MODELING AND SIMULATION OF FULL-COMPONENT INTEGRATED CIRCUITS IN TRANSIENT ESD EVENTS

BY

KUO-HSUAN MENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Elyse Rosenbaum, Chair
Associate Professor Deming Chen
Professor Jose E. Schutt-Aine
Professor Martin D. F. Wong

Abstract

This thesis presents a methodology to model and simulate transient electrostatic discharge (ESD) responses of integrated circuits (IC). To obtain valid simulation results, the IC component must be represented by a circuit netlist composed of device models that are valid under the ESD conditions. Models of the nonlinear devices that make up the ESD protection network of the IC must have transient I-V responses calibrated against measurements that emulate ESD events. Interconnects, power distribution networks, and the silicon substrate on the chip die as well as on the IC package must be faithfully constructed to emulate the fact that ESD current flows in a distributed manner across the entire IC component. The resultant equivalent circuit model therefore contains a huge number of nodes and devices, and the simulation runtime may be prohibitively long. Techniques must be devised to make the numerical simulation process more efficient without sacrifice of accuracy. These techniques include reasonable abstraction of the distributed full-component circuit netlist, dynamic piecewise-linear device models, and customized efficient transient circuit simulator. With the simulation streamlining techniques set up properly, comprehensive and predictive transient ESD simulation can be carried out efficiently to investigate the weakest link in the target IC, and the design can be fine-tuned to achieve optimal performance in both functionality and ESD reliability.

Acknowledgments

I am grateful for the following people who have helped me through my Ph.D. study.

My advisor, Professor Elyse Rosenbaum, has always been patient and knowledgeable to guide me through research and publication.

Current students in the group—Min-Sun Keel, Robert Mertens, Nick Thomson, Zaichen Chen, and Yang Xiu—have been my best resources for brainstorming.

Former students—Vrashank Shukla, Nathan Jack, and Nick Olson—provided training on lab equipment and technical discussions.

Finally, my utmost gratitude is to my parents and family, for reasons beyond words.

Contents

Chapter 1. Introduction 1

1.1 Common Component-Level ESD Test Standards 1

1.1.1 Human Body Model (HBM) Test Standard 1

1.1.2 Field-Induced Charged Device Model (FICDM) Test Standard 2

1.2 Motivation for Modeling and Simulation of ESD Events 2

1.3 Challenges in Modeling and Simulation of IC Devices 2

1.3.1 Modeling and Simulation of Devices under ESD Conditions 2

1.3.2 Modeling and Simulation of Full IC Component 3

1.3.3 Modeling and Simulation of ESD Test Environment 3

1.4 Table and Figures 4

Chapter 2. Modeling Transient Dynamics of Snapback Devices 6

2.1 Motivation 6

2.2 Comparison of Pulse I-V and Transient I-V Techniques 8

2.3 GGNMOS Model 10

2.3.1 MOSFET ESD Snapback Model 10

2.3.2 The Holding Point 11

2.4 Investigation of Device-Tester Interactions Using Circuit Simulation 13

2.4.1 Overvoltage Stress due to Snap-up 13

2.4.2 Relaxation Oscillator 14

2.4.3 Instability during Turn-off in HBM Testing 17

2.5 Summary 18

2.6 Figures 19

Chapter 3. Distributed Multi-finger MOSFET ESD Models 31

3.1 Motivation 31

3.2 Model 32

3.3 Model Parameter Extraction 34

3.3.1 Measurement Setup 34

3.3.2 R_{body} Network Component Extraction 35

3.4 Model Verification 36

3.4.1 Gate Bias Effect 36

3.4.2 Trigger Voltage Vt1 36

3.4.3 Non-uniform Turn-on of Fingers 37

3.5 On-state Resistance and Self-heating 39

3.5.1 Single-finger On-state Resistance, R_{on1} 39

3.5.2 Modeling Non-uniform Self-heating 40

3.5.3 TCAD Simulation of Self-heating 42

3.6 Summary 43

3.7 Supplement: Derivation of Single-finger R_{on1} Model 43

3.8 Supplement: Analysis of Finger Current Distribution 46

3.9 Figures 47

Chapter 4. Piecewise-Linear Model with Transient Relaxation for Circuit-level ESD Simulation 60

