UNCALIBRATED TCAD METHODOLOGY FOR ANALYSIS OF ESD PROTECTION DEVICES

BY

ZAICHEN CHEN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Professor Elyse Rosenbaum

ABSTRACT

In this work, an uncalibrated TCAD methodology for simulation of electrostatic discharge (ESD) devices is presented. The methodology addresses TCAD setup issues including device construction, boundary conditions, and choosing a physical model and parameters. A major trade-off between computation complexity and accuracy, 2D vs. 3D simulations, is examined in detail. TCAD simulation results for the GGNMOS in 32 nm CMOS technology is compared with published measurement results for methodology validation. The established TCAD methodology is then applied to ESD protection silicon controlled rectifier (SCR) devices to identify physical causes for high overshoot of a certain SCR layout, and to verify proposed improvements. The performance of the SCR with the improved layout structure is characterized in silicon to prove its consistency with TCAD prediction.

ACKNOWLEDGMENT

I would like to thank my research advisor, Professor Elyse Rosenbaum, for her hearty guidance and support, which helped me become a better researcher and engineer in general. I would also like to thank all members of my research group, especially Dr. Kuo-Hsuan Meng and Dr. Robert Mertens, for their friendly help in my research.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1
1.1. Background and Motivation 1
1.2. Literature Review 2
CHAPTER 2. UNCALIBRATED TCAD METHODOLOGY FOR ESD PROTECTION NMOSFETS 4
2.1. Simulation Setup 4
2.1.1. Device 4
2.1.2. Tester 6
2.1.3. Thermal Boundary Conditions 6
2.1.4. Physical Models 8
2.2. Comparison between 2D and 3D TCAD 10
2.3. Conclusion 16
CHAPTER 3. TCAD ASSISTED LAYOUT IMPROVEMENT OF GGNMOS TRIGGERED SILICON CONTROLLED RECTIFIER (GGSCR) DEVICE 17
3.1. Introduction 17
3.2. Current Uniformity and GGSCR Layout 20
3.3. Test Structures 22
3.4. Measurement Results 23
3.4.1. Stripe vs. Point Trigger Layout 23
3.4.2. Effect of Layout Spacings 25
3.5. Discussion 28
3.6. Comparison of P-Well and N-Well Triggered GGSCR 32
3.7. Trigger GGNMOS Design Considerations 34
3.8. Conclusions 38
CHAPTER 4. CONCLUSIONS 39
REFERENCES 40

CHAPTER 1. INTRODUCTION

1.1. Background and Motivation

Electrostatic discharge (ESD) phenomena have been identified as one of the primary factors for yield degradation and field return of semiconductor integrated circuit (IC) products. Consequently, on-chip ESD protection circuitry has always been an essential component in modern ICs. Since transistors become more susceptible to ESD damage with each technology node advance, ESD design has been an active research area in both academia and industry for the past few decades. The recent introduction of silicon-on-insulator (SOI) and FinFET technologies brought even more challenges to ESD design.

One major challenge in ESD design is the incompatibility between ESD devices and general-purpose device models. The accuracy of circuit simulation is based on accurate modeling of semiconductor devices, e.g. diodes, MOSFETs, bipolar junction transistors (BJTs). However, an ESD device often works in very high current regime in which an established device model is unable to describe. Therefore, ESD designers have to seek different design approaches. One method is to develop specific device models for ESD devices. However, most semiconductor foundries do not provide ESD device models to fabless IC designers. In practice, IC designers who opt for custom ESD design usually need to run multiple iterations of testing in silicon, which is very costly.

Given the rising cost of test wafers, TCAD becomes an increasingly attractive option for evaluating the inherent ESD reliability for new technologies and optimizing ESD device design. TCAD simulation takes the structure and boundary condition of a physical device as the input, solves the Poisson's and continuity equations inside the device and gives terminal

voltage/current relationship as output. In order to perform quantitatively accurate TCAD simulations, an exact description of the device structure, e.g. dimensions and doping profiles, as well as an accurate modeling of the physical properties of the materials, e.g. mobility, generation-recombination rate, impact ionization rate, are essential. The process and physical parameters can only be obtained through calibration against measurement data from physical devices. Since IC designers usually do not have access to process information, a quantitatively accurate TCAD simulation is impractical. However, even without accurate device structure and physical models, TCAD is capable of performing a qualitative evaluation of ESD devices and those results may be useful for IC designers to improve their ESD design, and reduce the number of test wafers required in the design process. Since no parameter calibration is involved in this type of TCAD, it is referred to as uncalibrated TCAD.

In this work, a methodology of using uncalibrated TCAD simulation to evaluate ESD devices is developed and the capabilities and limitations of the methodology are evaluated. An example of utilizing TCAD to improve ESD device design is demonstrated.

1.2. Literature Review

The pioneering work on pre-silicon ESD TCAD was published in 2000 [1]. In [1] the authors concluded that 2D simulation can generate reasonably accurate predictions for the performance of an ESD protection NMOS. However, in [1] TCAD was evaluated based on 0.18 μm technology, in which the local E-field intensity and current density in a device are not so high that the uncalibrated physical models will give massively inaccurate results. Whether or not this conclusion still holds for 32 nm technology and beyond must be reevaluated. Since [1] there has been no complete work on ESD TCAD methodology, although some new ideas have been introduced. In [2] an innovative method of using 2D TCAD to determine the failure current of

ESD devices is proposed. In [3] a novel algorithm to simulate devices with snapback behavior is introduced. Those studies give insights on potential improvements that could be made to ESD TCAD methodology, although their practicalities are very limited. This work consists of a complete TCAD methodology suitable for the simulation of 32 nm and beyond technologies.

CHAPTER 2. UNCALIBRATED TCAD METHODOLOGY FOR ESD PROTECTION NMOSFETS

The TCAD methodology was developed under the context of simulation of ESD protection grounded-gate NMOSFETs (GGNMOS) in 32 nm CMOS technology. MOSFETs are chosen since it is expected that the simulation methodology developed based on them can be directly adapted to other ESD protection devices: diodes, BJTs and SCRs. The 32 nm CMOS is the target technology since it is the most advanced technology node for which a significant amount of data is available in the open literature.

2.1. Simulation Setup

The Sentaurus TCAD software is used in this work. The 2D simulations are run on a desktop computer with a 2.67 GHz Intel i5-750 processor and 16 GB of memory. The 3D simulations are run on a cluster in which each node has two 2.66 GHz Intel Xeon X5650 processors and 24 GB of memory. In all cases, the numerical solver runs parallel on 4 CPU cores.

The device is simulated in a transmission line pulse (TLP) simulation setup, since TLP is the most common ESD device characterization method. The finite element mesh of the device is inserted into a circuit model of the TLP tester, and transient simulation is run for the combined system. Also, heating is considered for the device since failures of ESD devices are usually caused by overheating.

2.1.1. Device

As uncalibrated simulation, all knowledge about the simulation structure needs to come from the open literature. The doping profile of the NMOS transistor is tuned so that the simulated V_T and I_{DSAT} are representative of those for 32 nm devices [4], [5], [6] (see Table 2.1). A sample device

cross-section is shown in Figure 2.1. Since details of the high-K/metal gate stack are not provided in publications (nor in a PDK), a metal-SiO_2 gate stack used in simulation. The work function of the gate metal and thickness of the gate oxide (T_{ox}) are adjusted so that simulated V_T and effective oxide thickness (EOT) matches the published data. In particular, the simulated EOT is obtained through C-V simulation: $\mathrm{EOT} = \frac{\epsilon_{SiO2}}{C_G(V_G=V_{DD})}$. Notice that T_{ox} is not directly set equal to EOT, one major reason being that in strong inversion, quantum confinement effect pushes the peak concentration of the inversion charge from the silicon-to-dielectric boundary, so EOT becomes larger than T_{ox}. Consequently, to determine T_{ox} from C-V simulation, corrections accounting for the quantum effect must be applied.

Simulation and measurement results are compared in Table 2.1. The simulated I_{DSAT} is a bit smaller than found in measurements, but this is expected since strained silicon effects were not included. DIBL and I_{off} are also smaller in simulation; this is the result of using a SiO_2 gate dielectric [7] and will not affect ESD simulations. The overall agreement between measurement and simulation is very good, given the lack of process information.

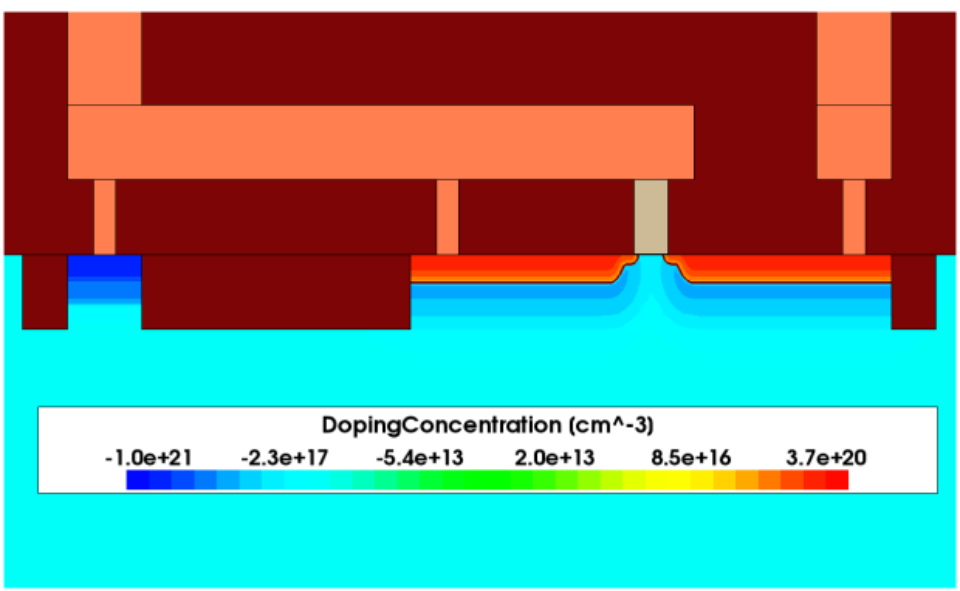


Figure 2.1. Cross-section of device, L_{gate} = 45 nm.

Table 2.1. Comparison of simulated NMOS metrics with those for devices from three different foundries. In the simulations, the quantum correction is enabled and L_{gate} = 30 nm.

	[4]	[5]	[6]	Simulation
V_{DD} (V)	1.0	1.0	0.9	1.0
EOT (nm)	0.9	1.0	1.2	1.16
$V_{T,lin}$ (V)	0.34	0.23	0.52	0.33
$V_{T,sat}$ (V)	0.14	0.09	0.3	0.25
I_{DSAT} (μA/μm)	1550	1340	1250	900
I_{off} (nA/μm)	100	100	100	24

2.1.2. Tester

The TLP tester model consists of a step voltage source with output resistance R_S in series with a rise-time filter [8]. Increasing R_S leads to better resolution of the TLP I-V curve near the holding point, due to a reduced slope of the tester load line. Also, if the ratio between R_S and the device under test (DUT) on-resistance (R_{on}) is too small, the negative slope in simulated current waveform (due to device self-heating) will become arbitrarily large, as is shown in Figure 2.2. Simulation results presented in this work are for single-finger devices with a maximum width of 20 μm; these relatively small devices have an input resistance that is much larger than a typical protection device. Thus, to achieve a reasonable match to measurement data, the tester output resistance is set to 500 Ω in the simulations.

2.1.3. Thermal Boundary Conditions

To simulate a MOSFET under normal operating conditions, the device structure may consist of just the active region, gate and contacts. If self-heating is important, resistive thermal boundary conditions can be applied at the silicon and electrode boundaries. However, this approach should not be used for ESD simulations. Since the device does not reach a thermal steady-state condition during the event, both the thermal capacitance and thermal resistance of the materials must be simulated. To accurately simulate the temperature rise, a sufficiently large “thermal

buffer" is placed around the active device as shown in Figure 2.3. The mesh in the buffer regions can be very sparse to minimize the impact on the simulation time. To determine the proper size of the buffers, a device is simulated with perfect heat sinking (T = 300 K) and insulating ($dT/d\vec{n} = 0$) boundary conditions. If the two results are nearly identical, then the buffer is sufficiently large. Based on the results shown in Figure 2.3, it is concluded that a buffer thickness (W_{buf}) of 5 µm is sufficient for 100 ns TLP simulations.

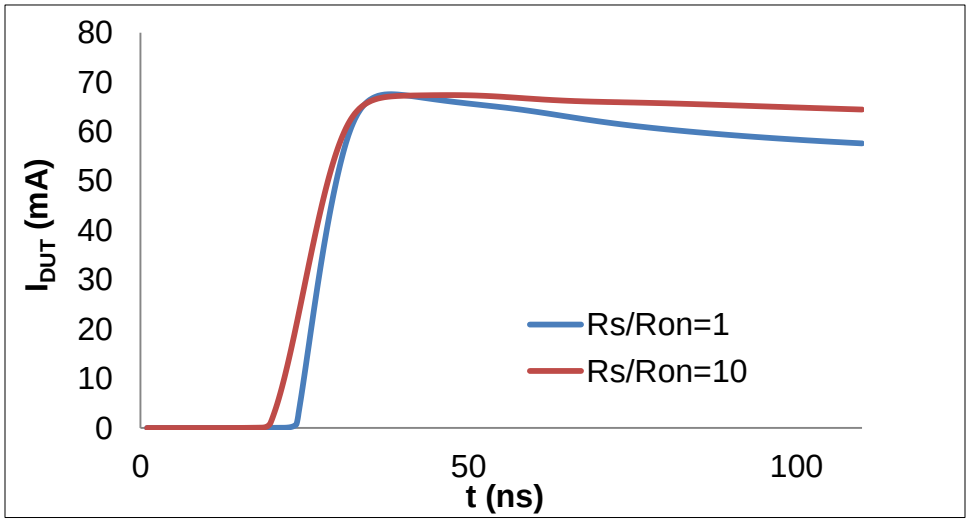


Figure 2.2. Simulated current for different source impedances. 2D TCAD. DUT: GGNMOS, W = 10 µm.

