée à la résistance qui est connectée à la grille de la diode, la valeur de cette résistance n'a pas d'impact sur la conduction des électrons dans la partie NMOS du composant. Tout ceci explique pourquoi V_{T1} du GDNMOS ne change pas avec la valeur de résistance si seule la grille de la diode est branchée à celle-ci. V_{T1} vaut alors environ 4,8 à 5,0 V. Le fait que V_{T1} ne soit pas modifiable implique qu'il n'est pas possible d'utiliser la valeur de résistance pour faire coïncider les caractéristiques du GDNMOS à différentes fenêtres de conception si besoin. Cependant, l'idée que la résistance sur la grille de la diode aide à rendre le GDNMOS plus résistif avant son déclenchement est intéressante et peut être utilisée pour réduire le courant de fuite. Si la grille de la diode et celle du NMOS sont branchées ensemble et connectées à une résistance, V_{T1} d'un tel GDNMOS est plus faible que pour un GDNMOS dont la grille du NMOS est mise à la masse ; et plus la valeur de la résistance branchée aux deux grilles augmente, plus V_{T1} du GDNMOS baisse (Figure 53). C'est parce que les capacités parasites des deux grilles aident à augmenter la tension sur les grilles grâce à la résistance. Une tension plus haute sur les grilles permet à la partie NMOS du composant de commencer à conduire du courant pour une tension plus faible d'anode. Le décalage de V_{T1} rend possible l'utilisation de la protection dans une autre fenêtre ESD, en ajustant simplement la valeur de la résistance externe de polarisation.

Augmenter la tension sur la grille arrière conduit à baisser la tension de déclenchement V_{T1} (Figure 62). Ainsi la grille arrière est un bon moyen de moduler V_{T1} afin d'être compatible avec la fenêtre ESD désirée. Une augmentation du courant de fuite pour une tension positive de grille arrière plus grande est observée (Figure 63). Cette augmentation de courant de fuite reflète l'activation graduelle du canal du NMOS face arrière près de l'interface film-BOX. Une solution pour améliorer le composant est de connecter la grille arrière aux grilles avant (qui sont connectées à la résistance), afin qu'elle contribue à un courant plus grand circulant dans le composant en mode ouvert, et à un V_{T1} plus petit, tout en gardant un faible courant de fuite avant le déclenchement. Il s'agit d'une solution élégante où le composant est auto-polarisé, sans avoir besoin de tensions externes

pour le contrôler. Ainsi le composant peut protéger le circuit intégré même s'il n'y a pas de tension d'alimentation. Dans notre contexte, étant donné que le BOX est bien plus épais que l'oxyde des grilles avant, le bénéfice de connecter la grille arrière aux grilles avant n'est pas marquant. En effet, la tension maximale appliquée aux grilles n'excède jamais les 1 V.

Les GDNMOS ont le même V_{T1} que les GDBIMOS (pour une même valeur de résistance branchée aux deux grilles) (Figure 57). A cause du multi-déclenchement (qui est normal lorsqu'il y a plusieurs doigts) il est difficile de comparer expérimentalement le R_{ON} de ces deux composants. En théorie, R_{ON} est amélioré dans le GDBIMOS par rapport au GDNMOS, surtout si la valeur de la résistance branchée sur les grilles est grande. C'est parce que la tension des grilles ne redescend pas à 0 V avec le temps comme dans le GDNMOS (Figure 55 et Figure 58), donc la tension de la grille aide à la conduction de courant dans le canal de la partie NMOS du GDBIMOS. C'est la partie BIMOS du composant qui maintient la tension positive de grille permanente, puisqu'elle bénéficie d'une boucle de rétroaction positive : lorsqu'il y a un courant d'électrons qui circule à travers le canal du BIMOS, l'ionisation par impact près du train produit des trous qui sont attirés par le contact de canal dopé P^+. Le courant de trous dans le contact de canal aide la tension de grille et de contact de canal à être augmentée, de par la résistance. Avec l'augmentation de la tension de grille, le courant dans le canal du BIMOS est augmenté grâce à l'effet MOS. Ce courant d'électron plus important dans le canal produit plus de trous par ionisation par impact, etc. Ce sont les capacités parasites qui provoquent l'augmentation de la tension de grille lorsque la tension d'anode augmente, rendant ainsi le composant passant, et c'est le contact de canal du BIMOS qui maintient cette tension de grille lorsque le composant est passant, augmentant ainsi la résistance à l'état passant R_{ON}.

Toutes les structures mentionnées précédemment avaient le drain flottant. Dans ce paragraphe, différents composants avec le drain connecté sont explorés. Les composants avec un drain flottant avaient tous un faible courant de fuite autours de 2 nA à 1 V. L'idée de décroitre le courant de fuite de façon plus poussée est intéressante, car les GDNMOS et GDBIMOS avec un dopage conventionnel dans le drain peuvent potentiellement être utilisés pour des applications à faible consommation, à condition que leur V_{T1} soit contrôlé par un circuit de déclenchement (situé par exemple sur la grille arrière). Le premier composant étudié (numéro 19 dans la Figure 64) est un GDNMOS dont le drain est connecté à la grille de la diode. Le but de cette connexion est d'augmenter la barrière d'énergie pour les trous entre l'anode et le canal en augmentant la tension sur la grille de la diode. En effet, à partir de 0,6 V entre l'anode et le drain, la diode commence à conduire du courant. Lorsque la diode est conductrice, la tension s'élève à la fois sur le drain et sur la grille de la diode, donc, la diode devient moins conductrice. Un équilibre est atteint dans cette boucle avec la grille de la diode qui est ainsi auto-polarisée. Comme prévu, le courant de fuite du composant 19 est réduit (70 pA) mais son V_{T1} est plus adapté pour une protection haute tension (Figure 65 et Figure 66). Le composant 20, pour lequel le drain est connecté à l'anode, a été conçu pour diminuer le courant de fuite d'un GDNMOS avec un drain faiblement dopé et non pour un GDNMOS avec un drain qui a un dopage conventionnel. Cependant la structure avec drain conventionnel reste intéressante à étudier puisqu'elle donne de nouvelles options en termes de fenêtre ESD. Avec le drain conventionnel, il ne peut pas y avoir de comportement de type SCR et le composant agit comme une diode en série avec un MOS. Le composant 20 se déclenche avant un NMOS classique, mais possède une tension de maintien et un courant de fuite plus élevés. La tension de déclenchement plus basse est liée au « boost » capacitif de la

structure, et la tension de maintien plus haute est due à la grille de la diode qui induit une barrière qui lutte contre la circulation des porteurs dans le composant. La tension de déclenchement du composant 20 est plus faible également que pour un GDNMOS dont le drain est flottant grâce au fait que les deux capacités parasites C_{AGD} et C_{DGM} sont actives et presque tout le courant peut circuler par le chemin le plus facile (*c.a.d.* la partie NMOS du composant) ; la tension de maintien est faible et le courant de fuite est le même. Le même comportement que celui des composants 19 et 20 est attendu pour des GDBIMOS qui auraient la même configuration de drain. Le composant 21 est un GDBIMOS pour lequel l'un des deux contacts de canal est connecté au drain. Sa tension de déclenchement est réduite et son courant de fuite est augmenté pour V_{DD} = 1 V par rapport à un GDBIMOS avec drain flottant. Dès que la diode est conductrice, le courant va dans le premier contact de canal à travers la connexion du drain. Le transistor PMOS parasite (constitué par les deux contacts de canal du BIMOS et son canal) s'enclenche très rapidement, ce qui mène à la polarisation de la grille du BIMOS. Ceci enclenche le BIMOS dans sa totalité, ainsi tout le GDBIMOS devient conducteur. Avec un drain flottant, le GDBIMOS s'enclenche dynamiquement principalement grâce à ses capacités parasites, tandis qu'en mode statique, il est activé grâce à l'ionisation par impact et à l'effet tunnel de bande à bande. Le PMOS parasite étant bien plus rapide que le circuit RC formé par les capacités parasites et que les effets de l'ionisation par impact et du tunnel de bande à bande, le composant 21 s'active bien plus tôt, que ce soit en mesure TLP ou DC. Son courant de fuite à V_{DD} = 1 V est relativement haut (30 nA), mais si la protection est conçue pour des applications basse consommation (avec $V_{DD} \leq 0{,}8$ V), le courant de fuite devient acceptable ($I_{leak} < 2$ nA). En fait, les courbes DC du composant 21 et celle du GDBIMOS avec drain flottant commencent à diverger après 0,6 V, ce qui correspond à la tension pour laquelle la diode commence à conduire du courant, et ensuite le BIMOS est activé par un moyen qui dépend du composant et de sa connectique.

2. Le GDxMOS en tant que protection très basse tension

En théorie, le GDNMOS peut être considéré comme un SCR puisqu'il s'agit d'une structure P/N/P/N. Cependant le dopage du drain étant trop haut (N^+), le comportement du GDNMOS ne peut pas être basé sur les transistors bipolaires parasites (comme pour un SCR), puisque l'efficacité de l'émetteur du transistor bipolaire parasite PNP composé de la diode et du canal du NMOS est dégradé par le fort dopage. Si on veut obtenir un comportement de type SCR, il faut que l'efficacité de l'émetteur de chaque transistor bipolaire soit améliorée, soit en diminuant la longueur de la base ou en baissant le dopage de l'accepteur dans la base. La longueur de la base du PNP parasite – il s'agit du drain - est diminuée au maximum selon les règles de dessin, ce qui réduit la tension de maintien V_H et la tension de déclenchement V_{T1}, comme prévu (Figure 69). Afin d'obtenir un SCR, nous proposons de doper le drain avec un dopage plus faible (N-LDD au lieu de N^+). Cette modification n'augmente pas le coût du composant puisque le nombre d'étapes requises pour le fabriquer est le même. Le dopage dans la base du PNP étant réduit, V_H et V_{T1} sont fortement réduits par rapport à ceux du composant dopé N^+, comme prévu également (Figure 68). Le fait de brancher une résistance sur les grilles de la structure (au lieu de les mettre à la masse) permet de diminuer encore plus V_{T1} (Figure 70, Figure 71, Figure 72, et Figure 73). Si la résistance est plus grande, cela aide le composant à se déclencher plus vite grâce aux

capacités parasites, puis les transistors bipolaires vont prendre le relai pour enclencher tout le SCR.

Des simulations TCAD montrent que le GDBIMOS conçu avec un drain faiblement dopé (N-LDD) est plus efficace que le GDNMOS pour une fenêtre ESD réduite (Figure 74). D'autres simulations montrent que plus la valeur de résistance branchée sur les grilles du GDBIMOS est grande, plus V_{T1} et I_{T1} sont faibles (comme pour le GDNMOS) mais également plus R_{ON} est grand (Figure 75). Ces résultats doivent être confirmés par des mesures.

Le GDNMOS et le GDBIMOS avec drain faiblement dopé ont un fort courant de fuite à 1 V. Leur courant de fuite est raisonnable à 0,6 V (8.10^{-11} A/µm), ainsi ce sont des bonnes protections pour les applications très basse tension. Une solution pour que ces protections soient utilisables à 1 V, est de les brancher en série ou de changer leur connectivité de drain par exemple.