4.1 Motivation 60

4.2 Piecewise-Linear Model with Transient Relaxation 61
4.2.1 Piecewise-linear I-V Branches 61
4.2.2 Transient Relaxation 61
4.2.3 Application to Non-snapback Devices 63
4.3 Simulation Results and Discussion 63
4.3.1 Diode 63
4.3.2 Silicon-controlled Rectifier (SCR) 64
4.4 Summary 64
4.5 Figures 65
Chapter 5. Full-Component Modeling and Simulation of Charged Device Model ESD 68
5.1 Motivation 68
5.2 Analysis of Cross-Domain Stress 69
5.3 Full-Component Model for CDM Simulation 71
5.3.1 FICDM Tester Model 71
5.3.2 Package Interconnect Model 71
5.3.3 Pad-ring Model 72
5.3.4 Core Power Distribution Network Model 72
5.3.5 Die Substrate Model 74
5.3.6 On-Die Decoupling Capacitance Model 74
5.3.7 Tester-Die Coupling Model 75
5.3.8 P-bulk Network Z-directional Meshing 76

5.4 Simulation Results and Discussion 77
5.4.1 Current Bypass Effect by Core Power Distribution Network 77
5.4.2 Effect of Decoupling Capacitors 78
5.4.3. Effect of Package-level Ground Network 80
5.4.4 Effect of Tester-Die Coupling 80
5.5 Localized Simulation of Domain-Crossing Circuit 81
5.6 Summary 82
5.7 Supplement: Full-component Netlist Extraction 82
5.8 Table and Figures 84
Chapter 6. Fast Circuit Simulator for Transient Analysis of CDM ESD 98
6.1 Motivation 98
6.2 Computational Complexity of CDM ESD Simulation 99
6.3 Problem Formulation 101
6.3.1 Transient Nodal Analysis 101
6.3.2 Linear Elements 102
6.3.3 PWL-TR Models 104
6.3.4 State-based Newton-like Iteration 104
6.4 Preconditioned Conjugate Gradient Method 105
6.4.1 Choice of Preconditioner 105
6.4.2 Convergence Criteria 106
6.5 Speed-Up Techniques 106

6.5.1 Parallel Programming 106

6.5.2 Speed-up from Customized Linear System Solver 108

6.5.3 Speed-up from PWL-TR Modeling 110

6.5.4 Speed-up from Built-in Models 111

6.6 Full-chip Simulation Example 112

6.6.1 Cluster Computing 112

6.6.2 Simulator-dependent Run-times 113

6.7 Summary 114

6.8 Tables and Figures 115

Chapter 7. Conclusions and Future Works 127

References 130

Chapter 1. Introduction

Electrostatic discharge (ESD) is defined as a sudden transfer of static charges between objects at different electric potentials [1]. Static charges can be generated by direct charging, ionic charging, tribo-electrical charging, or field-induced charging [2]. ESD can take place upon an integrated circuit (IC) in any of the following ways: a charged body touches the IC, the charged IC touches a grounded surface, or an electrostatic field induces a voltage across a dielectric sufficient to break it down. When these static charges flow across the victim IC, they form an electrical current that may damage gate oxides, metallization, and junctions along its path. Because ESD impacts production yields and product quality, there is continuous growth in research into the effects of ESD on the performance of semiconductor integrated circuits (ICs). ESD problems are increasing in the electronics industry because of the trends of technology scaling toward smaller device sizes in order to achieve faster operation and lower power. ESD therefore becomes a major consideration in the design and manufacture of IC components.

1.1 Common Component-Level ESD Test Standards

To quantify an IC component's robustness against ESD, several ESD models have been proposed. Each aims to emulate a particular type of real-world ESD event with specifications on testing setup and procedures. Among them are two test standards that the industry has adopted to qualify products before release; they are the human-body model (HBM) [3] and the field-induced charged device model (FICDM) [4],[5].

1.1.1 Human Body Model (HBM) Test Standard

HBM is one of the most commonly required qualification standards for characterizing the susceptibility of an IC component to damage from ESD [3]. The model is a simulation of the discharge which might occur when a human touches an electronic device. Figure 1.1 shows the schematic of the HBM test setup. Table 1.1 specifies all the required combinations of zap-pins of HBM. According to Table 1.1, almost all the combinations of any two different pins are required to be tested in order to fulfill

the standard. To locate the weakest spot in the IC that is most susceptible to HBM ESD, comprehensive simulation of HBM discharge on all combinations of zap-pins is required.