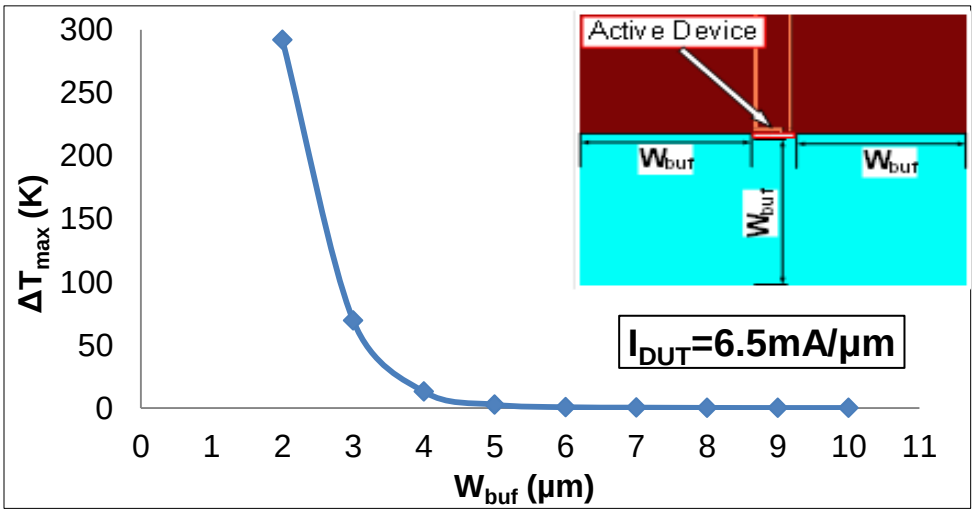
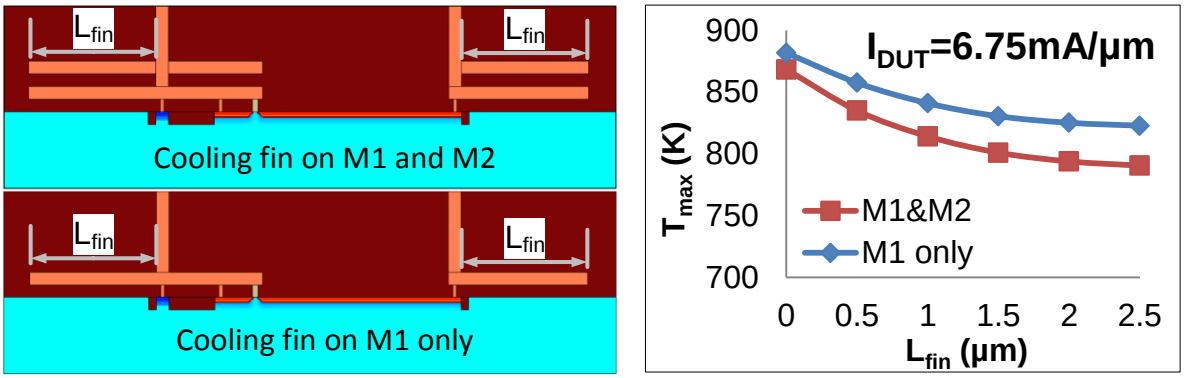


Figure 2.3. The figure inset shows a device with a thermal buffer. $\Delta T_{max} = T_{max}$("insulating boundary") − T_{max}("heat sink"), where max is the maximum temperature in the device during the TLP pulse. ΔT_{max} decreases as the thermal buffer size, W_{buf}, increases.

Despite the high thermal conductivity of metal, an ESD event is too brief for heat generated in the device to propagate to peripheral heat sinks, e.g. pads and bonding wires. However, the lower-level interconnects work as cooling fins which diffuse heat into the inter-layer dielectrics. The effect of the lower-level routing was simulated using the structures shown in Figure 2.4(a), and the results are plotted in Figure 2.4(b). For ESD TCAD simulations, it is advised to include at least the M1 layer in the device structure. Also, a sufficiently thick passivation oxide must be placed on top of the device to ensure heat diffusion in this layer is captured correctly.



(a) (b)

Figure 2.4. (a) Device cross-section including M1 and M2 interconnects. (b) Effect of interconnect on simulated T_{max}.

2.1.4. Physical Models

For ESD simulation of MOS devices, the following models should be activated: bandgap narrowing; doping dependent Lombardi mobility model with velocity saturation; SRH recombination; impact ionization; band-to-band tunneling (B2B); and Auger recombination.

Although the quantum correction model was enabled when simulating the NMOSFET under normal operating conditions, it causes convergence problems in ESD simulations of the GGNMOS. As a solution it is turned off, and the cost in accuracy should be negligible since no inversion layer is formed in a GGNMOS during ESD operation. The strained silicon model is

disabled since it is futile to attempt a highly accurate modeling of the channel mobility in a pre-silicon simulation. The gate tunneling model is also disabled since high-K dielectric makes the gate current small, and no information of a realistic high-K gate stack could be obtained.

The default model parameters in the TCAD software do not yield qualitatively correct ESD simulation results, thus modifications are necessary. For pre-silicon TCAD, the best model parameters will be those calibrated for a recent technology generation. These may be obtained using the procedures described below. Alternatively, one may use the parameters given in Table 2.2.

In [1], it was reported that activating the B2B model will lead to convergence problems in ESD simulations and this was indeed found to be the case with some of the provided models. Neglecting B2B is not an option today as it determines the value of V_{t1} in modern GGNMOS [9]. Fortunately, the "simple" B2B model [10] embedded in the TCAD software gives the desired numerical stability. The model equation for the B2B generation rate is:

$$\mathrm{G_{B2B}} = A_{B2B} E_{\parallel}^{2} \exp\left(-\frac{B_{B2B}}{E_{\parallel}}\right) \tag{2.1}$$

Using the default parameters, the simulated V_{t1} is too small. B2B parameters for ESD NMOS simulations were obtained by measuring the B2B current in some available 65 nm GGNMOS test structures and then adjusting the parameters until TCAD replicated the results. Using the extracted parameters, the simulated V_{t1} for the 32 nm GGNMOS is close to the value reported in [11].

The van Overstraeten impact ionization model [12], in Equation (2.2), is used to simulate avalanche generation:

$$\alpha_{II} = \gamma A_{II} \exp\left(-\frac{\gamma B_{II}}{E_{\parallel}}\right), \text{ where } \gamma = \frac{\tanh\left(\frac{h\omega_{op}}{kT}\right)}{\tanh\left(\frac{h\omega_{op}}{kT_0}\right)} \tag{2.2}$$

and T_0 = 300 K. The default parameter set was extracted for electric fields up to 0.6 MV/cm, whereas simulation shows that the maximum electrical field in a 32 nm GGNMOS exceeds 1 MV/cm. Furthermore, the parameter for temperature dependent impact ionization, $h\omega_{op}$, needs to be modified to generate realistic results under ESD conditions [1]. The model parameters were adjusted such that the simulated V_{hold} is close to the value in [11].

Thermal generation/recombination via mid-band traps in silicon is governed by the SRH rate formula:

$$R_{net} = \frac{np - n_{i,eff}^2}{\tau_p(n + n_{i,eff}) + \tau_n(p + n_{i,eff})}; \; \tau_{n,p} = \frac{\tau_{n,p}^{max}}{1 + \frac{N_A + N_D}{N_{ref}}} \tag{2.3}$$

In 3D simulation using the default parameters, the thermal generation rate is too small to support the formation of a current filament. In previous works on *calibrated* ESD TCAD, the lifetime parameters used were more than an order of magnitude smaller than the defaults [3], [13]. This is attributed to a higher density of defects in the heavily implanted drain region than in bulk silicon, exacerbated by the reduced thermal budget for annealing [13].

The modified parameter set is listed in Table 2.2. Table 2.3 benchmarks the simulation results against published data. The simulated I_{t2} in Table 2.3 is from 2D simulation.

2.2. Comparison between 2D and 3D TCAD

Strictly speaking, 2D TCAD simulation should be limited to devices in which current density and temperature are uniform across the device width. Since 3D TCAD requires more computational resources, it is very important to determine the applicability of 2D TCAD. Even in

cases where clear 3D effects exist, it may be possible to establish a correlation between 2D and 3D simulation results. Table 2.4 compares the runtime for 2D and 3D TCAD simulation.

Table 2.2. Parameters for ESD NMOS simulations.

Parameter	Default value	Modified value	Unit
A_{B2B}	3.50×10^{21}	3.47×10^{21}	$cm^{-1}s^{-1}V^{-2}$
B_{B2B}	2.25×10^{7}	3.28×10^{7}	Vcm^{-1}
$A_{II,n}$	7.03×10^{5}	5.5×10^{6}	cm^{-1}
$B_{II,n}$	1.231×10^{6}	6.0×10^{6}	Vcm^{-1}
$A_{II,p}$	6.71×10^{5}	4.4×10^{6}	cm^{-1}
$B_{II,p}$	1.693×10^{6}	7.2×10^{6}	Vcm^{-1}
$h\omega_{op}$	0.063	0.126	eV
τ_n^{max}	1×10^{-5}	1×10^{-7}	s
τ_p^{max}	3×10^{-6}	1×10^{-7}	s

Table 2.3. Comparison between measurement data and 2D simulation results (2D I_{t2} will be conservative).

	[11]	[11]	[14]	Simulation
CMOS node	32 nm	32 nm	28 nm	(32 nm)
L_{gate} (nm)	40	60	42	45
V_{t1} (V)	3.3	4.0	3.3	3.5
V_{hold} (V)	3.3	3.8	3.1	3.5
I_{t2} (mA/μm)	8-9	8-9	-	7.0

Table 2.4. Time consumption for 2D and 3D simulations *run on the same machine*.

	2D	3D (W = 1 μm)	3D (W = 10 μm)
Mesh count	2500	14000	51000
Time *	5-7 min	20-25 min	150-200 min
* time for simulating one 100 ns TLP pulse			

It is well known that GGNMOS often suffers thermal failure resulting from current filamentation, a 3D effect [15]. A comparison is made between the I_{t2} values obtained from 2D and 3D TCAD, as a function of the drain silicide-blocking length (L_{dblk}). Obtaining I_{t2} from 3D simulation is straightforward: $T_{max} > T_{melt,Si}$ is used as the failure criterion; this condition is met very soon (e.g., 1ns) after filamentation. In 2D simulation, change of dV/dt from positive to negative in the

quasi-steady-state condition is used as the failure criterion, as is suggested in [13]. It is observed that at current levels high enough for 2D failure to be identified, 2D and 3D simulation results significantly differ. The question to be investigated is whether this current level is close to I_{t2}.

Figure 2.5 shows the simulated I_{t2} obtained from 2D and 3D simulations. I_{t2} is the highest I_{DUT} for which the device does not reach the failure criterion. The 2D I_{t2} is less than the 3D I_{t2} in all cases, but both values have the same qualitative dependence on L_{dblk}. The 2D I_{t2} is expected to be smaller because only 3D simulation captures heat diffusion in the transverse direction, i.e., along the device width. This effect is more pronounced for devices with small W, as clearly seen in the 3D TCAD results (Figure 2.5). To investigate whether this is the only reason for the discrepancy between 2D and 3D simulation results, we eliminate transverse heat diffusion from the 3D structure by replacing the thermal buffers in the transverse direction with thermally insulating walls. Figure 2.6 shows the results of this exercise. The 3D I_{t2} is almost identical to the 2D value for small L_{dblk}, but the discrepancy grows with L_{dblk}. This can only be explained by the current ballasting effect, i.e. a localized increase in the current density increases the local temperature thereby providing negative feedback, which is inherently a 3D effect.

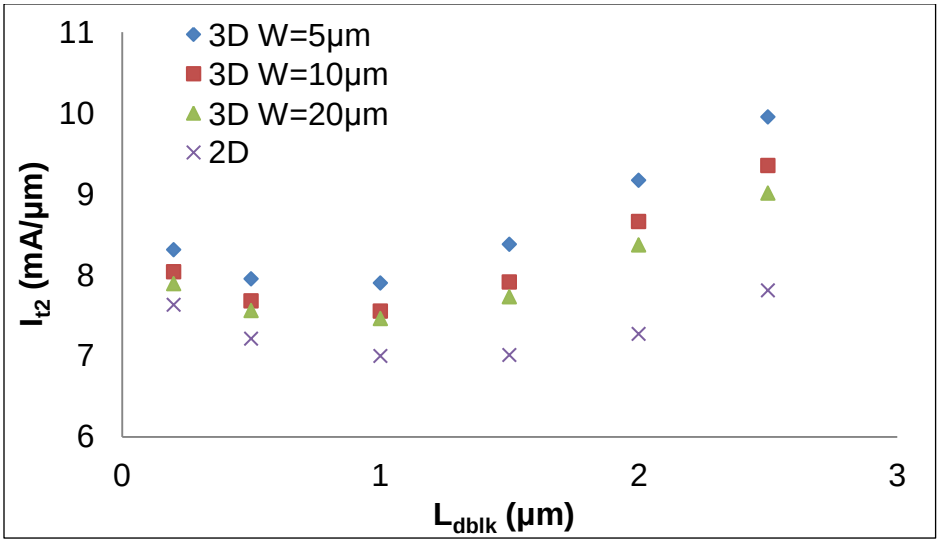


Figure 2.5. Simulated I_{t2} vs. L_{dblk} for 2D and 3D GGNMOS.