3. Gestion du siliciure dans le GDxMOS

Dans les deux sections précédentes, le GDxMOS était présenté comme une protection haute tension (4 V) avec un drain dopé N^+, et comme une protection très basse tension (0,6 V) avec un drain dopé N-LDD. Le but principal de cette section est d'obtenir une protection basse tension (1 V) grâce à du dopage N-LDD, la connectivité du drain, et la gestion du siliciure.

Pour avoir du dopage N-LDD dans le drain il faut un masque qui empêche le dopage N^+ d'être rajouté, mais également un autre masque, celui qui empêche la siliciuration du drain. Normalement les protections ESD sont dessinées dans le substrat et dans ce cas, rajouter ou enlever le siliciure ne change pas grand-chose, puisque le courant est volumique et que le siliciure est seulement présent à la surface. Dans le film mince en revanche, le siliciure prend presque toute l'épaisseur du film mince, et agit comme si le silicium était fortement dopé ; c'est pourquoi il faut l'enlever si on veut obtenir un SCR basse tension. En fait la présence du siliciure provoque la recombinaison des porteurs dans la zone du drain, ce qui fait que seul un très faible courant de trous passe entre l'anode et le canal du NMOS (Figure 85). Ainsi le transistor PNP ne peut pas fonctionner ce qui tue le comportement SCR de la structure. Dans ce cadre on peut imaginer quatre protections ESD différentes : le GDxMOS haute tension (avec siliciure et dopage N^+), le GDxMOS basse tension (avec au choix siliciure ou dopage N^+), et le GDxMOS très basse tension (sans siliciure et avec dopage N-LDD) (Figure 86).

Un moyen d'obtenir une protection basse tension (1 V) avec un GDxMOS qui a un drain dopé N-LDD, est de changer la connectivité du drain, en le branchant sur l'anode ou sur la grille de la diode par exemple. La connexion entre le drain et la grille de la diode améliore le courant de fuite, qui est lié au courant de trous provenant de la diode (et qui ne sont pas recombinés dans le drain). En effet la tension sur la grille de la diode va augmenter lorsque les trous vont passer dans le drain (auto-polarisation de la grille), ainsi la barrière pour les trous entre l'anode et le canal de la diode sera agrandie et moins de trous vont passer, et ceci tant que la tension d'anode n'est pas suffisamment haute. La connexion entre le drain et l'anode améliore également le courant de fuite, puisque le NMOS et le système

« diode + NMOS = SCR » se retrouvent en parallèle, ce qui fournit une alternative de passage pour le courant, qui n'est plus obligé de s'écouler intégralement par la diode. En réduisant le courant qui circule dans la diode, on réduit le nombre de trous qui sont envoyés vers la masse et qui créent le courant de fuite. Cependant ces connexions de drain ne peuvent pas être réalisées sur une zone sans siliciure (mauvais rendement, risque que le silicium soit transpercé par le contact en tungstène, contact Schottky au lieu d'un contact ohmique). Une solution est de ne pas enlever le siliciure partout sur le doigt, le but étant de pouvoir placer les connexions métalliques au-dessus du siliciure, et d'avoir une grande largeur de doigt sans siliciure afin de garantir un faible dopage de celui-ci.

Une siliciuration partielle du doigt peut aussi être utilisée pour décaler V_{T1} et le courant de fuite de la structure GDxMOS avec drain flottant (en aillant une proportion du doigt plus ou moins siliciurée), afin d'obtenir un jeu de protections pouvant parer à différentes fenêtres de conception. Plus la proportion de siliciure est grande, plus V_{T1} augmente et I_{Leak} diminue, et R_{ON} est également impacté puisque la partie qui a le siliciure va moins participer à la conduction de courant que la partie siliciurée (Figure 90). Cet effet n'est pas linéaire : jusqu'à 50% du doigt peut être siliciuré sans impacter la performance ACS du GDxMOS (V_{T1} n'augmente pas trop) (Figure 91). Cela montre d'un côté que l'idée de mettre du siliciure sur un bout du doigt pour rajouter un contact en tungstène est valide et ne va pas empêcher la protection d'avoir un faible V_{T1}. D'un autre côté, les concepteurs doivent faire attention de garder un pourcentage du siliciure élevé s'ils veulent avoir une action sur les caractéristiques ACS et AVS de la protection.

Comme prévu, les caractéristiques ACS et AVS du GDxMOS N-LDD avec drain connecté à la grille de la diode sont décalées par rapport à celles du GDxMOS N-LDD avec drain flottant, et la siliciuration partielle du doigt peut être utilisée pour décaler les paramètres électriques même si la connectivité du drain est différente (Figure 94). Les courbes présentent deux déclenchements : l'un lorsque le courant commence à circuler dans la partie du doigt qui n'a pas d'électrode (drain flottant et sans siliciure), et l'autre lorsque le courant circule également dans la partie qui a une électrode (connexion à la grille de la diode et siliciure) (Figure 96). La densité de courant circulant dans la partie du drain qui a l'électrode est toujours plus faible que dans l'autre partie, même après le deuxième déclenchement. La zone intermédiaire, où l'influence de l'électrode s'exerce dans la partie sans électrode, peut s'étendre jusqu'à 1 µm grâce à la résistivité du silicium sans siliciure (Figure 97).

Le courant de fuite est effectivement réduit dans le GDxMOS N-LDD avec drain connecté à l'anode par rapport à la structure avec drain flottant (Figure 102). La caractéristique ACS est aussi améliorée pour les forts courants (Figure 101) (sauf si le siliciure est déposé sur tout le doigt), ceci grâce au second déclenchement. Le premier déclenchement est dû à la partie SCR du dispositif (celle sans siliciure), et le second déclenchement à la partie NMOS (celle avec siliciure et une connexion entre le drain et l'anode) (Figure 103). Plus la partie siliciurée est large, plus le second déclenchement se produit pour un courant faible. En effet, la partie SCR devient moins large et ne peut plus conduire un aussi grand courant que si sa largeur était plus importante. Lorsque le SCR est saturé de courant, il devient plus facile pour le courant de passer dans le MOS. Le R_{ON} entre les deux déclenchements est réduit si la partie SCR est moins large, puisque cela réduit sa largeur de conduction. Des mesures doivent être faites pour vérifier que le second déclenchement se produise avant le courant de casse du dispositif. Il faut également vérifier

par mesures qu'une taille de SCR trop petite ne cause pas de problème de robustesse. Lorsque le siliciure est présent sur tout le drain (Figure 105), il n'y a pas de conduction de type « SCR », étant donné qu'il est court-circuité (connexion entre l'anode et le drain), donc il n'y a que la conduction du NMOS ; la structure peut être vue également comme un SCR pour lequel la base N du transistor bipolaire PNP est branché à un potentiel haut, ce qui bloque le SCR. Ainsi lorsque le drain est connecté à l'anode avec du siliciure sur tout le doigt, le composant a le même V_{T1} et R_{ON} qu'un MOS, et le même courant avant déclenchement que le GDxMOS avec drain flottant (puisque cela correspond aux trous qui viennent de l'anode).

On peut imaginer plusieurs parties siliciurées et non siliciurées sur un même doigt. Un plus grand nombre d'électrodes mène à de plus grands V_{T1} et R_{ON} et à un I_{Leak} plus faible (Figure 107). En effet chaque électrode influence le potentiel dans le drain qui n'a pas d'électrode et qui est suffisamment proche de l'électrode. Plus d'électrodes signifie plus de frontières entre siliciuration et non-siliciuration, ainsi plus de pourcentage du doigt qui est influencé par les électrodes.

Ainsi, il est possible de modifier le nombre d'électrodes, leur taille, et leur connectivité, afin de décaler les courbes ACS et AVS pour rentrer dans une fenêtre ESD donnée. Cette technique n'a pas de coût de fabrication supplémentaire, puisqu'il suffit d'utiliser des masques et des composants qui existent déjà, et que cela demande peu ou aucune place de silicium supplémentaire pour réaliser le composant.

Pour conclure le chapitre, les GDNMOS et GDBIMOS sont plus robustes si leur grille de diode est connectée à une résistance suffisamment grande. La robustesse est améliorée si la grille du NMOS est également connectée à cette résistance. Pour correspondre à la bonne fenêtre ESD, il faut savoir que connecter la grille du NMOS à une résistance aide à décroitre V_{T1}. Connecter la grille arrière à cette résistance aide également à réduire V_{T1}. Un BIMOS peut être fusionné dans le GDNMOS pour réduire R_{ON}. Décroitre le dopage dans le drain est obligatoire pour obtenir une protection très basse tension. Il est nécessaire d'enlever le siliciure pour empêcher la recombinaison des porteurs dans le drain faiblement dopé. Du siliciure partiel peut être utile pour pouvoir mettre des contacts sur le drain et connecter celui-ci à d'autres terminaux. Le siliciure partiel peut également servir pour décaler V_{T1} et I_{Leak} du dispositif. Toutes les solutions sont conçues avec le procédé standard et ne présentent pas de coût de fabrication supplémentaire. Les protections sont flexibles pour différentes applications.

Chapitre 3 : Matrices de BIMOS

I. La topologie BIMOS dot

1. BIMOS dot en 1D

Différentes topologies de BIMOS sont disponibles dans le film mince en technologie 28 nm FD-SOI (Figure 110 et Figure 111) : (i) le BIMOS « classique », où la grille est une ligne et le film de silicium est coupé de sorte à séparer les contacts de canal de la source ; et (ii) le BIMOS T-gate qui a une grille en forme de T. Le BIMOS classique a une longueur de grille importante (108 nm) et sa version en matrice n'est pas optimisée. Le BIMOS T-gate a une longueur de grille faible (48 nm), ce qui est un atout, mais il est difficilement portable en matrice. C'est ce qui a motivé la création d'une nouvelle topologie de BIMOS, le BIMOS dot, qui a une longueur de grille faible (48 nm) et qui peut être efficacement porté en matrice, en plus d'avoir des performances ESD améliorées.

La topologie BIMOS dot est appelée ainsi car son contact de canal est placé au milieu de la grille (Figure 112) et ressemble à un gros point (point = « dot » en anglais). Afin de réduire la longueur de grille, celle-ci peut être affinée loin du contact de canal (Figure 113).

Dans une première étude, trois composants sont comparés (Figure 114) : le BIMOS classique, et deux BIMOS dot (avec grille épaisse) avec un ou deux contacts de canal de 500 nm de large (chacun). La longueur de grille de chaque composant est de 300 nm (pour des soucis de comparaison), et leur largeur de doigt est de 5 µm (dot compris).

Dans les mesures TLP, la tension de déclenchement V_{T1} des trois dispositifs est similaire (~ 4,2 V), mais dans les mesures VF-TLP, elle diminue avec le nombre de « dots » (~ 3,7 V pour le BIMOS classique, ~ 3,5 V pour le BIMOS avec un « dot », et ~ 3,4 V pour le BIMOS avec deux « dots ») (Figure 115). Cela suggère que la topologie BIMOS dot est avantageuse en tant que protection CDM.