1.1.2 Field-Induced Charged Device Model (FICDM) Test Standard

FICDM is another commonly required qualification standard for characterizing the susceptibility of an IC component to damage from ESD [4],[5]. The model is a simulation of discharge of electrostatically induced charge stored in an IC component through a metallic contact. Figure 1.2 shows the schematic of the FICDM test setup. FICDM test standard requires that each pin be tested. To locate the weakest spot in the IC that is most susceptible to FICDM ESD, comprehensive simulation of FICDM discharge on each zap-pin is required.

1.2 Motivation for Modeling and Simulation of ESD Events

A CAD tool can be employed in the development of the ESD protection networks of IC products. It can also help analyze the details of failure mechanisms after the product undergoes a catastrophic ESD event. One of the most versatile tools is circuit simulator to simulate the transient ESD responses. Through circuit simulation, the ESD protection network can be designed and optimized, and the robustness of the IC can be evaluated before the IC is fabricated. Failure analysis can also be carried out through transient simulation of ESD events on an IC component [6],[7],[8].

1.3 Challenges in Modeling and Simulation of IC Devices

1.3.1 Modeling and Simulation of Devices under ESD Conditions

ESD may inflict strong overstress upon the internal devices of an IC component in the form of voltage and current. The voltage and current overstress experienced by an internal device, such as a MOSFET, can be much higher than its nominal operating conditions. To simulate authentic ESD responses, the circuit simulation must incorporate the device models that are calibrated over a wide range of I-V conditions covering the normal operation as well as ESD conditions. Moreover, ESD is an

extremely fast and short event. The duration of an ESD event ranges from 1ns to 100ns; the overstress inflicted upon the internal devices can reach its full magnitude within an order of 100ps to 10ns. The device models must be calibrated to reproduce the correct transient dynamics under ESD transients [9]–[31].

1.3.2 Modeling and Simulation of Full IC Component

To faithfully emulate the fast transient of charge traversing across the IC victim during the simulated event, a detailed construct of parasitic capacitances, inductances, and resistances of each I/O pin must be present in the simulation setup. ESD failure can occur along the pad-ring circuitry as well as deep within the internal circuit of an IC component. Therefore, to reproduce the corresponding ESD failure, a full distribution of the pad-ring and core circuits must be reasonably constructed to simulate an authentic ESD failure [6].

1.3.3 Modeling and Simulation of ESD Test Environment

A real-world ESD event involves charge transfer between the victim IC and the surrounding environment. Accurate simulation of the ESD event therefore involves accurate modeling of the victim IC as well as the surroundings. In the context of predicting ESD test failure under a given ESD test standard, such as HBM or FICDM, the IC as well as the FICDM tester must both be represented by appropriate circuit models in the simulation netlist. Prior works have shown that it is critical to properly include the tester models in order to obtain correct simulation results [6],[7],[8].

Based on the above observation, transient ESD simulation of the full IC component necessitates a distributed model that includes the full pad-ring circuitry and core circuit with all the package pins, interconnects, power networks, and environmental parasitics. Although detailed and thus accurate, such a large circuit netlist inevitably requires a certain simulation runtime. Moreover, all the building blocks of the ESD protection networks are inherently nonlinear devices; simulation of nonlinear devices further exacerbates the overall simulation runtime. The challenge in the transient ESD simulation therefore lies in

how to reasonably abstract the distributed circuit model as well as simplify the nonlinear device models so as to speed up the simulation runtime.

1.4 Table and Figures

Table 1.1 Pin Combinations Specified in the HBM test Standard

Pin Combi-nation Set	**Connect Individually to Terminal A**	**Connect to Terminal B (Ground)**	**Floating Pins (unconnected)**
1	All pins one at a time, except the pin(s) connected to terminal B	Vps(1) [First power pin(s)]	All pins except pin under test (PUT) and Vps(1) [First power pin]
2	All pins one at a time, except the pin(s) connected to terminal B	Vps(2) [Second power pin(s)]	All pins except PUT and Vps(2) [Second power pin]
i	All pins one at a time, except the pin(s) connected to terminal B	Vps(i) [i^{th} power pins] [1,2,…,i]	All pins except PUT and Vps(i)
n-1	All pins one at a time, except the pin(s) connected to terminal B	Vps(n-1)	All pins except PUT and Vps(n-1)
n	All non-Vps(i) pins one at a time	All other non-Vps(i) pins, except the pin connected to terminal A	All Vps(i) pins