In Figure 2.5 and Figure 2.6, for L_{dblk} < 1µm, I_{t2} appears to be negatively correlated with L_{dblk}. Although surprising, these results do not appear to contradict any measurement data. I_{t2} is often observed to roll-off for small L_{dblk} (e.g. in [11]), but that well-known result is for multi-finger devices and results from non-uniform triggering among the fingers. A reduced I_{t2} for small L_{dblk} was reported in one study of single finger devices [1], but the test device was very wide (W > 50 µm), and current flow across such a wide finger is intrinsically non-uniform [16]. Moderately high I_{t2} are observed in measurements of fully salicided devices with smaller widths [17]. Accepting the results of Figure 2.5 as physically correct, further scrutiny of the simulation results is performed to illuminate the underlying physics.

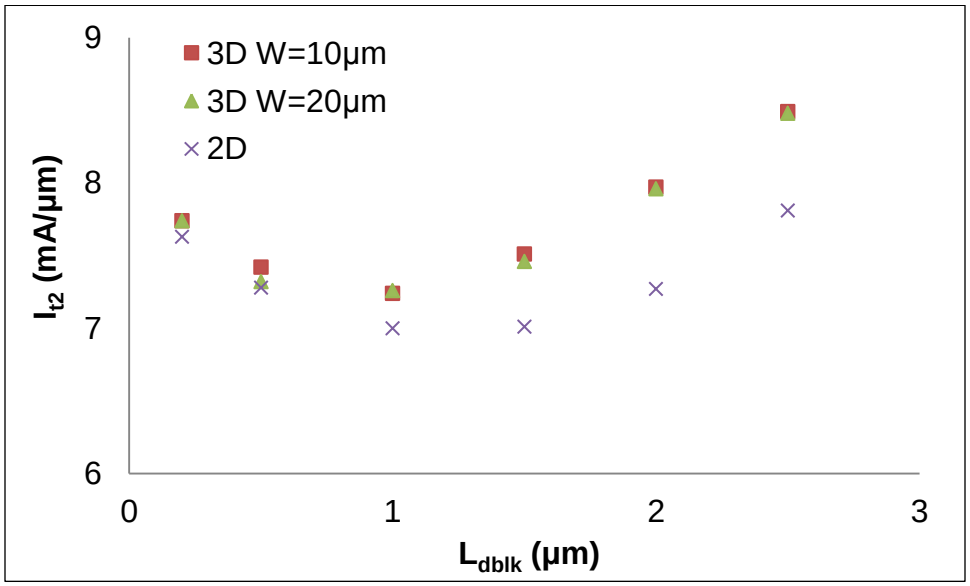


Figure 2.6. Simulated I_{t2}, the 3D structures have thermally insulating side-walls.

Initially, I_{t2} is a decreasing function of L_{dblk} because the distance from the hot spot to the heat sink is increasing; see Figure 2.7. The drain silicide-blocked region cannot act as a heat sink since there is significant Joule heating in this region. For a device with longer L_{dblk}, the distance

from the hot spot to the drain side bulk silicon is longer, thus the heat flux in the drain direction is smaller.

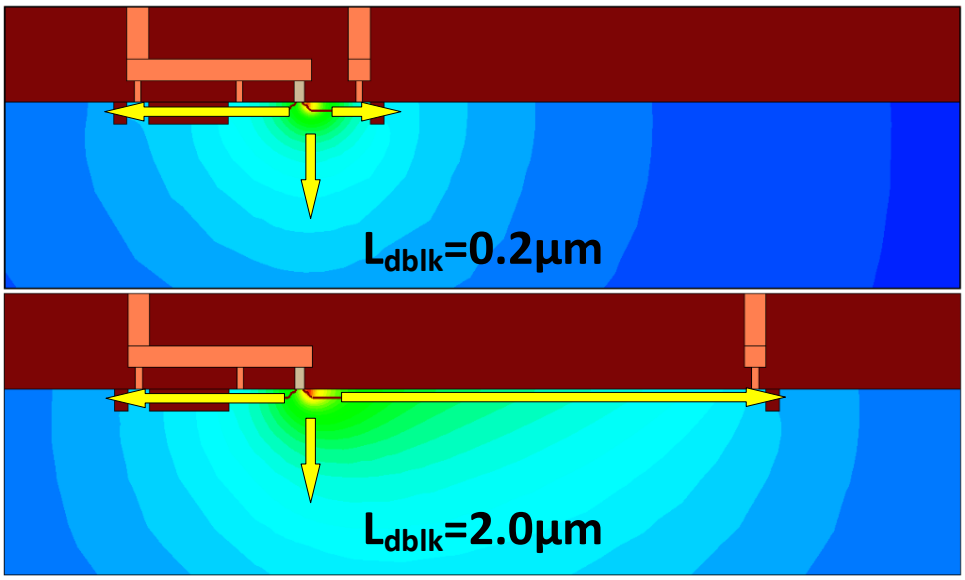


Figure 2.7. Illustration of temperature profile (blue = cold, red = hot) in the device and heat flow from the hot spot.

For L_{dblk} > 1 µm, I_{t2} is positively correlated to L_{dblk}. This trend is consistent with published data [17], [18]. Although this trend is often attributed to a ballasting effect and such ballasting was surmised from the 2D vs. 3D comparison, it is worth noticing that the same I_{t2} vs. L_{dblk} trend is predicted by 2D simulation. The 2D simulation shows that as L_{dblk} is made longer, a reduced fraction of the current will flow through the high-field region in the drain-extension near the surface, as demonstrated in Figure 2.8. The reduced current density near the surface reduces the device self-heating, resulting in higher I_{t2}. The 3D simulations show a similar dependence of the current distribution on L_{dblk}.

Having established that 2D TCAD can provide reasonable, conservative estimates of the ESD robustness of *planar* ESD devices, it is worth considering exactly when 3D TCAD is needed. Besides the obvious case of non-planar devices (e.g., FinFETs), it will be demonstrated that 3D

TCAD can be used for mitigating layout-dependent non-uniform current flow in SCRs, in the Chapter 3. Here, 3D TCAD is used to investigate the effect of body pickup layout on MOSFET V_{t1}. Figure 2.9 demonstrates that a ring-type body pickup causes the GGNMOS V_{t1} to shift relative to the case of a stripe body pickup. The 2D TCAD and 3D TCAD are seen to be equally well suited for simulating the stripe case.

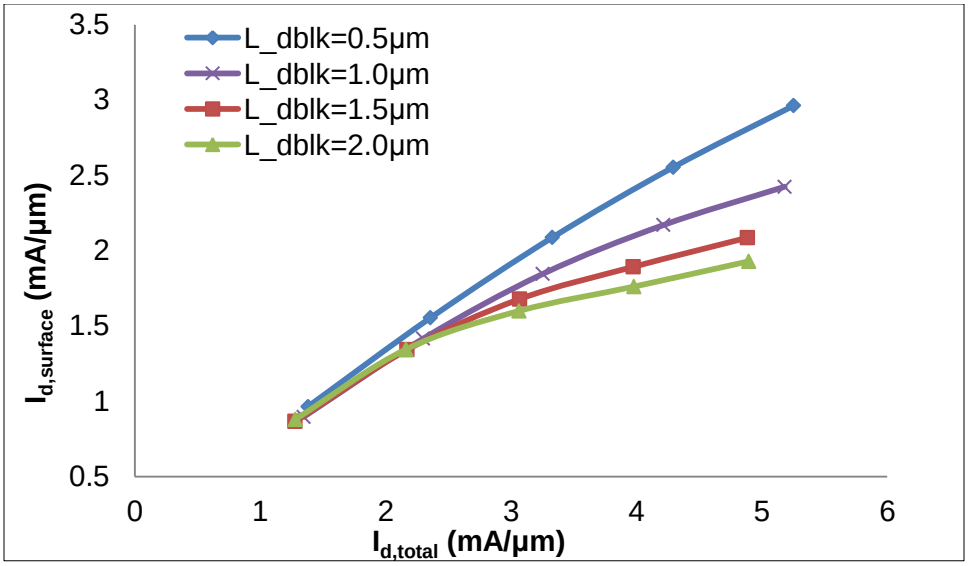


Figure 2.8. Surface current vs. total current for varying L_{dblk}. Surface current is the integral of current density from 0 to 10 nm under the surface at the edge of the drain side gate spacer.

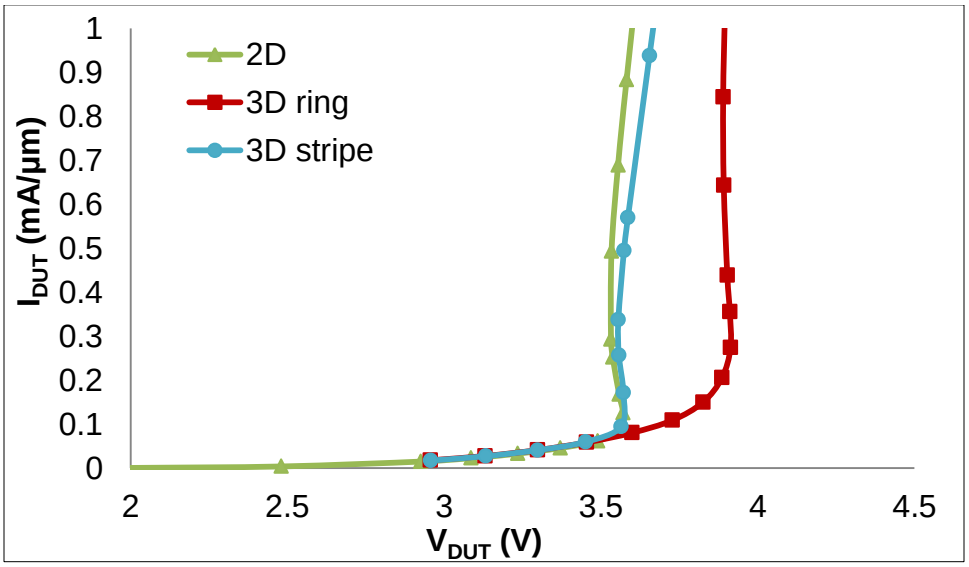


Figure 2.9. TLP I-V zoomed in to show turn-on behavior, which varies with the body-pickup layout.

2.3. Conclusion

A guide for uncalibrated ESD TCAD has been presented in this chapter. The importance of including a Si thermal buffer, the M1 interconnect and the passivation oxide was established. A parameter set for ESD simulations of NMOS devices is provided. Guidelines for constructing the gate stack in the absence of material information are provided. The 2D simulation is demonstrated to provide a conservative estimate of the real I_{t2}. Layout optimization via TCAD will generally require 3D simulation.

CHAPTER 3. TCAD ASSISTED LAYOUT IMPROVEMENT OF GGNMOS TRIGGERED SILICON CONTROLLED RECTIFIER (GGSCR) DEVICE

The TCAD methodology described Chapter 2 is applied to evaluate the performance of another commonly used ESD device: silicon controlled rectifier (SCR). Using 3D TCAD simulation, it is demonstrated that the traditional layout of the GGSCR device leads to non-uniform conduction and high voltage overshoot. An improved layout is proposed and evaluated in TCAD, then validated in 65 nm CMOS technology. Additionally, an in-depth study of the proposed device is carried out with measurement [19].

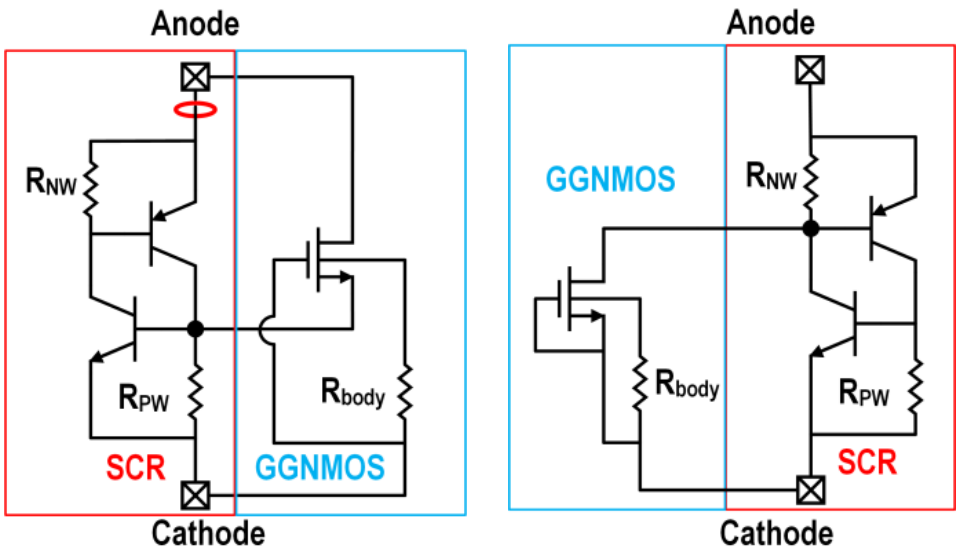
3.1. Introduction

ESD protection devices based on the SCR are suitable for signal pin protection due to the high level of current shunting they provide per unit capacitance. In this regard, the SCR is surpassed only by the dual-diode circuit [20], which is not suitable for use at certain types of signal pins, e.g. those which are required to be high voltage tolerant. A trigger circuit is used to turn on the SCR at a sufficiently low voltage such that it can protect the transistors in a low voltage CMOS circuit. The diode-triggered SCR (DTSCR) is attractive due to the ease of adjusting its trigger voltage, V_{t1}, by design. However, the DTSCR may be unsuitable for low power products due to its high leakage current [21]. As an alternative, the grounded-gate NMOS triggered SCR (GGSCR) was proposed [22]; it has much lower leakage current than the DTSCR [23]. A circuit representation of the original GGSCR [22] is shown in Figure 3.1(a); note that the SCR is triggered on by current injection from the GGNMOS into the SCR p-well.