Le fait que la tension de déclenchement des dispositifs soit plus faible en VF-TLP qu'en TLP est dû au temps de montée plus court du VF-TLP. Pour une impulsion de tension donnée, un temps de montée plus court va emmener le nœud de la grille à une tension plus haute grâce aux capacités parasites entre le drain et la grille (C_{DG}) et à la résistance externe de polarisation (qui forment un circuit RC de déclenchement qui est un filtre passe-haut : avec un signal d'entrée rapide, la capacité permet d'augmenter très vite la tension sur la grille, tandis que les porteurs n'ont pas le temps d'être évacués par la résistance). En conséquence, la tension de déclenchement (liée à la tension de seuil du MOS) est atteinte pour une valeur plus faible de tension de drain.

La différence de tension de déclenchent entre les topologies peut être due à différents phénomènes : (i) les courants de conduction, (ii) les résistances et capacités parasites, (iii) la diode parasite.

Des simulations ACS permettent d'observer les différents courants de conduction dans les BIMOS dot. Lorsque la tension de drain augmente rapidement, la capacité parasite C_{DG} augmente la tension sur la grille. C'est ce qui se passe avant le premier déclenchement (un peu après 1.10^{-10} s de simulation) (Figure 117). La variation de tension en fonction du temps est rapide jusqu'à 2.10^{-10} s ; c'est à ce moment que les effets de capacité parasite jouent un rôle, et après le premier déclenchement ils n'en jouent plus vraiment. Ainsi, la tension de la grille augmente jusqu'au premier déclenchement par effet capacitif, et cette tension active la conduction de courant sous le seuil dans le NMOS (Figure 118). Avec l'augmentation du courant de drain et de tension de grille, l'ionisation par impact près du drain augmente, fournissant ainsi des électrons et des trous. Après ce premier déclenchement, les trous qui sont créés par ionisation par impact circulent dans le contact de canal. Les trous sont attirés là où la tension est la plus faible et où le dopage leur est favorable, c'est pourquoi ils sont attirés par la source et le contact de canal. Cependant plus de trous (au moins une décade de différence) vont vers la source que vers le contact de canal. Ainsi seule une petite partie des trous disponibles contribue au courant de trous dans le contact de canal tandis qu'une grande partie est perdue dans la source. Ce courant de contact de canal est très important puisque c'est lui qui aide à augmenter la tension sur la grille. De plus en plus de trous sont créés par ionisation par impact, augmentant ainsi progressivement la tension de la grille, jusqu'au second déclenchement, qui correspond au seuil du NMOS (pour un certain courant de trous dans le contact de canal). Ensuite les électrons commencent à circuler dans le contact de canal, et donc grâce à la résistance externe de polarisation, la tension du contact de canal est de plus en plus haute, et toute la conduction NMOS est active. Plus de trous peuvent être aspirés par les zones de contact de canal entre le premier et le second déclenchement dans le cas du BIMOS deux « dots » (par rapport au BIMOS qui n'a qu'un seul dot), puisque les différentes zones dans le canal sont plus proches d'un dot (qui attire les trous) s'il y en a deux. Cette différence de courant de trous explique pourquoi le BIMOS qui a deux dots se déclenche (second déclenchement sur la Figure 116) avant celui qui n'en a qu'un.

Sur la Figure 117 on voit que la tension de grille du BIMOS à un « dot » et celle du BIMOS à deux « dots » sont superposées avant le premier déclenchement. Cela correspond au moment où le couplage capacitif parasite joue un rôle, ce qui signifie qu'à priori, il n'y a pas de différence de capacité parasite entre le BIMOS à un ou à deux dots. Et pourtant, la différence de tension de déclenchement se creuse entre le TLP et le VF-TLP (Figure 115), et une valeur plus haute de C_{DG} dans le cas du BIMOS deux dots expliquerait cette augmentation de différence de tension de déclenchement. Une revue des différents parasites présents dans le composant est effectuée. Dans notre hypothèse, C_{DG} est plus grand dans le BIMOS deux dots grâce à l'augmentation de la capacité parasite entre le contact de canal et le drain, de par la jonction N^+/P (diode parasite dont la zone de charge d'espace agit comme une capacité). La surface de cette jonction est deux fois plus grande dans le cas de deux dots, ainsi la capacité de la jonction est supposée être environ deux fois plus grande. Le BIMOS classique ne bénéficie pas de cette capacité directe additionnelle entre le drain et le contact de canal (par topologie), ce qui explique pourquoi sa tension de déclenchement est encore plus grande que celle du BIMOS à un dot en VF-TLP. Il y a également une capacité parasite due à la jonction drain-canal. Le canal est lié au contact de canal puisqu'ils se touchent et qu'ils sont dopés du même type (P). Le canal étant dopé Pint (intrinsèque), il est très résistif, alors le chemin drain-canal-contact de canal comprend une capacité de jonction suivie d'une grande résistance (Figure 119). Une résistance de canal

plus faible (dans le cas de deux dots) permet au composant de se déclencher plus uniformément (c'est-à-dire que tout le doigt est impliqué), tandis qu'une résistance plus grande dans le canal peut induire des différences de tension et des décalages temporels dans le canal. La Figure 120 montre l'influence du temps de montée d'une mesure TLP. Le BIMOS à deux dots est plus sensible au temps de montée que le BIMOS à un dot (sa tension de déclenchement et celle de maintien changent de façon significative avec la modification du temps de montée), qui est plus sensible que le BIMOS classique. Ceci va en faveur de la théorie comme quoi le temps de montée plus court du VF-TLP par rapport au TLP implique un plus fort changement de tension de déclenchement dans le BIMOS deux dots que dans les autres dispositifs (grâce à ses capacités parasites).

Le dernier élément à considérer est la diode parasite $P^+/P/N^+$ située entre le contact de canal et la source. Cette diode est polarisée en direct, ce qui induit un courant de trous qui va du contact de canal à la source. Ce courant réduit le courant total qui va du composant au contact de canal, et qui contribue à l'augmentation de la tension de contact de canal. Ainsi cette diode parasite (qui est plus large dans le cas du BIMOS dot) contrebalance l'augmentation de la tension de grille. La tension de grille augmente tout de même avec le temps, mais moins que s'il n'y avait pas de diode parasite.

Le BIMOS classique casse pour un I_{T2} et un V_{T2} bien plus faible que le BIMOS dot (Figure 115). La robustesse de cette topologie est probablement plus faible à cause du pont de silicium.

Pour conclure cette étude, lorsque l'on compare les données TLP du BIMOS dot à celles du BIMOS classique (Table 2), on voit qu'il a la même tension de déclenchement V_{T1}, un courant de maintien I_H plus faible, et un courant de casse I_{T2} plus élevé (ce qui est très important, puisque ça montre que le composant est plus robuste et a une meilleure propagation du courant). En VF-TLP, le BIMOS dot a une tension de déclenchement V_{T1} plus faible (particulièrement s'il a deux dots), et un courant de casse I_{T2} plus élevé. Des mesures DC (Figure 122) montrent que le courant de fuite des structures est compatible avec la technologie 28 nm FD-SOI (quelques dizaines de pA à 1 V).

2. Matrice de BIMOS dot

Dans cette nouvelle étude, quatre composants sont comparés : le BIMOS classique, le BIMOS T-gate, le BIMOS dot avec grille épaisse, et le BIMOS dot avec grille affinée. Les dimensions minimales ont été sélectionnées afin que chaque dispositif ait des performances maximales (Figure 124). La surface du BIMOS dot est bien plus grande que celle du BIMOS classique et du T-gate (Figure 125), mais ce problème de surface peut être éludé grâce à la topologie en matrice.

Des simulations ACS (Figure 126) permettent de comparer les composants. La haute tension de déclenchement du BIMOS classique est attribuée à sa longue grille (108 nm). Le T-gate et le BIMOS dot affiné ont la même longueur de grille (48 nm) ce qui permet de comparer leur performance. Le BIMOS dot affiné a la tension de déclenchement la plus faible en ACS (c'est ce que l'on cherche à obtenir). Cependant, les mesures montrent que la différence entre le T-gate et le BIMOS dot affiné est négligeable (Figure 127). Les capacités

parasites des lignes de connexion métallique masquent probablement les améliorations capacitives du BIMOS dot. A part cela les tendances sont confirmées : les composants avec une grille plus longue ont une tension de déclenchement plus haute.

Considérons à présent une matrice du BIMOS classique et du BIMOS dot affiné (le T-gate n'étant pas portable en matrice) (Figure 128). Le pont de silicium du BIMOS classique induit de grosses contraintes de dimensionnement : non seulement sa grille est plus longue (108 nm au minimum), mais également d'autres espacements entre certains éléments doivent être pris en compte (par exemple la distance minimale entre drain/source et le contact de canal, qui vaut 147 nm). Ainsi la présence du contact de canal dans la matrice fait perdre 672 nm, au lieu de 414 nm dans le cas du BIMOS dot, qui a moins de contraintes. Chaque matrice a été optimisée par calcul en termes de surface de silicium consommée en fonction de la taille de son contact de canal (déterminé par les règles de dessin). Le nombre de doigts a été adapté pour obtenir une largeur de conduction totale de 100 µm dans les deux cas (matrice de BIMOS classique et de BIMOS dot). Les dimensions obtenues sont résumées dans la Table 3. La matrice de BIMOS dot a une surface totale 39% moins grande que celle de la matrice de BIMOS classique.

Les mesures de ces dispositifs (Figure 129) montrent que la matrice de BIMOS dot a de meilleures performances (tension de déclenchement plus faible) que la matrice de BIMOS classique, grâce à sa grille plus courte. De plus sa topologie de grille lui permet d'avoir une meilleure intégration que le pont de silicium de la matrice de BIMOS classique : la matrice de BIMOS dot est plus robuste, plus conforme aux règles de dessins, et sa surface est réduite (ce qui lui donne de la marge pour augmenter son nombre de doigts et donc son courant de casse I_{T2}).

II. Comparaison de différents BIMOS

Dans la section précédente, la topologie de BIMOS dot s'est montrée avantageuse lorsqu'elle est portée en matrice, en partie grâce à la surface de silicium gagnée à chaque contact de canal de la matrice. Mais est-ce que la matrice de BIMOS a vraiment besoin d'un contact de canal à côté de chaque petit NMOS (un drain et une source) ? Cette section va répondre à la question en comparant plusieurs topologies de BIMOS dans le film mince, parmi lesquelles, une matrice 2D de BIMOS qui n'a qu'un seul contact de canal en périphérie.

Les composants 1 à 6 (Table 4 et Figure 130) sont des matrices 2D de NMOS entourées par un contact de canal. Les matrices 1 à 4 sont conçues avec les dimensions minimales compatibles avec les règles de dessin de la technologie considérée : il n'y a qu'un seul contact en tungstène par source ou drain (« petite surface »). Les matrices 5 et 6 sont plus larges car il y a cette fois ci deux contacts en tungstène (dans chaque direction) par source/drain (« grande surface »). Les composants 7 à 12 sont des BIMOS en 1D conventionnels. Ils ont chacun été conçus avec les dimensions minimales (longueur de grille, etc.) en fonction de leur topologie. Leur nombre de doigts a été adapté de sorte à avoir une surface de silicium similaire à celle des matrices (« petite » et « grande » surface), pour pouvoir les comparer avec une même référence de surface. Les composants 9 et 10 ont une

grille en forme de π (il s'agit des « BIMOS T-gate »), les composants 7 et 8 ont une grille en forme de H (« H-gate » : il s'agit d'une petite variante au T-gate), et les composants 11 et 12 sont des « BIMOS classique ». Tous les composants ont une grille de 48 nm sauf les BIMOS classiques (108 nm). Dans les mesures effectuées, la grille arrière de tous les dispositifs est mise à la masse.