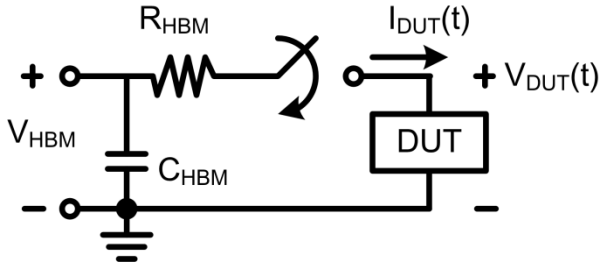


Figure 1.1 Schematic of the HBM test setup. V_{HBM} is the pre-charge voltage and usually ranges from 500V to 8kV. C_{HBM}=300pF. R_{HBM}=1.5kΩ. DUT stands for the device under test.

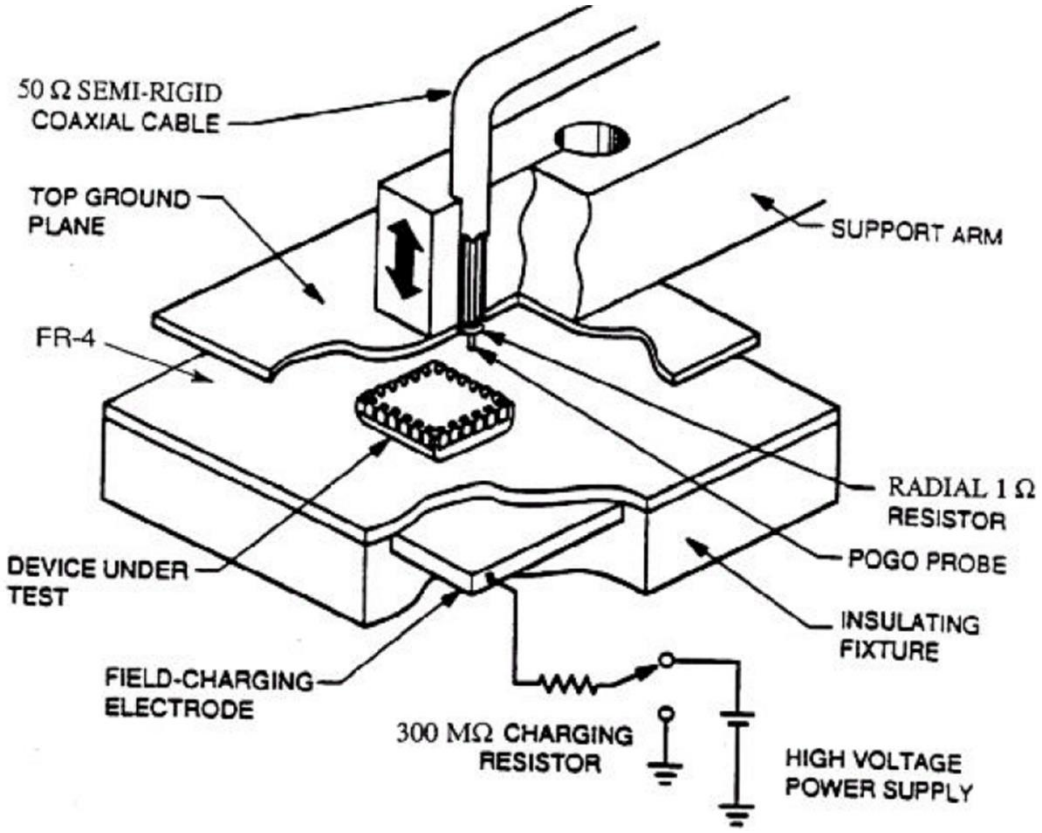


Figure 1.2 Schematic of the FICDM test setup [4][5][104].

Chapter 2. Modeling Transient Dynamics of Snapback Devices

Tuning the parameters of an ESD device compact model to match the results of pulsed I-V measurements does not guarantee that the model will correctly reproduce the device response to arbitrary ESD waveforms. Transient I-V measurement data are demonstrated to improve the completeness of the device model assessment process, as demonstrated in this chapter for the case of the grounded-gate MOSFET protection device [9].