An SCR cannot turn on and reach its final on-state voltage instantly; this becomes noticeable on a sub-ns timescale, i.e., the timescale of a CDM-ESD transient. As a result, the voltage across the

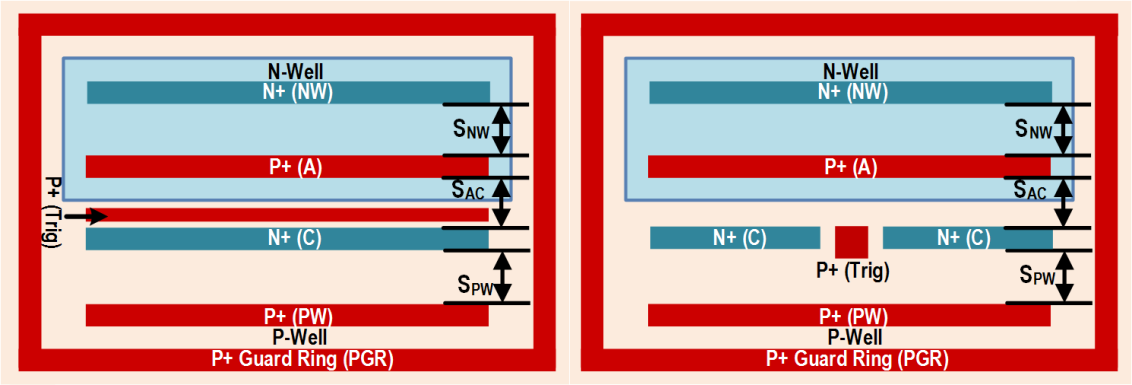
SCR terminals will temporarily overshoot its trigger voltage, potentially compromising the reliability of the transistors it is protecting, especially the gate dielectric reliability [24]. Therefore, a critical design challenge for SCR-based ESD protection devices is to limit the transient voltage overshoot. The original layout for the GGSCR [22] is illustrated in Figure 3.2(a); observe that the trigger current is injected into the SCR p-well by a "point-trigger tap." In [25], it was conjectured that the point-trigger tap layout exacerbates non-uniform current conduction in the SCR. In this work, it is demonstrated that the point-trigger tap layout also increases the voltage overshoot, and a modified GGSCR layout is proposed for transient voltage overshoot reduction. The new layout is validated in 65 nm CMOS technology. The efficacy of the layout modification is evaluated as a function of critical layout spacings, e.g., the well-tap spacing.

In principle, the GGNMOS trigger device could be connected to the SCR n-well rather than to its p-well; this is illustrated in Figure 3.1(b). The n-well triggered configuration would present a smaller capacitance at a signal pin. Despite this benefit, the p-well triggered device is often favored due to the ease of increasing the holding voltage V_{hold} by inserting one or more diodes in series with the SCR anode, at the position labeled in Figure 3.1(a), without compromising the low trigger voltage. This work investigates which configuration is more favorable from the viewpoint of voltage overshoot.

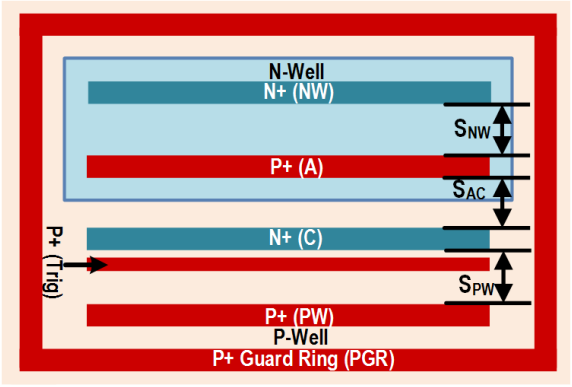


(a) (b)

Figure 3.1. Circuit schematics of (a) p-well triggered GGSCR [22] and (b) n-well triggered GGSCR. Holding voltage of the p-well triggered GGSCR can be tuned by adding series diodes at the position labeled with a red circle.



(a) (b)



(c)

Figure 3.2. SCR layout for (a) point-triggered GGSCR and (b) center-stripe-triggered GGSCR, and (c) side-stripe-triggered GGSCR.

3.2. Current Uniformity and GGSCR Layout

It was reported in [25] that the pulsed I-V characteristic of a point-triggered GGSCR varies with the pulse-width and rise-time. It was hypothesized that the initial current flow between the SCR anode and cathode occurs across only a small fraction of the device width, near the center where the trigger current is injected [25]. In this scenario, the device on-resistance would decrease as the current gradually spreads across the entire device width and I-V curves obtained using a longer pulse-width would differ from those obtained using very short pulse-widths [26]. The validity of these conjectures may be evaluated using TCAD simulation.

In the interest of computational efficiency, mixed-mode TCAD simulation was utilized, in which the SCR is represented in a finite element structure and the trigger GGNMOS is modeled as a voltage controlled switch which has the I-V characteristic shown in Figure 3.3. Since the SCR doping profiles were generated without the benefit of accurate process simulation, the TCAD simulation results should only be used for qualitative analysis. As shown in Figure 3.4(a), 3D TCAD simulation confirms that the point-triggered GGSCR will undergo non-uniform turn-on followed by current spreading.

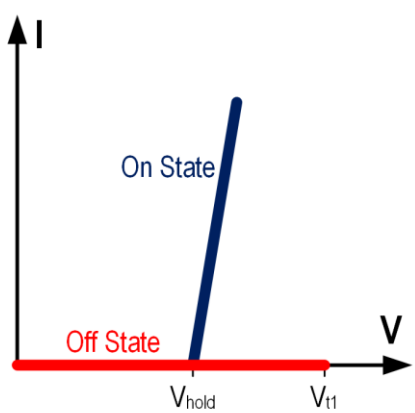
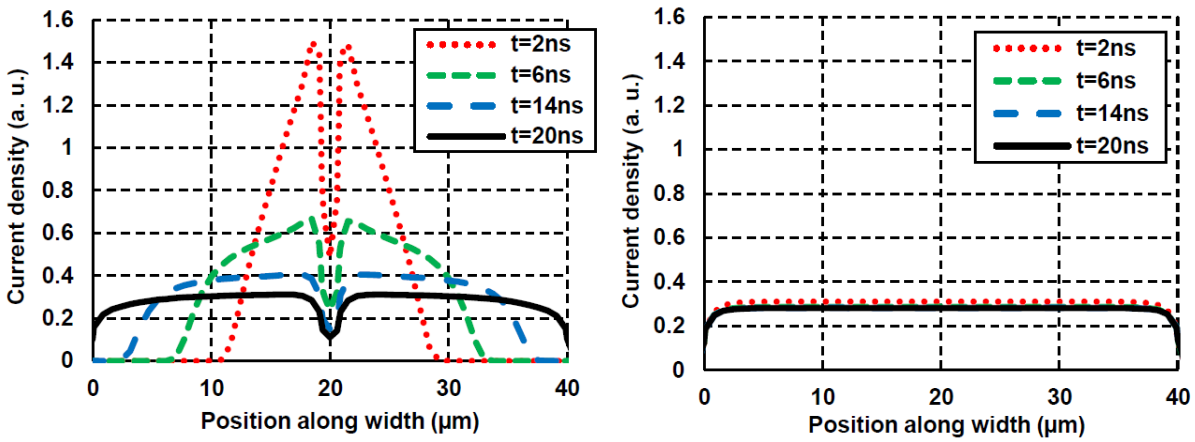


Figure 3.3. I-V characteristic of the "GGNMOS" used in TCAD simulations, with V_{t1} = 8 V, V_{hold} = 5 V. The device will switch from its off state to its on state if $V > V_{t1}$, and from the on state to the off state if $V < V_{hold}$.



(a) (b)

Figure 3.4. 3D TCAD transient simulation of current distribution in (a) a point-triggered GGSCR, the dip at the center is due to the discontinuity in the cathode (see Figure 3.2(a)); (b) a stripe-triggered GGSCR. I_{DUT} = 17.5 mA/μm.

If the non-uniform conduction is the result of using a point-trigger, it follows that a modified layout that promotes uniform conduction may reduce the device on-resistance obtained in response to a short duration current pulse. Thus it is proposed that the trigger tap be changed to a stripe that extends across the full width of the device. Two such layout options are investigated: the "center-stripe" in which the trigger diffusion lies between the anode and cathode diffusions, and the "side-stripe" in which the trigger diffusion lies between the p-well tap and the cathode; these are illustrated in Figure 3.2(b) and (c), respectively.

The 3D TCAD is used to predict whether the modified layouts of Figure 3.2(b) and (c) will impact the uniformity of SCR current conduction; the results shown in Figure 3.4(b) indicate that non-uniform conduction will be greatly suppressed in a stripe-triggered GGSCR. Furthermore, the TCAD simulation results shown in Figure 3.5 indicate that a stripe-triggered GGSCR should also have significantly smaller transient voltage overshoot than the point-triggered GGSCR. The simulation results further suggest that the center-stripe layout will result in lower overshoot than

the side-stripe layout, if the anode to cathode spacing (S_{AC}) is held constant. However, the minimum achievable S_{AC} is smaller for the side-stripe structure, potentially making it preferable.

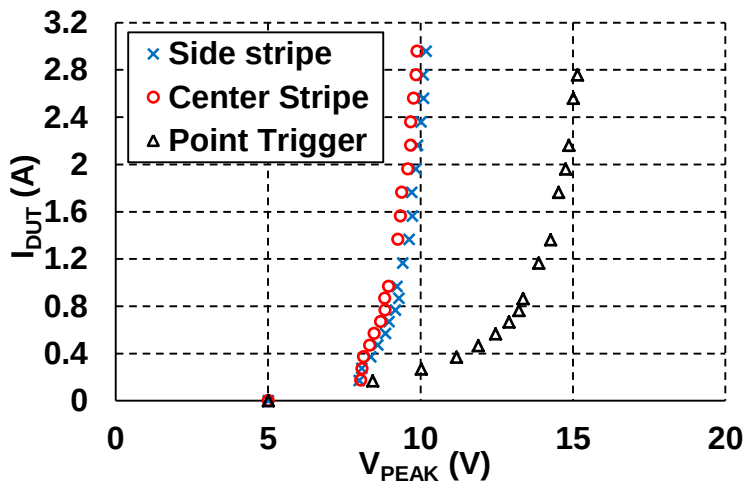


Figure 3.5. Peak voltage (overshoot) for point and stripe-triggered GGSCRs, obtained from 3D TCAD. t_{rise} = 300 ps. All SCRs have W = 40 µm, S_{AC} = 0.6 µm, S_{PW} = 1.5 µm, and S_{NW} = 0.3 µm.

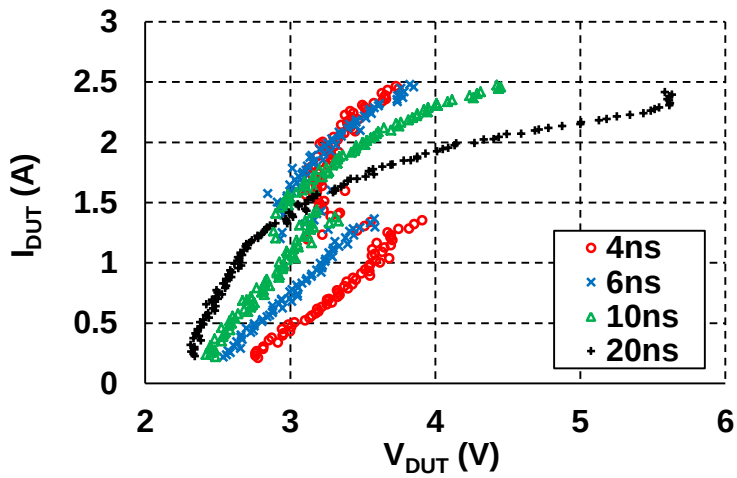


Figure 3.6. Pulsed I-V curves of point-triggered GGSCR fabricated in 65 nm CMOS, curves obtained for different pulse-widths. t_{rise} = 100 ps.

3.3. Test Structures

Point-triggered and stripe-triggered GGSCR test structures were fabricated in a 65 nm CMOS process. The trigger GGNMOSs are thick oxide I/O devices with eight fingers and a total width

of W_{total} = 160 µm; the GGNMOSs have a trigger voltage (V_{t1}) of 7.8 V when measured with pulses of 300 ps rise time. All SCRs are single finger devices with W_{SCR} = 40 µm. The pulsed I-V characteristics of the various test structures are obtained using the very fast TLP (vfTLP) method [27].