Le but de ce plan d'expérience est de savoir quelle topologie (matrice, H-gate, T-gate ou classique) est la plus adaptée pour être utilisée en tant que dispositif de protection ESD pour une technologie 28 nm FD-SOI, pour une surface donnée (petite ou grande). Les structures aillant une petite ou une grande surface sont comparées de sorte à faire ressortir l'information suivante : la petite matrice 2D (qui n'a qu'un contact par source/drain) peut elle aussi être utilisée comme protection ESD ? Dans les composants aillant une petite surface, les dimensions ont été poussées à la limite de la technologie, et la robustesse de la petite matrice 2D pourrait être encore plus impactée (puisque chaque carré de NMOS doit être parfaitement fabriqué afin que le composant soit opérationnel).

L'effet de la résistance connectée à la grille et au contact de canal du BIMOS est montré dans la Figure 131 et la Figure 132. Comme expliqué dans le chapitre 2, la résistance aide à contrôler la tension de déclenchement. Une forte valeur de résistance permet de polariser à la fois le contact de canal et la grille à une tension plus élevée en un temps plus court, ce qui diminue la tension de déclenchement. La résistance externe de polarisation et la capacité parasite C_{DG} (c'est la capacité du séparateur entre le drain et la grille, celle de l'oxyde qui sépare la grille du canal au niveau du chevauchement du drain, et celle de la jonction drain/canal) forment un système RC de déclenchement. Etant donné que C_{DG} vaut environ 4.10^{-14} F, la constante de temps τ de ce système vaut 0,4 ns pour la résistance R_1, 2 ns pour R_2 et 4 ns pour R_3. Ceci explique que la différence de tension de déclenchement soit observée entre R_2 et R_3 en TLP (mesure qui dure 100 ns) et non en VF-TLP (qui ne dure qu'une nanoseconde) (Figure 131). En mesure quasi-statique en revanche, la différence de tension de déclenchement pour différentes valeurs de résistance est systématiquement observée puisque la mesure est suffisamment longue (Figure 132). Il faut brancher une résistance la plus élevée possible afin de diminuer la tension de déclenchement pour pouvoir rentrer dans la fenêtre ESD de la technologie 28 nm FD-SOI. La grille du dispositif numéro 4 est laissée flottante, ainsi le composant se comporte comme si une résistance infinie était branchée à sa grille et il se déclenche très tôt en DC. A cause de cela, le dispositif a un courant de fuite élevé pour des tensions plutôt faibles : si on accepte 10 nA de courant de fuite au maximum, alors il ne faut pas utiliser ce composant à une tension supérieure à 1,1 V, ce qui laisse peu de marge par rapport à une technologie fonctionnant à 1 V. Il est donc recommandé de brancher une résistance à la grille, même si cette résistance prend de la surface de silicium.

Lorsque l'on compare différentes topologies (matrice, H-gate, T-gate, classique) en TLP (Figure 133), on voit que la courbe du BIMOS classique est séparée des autres, qui sont plutôt regroupées. La tension de déclenchement V_{T1} et la résistance dynamique à l'état passant R_{ON} du BIMOS classique sont plus hautes que pour les autres topologies, ce qui est problématique si le but est de protéger des transistors qui ont un oxyde de grille fin. En revanche le BIMOS classique a beaucoup moins de courant de fuite que les autres topologies (Figure 134). Ces différences sont attribuées à la longue grille du BIMOS classique. La comparaison de courant de fuite est prévue pour vérifier que chaque topologie (dessinée avec les dimensions minimales) atteint les objectifs de courant de fuite pour une technologie

donnée et non pour pouvoir vraiment comparer les dispositifs entre eux ; cependant il est important de savoir que le courant de fuite des autres topologies peut aussi être relaxé si leur longueur de grille est augmentée. Le courant de casse I_{T2} du BIMOS classique est également plus bas que pour les autres composants (Figure 133), étant donné que moins de doigts pouvaient être inclus dans la structure (et ce pour une même surface). Ce BIMOS est intéressant puisque si un courant de fuite plus faible est demandé, la grille des autres topologies devrait être allongée, et si tous les composants ont la même de longueur de grille le BIMOS classique pourrait redevenir compétitif à nouveau. L'éventuel bénéfice du BIMOS classique est qu'il y a moins de conduction par la diode parasite entre le contact de canal et la source (où les trous sont attirés par la source). Ceci doit être confirmé dans une nouvelle étude où tous les dispositifs ont la même longueur de grille.

Il n'y a pas de différence visible entre le H-gate et le T-gate, ce qui montre que les concepteurs peuvent choisir l'un ou l'autre.

Quand la matrice a une « grande » surface, son comportement est le même que celui du H-gate et du T-gate. Cependant, lorsqu'elle a une « petite surface », son R_{ON} est affecté (Figure 133). Cela peut s'expliquer par (i) l'arrondi des grilles qui prend une portion trop grande de la largeur de chaque MOS, ou par (ii) la forte résistance d'accès due à l'unique contact en tungstène par MOS. Les mêmes raisons peuvent expliquer pourquoi le courant de fuite à 1 V de la matrice est plus faible que celui des H-gate et T-gate (Figure 134). Il est intéressant de noter que même si le contact de canal se situe loin de chaque MOS, la matrice de BIMOS marche de façon similaire aux autres topologies de BIMOS. Ce résultat est plutôt important puisque les matrices 2D peuvent être inspirantes pour créer des matrices 3D. La simplification de la topologie (en plaçant le contact de canal à la périphérie de la matrice) serait probablement requise pour imaginer une matrice de BIMOS en 3D, où les grilles contrôleraient des canaux situés au-dessus et en dessous (Figure 135), voire même latéralement.

En conclusion, il a été montré que la matrice 2D était aussi performante que les topologies classiques en 1D. Celle-ci se comportait de façon similaire même si son contact de canal était situé à la périphérie du dispositif et non à chaque nœud de NMOS comme dans le BIMOS dot. Des études supplémentaires sont requises afin de déterminer à quel point le contact de canal peut être éloigné du centre de la matrice sans observer d'effets tels que le multi-déclenchement. La suppression de ces petits contacts de canaux à chaque nœud permet de gagner une surface de silicium considérable (tellement, que les matrices de BIMOS deviennent à nouveau compétitives face aux composants 1D en termes de surface).

Chapitre 4 : Protections ESD 3D en technologie FD-SOI avec continuité de silicium

Le fait de fusionner deux composants ESD, ou de fusionner le composant de protection avec son circuit de déclenchement peut présenter de nombreux avantages. Des protections ESD fusionnées avec leur circuit de déclenchement existent déjà dans le substrat ou bien dans le film mince. Cependant il n'existe pas encore de protection dans le substrat (qui bénéficierait ainsi d'une conduction volumique) compactée avec son circuit de déclenchement dans le film mince. Nous nous sommes inspirés de l'idée de composant-sur-composant [118] (Figure 138), où un premier composant se trouve sous le BOX (dans le substrat) et un autre se trouve dessus (dans le film mince), afin de proposer la fusion du « composant-sur-composant », en creusant un trou dans le BOX au milieu du composant (Figure 139). Cela constitue un exemple d'intégration 3D à partir d'une technologie 2D. Avec la fusion du composant-sur-composant, avec continuité de silicium entre le composant du dessus et celui du dessous, de nouvelles topologies et de nouveaux composants sont permis avec de nouveaux degrés de courant de conduction.

Cette fusion peut être réalisée grâce à une brique de procédé de STMicroelectronics, ainsi elle ne présente pas de surcoût. Le procédé NOSO fonctionne avec deux masques (Figure 140). En intervertissant la taille des deux masques à utiliser, on peut obtenir une continuité de silicium entre le film mince et le substrat (Figure 142). La frontière entre le silicium qui remplit le trou dans le BOX et le silicium du film mince n'est peut-être pas de qualité, ce qui implique des distances minimales de marge à respecter dans le dessin des composants. Dans ce chapitre le procédé NOSO est utilisé pour réaliser un BIMOS fusionné avec un SCR en 3D, et ensuite un BIMOS fusionné avec sa résistance. Seules des simulations TCAD font office de preuve de concept puisque les mesures électriques seront réalisées après la fin de cette thèse.

I. BIMOS fusionné avec un SCR en 3D

Un exemple de BIMOS (circuit de déclenchement dans le film mince) fusionné avec un SCR (dans le substrat pour bénéficier d'une conduction volumique) est donné (Figure 143, Figure 144, Figure 145). Une fusion est réalisée dans le contact de canal du BIMOS (qui est aussi la gâchette Gp du SCR), et également dans la cathode du SCR (source du BIMOS). La gâchette dopée P a été choisie pour contrôler l'activation du SCR car le transistor bipolaire NPN est plus rapide à s'activer que le PNP.

Des simulations (Figure 146) comparent un BIMOS, un SCR avec gâchettes flottantes, le BIMOS et le SCR connectés ensembles, et le BIMOS et le SCR fusionnés. Le courant de fuite des quatre solutions est acceptable. Il faut éventuellement faire attention à ce que le SCR ne soit pas sujet au « Latch Up ». Le SCR a une bonne tension de maintien V_H et résistance dynamique à l'état passant R_{ON}, mais il a une trop haute tension de déclenchement V_{T1}, donc il a besoin d'un circuit de déclenchement. Le BIMOS pourrait être le circuit de

déclenchement du SCR car son V_{T1} est plus faible. Le BIMOS connecté au SCR est donc une bonne solution, car il a l'avantage d'un V_{T1} plus faible grâce au BIMOS et également d'un R_{ON} plus faible grâce au SCR. Le SCR fusionné au BIMOS est la meilleure solution. Son V_{T1} est plus bas grâce aux porteurs du premier composant qui aident le deuxième à se déclencher. Aussi le contact de canal du BIMOS agit comme une grille arrière polarisée positivement pour le BIMOS, puisque le BOX est ouvert, ce qui aide à réduire V_{T1}. De plus, le composant fusionné prend 16% moins de surface que le composant connecté (Table 5).

La Figure 147 et la Figure 148 expliquent la forme de la courbe du composant fusionné, avec son double déclenchement, en montrant la contribution du SCR et du BIMOS. Le principe est le suivant : la tension de déclenchement du composant fusionné correspond à celle du BIMOS. Dès que le BIMOS entre en conduction, le potentiel augmente sur le système grille-contact de canal, donc la gâchette Gp passe d'une basse tension (SCR bloqué) à une haute tension (SCR ouvert) (Figure 150). Le composant atteint sa tension de maintien lorsque le SCR et le BIMOS sont en conduction.

Le même schéma électrique et principe de fonctionnement peut être utilisé mais avec une autre topologie. Par exemple, dans la topologie présentée en Figure 151, le SCR est plus long, la topologie du BIMOS est différente (grille avec une forme en T), et le BIMOS est vraiment au-dessus du SCR.