Given an accurately calibrated ESD compact model, circuit simulations may be used to identify device-tester interactions that have been reported to cause unexpected damage during ESD testing. The interaction between a snapback device and an ESD tester can be understood in the context of a relaxation oscillator [10].

2.1 Motivation

The ESD response of a semiconductor device is often characterized using pulsed I-V measurement data obtained using a transmission-line pulse (TLP) [32] or very-fast transmission-line pulse (VFTLP) tester [33]. These I-V data are widely used as the basis for ESD compact model development and verification [9]–[31]. The resulting models have been demonstrated to reproduce the device's pulsed I-V characteristic. However, pulsed I-V characterization has been demonstrated to be insufficient to guarantee that the device response to arbitrary waveforms, such as HBM and system-level ESD, will be correctly reproduced by circuit simulation.

For snapback-type devices, accurate measurement, and subsequent modeling, of the holding voltage V_{hold} and current I_{hold} may be important; for example, this is needed to predict the response of a protection device to a secondary HBM pulse [34],[35]. Accurate measurement of the holding point is generally not possible using a 50-Ω TLP system. The components of a TLP or VFTLP system are often matched to the 50-Ω transmission line impedance to preserve the pulse integrity [36]; however, the 50-Ω load-line of the system limits its ability to pinpoint the holding point (V_{hold}, I_{hold}). This shortcoming motivated the

development of high-impedance TLP systems [37],[38], multi-level TLP systems [39]–[41], and transient I-V measurement techniques [42]–[45].

Commercial, semiconductor parameter analyzers and curve tracers provide an I-V sweep with high source impedance, but the measurement duration is often too long for on-chip snapback devices to survive the stress. High-impedance TLP provides the necessary short pulse-width, but the pulse quality tends to be degraded relative to 50-Ω TLP [36].

Multi-level TLP testers maintain a 50Ω environment for pulse waveform integrity. These testers first drive the device into the on-state by applying a high amplitude current pulse; then the pulse level is stepped down and the voltage across the device is sampled. High-low pulses with variable low level are applied to the device and then the holding point is identified from an examination of $V_{low}(I_{low})$. This is a relatively slow measurement technique for finding the holding point.

Transient I-V curves have been obtained using wafer-level HBM testers [42]–[44]. Like multi-level TLP, this technique provides a means to obtain the holding point and, furthermore, it characterizes the device's response to a gradually decaying stress current. Once the stress current falls below I_{hold}, the snapback-type device under test (DUT) returns to its high impedance off-state. Any remaining charge on the source capacitor, C_{HBM}, will charge up the DUT and the test board capacitance (C_{TB}) to a potential above V_{hold}. V_{DUT} can be much larger than V_{hold} [46] and the consequent, long-lasting overvoltage stress may significantly degrade IC reliability [46]–[49]. However, the time-constant of the decaying pulse tail is constant because C_{HBM} of a standardized HBM tester is fixed at 300pF.

In the following sections, transient I-V curves are obtained using an exponential-edge TLP tester (EETLP) [45]. With EETLP, the time constant of the decaying pulse tail can be varied, allowing potentially thorough characterization of the device's turn-off behavior. It will be shown that a model may accurately reproduce the device pulsed I-V characteristic without necessarily being able to reproduce its ESD response. Transient IV measurement techniques are thus needed to complete the model verification

process. Models that correctly capture the device transient response can be used to simulate device-tester interactions that are commonly observed in practice. The device-tester combination is modeled as a relaxation oscillator to provide a theoretical framework for understanding the simulation results.

2.2 Comparison of Pulse I-V and Transient I-V Techniques

Figure 2.1 shows the DC and pulsed I-V characteristics of a 5-V, 90-nm grounded-gate NMOS transistor (GGNMOS). Several important parameters can be extracted from these data: failure current (I_{t2}), trigger voltage (V_{t1}), and on-state resistance (R_{ON}). Each I-V point is obtained by averaging the current and voltage during a measurement window located in a timeslot during which the current and voltage across the device are fairly constant; for 100-ns pulses, the measurement window might extend from 70 to 90 ns.