3.4. Measurement Results

3.4.1. Stripe vs. Point Trigger Layout

The pulsed I-V characteristic of a point-triggered GGSCR is plotted in Figure 3.6 for a variety of different pulse widths. Consistent with measurement results reported in [25], the low current portion of the I-V curve moves to the left as the pulse width increases. This behavior is attributed to non-uniform turn-on followed by gradual current spreading across the device width. Figure 3.7 shows the voltage waveforms measured across the terminals of point and stripe-triggered GGSCRs in response to current pulses with an amplitude of 2 A and a rise time of 300 ps. The stripe-triggered GGSCRs approach the steady-state on-voltage more quickly than the point-triggered GGSCR, which suggests that the stripe-triggered devices achieve uniform conduction faster than the point-triggered device. In Figure 3.8, the GGSCR pulsed I-V curves are compared with those obtained by plotting the quasi-steady-state current versus the peak voltage (i.e., the overshoot). The measurement results show that the stripe-triggered devices have less overshoot, consistent with the TCAD results shown in Figure 3.5. Table 3.1 summarizes the performance metrics for point-triggered and stripe-triggered GGSCRs. In Table 3.1, the holding voltage (V_{hold}) was defined as the extrapolated x-intercept of the pulsed I-V curve, with the linear fitting performed over the range of 0.5 A < I_{DUT} < 1 A. The on-resistance (R_{on}) is calculated using linear regression over the same I_{DUT} range. The failure current (I_{t2}) is measured with a current pulse

width (t_{pw}) of 25 ns. By introducing the stripe-trigger, the overshoot is reduced by about 4 V, without any significant impact on the other performance metrics.

Table 3.1. Comparison between point- and stripe-triggered GGSCR.

	Point trigger	Center stripe	Side stripe
V_{peak} (V); t_r=300ps, I=2A	14.1	9.9	10.4
V_{t1} (V); t_r=300ps	7.9	7.9	7.9
V_{hold} (V)	2.15	1.99	1.92
R_{ON} (Ω-µm)	19.5	22.9	22.4
I_{t2} (mA/µm); t_{pw}=25ns	60.3	63.3	63.0

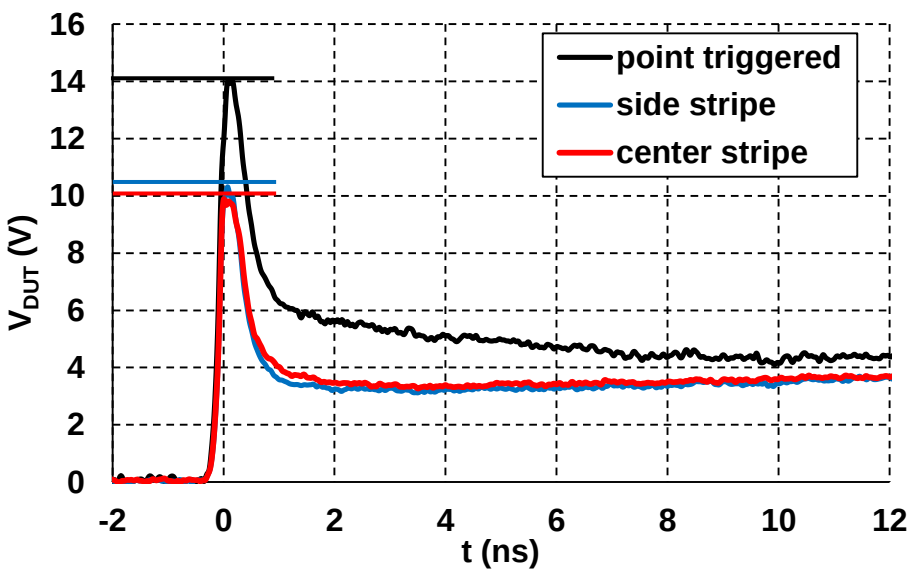


Figure 3.7. Measured voltage waveforms for point- and stripe-triggered GGSCRs; I = 2 A, t_{rise} = 300 ps. For all devices, S_{AC} = 0.65 µm, S_{PW} = 1.5 µm, S_{NW} = 1.5 µm.

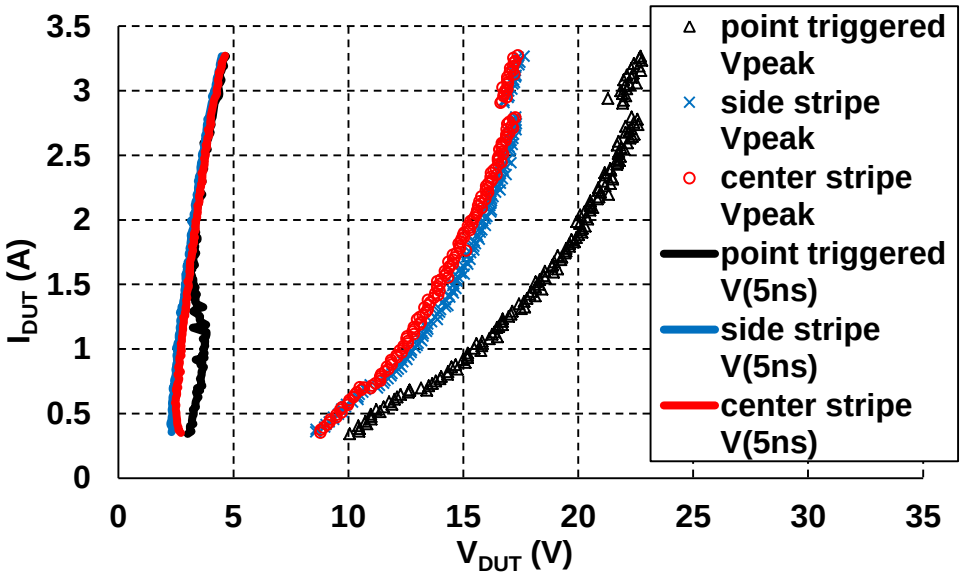


Figure 3.8. Pulsed I-V obtained using 5 ns pulses is plotted along with I vs. V_{peak}. Very short t_{rise} = 100 ps exacerbates overshoot. All layout spacings are the same as for Figure 3.7.

3.4.2. Effect of Layout Spacings

The key layout spacings for the SCR are the p-well tap spacing (S_{PW}), n-well tap spacing (S_{NW}), and anode to cathode spacing (S_{AC}), all of which are defined in Figure 3.2. The effect of these variables on the performance of a stripe-triggered GGSCR was quantified by measurements.

The peak voltage (V_{peak}) is plotted in Figure 3.9 as a function of S_{PW}; V_{peak} is observed to be a decreasing function of S_{PW}. Furthermore, the center-stripe-triggered GGSCR has a smaller overshoot than the side-stripe-triggered GGSCR, but the difference vanishes for large S_{PW}. Finally, the voltage overshoot of a side-stripe-triggered GGSCR is a stronger function of S_{PW} than for a center-stripe-triggered GGSCR.

In Figure 3.10, V_{peak} is plotted as a function of S_{NW}. S_{NW} apparently has no impact on the voltage overshoot.

V_{peak} is plotted as a function of quasi-steady-state current in Figure 3.11, for stripe-triggered GGSCRs with varying S_{AC}. The overshoot is observed to be an increasing function of S_{AC}. In this 65 nm process, the minimum achievable S_{AC} for the center-stripe-triggered GGSCR is 0.65 µm, and it is 0.32 µm for the side-stripe-triggered GGSCR. The data in Figure 3.11 shows that the side-stripe-triggered GGSCR with minimum S_{AC} has the smallest overshoot in this process.

The holding voltage (V_{hold}) values for the various test structures are listed in Table 3.2. V_{hold} is observed to be a decreasing function of both S_{NW} and S_{PW}, and an increasing function of S_{AC}. These results are consistent with those obtained in studies of the DTSCR [28], [29], and are unsurprising since the triggering mechanism should not affect the operation of the SCR once it is fully turned on. It is also observed that the value of V_{hold} is fairly insensitive to the trigger stripe placement.

Table 3.2. Extrapolated holding voltage for GGSCRs with varying layout parameters.

S_{PW} (µm)	S_{NW} (µm)	S_{AC} (µm)	V_{hold} (center stripe)	V_{hold} (side stripe)
1.5	1.5	0.65	1.99 V	1.92 V
0.6	1.5	0.65	2.58 V	2.52 V
1.5	**0.6**	0.65	2.43 V	2.33 V
1.5	1.5	**0.32**	N/A	1.61 V

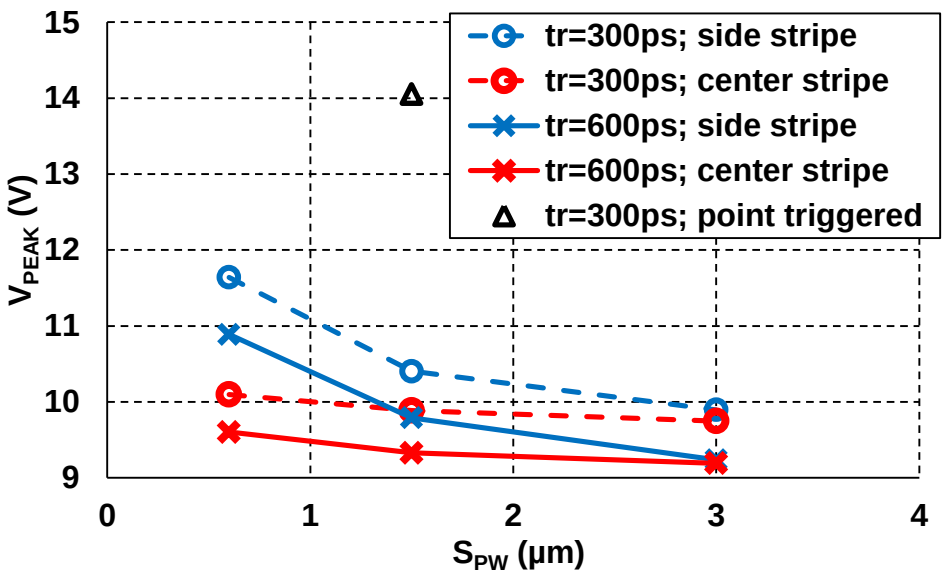


Figure 3.9. Measured V_{peak} of point and stripe-triggered GGSCRs with varying p-well tap spacing, I_{DUT} = 2 A. For all devices S_{AC} = 0.65 µm, S_{NW} = 1.5 µm.

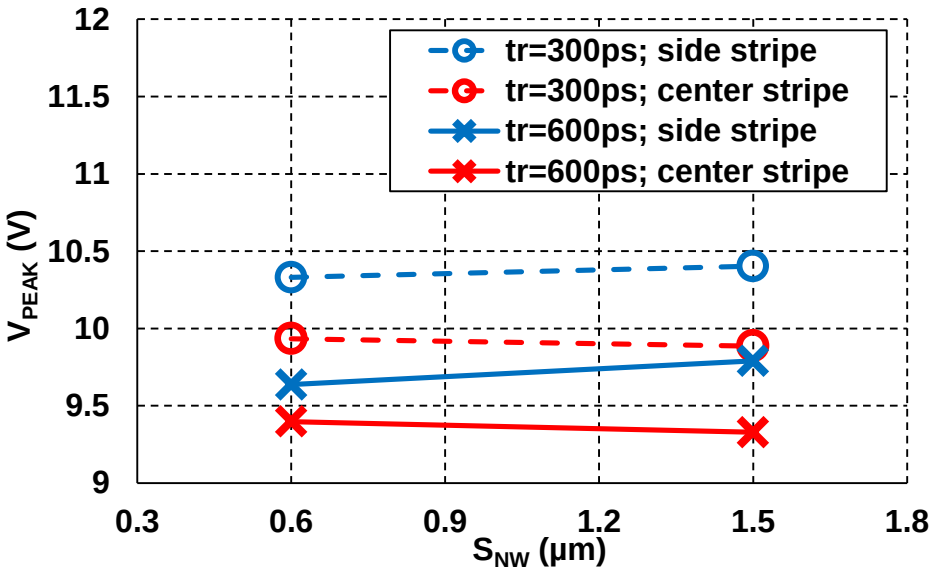


Figure 3.10. Measured V_{peak} of stripe-triggered GGSCRs with varying n-well tap spacing, I_{DUT} = 2 A. For all devices S_{AC} = 0.65 µm, S_{PW} = 1.5 µm.