La fusion peut également être faite dans la gâchette dopée N du SCR (Gn au lieu de Gp). Le schéma électrique et principe de fonctionnement de ce BIMOS fusionné avec un SCR en 3D est un peu différent du précédent (dans ce composant, le SCR est activé grâce à sa gâchette Gn). Un exemple de topologie de ce composant fusionné est donné en Figure 152 et Figure 153. Une fusion est faite dans le drain du BIMOS (qui est également la gâchette Gn du SCR), et une autre dans la cathode du SCR (source du BIMOS). Dès que la diode PN (dopage P de l'anode et dopage N de la gâchette Gn) laisse passer un courant suffisant (avec l'augmentation de la tension sur l'anode), la tension augmente sur le drain du BIMOS. Lorsque le BIMOS entre en conduction (pour une tension de drain suffisante), il passe d'un état très résistif à un état peu résistif. Dès lors, la tension du nœud Gn devient basse (car le NMOS agit comme une petite résistance), ce qui ouvre le SCR. La Figure 154 montre le comportement du composant fusionné. C'est la tension sur le drain du BIMOS qui diminue drastiquement juste après avoir atteint 1,5 V qui active le SCR. Dans le tableau de la Figure 154, les performances du composant fusionné sont comparées à celles d'un simple BIMOS et d'un simple SCR qui auraient les mêmes dimensions et topologie.

Pour résumer, le BIMOS est un bon candidat pour jouer le rôle de circuit de déclenchement pour un SCR de puissance. Avec l'utilisation du procédé NOSO il est possible de fusionner le BIMOS dans le film mince avec le SCR dans le substrat. De multiples topologies et schémas électriques de BIMOS fusionné avec un SCR peuvent être imaginés. Les avantages de la structure fusionnée sont les suivants : (i) le composant est compact, donc grâce à la fusion il y a moins de résistivité dû à la connectique, et d'autres problèmes liés à la distance entre le film mince et les composants dans le substrat sont évités ; (ii) on gagne de la surface de silicium ; (iii) la conduction de courant entre le film mince et le substrat est permise, ainsi que le contrôle direct du potentiel, une homogénéité thermique, etc. ; ce qui mène à (iv) des améliorations pour la solution ESD (qui lui permettent de rentrer dans de nouvelles fenêtres de conception).

II. Résistance fusionnée

1. Résistance fusionnée dans le film mince

L'approche classique pour utiliser un BIMOS en tant que protection ESD consiste à brancher le contact de canal à la grille, et relier ce nœud à la masse par l'intermédiaire d'une résistance externe en poly-silicium. Ceci permet à une impulsion ESD positive de s'écouler du drain à la source du BIMOS. Pour l'impulsion ESD négative, une diode externe est branchée entre la source et le drain. Une nouvelle façon d'utiliser le BIMOS serait de remplacer la résistance externe de polarisation par une résistance intégrée dans le canal (Figure 155). Pour ce faire, un premier contact de canal (à un bout du canal) est branché à la grille (et ce nœud est laissé flottant), et un deuxième contact de canal (à l'autre bout du canal) est mis à la masse. Cette configuration permet une diode parasite entre le deuxième contact de canal (dopé P^+) et le drain (dopé N^+), ainsi qu'une résistance parasite entre les deux zones de contact de canal. En conséquence, à la place de trois composants (le BIMOS, la résistance et la diode), il n'y aurait plus qu'un seul composant fusionné. Le but principal de l'étude est de réduire la surface. Cependant, ces éléments parasites ne sont pas aussi facile d'utilisation que les composants externes. Cette section se propose de fournir une preuve de concept à propos du bon fonctionnement de ce BIMOS fusionné en tant que protection ESD.

La résistance du canal (dopé Pint) est très grande (quelques MΩ). Sa valeur est difficile à calculer avec précision puisqu'il s'agit d'une résistance distribuée le long du canal. En effet, il est possible de décrire le BIMOS avec résistance fusionnée comme une multitude de « petits » NMOS qui ont une « petite » largeur et qui sont chacun une section du BIMOS (en Figure 155 par exemple, trois NMOS ont été utilisés pour décrire schématiquement le BIMOS). De plus, la valeur de la résistance du canal change avec la conduction à l'intérieur de celui-ci.

Une forte valeur de résistance est bénéfique pour le comportement du BIMOS en ACS, car cela réduit sa tension de déclenchement ; cependant cela tend également à l'activer trop tôt en AVS, alors son courant de fuite pourrait être trop élevé. Il faut donc réduire la valeur de la résistance, et pour ce faire, on peut modifier différents paramètres :

- La largeur W du MOS, qui doit être la plus faible possible afin de réduire la valeur de la résistance fusionnée (Figure 156). Comme la largeur doit être très faible pour pouvoir rester dans la fenêtre ESD, le composant serait un bon candidat pour être utilisé en tant que protection CDM ou en tant que circuit de déclenchement, mais pas en tant que protection HBM. Quelques doigts peuvent être ajoutés afin d'agir comme des résistances en parallèle.
- La longueur L du NMOS doit être la plus longue possible afin de réduire la valeur de la résistance, mais cela dégrade l'ACS (V_{T1} plus haut).
- Une tension négative peut être appliquée sur la grille arrière pour augmenter la tension de déclenchement dans la caractéristique AVS (cela change la résistivité du silicium) (Figure 157).

Des simulations 3D ont été effectuées afin de trouver numériquement une dimension (L et W) de sorte à obtenir une valeur de résistance similaire sur un BIMOS T-gate qui a une

résistance externe classique adéquate ou bien une résistance distribuée dans son canal. Dans cette étude, la grille arrière est mise à la masse. La Figure 158 montre qu'un BIMOS T-gate qui a une largeur de 114 nm (dimensions minimales) peut avoir un comportement similaire selon que sa résistance soit externe ou interne, si sa longueur de grille vaut 300 nm.

2. Résistance fusionnée dans le substrat

Dans la section précédente, le canal du BIMOS était utilisé comme une résistance pour relier le nœud contact de canal – grille à la masse. Dans cette section, c'est le Pwell qui est utilisé comme résistance, en ouvrant le BOX grâce au procédé NOSO à l'intérieur des zones de contact de canal (Figure 159 et Figure 160). La résistance du Pwell est donc mise en parallèle avec celle du canal, et comme sa valeur est bien moins grande, on peut négliger la résistance du canal. Vu que la résistance du Pwell est située au niveau de la grille arrière, la tension du premier contact de canal (celui branché à la grille) va également aider le BIMOS à se déclencher.

Une attention spéciale est portée aux dimensions du BIMOS. Etant donné que le dopage du Pwell est bien plus fort que dans le canal, cette fois ci la résistance distribuée tend à être trop faible (la résistance externe de polarisation typiquement branchée sur le BIMOS vaut quelques dizaines de kΩ). Ainsi, il faut utiliser un nombre minimum de doigts (idéalement un seul), et le STI à côté du drain doit être rapproché au maximum du STI à côté de la source de sorte à ne pas avoir une résistance trop large (cela implique des dimensions minimales pour les longueurs de drain, de grille et de source). La longueur de la résistance doit aussi être maximisée (grande largeur W du BIMOS). Une protection CDM pourrait typiquement avoir un doigt et une petite longueur de grille, mais sa largeur est de l'ordre de 1 µm. Des BIMOS aillant différentes largeurs ont été simulés numériquement : 500 nm (Figure 161) et 2 µm (Figure 162). Sur ces graphes, le composant avec une résistance distribuée dans le Pwell est comparé aux composants qui ont différentes valeurs de résistance externe de polarisation (R_1, R_2 et R_3 dans l'ordre de valeur croissante). A cause de sa largeur, le BIMOS de 500 nm de large avec résistance distribuée a un comportement électrique similaire à celui du BIMOS avec une résistance externe de polarisation R_1. Celui de 2 µm a un comportement plus proche du BIMOS avec résistance externe R_2. A titre de comparaison, R_3 est la valeur qui devrait typiquement être utilisée dans un circuit classique de BIMOS avec résistance externe de polarisation. Il est possible de réduire d'avantage la tension de déclenchement du BIMOS avec résistance distribuée en augmentant sa largeur, mais ce ne serait pas raisonnable en termes de surface de silicium (pour être utilisé en tant que protection CDM).

Une remarque par rapport à la Figure 161 et la Figure 162 lorsque l'on compare un BIMOS branché à une résistance externe de polarisation R_3 et le même BIMOS branché à R_3 sauf que des ouvertures ont été faites dans son BOX au niveau des zones de contact de canal (ce qui connecte la grille avant à la grille arrière) : avec les ouvertures dans le BOX, le composant se déclenche plus tard. Ceci est en contradiction avec le fait que la grille arrière aide à réduire la tension de déclenchement. Une explication peut être que si l'on ouvre le BOX, alors la tension sur le contact de canal monte plus doucement puisque celui-ci est

connecté à un gros volume (le substrat), et donc il lui faut un nombre plus élevé de porteurs pour monter son potentiel. Cette hypothèse doit être vérifiée dans une prochaine étude.

La tension de déclenchement du BIMOS avec résistance distribuée est plus faible que celle du BIMOS classique en ACS négatif (Figure 163), grâce à la diode parasite entre le contact de canal relié à la masse et le drain. Il n'y a pas de risque de « Latch Up » en ACS négatif puisque le circuit n'est pas supposé être polarisé inversement.

Des mesures doivent être faites afin de trouver la vraie tension de déclenchement des BIMOS avec résistance distribuée, et pour s'assurer que ces dispositifs peuvent être utilisés en tant que protection CDM. Le BIMOS de 2 µm de largeur semble être un bon candidat pour de futures investigations.

D'autres composants avec résistance distribuée (grâce à l'ouverture du BOX) peuvent être imaginés. Par exemple, une résistance distribuée peut être ajoutée au BIMOS fusionné avec SCR de la section précédente (Figure 164 et Figure 165).

Pour conclure le chapitre, de nouvelles variantes de protections ESD ont été découvertes. Leur avantage repose sur l'idée de conduction 3D et de fusion de composants avec continuité de silicium (entre la couche supérieure et inférieure autours du BOX). Il a été montré qu'il était possible de créer une protection ESD en 3D même avec une technologie 2D. Les simulations ont apporté une preuve de concept, et des mesures doivent être effectuées afin de prouver la faisabilité et d'évaluer la performance des composants plus précisément.

Conclusions générales

Le contexte de ce manuscrit est l'étude de protections ESD existantes et leur amélioration, en ligne avec le concept de technologie 3D, comme par exemple les composants 3D homothétiques et polymorphiques sans interconnexions métalliques.

Il a été décidé de commencer par l'étude de composants ESD dans le film mince dans la technologie FD-SOI. Il est utile d'améliorer des composants ESD dans le film mince car le niveau de protection pouvant être actuellement atteint est réduit (à cause de leur conduction surfacique). Il est également utile d'élargir les possibilités de protections pour pouvoir fournir des protections pour différentes fenêtres de conceptions. La thèse inclus une solution de « boost » capacitif afin de réduire la tension de déclenchement, qui consiste à ajouter un doigt connecté au drain dans un NMOS ou un BIMOS, afin d'ajouter une capacité parasite qui aide le composant à se déclencher plus tôt. Cette technique peut aussi être utilisée pour d'autres composants, par exemple pour un GDNMOS. Le GDNMOS est une diode à grille fusionnée avec un transistor NMOS ; il a intensivement été étudié dans la thèse, ce qui a mené à des améliorations dans la compréhension du mécanisme d'activation. En fait, le GDNMOS, le NMOS et le BIMOS bâtis dans le film mince sont principalement déclenchés au travers de leurs capacités parasites. Ce phénomène prend le pas, bien plus que les autres phénomènes tels que l'effet tunnel de bande à bande ou l'ionisation par impact. Ces mécanismes jouent encore un rôle, mais ils ne sont pas prédominants comme c'est le cas dans les mêmes composants bâtis dans le substrat.

Le but principal de l'étude sur le GDNMOS était d'obtenir un comportement de type SCR. Il a été souligné que le drain était trop fortement dopé pour ce faire. En abaissant le dopage, ainsi qu'en enlevant le siliciure présent sur le drain, la tension de déclenchement du GDNMOS était effectivement abaissée et le composant se comportait comme un SCR. Pourtant le courant de fuite dans le GDNMOS avec drain faiblement dopé était trop haut. Une nouvelle idée a été proposée pour résoudre ce problème. Cela consiste à enlever partiellement le siliciure le long de la largeur du drain. Cette solution permet de décaler les paramètres électriques. Ainsi tout un ensemble de composants sont disponible et utilisables pour différentes tensions d'alimentations, donc pour différentes technologies. Lors de l'étude des GDNMOS, l'importance de brancher des résistances externes sur les grilles des composants a été soulignée. En fait, non seulement la résistance aide à décaler les paramètres électriques, mais elle est aussi vitale pour des questions de robustesse. Un tout nouveau dispositif a été créé au cours de ces études sur les composants dans le film mince : le GDBIMOS, qui est un GDNMOS fusionné avec un BIMOS. En théorie le BIMOS devrait améliorer la résistance dynamique à l'état passant du GDNMOS. Ce nouveau composant permet également des connectiques additionnelles. On pourrait par exemple imaginer de connecter l'un des contacts de canal à la grille de la diode, et l'autre quelque part ailleurs, afin de créer un transistor PMOS parasite dans le GDBIMOS. Certaines connectivités ont été investiguées dans ce manuscrit, mais il reste toujours des possibilités intéressantes de connexions à trouver, puisque ce composant a l'avantage d'être très reconfigurable.

L'analyse de différentes connectivités, l'addition de nouveaux doigts, et la fusion de composants ne sont pas les seuls leviers afin d'améliorer les performances des composants et de mieux comprendre leur physique. Il y a également l'investigation de leur topologie.

Dans ce cadre, des composants 2D ont été étudiés (avec dans l'idée de pouvoir les transformer en composants 3D par après). Par 2D on entend une conduction en 2D de courant, et on pense donc typiquement à des matrices. Les plus simples à fabriquer sont les matrices de NMOS ou de BIMOS, puisqu'elles ont moins de terminaux qu'un GDNMOS par exemple. La topologie de « BIMOS dot » a été proposée et a menée à une matrice de BIMOS dot qui est plus efficace qu'une matrice de BIMOS classiques avec un contact de canal à chaque coin de source ou de drain. Des topologies de BIMOS en 1D différentes ont aussi été comparées à une matrice de NMOS avec un anneau externe de contact de canal. Les résultats ont montré que les carrés de sources ou de drains doivent être les plus petits possibles afin de réaliser de meilleures performances que les composants en 1D, bien que les contraintes de procédé ne permettent actuellement pas de réduire suffisamment les dimensions. Ce résultat doit être gardé en tête, puisque ce n'est probablement qu'une question de temps avant que les contraintes de procédé ne soient relâchées. Un jour, on pourra même penser à une matrice 3D de BIMOS.

Fusionner différents composants entre eux a également été inspirant pour trouver une nouvelle structure avec une conduction en 3D. L'idée était de creuser un trou dans le BOX au milieu du composant dans le film mince, afin de le fusionner avec un autre composant présent dans le substrat. Un BIMOS fusionné avec un SCR en 3D a été présenté. Il présentait l'avantage d'utiliser la conduction surfacique pour le circuit de déclenchement (le BIMOS) et la conduction volumique pour le SCR, afin d'obtenir une grande robustesse et atteindre un bon niveau de protection ESD. La même technique d'ouverture du BOX peut aussi être utilisée pour ne plus avoir besoin d'une résistance dans un BIMOS.

Tous les composants étudiés durant la thèse ont été dessinés et simulés. Une attention particulière a été portée sur la faisabilité de la fabrication, même dans le cas des structures 3D. La compatibilité avec le procédé était un critère important pour pouvoir utiliser les dispositifs de façon relativement immédiate. Certains composants étaient fabriqués et mesurés. D'autres, comme le BIMOS fusionné avec le SCR en 3D ou le BIMOS avec la résistance fusionnée doivent encore être fabriqués. La prochaine personne qui va s'occuper d'un sujet similaire aura encore de la marge pour une meilleure compréhension et des découvertes supplémentaires. Le travail de vérification des résultats de simulation grâce aux mesures doit être continué. De plus, d'un point de vue industriel, des degrés supplémentaires de maturité doivent être atteints pour permettre la commercialisation, tels que s'assurer d'un bon rendement et réduire le décalage des comportements électriques dus à la variabilité du procédé.

Ce travail ouvre la voie à des protections ESD qui n'existent pas encore dans l'industrie. Quelques composants innovants ont été proposés, et certaines problématiques de 3D ont été abordées.

References

[1] D. Kahng, "A historical perspective on the development of MOS transistors and related devices," *IEEE Transactions on Electron Devices,* vol. 23, no. 7, pp. 655-657, July 1976.

[2] S. Cristoloveanu and S. S. Li, Electrical characterization of Silicon-On-Insulator materials and devices, Springer Science + Business Media, 1995.

[3] Z. H. Liu et al., "Threshold voltage model for deep-submicrometer MOSFETs," *IEEE Transactions on Electron Devices,* vol. 40, no. 1, pp. 86-95, January 1993.

[4] M. Lundstrom, "Elementary Scattering Theory of the Si MOSFET," *IEEE Electron Device Letters (EDL),* vol. 18, no. 7, pp. 361-363, 1997.

[5] K. Y. Toh, P. K. Ko, and R. G. Meyer, "An Engineering Model for Short-Channel MOS Devices," *IEEE Journal of Solid-State Circuits,* vol. 23, no. 4, pp. 950-958, 1988.

[6] D. J. DiMaria, E. Cartier, and D. Arnold, "Impact ionization, trap creation, degradation, and breakdown in silicon dioxide films on silicon," *Journal of Applied Physics (JAP),* vol. 73, no. 7, pp. 3367-3384, 1993.

[7] C. Hu et al., "Hot-electron-induced MOSFET degradation - Model, monitor, and improvement," *IEEE Transactions on Electron Devices (TED),* vol. 32, no. 2, pp. 375-385, 1985.

[8] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Letters (EDL),* vol. 8, no. 11, pp. 515-517, 1987.

[9] S. Tam, P. K. Ko and Chenming Hu, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Transactions on Electron Devices (TED),* vol. 31, no. 9, pp. 1116-1125, 1984.

[10] M. Vinet et al., "Opportunities brought by sequential 3D CoolCubeTM integration," in *46th European Solid-State Device Research Conference (ESSDERC)*, Lausanne, 2016.

[11] S. Datta, “The Era of Hyperscaling in Electronics,” in *IRPS 2019: IEEE International Reliability Physics Symposium*, Monterey CA USA, April 2019.

[12] C. Maleville and C. Mazure, "Smart-Cut technology: from 300 mm ultrathin SOI production to advanced engineered substrates," *Solid-State Electronics,* vol. 48, no. 6, pp. 1055-1063, June 2004.

[13] W. Schwarzenbach et al., "Excellent silicon thickness uniformity on Ultra-Thin SOI for controlling Vt variation of FDSOI," in *IEEE International Conference on IC Design & Technology (ICICDT)*, Kaohsiung, 2011.

[14] C. Gallon, "Architectures avancées de transistors CMOS SOI pour le noeud 32 nm et en deçà : films ultra-fins, contraintes mécaniques, BOX mince et plan de masse," PhD manuscript, Grenoble, 2007.

[15] F. Arnaud et al., "Competitive and Cost Effective high-k based 28nm CMOS Technology for Low Power Applications," in *IEEE International Electron Devices Meeting (IEDM)*, Baltimore, 2009.

[16] N. Planes et al., "28nm FD-SOI technology platform for high-speed low-voltage digital applications," in *Symposium on VLSI Technology (VLSIT)*, Honolulu, 2012.

[17] C. Fenouillet-Beranger et al., "Low power UTBOX and back plane (BP) FDSOI technology for 32nm node and below," in *IEEE International Conference on IC Design & Technology (ICICDT)*, Kaohsiung, 2011.

[18] STMicroelectronics SA, "CMOSM28 Design Rules Manual," 2018.

[19] J. Mazurier, "Etude de la variabilité en technologie FDSOI : du transistor aux cellules mémoires SRAM," PhD manuscript, Grenoble, 2012.

[20] ESDA (EOS/ESD Association), "ESD ADV1.0-2012 ESDA Advisory for Electrostatic Discharge Terminology - Glossary," 2012.

[21] B. Viale, "Development of predictive analysis solutions for the ESD robustness of integrated circuits in advanced CMOS technologies," PhD manuscript, Lyon, 2017.

[22] A. Amerasekera and C. Duvvury, "ESD in silicon integrated circuits," Second ed., John Wiley & Sons, Ltd, 2002.

[23] K. T. Kaschani, "What is Electrical Overstress? Analysis and Conclusions," *Microelectronics Reliability,* vol. 55, no. 6, pp. 853-862, 2015.

[24] M. K. Radhakrishnan, K. L. Pey, C. H. Tung, and W. H. Lin, "Physical analysis of hard and soft breakdown failures in ultrathin gate oxides," *Microelectronics Reliability,* vol. 42, no. 4-5, pp. 565-571, 2002.

[25] S. Voldman, R. Gauthier, D. Reinhart and K. Morrisseau, "High-current transmission line pulse characterization of aluminum and copper interconnects for advanced CMOS semiconductor technologies," in *IEEE International Reliability Physics Symposium Proceedings (IRPS)*, Reno, NV, USA, 1998.

[26] P. Olivo, T. N. Nguyen and B. Ricco, "High-Field-Induced Degradation in Ultra-Thin SiO2

Films," *IEEE Transactions on Electron Devices (TED),* vol. 35, no. 12, 1988.

[27] B. Schlund, C. Messick, J. Suehle and P. Chaparala, "A new physics-based model for time-dependent-dielectric-breakdown," in *IEEE International Integrated Reliability Workshop*, Lake Tahoe, CA, USA, 1995.

[28] J. S. Smith, "General EOS/ESD equation," in *EOS/ESD Symposium Proceedings*, Santa Clara, CA, USA, 1997.

[29] A. Griffoni et al., "Electrical-Based ESD Characterization of Ultrathin-Body SOI MOSFETs," *IEEE Transactions on Device and Materials Reliability,* vol. 10, no. 1, pp. 130-141, 2010.