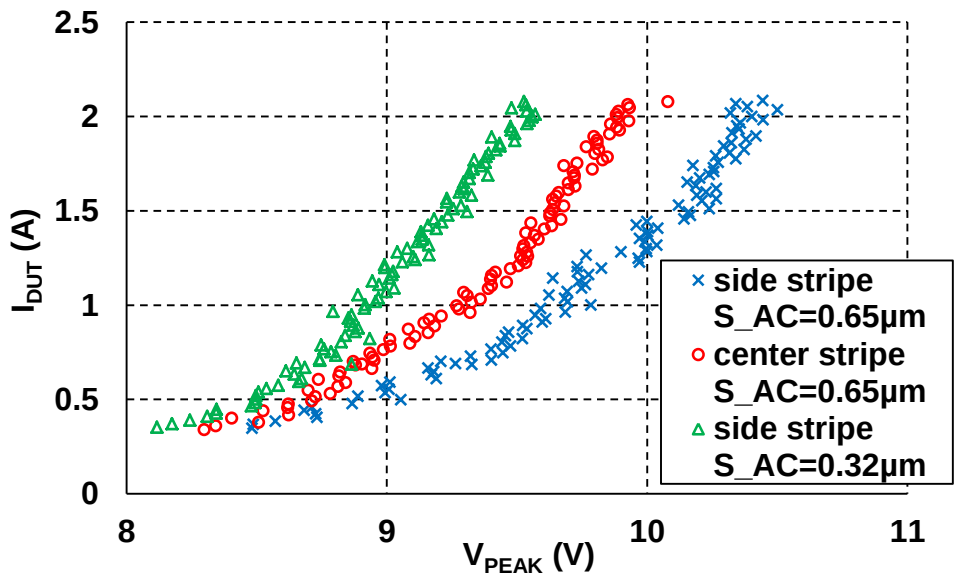


Figure 3.11. I vs. V_{peak} for stripe-triggered GGSCRs with varying anode to cathode spacing, t_{rise} = 300 ps. For all devices $S_{PW} = S_{NW}$ = 1.5 µm.

3.5. Discussion

In this section, physical explanations are provided for the measurement results of Section 3.4. The voltage across a GGSCR usually reaches its peak value at the end of the rising edge of the current pulse. A simple circuit representation of the pulse tester and the DUT (device under test) at the time of the peak voltage can be constructed as shown in Figure 3.12. This model will be used to explain why the GGSCR overshoot voltage varies with the layout spacings. In Figure 3.12, V_{TLP} and R_{TLP} represent the pulse source and the associated source impedance, usually 50 Ω. V_{hold} and R_{GGNMOS} represent the holding voltage and on-resistance, respectively, of the trigger GGNMOS operating in snapback mode. R_{PW} and R_{NW} are the series resistances associated with the p-well tap and n-well tap, respectively. $V_{BE,NPN}$, R_B and I_{link} together represent the NPN formed between the n+ cathode, p-well and n-well. $V_{BE,NPN}$ is the voltage drop across the junction between the p-well and the cathode; R_B is the resistance in series with this PN junction, and is also the NPN base resistance. $I_{link} = \beta \cdot I_{BN}$ is the NPN link current which flows from the n-

well into the p-well. This model does not include the PNP formed by the p+ anode, n-well and p-well because for pulses with a short rise-time—i.e., less than 1 ns—the SCR will not yet be triggered on at the end of the rising edge, thus the PNP will be off. Taking into account that the current through R_{PW} is negligible after the cathode PN junction turns on, the peak voltage across the DUT can be expressed as

$$V_{peak} = V_{hold} + V_{BE,NPN} + \frac{I_{TLP}}{\beta + 1}(R_{GGNMOS} + R_B) \tag{3.1}$$

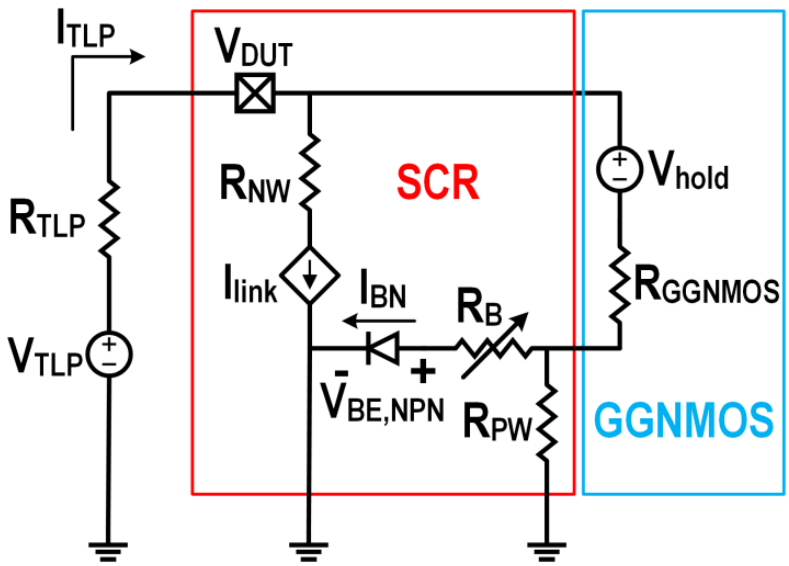


Figure 3.12. An equivalent circuit diagram of the GGSCR at the time of V_{peak}. $V_{BE,NPN} \cong V_{BE,on}$. The SCR has not yet been latched on into the low-impedance state.

For the point-triggered GGSCR, the current I_{BN} will be concentrated near the trigger tap at the center of the device width, and only a limited portion of the cathode junction will be conducting. Therefore, the effective R_B will be much larger than for a stripe-triggered GGSCR, in which the entire cathode junction will be conducting current. From Equation (3.1), it follows that the point-triggered GGSCR should have a higher overshoot voltage than the stripe-triggered GGSCR.

Increasing S_{AC} will cause the NPN gain, denoted by the variable β, to decrease. Therefore, according to Equation (3.1), V_{peak} should be an increasing function of S_{AC}, consistent with the measurement results.

To understand why V_{peak} is a function of S_{PW}, it is very important to note that R_B is time varying due to conductivity modulation. With no current flowing through the cathode junction, R_B stays at its equilibrium value, $R_{B,max}$. As soon as I_{BN} becomes non-zero, electrons start to be injected into the p-well and, to maintain charge neutrality, the hole concentration also increases. The increased concentration of majority carriers reduces R_B; this, in turn, reduces the net impedance seen by the pulse source, which increases I_{TLP} and, proportionately, the current through the cathode junction, thereby further reducing R_B. Thus, if the portion of the trigger current that flows across the cathode junction is large, conductivity modulation will occur rapidly and the voltage overshoot will be limited. In other words, the larger the value of I_{BN} prior to triggering, i.e. during the pulse rising edge, the smaller will be the overshoot. Using the resistive model shown in Figure 3.13, it can be demonstrated that the value of I_{BN} prior to triggering is modulated by the p-well tap spacing. Figure 3.13 encompasses both the center-stripe and side-stripe cases. In the figure, the pulse source is represented by its 50 Ω output resistance and its amplitude V_{TLP}; the trigger GGNMOS, operating in snapback, is represented as an ideal voltage source V_{hold} and a series resistance R_{GGNMOS}. Initially, the voltage across the cathode junction is zero. Thus, an expression for the initial I_{BN} can be derived based on the parameters labeled in Figure 3.13:

$$I_{BN}^{\text{side stripe}} = \frac{V_{TLP} - V_{hold}}{R_1 + R_T' + R_C + \frac{R_T' R_1 + R_T' R_C}{R_{PW}}} \tag{3.2}$$

$$I_{BN}^{\text{center stripe}} = \frac{V_{TLP} - V_{hold}}{R_1 + R'_T + R_C + \frac{R_C R_1 + R'_T R_C}{R_{PW} + R_1}} \tag{3.3}$$

where $R'_T = R_T + R_{GGNMOS} + 50\Omega$. It is important to note that the model of Figure 3.13 and the Equations (3.2) and (3.3) are valid only prior to triggering; in fact, the trigger GGNMOS will usually turn off once the SCR is triggered on.

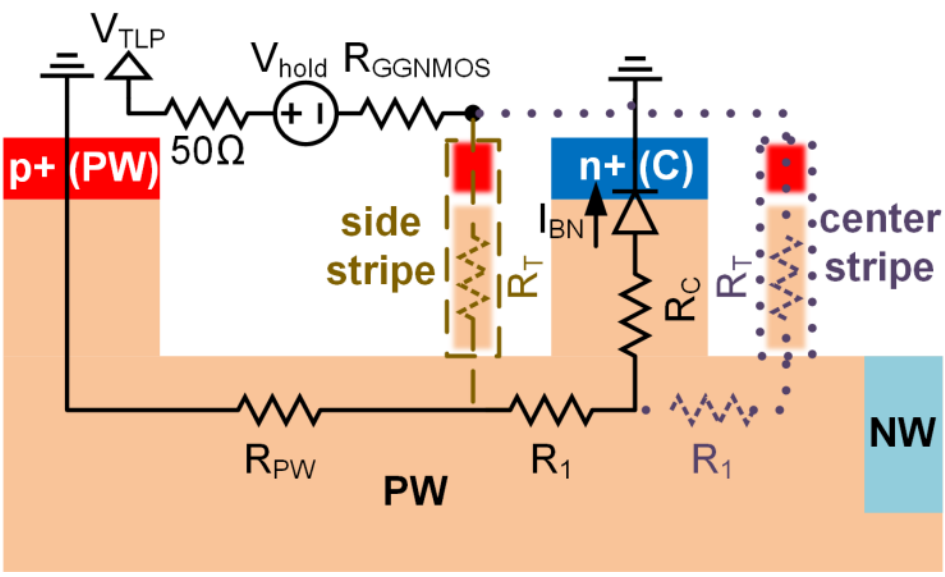


Figure 3.13. Resistive network through which trigger current flows in a stripe-triggered GGSCR. Only one of the two trigger stripes will be present. Dashed line: side stripe. Dotted line: center stripe.

Equations (3.2) and (3.3) show implicitly that I_{BN} is a function of S_{PW}, since R_{PW} is linearly proportional to S_{PW}. Equations (3.2) and (3.3) predict that I_{BN} is an increasing function of R_{PW}, and that the side-stripe layout has the greater sensitivity to this parameter. The equations also indicate that the center-stripe layout will yield a larger value of I_{BN} than the side-stripe layout (for fixed V_{TLP}), with the difference disappearing in the limit of very large R_{PW}. All conclusions drawn from the equations are consistent with the measurement results shown in Figure 3.9.

The value of the resistor R_{NW} in Figure 3.12 is linearly proportional to S_{NW}. However, a change in S_{NW} will not affect the overshoot, since V_{DUT} is determined by the voltage drop along the

parallel current path through the GGNMOS. This conclusion is consistent with the measurement results presented in Figure 3.10.

3.6. Comparison of P-Well and N-Well Triggered GGSCR

Although the original GGSCR was a p-well triggered device, it is also possible to trigger the GGSCR by current injection into the n-well. The circuit diagram of such an n-well triggered GGSCR was shown in Figure 3.1(b). This work only investigates n-well stripe-triggered GGSCR devices since the point-trigger layout has been shown to be inferior. The trigger stripe is an n+ diffusion in the n-well, either placed between the anode and cathode (center-stripe), or between the anode and n-well tap (side-stripe), as shown in Figure 3.14. The measurement data in Figure 3.15 shows that n-well triggered devices suffer more overshoot than p-well triggered devices. TCAD simulation results, shown in Figure 3.16, show the same trend, suggesting that the larger overshoot of the n-well triggered GGSCR is a general result, rather than one specific to the particular CMOS technology utilized in this work. A possible physical explanation for the large overshoot associated with the n-well triggered GGSCR is given in Figure 3.14.

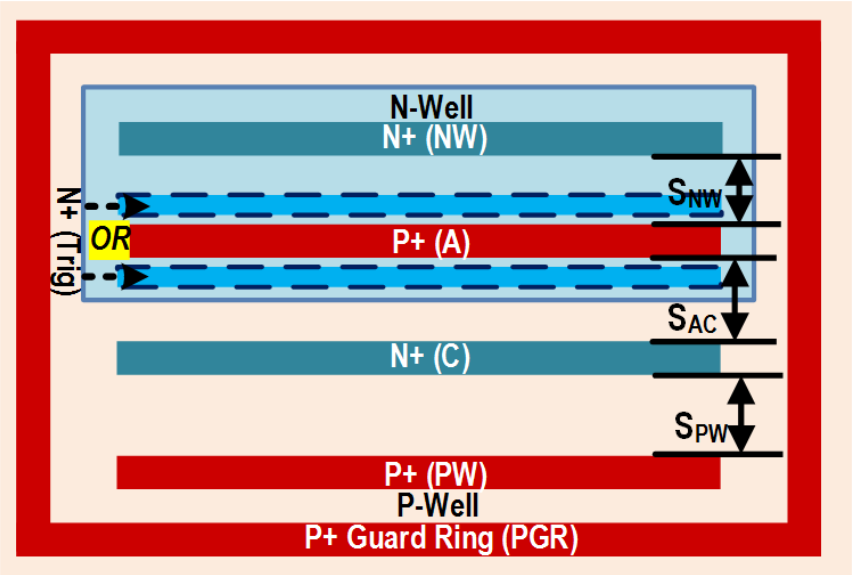


Figure 3.14. N-well stripe-triggered GGSCR layout, showing two stripe placement options.