[30] D. J. Dumin, J. R. Maddux, R. S. Scott and R. Subramoniam, "A model relating wearout to breakdown in thin oxides," *IEEE Transactions on Electron Devices,* vol. 41, no. 9, pp. 1570-1580, 1994.

[31] R. Degraeve et al., "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Transactions on Electron Devices (TED),* vol. 45, no. 4, pp. 904-911, 1998.

[32] C. Duvvury and A. Amerasekera, "ESD: a pervasive reliability concern for IC technologies," *Proceedings of the IEEE,* vol. 81, no. 5, pp. 690-702, 1993.

[33] J. E. Vinson and J. J. Liou, "Electrostatic discharge in semiconductor devices: an overview," *Proceedings of the IEEE,* vol. 86, no. 2, p. 399–420, 1998.

[34] Ph. Galy et al., "Inventory of silicon signatures induced by CDM event on deep sub-micronic CMOS–BICMOS technologies," in *21st European Symposium on the Reliability of Electron Devices, Failure Physics and Analysis (ESREF)*, 2010.

[35] M. Scholz et al., "Miscorrelation between IEC61000-4-2 type of HMM tester and 50 Ω HMM tester," in *EOS/ESD Symposium Proceedings*, Tucson, AZ, 2012.

[36] K. Muhonen et al., "HMM round robin study: What to expect when testing components to the IEC 61000-4-2 waveform," in *EOS/ESD Symposium Proceedings*, Tucson, AZ, 2012.

[37] M. Hopkins, "Cable Discharge Events," in *Part I & II, ESD Open Forum 2009 (ESDA)*, 2009.

[38] Texas Instrument, "AN-1511 Cable Discharge Event," in *Application Report*, 2006 (revised 2013).

[39] W. Huang et al., "An Ethernet Cable Discharge Event (CDE) test and measurement system," in *IEEE International Symposium on Electromagnetic Compatibility (EMC)*,

Raleigh, NC, 2014.

[40] ANSI (American National Standard Institute) / ESDA (EOS/ESD Association) / JEDEC (Joint Electron Device Engineering Council), "JS-001-2017 Joint Standard for Electrostatic Discharge Sensitivity Testing - Human Body Model (HBM) - component Level," 2017.

[41] ESDA (EOS/ESD Association), "Electrostatic Discharge (ESD) Technology Roadmap – Revised May 2016," 2016.

[42] JEDEC (Joint Electron Device Engineering Council), "JESD22-A115C Electrostatic Discharge (ESD) Sensitivity Testing, Machine Model (MM)," 2010.

[43] ESDA (EOS/ESD Association) / JEDEC (Joint Electron Device Engineering Council), "JEP172A Discontinuing Use of the Machine Model for Device ESD Qualification," 2015.

[44] ANSI (American National Standard Institute) / ESDA (EOS/ESD Association) / JEDEC (Joint Electron Device Engineering Council), "JS-002-2018 Joint Standard for Electrostatic Discharge Sensitivity Testing - Charged Device Model (CDM) - Device Level," 2018.

[45] C. Goeau, C. Richier, P. Salome, J. Chante and H. Jaouen, "Impact of the CDM tester ground plane capacitance on the DUT stress level," in *EOS/ESD Symposium Proceedings*, Tucson, AZ, 2005.

[46] J. Di Sarro, B. Reynold and R. Gauthier, "Influence of package parasitic elements on CDM stress," in *EOS/ESD Symposium Proceedings*, Las Vegas, NV, 2013.

[47] T. Lim, "Dispositifs de protection contre les décharges électrostatiques pour les applications radio fréquences et milimétriques," PhD manuscript, Grenoble, 2013.

[48] T. Benoist et al., "Experimental investigation of ESD design window for fully depleted SOI N-MOSFETs," in *Microelectronic Engineering, Proceedings of the 17th Biennial International Insulating Films on Semiconductor Conference*, 2011.

[49] C. Duvvury, "ESD qualification changes for 45nm and beyond," in *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2008.

[50] S. Cao, T. W. Chen, S. G. Beebe and R. W. Dutton, «ESD design challenges and strategies in deeply-scaled integrated circuits,» chez *IEEE Custom Integrated Circuits Conference*, Rome, 2009.

[51] C. Lin, M. Ker, P. Chang and W. Wang, "Study on the ESD-induced gate-oxide breakdown and the protection solution in 28nm high-k metal-gate CMOS technology," in *IEEE Nanotechnology Materials and Devices Conference (NMDC)*, Anchorage, AK, 2015.

[52] T. J. Maloney, "Designing power supply clamps for electrostatic discharge protection of integrated circuits," *Microelectronics Reliability,* vol. 38, no. 11, p. 1691–1703, 1998.

[53] M. Stockinger et al., "Boosted and distributed rail clamp networks for ESD protection in advanced CMOS technologies," in *EOS/ESD Symposium Proceedings*, Las Vegas, NV, 2003.

[54] T. J. Maloney and N. Khurana, "Transmission Line Pulsing techniques for circuit modeling of ESD phenomena," in *EOS/ESD Symposium*, 1985.

[55] J. Manouvrier, P. Fonteneau, C. A. Legrand, P. Nouet and F. Azais, "Characterization of the transient behavior of gated/STI diodes and their associated BJT in the CDM time domain," in *EOS/ESD Symposium Proceedings*, Anaheim, CA, 2007.

[56] H. Wolf et al., «Transient analysis of ESD protection elements by time domain transmission using repetitive pulses,» chez *EOS/ESD Symposium Proceedings*, Anaheim, CA, 2006.

[57] J. C. Lee et al., "A method for determining a transmission line pulse shape that produces equivalent results to human body model testing methods," in *EOS/ESD Symposium Proceedings*, Anaheim, CA, USA, 2000.

[58] H. Hyatt, J. Harris, A. Alonzo and P. Bellew, "TLP measurements for verification of ESD protection device response," in *IEEE Transactions on Electronics Packaging Manufacturing*, 2001.

[59] H. Gieser and M. Haunschild, "Very-fast transmission line pulsing of integrated structures and the charged device model," in *EOS/ESD Symposium Proceedings*, Orlando, FL, USA, 1996.

[60] Synopsis Sentaurus TCAD, version N-2017.09.

[61] Ph. Galy, V. Berland, B. Foucher, and A. Guillaume, "Numerical evaluation between Transmission Line Pulse (TLP) and Average Current Slope (ACS) of a submicron gg-nMOS transistor under Electrostatic Discharge (ESD)," in *EOS/ESD/EMI Workshop*, 2002.

[62] T. Benoist, "Conception de protections contre les décharges électrostatiques sur technologies avancées silicium sur isolant," PhD manuscript, 2012.

[63] C. Fenouillet-Beranger et al., "Hybrid FDSOI/bulk High-k/metal gate platform for low power (LP) multimedia technology," in *IEEE International Electron Devices Meeting (IEDM)*, Baltimore, MD, 2009.

[64] D. Golanski et al., "First demonstration of a full 28nm high-k/metal gate circuit transfer from Bulk to UTBB FDSOI technology through hybrid integration," in *Symposium on*

VLSI Circuits, Kyoto, 2013.

[65] T. Benoist et al., "Improved ESD protection in advanced FDSOI by using hybrid SOI/bulk Co-integration," in *EOS/ESD Symposium Proceedings*, Reno, NV, 2010.

[66] A. Dray et al., "ESD design challenges in 28nm hybrid FDSOI/Bulk advanced CMOS process," in *EOS/ESD Symposium Proceedings*, Tucson, 2012.

[67] S. Voldman, S. Geissler, J. Nakos, J. Pekarik and R. Gauthier, "Semiconductor process and structural optimization of shallow trench isolation-defined and polysilicon-bound source/drain diodes for ESD networks," in *EOS/ESD Symposium Proceedings*, Reno, NV, USA, 1998.

[68] S. Voldman et al., "CMOS-on-SOI ESD protection networks," in *EOS/ESD Symposium Proceedings*, Orlando, FL, USA, 1996.

[69] S. Ramaswamy, A. Amerasekera and M. Chang, "A unified substrate current model for weak and strong impact ionization in sub-0.25 /spl mu/m NMOS devices," in *International Electron Devices Meeting (IEDM)*, Washington, DC, USA, 1997.

[70] M. P. J. Mergens et al., "Multi-finger turn-on circuits and design techniques for enhanced ESD performance and width-scaling," in *EOS/ESD Symposium Proceedings*, Portland, OR, 2001.

[71] S. Verdonckt-Vandebroek, S. S. Wong, J. C. S. Woo, and P. K. Ko, "High-gain lateral bipolar action in a MOSFET structure," *IEEE Transactions on Electron Devices,* vol. 38, no. 11, pp. 2487-2496, November 1991.

[72] J. Olsson, B. Edholm, A. Soderbarg, and K. Bohlin, "High current gain hybrid lateral bipolar operation of DMOS transistors," *IEEE Transactions on Electron Devices (TED),* vol. 42, no. 9, pp. 1628-1635, 1995.

[73] Ph. Galy and V. Berland, "The ideal NPN vertical BIMOS transistor analytical model simulation and experimental results of the collector current," *International Journal of Electronics,* vol. 81, no. 5, pp. 501-516, 1996.

[74] Ph. Galy et al., "BIMOS transistor and its applications in ESD protection in advanced CMOS technology," in *IEEE International Conference on IC Design and Technology (ICICDT)*, Austin, TX, 2012.

[75] Ph. Galy et al, "Ultracompact ESD Protection With BIMOS-Merged Dual Back-to-Back SCR in Hybrid Bulk 28-nm FD-SOI Advanced CMOS Technology," *IEEE Transactions on Electron Devices (TED),* vol. 64, no. 10, pp. 3991-3997, 2017.

[76] J. P. Colinge, "An SOI Voltage-Controlled Bipolar-MOS Device," *IEEE Transactions on Electron Devices (TED),* vol. 34, no. 4, pp. 845-849, April 1987.

[77] S. Athanasiou, S. Cristoloveanu, and Ph. Galy, "Key parameters of BiMOS ESD protection device for UTBB FD-SOI advanced technology," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2015.

[78] S. Athanasiou, C. A. Legrand, S. Cristoloveanu, and Ph. Galy, "Novel Ultrathin FD-SOI BiMOS device with reconfigurable operation," *IEEE Transactions on Electron Devices (TED),* vol. 64, no. 3, pp. 916-922, March 2017.

[79] J. Bourgeat, "Etude du thyristor en technologie CMOS avancée pour implémentation dans des stratégies locale et globale de protection contre les décharges électrostatiques," PhD manuscript, Toulouse, 2011.

[80] B. Caillard, F. Azais, S. Dournelle, P. Salome and P. Nouet, "STMSCR: A new multi-finger SCR-based protection structure against ESD," in *EOS/ESD Symposium Proceedings*, Las Vegas, NV, 2003.

[81] J. D. Sarro, V. Vashchenko, E. Rosenbaum and P. Hopper, "A dual-base triggered SCR with very low leakage current and adjustable trigger voltage," in *EOS/ESD Symposium Proceedings*, Tucson, AZ, 2008.

[82] J. Bourgeat, C. Entringer, P. Galy, F. Jezequel and M. Bafleur, "TCAD study of the impact of trigger element and topology on silicon controlled rectifier turn-on behavior," in *EOS/ESD Symposium Proceedings*, Reno, NV, 2010.

[83] A. Chatterjee and T. Polgreen, "A low-voltage triggering SCR for on-chip ESD protection at output and input pads," in *Symposium on VLSI Technology*, Honolulu, Hawaii, USA, 1990.

[84] M. P. J. Mergens, O. Marichal, S. Thijs, B. Van Camp and C. C. Russ, "Advanced SCR ESD protection circuits for CMOS/SOI nanotechnologies," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, San Jose, CA, 2005.

[85] J. Li, J. Di Sarro and R. Gauthier, "Design and optimization of SCR devices for on-chip ESD protection in advanced SOI CMOS technologies," in *EOS/ESD Symposium Proceedings*, Tucson, AZ, 2012.

[86] J. Wan, S. Cristoloveanu, C. Le Royer, "Cellule mémoire dynamique munie d'un transistor à effet de champ à pente sous le seuil vertical / Z2FET Field-effect transistor with a vertical subthreshold slope and with no impact ionization," patent FR1103232, priority date 2011-10-21.

[87] J. Wan, C. Le Royer, A. Zaslavsky and S. Cristoloveanu, "A Compact Capacitor-Less High-Speed DRAM Using Field Effect-Controlled Charge Regeneration," *IEEE Electron Device Letters (EDL),* vol. 33, no. 2, pp. 179-181, 2012.

[88] J. Wan, "Dispositifs innovants à pente sous le seuil abrupte: du TFET au Z2-FET /

Innovative sharp switching devices: from TFET to Z2-FET," PhD manuscript, Grenoble, 2012.

[89] H. El Dirani et al., "A sharp-switching gateless device (Z3-FET) in advanced FDSOI technology," in *Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, Vienna, 2016.

[90] Y. Solaro, "Conception, fabrication et caractérisation de dispositifs innovants de protection contre les décharges électrostatiques en technologie FDSOI," PhD manuscript, Université de Grenoble, 2014.

[91] H. El Dirani, Y. Solaro, P. Fonteneau, P. Ferrari and S. Cristoloveanu, "Sharp-switching Z2-FET device in 14 nm FDSOI technology," in *European Solid State Device Research Conference (ESSDERC)*, Graz, 2015.

[92] S. Cristoloveanu et al., "A review of the Z2-FET 1T-DRAM memory. Operation mechanisms and key parameters.," *Solid State Electronics (SSE),* vol. 143, pp. 10-19, 2018.

[93] S. Cristoloveanu, S. Athanasiou, M. Bawedin and Ph. Galy, "Evidence of supercoupling effect in ultrathin silicon layers using a four-gate MOSFET," *IEEE Electron Device Letters (EDL),* vol. 38, no. 2, pp. 157-159, 2017.

[94] S. Athanasiou, "Conception, fabrication et caractérisation de nouveaux dispositifs FD-SOI avancés pour protection contre les décharges électrostatiques," PhD manuscript, Grenoble, 2017.

[95] Ph. Galy, C. Entringer, and J. Bourgeat, "Structure for protecting an integrated circuit against electrostatic discharges," patent 2010/0271741 US, 2009.

[96] J. Bourgeat, Ph. Galy, and B. Jacquier, "Beta-Matrix ESD network: Throughout end of placement rules?," in *IEEE International Conference on IC Design & Technology (ICICDT)*, Kaohsiung, 2011.

[97] Ph. Galy et al., "Beta-matrix concept for ESD power devices, demonstrators in C45nm & C32nm CMOS technology," in *EOS/ESD Symposium Proceedings*, Anaheim, CA, 2011.

[98] P. Batude et al., "3DVLSI with CoolCube process: An alternative path to scaling," in *Symposium on VLSI Technology (VLSI Technology)*, Kyoto, 2015.

[99] L. Brunet et al., "First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers," in *IEEE Symposium on VLSI Technology*, Honolulu, HI, 2016.

[100] B. Sklenard et al., "Low temperature junction formation by solid phase epitaxy on thin film devices: Atomistic modeling and experimental achievements," in *International*

Workshop on Junction Technology (IWJT), Shanghai, 2014.

[101] C. Fenouillet-Beranger et al., "Recent advances in low temperature process in view of 3D VLSI integration," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Burlingame, CA, 2016.

[102] L. Pasini et al., "High performance low temperature activated devices and optimization guidelines for 3D VLSI integration of FD, TriGate, FinFET on insulator," in *Symposium on VLSI Technology*, Kyoto, 2015.

[103] T. Bedecarrats, "Etude et intégration d'un circuit analogique basse consommation et à faible surface d'empreinte de neurone impulsionnel basé sur l'utilisation du BIMOS en technologie 28 nm FD-SOI," PhD manuscript, Grenoble, 2019.

[104] T. Benoist, Ph. Galy, J. Bourgeat, F. Jezequel, and N. Guitard, "Triggerable bidirectional semiconductor device / Dispositif semiconducteur bidirectionnel déclenchable utilisable sur silicium sur isolant," patent US8937334B2, priority date 2011-06-15.

[105] S. Athanasiou, C. A. Legrand, S. Cristoloveanu, and Ph. Galy, "GDNMOS: A new high voltage device for ESD protection in 28nm UTBB FD-SOI technology," in *EUROSOI-ULIS: Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon*, Vienna, 2016.

[106] S. Athanasiou, C. A. Legrand, S. Cristoloveanu, and Ph. Galy, "Reconfigurable ultra-thin film GDNMOS device for ESD protection in 28 nm FD-SOI technology," *Solid State Electronics,* vol. 128, pp. 172-179, February 2017.

[107] J. Coignus, "Etude de la conduction électrique dans les diélectriques à forte permittivité utilisés en microélectronique," PhD manuscript, Grenoble, 2010.

[108] H. S. P. Wong et al., "Metal-Oxide RRAM," in *Proceedings of the IEEE*, 2012.

[109] J. B. Baliga, Fundamentals of power semiconductor devices, Springer Publishing Company, Incorporated, 2008.

[110] J. P. Colinge and F. Van De Wiele, Physique des dispositifs semi-conducteurs, Département de Boeck Université, Paris, Bruxelles: De Boeck-Wesmael s. a., 1996.

[111] S. Athanasiou and Ph. Galy, "Structure de transistor / Transistor structure," French patent application no. FR1657587 filed in 2016-08-05, US publication number 20180012965 filed in 2017-02-08..

[112] George E. Smith, III, "T-Gate transistor with improved SOI body contact structure," patent US6316808B1, filed in 1998.

[113] Ph. Galy and S. Athanasiou, "Prise de contact substrat pour un transistor MOS dans un substrat SOI, en particulier FDSOI / Substrate contact land for a MOS transistor in a SOI

substrate, in particular an FDSOI substrate," French patent application no. FR1556515 filed in 2015-07-09, US patent number 10128242 filed in 2017-11-06.

[114] T. Bedecarrats, Ph. Galy, C. Fenouillet-Béranger and S. Cristoloveanu, "Investigation on built-in BJT in FD-SOI BIMOS," in *Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, Granada, 2018.

[115] Ph. Galy and J. Jimenez, "Method for producing an electronic device by assembling semi-conducting blocks and corresponding device," patent grant number 9230950, publication number 20130264677, 2013.

[116] J. Bourgeat, C. Entringer, Ph. Galy, and J. Jimenez, "Electronic device, in particular for protection against electrostatic discharges, and method for protecting a component against electrostatic discharges," patent US9019666B2, 2010.

[117] G. Troussier, N. Guitard, A. Dray, Ph. Galy, "Circuit électronique incluant un transistor MOS et des agencements pour resister aux décharges électrostatiques," patent application no. FR1250062, filed in January 2012.

[118] P. Fonteneau, Y. Solaro, D. Marin-Cudraz, C. Legrand and C. Fenouillet-Beranger, "Innovative high-density ESD protection device in state of the art UTBB FDSOI technologies," in *EOS/ESD Symposium Proceedings*, Reno, NV, 2015.

[119] O. Weber, E. Richard, and P. Boivin, "A method of making MOS and bipolar transistors," patent application no. FR3049111, filed in March 2016.

[120] R. Berthelon et al., "A novel dual isolation scheme for stress and back-bias maximum efficiency in FDSOI Technology," in *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2016.

[121] K. Chatty et al., "Study of factors limiting ESD diode performance in 90nm CMOS technologies and beyond," in *IEEE International Reliability Physics Symposium (IRPS)*, San Jose, CA, USA, 2005.

[122] H. Mathieu and H. Fanet, Physique des semiconducteurs et des composants électroniques, 6th ed., Paris: Sciences up, Dunod, 2009.

[123] S. M. Sze and K. K. Ng, Physics of semiconductor devices, third ed., John Wiley & Sons, Inc., 2007.

Abstract

The thesis objective was to design protection devices against electrostatic discharges (ESD) in the silicon thin-film using the 28 nm node ultra-thin Body and Buried Oxide (UTBB) Fully Depleted Silicon-On-Insulator (FD-SOI) technology with high-k dielectrics and metal gate. Existing devices were studied and new technological solutions were proposed to improve them. Besides, new devices were elaborated. 3D TCAD simulation was used for understanding their electrical behavior. Silicon characterization were performed to verify the response of devices to typical ESD tests. This work paves the way of innovative ESD protection devices built in the thin film with a special care given to 3D concerns, such as (i) the possibility of implementing the protection in a 3D monolithic integrated circuit, (ii) building a matrix as a protection device, and (iii) merging different devices such as benefiting from a 3D conduction of current.

Keywords: ESD, electrostatic discharges, thin-film, FD-SOI, 3D, TCAD, BIMOS, GDNMOS, GDBIMOS, BIMOS merged SCR.

Résumé

L'objectif de la thèse était de concevoir des composants de protection contre les décharges électrostatiques (ESD) sur film mince de silicium en technologie 28nm FD-SOI de chez STMicroelectronics (technologie silicium sur isolant « Silicon-On-Insulator » (SOI) entièrement déplété « Fully Depleted » (FD)). Cette technologie est caractérisée par un film de silicium, un oxyde enterré ultra minces (UTBB), et par une grille métallique avec oxyde à haute permittivité (high-k). En prenant en compte ces caractéristiques, des composants existants ont été étudiés et de nouvelles solutions technologiques ont été proposées pour les améliorer. De plus, de nouveaux composants ont été élaborés. Ils ont été simulés en 3D avec le logiciel TCAD afin de comprendre leur comportement électrique. Des plaques de silicium ont été mesurées afin de vérifier la réponse des composants lors de tests typiques pour les ESD. Ce travail ouvre la voie pour des composants de protection contre les décharges électrostatiques conçus dans le film mince avec une attention spéciale pour l'aspect 3D, tel que (i) la possibilité d'implémenter la protection dans un circuit intégré 3D monolithique, (ii) la conception de matrice en tant que composant de protection, et (iii) la fusion de différents composants pour bénéficier d'une conduction de courant en 3D.

Mots clés : Décharges électrostatiques, film mince, FD-SOI, 3D, TCAD, BIMOS, GDNMOS, GDBIMOS, BIMOS merged SCR.