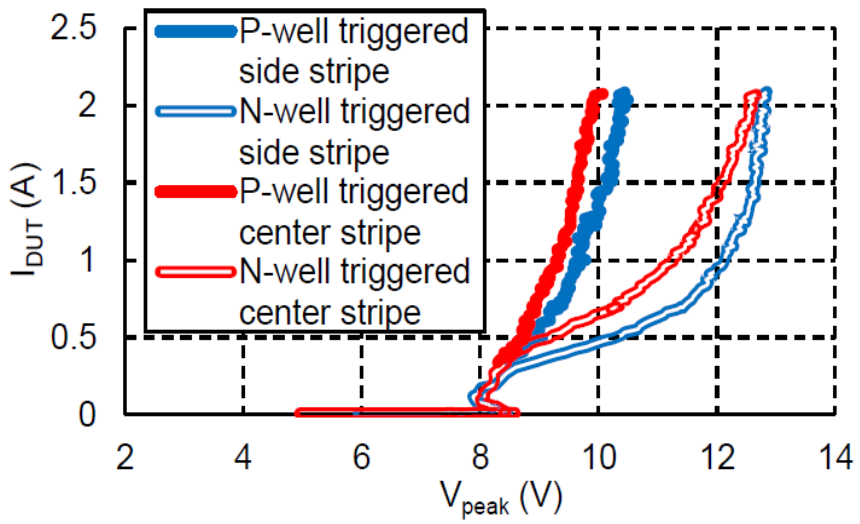


Figure 3.15. Voltage overshoot of p-well stripe-triggered and n-well stripe-triggered GGSCRs. t_{rise} = 300 ps. All spacings are the same as in Figure 3.7.

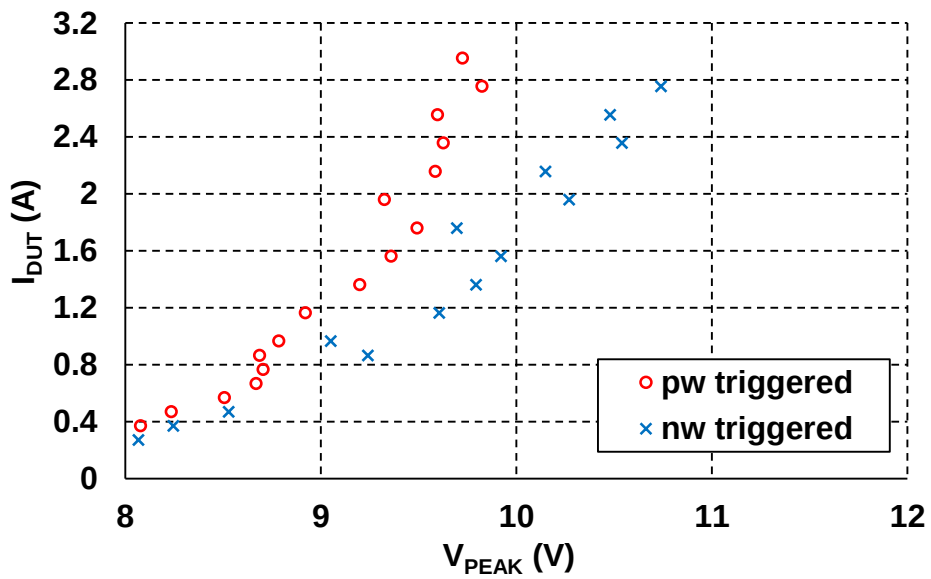


Figure 3.16. 3D TCAD simulation result of voltage overshoot in p-well and n-well triggered GGSCR devices; SCR width is 40 µm. The SCR p-well and n-well doping concentrations are equal.

As discussed in the Section 3.5, the peak voltage for the p-well triggered GGSCR is modulated by the resistance labeled R_B in Figure 3.12. In the n-well triggered GGSCR, the base resistance of the PNP plays an analogous role; this is the n-well resistance in series with the PN junction at the anode. As noted previously, the base resistance varies with the junction current due to

conductivity modulation. However, the rate at which minority carriers (electrons) are injected into the p-well of the p-well triggered GGSCR is larger than the rate at which minority carriers (holes) are injected into the n-well of the n-well triggered GGSCR, because electrons have higher diffusivity than holes. As a result, the NPN R_B in the p-well triggered GGSCR will undergo conductivity modulation at a faster rate than the corresponding PNP R_B in the n-well triggered GGSCR. In the TCAD simulation results shown in Figure 3.16, the n-well resistance has a smaller equilibrium value than the p-well resistance, yet the p-well triggered device has smaller overshoot, highlighting the critical importance of conductivity modulation. Therefore, it is concluded that the p-well triggered GGSCRs intrinsically have smaller overshoot than the n-well triggered GGSCRs.

3.7. Trigger GGNMOS Design Considerations

Some of the p-well triggered GGSCRs used in this work were observed to fail at very low stress currents, but were robust at higher current levels. These devices are said to have a “failure window”, the concept of which is illustrated in Figure 3.17. It is concluded that premature failure occurs in a p-well triggered GGSCR when the trigger GGNMOS gets damaged before the SCR turns on. The measurement results of Figure 3.18 show that the trigger GGNMOS is on and conducting current for a time duration that decreases as the current increases, which explains why the device only fails at low current levels. At low current levels, the SCR is slower to clamp the anode voltage below that needed to maintain the GGNMOS in the on-state, because the SCR cannot fully turn-on until a large number of carriers have been injected into its p-well and n-well.

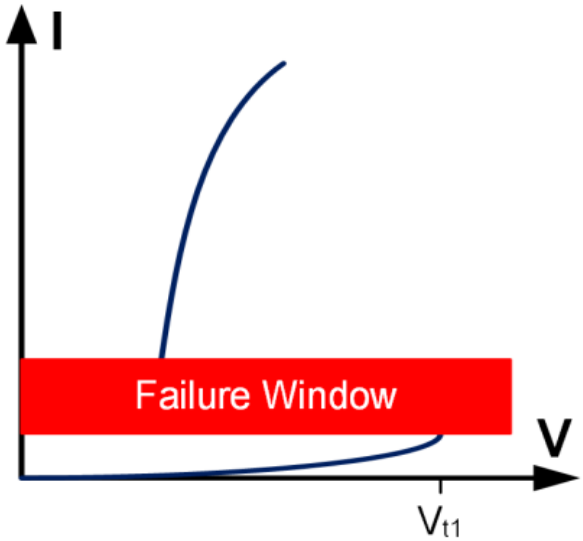


Figure 3.17. Illustration of failure window observed while measuring some p-well triggered GGSCR. The I-V curve in the region above the failure window is obtained by starting the TLP measurement sequence at a high precharge voltage and stepping down.

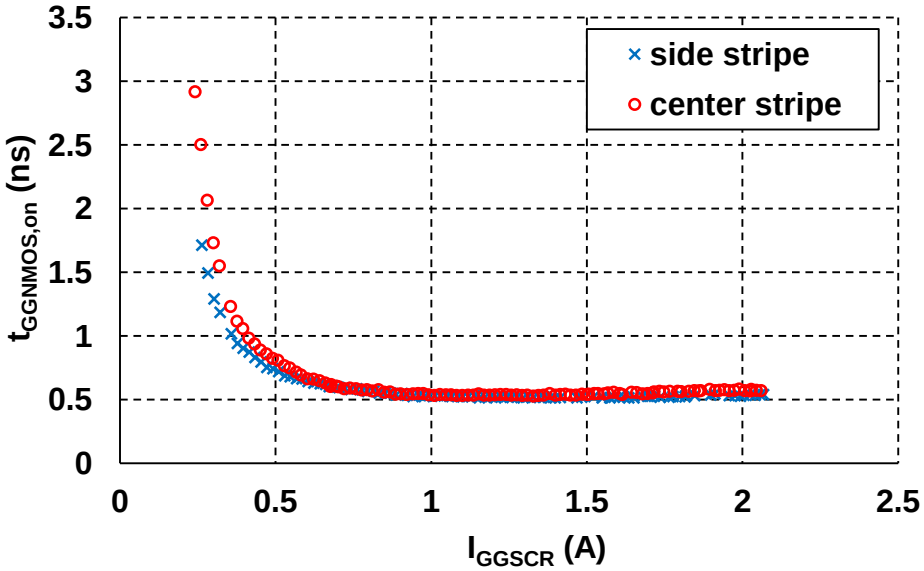


Figure 3.18. P-well triggered GGSCR: plot of the time duration during which the GGNMOS operates in snapback. t_{on} vs. I is extracted from the measured GGSCR voltage waveforms by assuming that the GGNMOS turns on at $V_{DUT} > V_{t1,GGNMOS}$ and off at $V_{DUT} < V_{hold,GGNMOS}$. t_{rise} = 300 ps. All spacings same as in Figure 3.7.

None of the n-well triggered GGSCRs suffers from premature failure. Furthermore, measurements performed on stand-alone GGNMOS test structures show that the GGNMOS

failure current is more than adequate to survive low current stress: I_{t2} = 6.75 mA/µm for t_{pw} = 25ns. The relatively poor robustness of the GGNMOS integrated into the p-well triggered GGSCR is attributed to its source connection. As illustrated in Figure 3.1(b), the GGNMOS of the n-well triggered GGSCR has its gate, source, and body all connected to ground, same as in the stand-alone test structures. In contrast, from Figure 3.1(a) it can be seen that the source node of the GGNMOS in a p-well triggered GGSCR is not grounded. TCAD simulation shows that a GGNMOS with an elevated source potential tends to have non-uniform conduction, which can cause premature failure. The 2D TCAD was used to examine the distribution of current among the fingers of an eight-finger GGNMOS. The results, shown in Figure 3.19, indicate that when a diode is connected between the source terminal and ground, the current is more crowded into the two center fingers relative to the source grounded case. However, in all cases, a uniform distribution of current between the fingers is obtained at sufficiently high current levels. Furthermore, 3D TCAD simulation results, shown in Figure 3.20, suggest that when the source potential is elevated, non-uniform conduction across the width of a single finger is exacerbated. The effect of elevated source potential on GGNMOS can be experimentally verified easily, by comparing measurement results of two GGNMOS structures with source grounded or elevated respectively. However, due to a lack of test structures, such experiment is not carried out.

A design strategy for obtaining a suitably reliable GGNMOS for the p-well triggered GGSCR is formulated with the aid of TCAD simulation. The simulation results shown in Figure 3.21 suggest that increasing the length of the drain-side silicide blocked region helps to alleviate non-uniform conduction between fingers in a GGNMOS with elevated source potential. The conjecture that higher R_{drain} will prevent early failure is supported by measurement data. In a previous work [29], p-well triggered GGSCR devices that utilized a trigger GGNMOS whose on-

resistance was approximately two times larger than the ones used in this work were measured, and early failures were not observed.

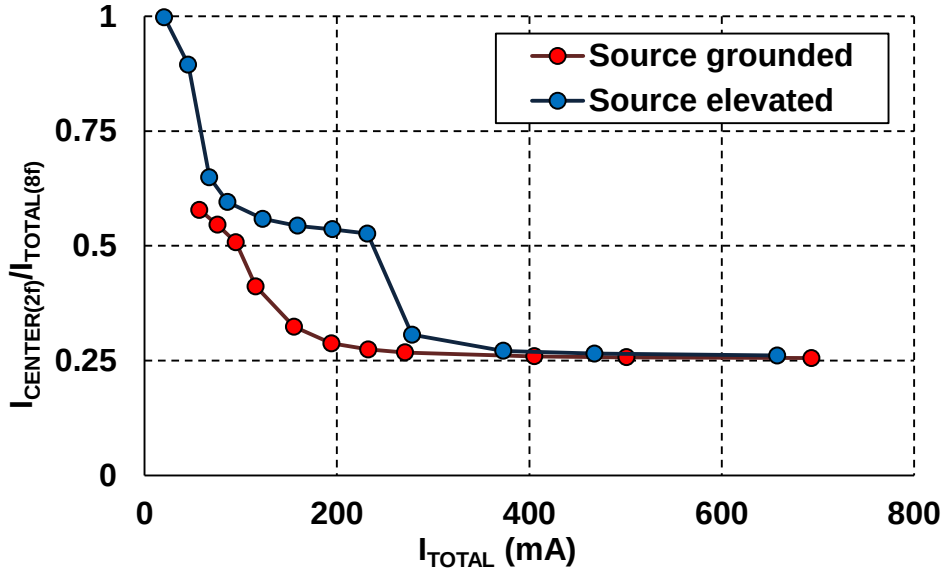


Figure 3.19. 2D TCAD simulation of an eight-finger GGNMOS, W_{total} = 160 µm. The vertical axis shows the ratio of current flowing through the two center fingers to the total current, evaluated at 10 ns after the rising edge.

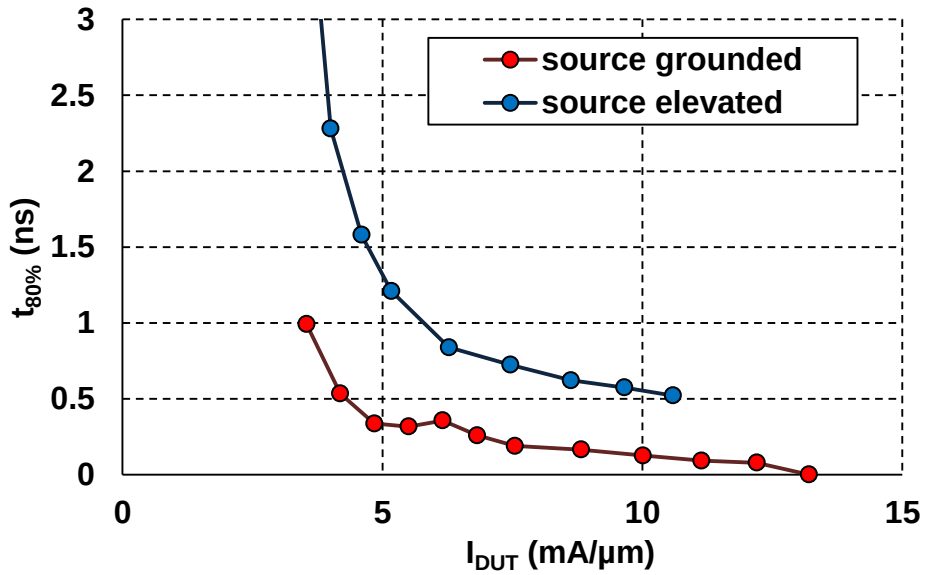


Figure 3.20. 3D TCAD simulation of a single-finger GGNMOS. The vertical axis shows the time it takes for 80% of the device width to turn on.

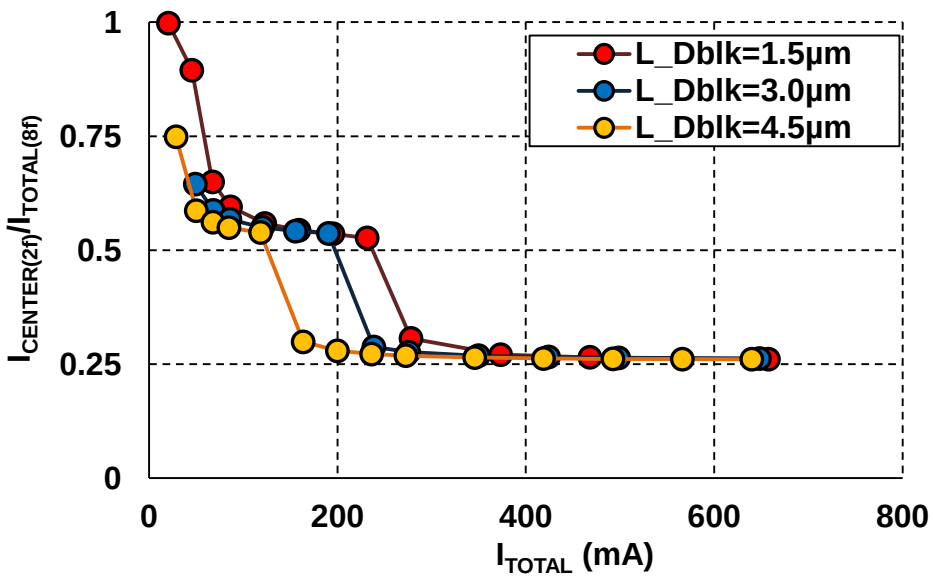


Figure 3.21. Same quantities plotted as in Figure 3.19 for source elevated GGNMOS, with varying drain-side silicide-blocking length (L_{dblk}).

3.8. Conclusions

The traditional point-triggered GGSCR layout is demonstrated to produce higher voltage overshoot than a stripe-triggered layout. The center-stripe-triggered GGSCR is found to have lower overshoot than the side-stripe-triggered GGSCR, provided that the anode to cathode spacing is kept constant. The overshoot voltage for stripe-triggered GGSCR devices is a decreasing function of the p-well tap spacing, and is independent of the n-well tap spacing. The trigger GGNMOS is susceptible to failing under low current stress conditions due to its elevated source potential; this problem can be alleviated by providing increased ballasting, e.g., by increasing the silicide block length on the drain-side. An n-well triggered version of the GGSCR may be constructed; however, this structure suffers from higher voltage overshoot than the usual p-well triggered structure. This last effect is attributed to a more rapid onset of conductivity modulation in the p-well than in the n-well.

CHAPTER 4. CONCLUSIONS

An uncalibrated ESD TCAD methodology is developed and evaluated in this work. In order to generate qualitatively accurate simulation results for ESD devices in very advanced processes (e.g. 32 nm and below), not only must there be a correct setup of the device structure and boundary conditions, the default physical model parameters embedded in TCAD software have to be modified. The physical models should be selected based on measurement results in the published literature. With a correct setup, TCAD is shown to generate qualitatively accurate simulation results for 32 nm ESD protection GGNMOS and SCR devices. For 32 nm GGNMOS, TCAD simulated failure current variation with respective to layout parameters matches qualitatively with published data; for the GGSCR device, TCAD predicted overshoot voltage variation due to layout change is validated by measurement.

The 2D simulation gives generally correct predictions for planar ESD devices. Compared with 3D simulation, 2D simulation generates more conservative predictions for the GGNMOS failure current and nearly identical results for the overshoot voltage of a SCR device. Due to the high computational resource required to perform 3D simulation, it is more justifiable to use 2D simulations as the major tool in ESD simulations. Additionally work on the methodology may be required to simulate an intrinsically 3D structure, e.g. FinFET devices.

REFERENCES

[1] K. Esmark, W. Stadler, M. Wendel, H. Gossner, X. Guggenmos, W. Fichtner, "Advanced 2D/3D ESD device simulation — A powerful tool already used in a pre-Si phase," in *Proc. EOS/ESD Symp*., 2000, pp. 420-429.

[2] C. Salamero, N. Nolhier, A. Gendron, M. Bafleur, P. Besse, M. Zecri, "TCAD methodology for ESD robustness prediction of smart power ESD devices," in *IEEE Trans. Dev. Mat. Rel.*, vol. 6, no. 3, pp. 399-407, Sept. 2006.

[3] J. A. Salcedo, J. J. Liou, Z. Liu, J. E. Vinson, "TCAD methodology for design of SCR devices for electrostatic discharge (ESD) applications," in *IEEE Trans. Electron Devices*, vol. 54, no. 4, pp. 822-832, Apr. 2007.

[4] S. Natarajan, M. Armstrong, M. Bost, R. Brain, M. Brazier, C.-H. Chang, V. Chikarmane, M. Childs, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He, R. Heussner, R. James, I. Jin, C. Kenyon, S. Klopcic, S.-H. Lee, M. Liu, S. Lodha, B. McFadden, A. Murthy, L. Neiberg, J. Neirynck, P. Packan, S. Pae, C. Parker, C. Pelto, L. Pipes, J. Sebastian, J. Seiple, B. Sell, S. Sivakumar, B. Song, K. Tone, T. Troeger, C. Weber, M. Yang, A. Yeoh, K. Zhang, "A 32 nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171 μm^2 SRAM cell size in a 291 Mb array," in *Proc. IEEE International Electron Devices Meeting*, 2008.

[5] C. H. Diaz, K. Goto, H. T. Huang, Y. Yasuda, C. P. Tsao, T. T. Chu, W. T. Lu, V. Chang, Y. T. Hou, Y. S. Chao, P. F. Hsu, C. L. Chen, K. C. Lin, J. A. Ng, W. C. Yang, C. H. Chen, Y. H. Peng, C. J. Chen, C. C. Chen, M. H. Yu, L. Y. Yeh, K. S. You, K. S. Chen, K. B. Thei, C. H. Lee, S. H. Yang, J. Y. Cheng, K. T. Huang, J. J. Liaw, Y. Ku, S. M. Jang, H. Chuang, M. S. Liang, "32 nm gate-first high-k/metal-gate technology for high performance low power applications," in *Proc. IEEE International Electron Devices Meeting*, 2008.

[6] F. Arnaud, J. Liu, Y. M. Lee, K. Y. Lim, S. Kohler, J. Chen, B. K. Moon, C. W. Lai, M. Lipinski, L. Sang, F. Guarin, C. Hobbs, P. Ferreira, K. Ohuchi, J. Li, H. Zhuang, P. Mora, Q. Zhang, D. R. Nair, D. H. Lee, K. K. Chan, S. Satadru, S. Yang, J. Koshy, W. Hayter, M. Zaleski, D. V. Coolbaugh, H. W. Kim, Y. C. Ee, J. Sudijono, A. Thean, M. Sherony, S. Samavedam, M. Khare, C. Goldberg, A. Steegen, "32 nm general purpose bulk CMOS technology for high performance applications at low voltage," in *Proc. IEEE Electron Devices Meeting*, 2008.

[7] D. J. Frank, Y. Taur, H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs," in *IEEE Electron Device Letters*, vol. 19, no. 10, pp. 385-387, Oct. 1998.

[8] A. R. Djordjevic and A. G. Zajic, "Low-reflection bandpass filters with a flat group delay," in *IEEE Trans. Microwave Theory and Techniques*, vol. 53, no. 4, pp. 1164-1167, Apr. 2005.

[9] V. Vassilev, M. Lorenzini, G. Groeseneken, M. Steyaert, H. Maes, "Analysis and improved compact modeling of the breakdown behavior of sub-0.25 micron ESD protection ggNMOS devices," in *Proc. EOS/ESD Symp*., 2001, pp. 61-69.

[10] J. J. Liou and J. S. Yuan, "An avalanche multiplication model for bipolar transistors," in *Solid-State Electronics*, vol. 33, no. 1, pp. 35-38, Jan. 1990.

[11] J. Li, K. Chatty, R. Gauthier, R. Mishra, C. Russ, "Technology scaling of advanced bulk CMOS on-chip ESD protection down to the 32 nm node," in *Proc. EOS/ESD Symp*., 2009, pp. 1-7.

[12] R. van Overstraeten and H. de Man, "Measurement of the ionization rates in diffused silicon p-n junctions," in *Solid-State Electronics*, vol. 13, no. 5, pp. 583-608, May 1970.

[13] K. Esmark, "Device simulation of ESD protection elements," Ph.D. dissertation, ETH Zurich, Switzerland, 2002.

[14] A. Dray, N. Guitard, P. Fonteneau, D. Golanski, C. Fenouillet-Beranger, H. Beckrich, R. Sithanandam, T. Benoist, C.-A. Legrand, P. Galy, "ESD design challenges in 28 nm hybrid FDSOI/Bulk advanced CMOS process," in *Proc. EOS/ESD Symp*, 2012, pp. 1-7.

[15] A. Amerasekera, W. van den Abeelen, L. van Roozendaal, M. Hannemann, P. Schofield, "ESD failure modes: Characteristics, mechanisms, and process influences," in *IEEE Trans. Electron Devices*, vol. 39, no. 2, pp. 430-436, Feb. 1992.

[16] K.-H. Oh, C. Duvvury, K. Banerjee, R. W. Dutton, "Impact of gate-to-contact spacing on ESD performance of salicided deep submicron NMOS transistors," in *IEEE Trans. Electron Devices*, vol. 49, no. 12, pp. 2183-2192, Dec. 2002.

[17] C.-S. Kim, H.-B. Park, B.-G. Kim, D.-G. Kang, M.-G. Lee, S.-W. Lee, C.-H. Jeon, W.-G. Kim, Y.-J. Yoo, H.-S. Yoon, "A novel NMOS transistor for high performance ESD protection devices in 0.18 µm CMOS technology utilizing salicide process," in *Proc. EOS/ESD Symp*, 2000, pp. 407-412.

[18] S. G. Beebe, "Methodology for layout design and optimization of ESD protection transistors," in *Proc. EOS/ESD Symp*, 1996, pp. 265-275.

[19] Z. Chen, R. Mertens, C. Reiman, E. Rosenbaum, "Improved GGSCR layout for overshoot reduction," in *Proc. IEEE Int. Rel. Phys. Symp*., 2015, pp. 3F.2.1-8.

[20] C. Richier, P. Salome, G. Mabboux, I. Zaza, A. Juge, P. Mortini, "Investigation on different ESD protection strategies devoted to 3.3 V RF applications (2 GHz) in a 0.18 µm CMOS process," in *Proc. EOS/ESD Symp*., 2000, pp. 251-259.

[21] J. Di Sarro, V. Vashchenko, E. Rosenbaum, P. Hopper, "A dual-base triggered SCR with very low leakage current and adjustable trigger voltage," in *Proc. EOS/ESD Symp*., 2008, pp. 242-248.

[22] C. C. Russ, M. P. J. Mergens, K. G. Verhaege, J. Armer, P. C. Jozwiak, G. Kolluri, L. R. Avery, "GGSCRs: GGNMOS-triggered silicon controlled rectifiers for ESD protection for deep sub-micron CMOS processes," in *Proc. EOS/ESD Symp*., 2001, pp. 23-31.

[23] J. Di Sarro, K. Chatty, R. Gauthier, E. Rosenbaum, "Evaluation of SCR-based ESD protection devices in 90 nm and 65 nm CMOS technologies," in *Proc. IEEE Int. Rel. Phys. Symp*., 2007, pp. 348-357.

[24] J. Wu, P. Juliano, E. Rosenbaum, "Breakdown and latent damage of ultra-thin gate oxides under ESD stress conditions," in *Proc. EOS/ESD Symp*., 2000, pp. 26-28.

[25] R. Mertens and E. Rosenbaum, "Separating SCR and trigger circuit related overshoot in SCR-based ESD protection circuits," in *Proc. EOS/ESD Symp*., 2013, pp. 1-8.

[26] K. Esmark, H. Gossner, S. Bychikhin, D. Pogany, C. Russ, G. Langguth, E. Gornik, "Transient behavior of SCRs during ESD pulses," in *Proc. IEEE Int. Rel. Phys. Symp*., 2008, pp. 247-253.

[27] H. Gieser and M. Haunschild, "Very-fast transmission line pulsing of integrated structures and the charged device model," in *Proc. EOS/ESD Symp.*, 1996, pp. 85-94.

[28] J. P. Di Sarro and E. Rosenbaum, "A scalable SCR compact model for ESD circuit simulation," *IEEE Trans. Electron Devices*, vol. 57, no. 12, pp. 3275-3286, Dec. 2010.

[29] R. Mertens and E. Rosenbaum, "A physics-based compact model for SCR devices used in ESD protection circuits," in *Proc. IEEE Int. Rel. Phys. Symp*., 2013, pp. 2B.2.1-7.