One may also examine the transient response of a device to a single pulse. Figure 2.2 shows an exemplar voltage waveform, $V_{DUT}(t)$, obtained using VFTLP at a 125-mA current level. The voltage across the device quickly snaps back after an initial spike. It is important to note that the data of Figure 2.2 cannot be used to obtain the value of V_{t1}. The device turns on extremely quickly and very few voltage samples can be captured by the oscilloscope during this interval, practically ensuring that the maximum point will be missed, even using a 12 GHz oscilloscope. This is why the maximum V_{DUT} seen in Figure 2.2 is less than the VFTLP V_{t1} plotted in Figure 2.1. In contrast, SCR-based ESD protection devices turn on slowly and transient voltage overshoot above the DC V_{t1} often is observed. If any such overshoot takes place in these GGNMOS devices, it occurs on a timescale too short to observe.

The components of a TLP or VFTLP system are often matched to the 50-Ω transmission line impedance to preserve the pulse integrity [36]; however, the 50-Ω load-line of the system limits its ability to pinpoint the holding point (V_{hold}, I_{hold}). A pulsed I-V curve represents a device's quasi-static response. The device transient response must be obtained from AC or transient measurements. In principle, one may capture the device waveform during pulse testing, especially VFTLP. The pulses produced by these

testers have a cleaner rising edge than falling edge, and thus one would attempt to capture the device turn-on transient. However, some devices turn on so rapidly in response to an incoming pulse that an accurate measurement cannot be made [9],[10]. It may be easier to capture a device's transient turn-off response, since the current sourced from the tester is greatly reduced. As indicated in the section 2.1, this chapter uses EETLP to obtain transient I-V curves. A schematic representation of the EETLP tester is shown in Figure 2.3. It first produces a square pulse, similar to TLP, followed by an exponentially decaying edge whose decay time constant is set by the value of the termination capacitor (C_{term}). Current pulses produced by the EETLP tester may be seen in Figure 2.4(a). The GGNMOS whose quasi-static I-V is shown Figure 2.1 was subjected to EETLP pulses with a variety of decay time constants. Sample current and voltage waveforms are plotted in Figure 2.4. When the current $I_{DUT}(t)$ becomes very small, V_{DUT} is observed to "snap-up," except for the case of the 15ns decay time constant (C_{term} = 82pF). The data of Figure 2.4 are used to obtain the transient I-V curve of Figure 2.5(a); this is a plot of (V_{DUT}, I_{DUT}) measured at each time step. This figure was generated using the C_{term}=4.3nF dataset; Figure 2.5(b) provides a zoomed in view of the snap-up portion of the I-V curve. It is observed that EETLP provides a far more accurate measure of the holding point than is obtained from the TLP, VFTLP or DC IV (see Figure 2.1).

Figure 2.6 shows the EETLP measurement results for three GGNMOS devices with different drain-side silicide-block length (L_{DSBLK}). The direct consequence of increased L_{DSBLK} is an increased on-state resistance, R_{ON}. Based on this data plot, L_{DSBLK} appears not to affect V_{hold} or $V_{snap-up}$, an observation consistent with previous works [50],[51]. However, if these same devices were characterized using only TLP, a different conclusion might be reached. To illustrate, two of the same devices used for Figure 2.6 were characterized using a 50-Ω TLP tester. As shown in Figure 2.7(a), if V_{hold} is defined as the minimum voltage point on the on-branch of the pulsed I-V curve, then the two devices appear to have different values of V_{hold}. The cause is the load-line effect, illustrated in Figure 2.7(b).

One might attempt to address the discrepancy between the two holding points by instead defining the holding voltage as the x-intercept of the I-V curve on-branch. However, non-uniform turn-on among the

fingers of a multi-finger device, such as those used here, causes scatter in the I-V plot [52]–[56]; in Figure 2.7(a), the data-points are scattered among multiple on-state branches due to the stochastic nature of snapback. Given the non-smooth I-V curve, it is difficult to use linear extrapolation to find the holding point and, furthermore, extrapolation does not reveal the true holding point, especially in regards to I_{hold}. In contrast, a high amplitude pulse from a transient I-V tester will turn on all the device fingers, and all the fingers stay latched on during the falling edge until the holding point is reached. The true V_{hold} of the device can be easily extracted, as was shown in Figure 2.6.

2.3 GGNMOS Model

2.3.1 MOSFET ESD Snapback Model

A schematic representation of the GGNMOS model used in this chapter is shown in Figure 2.8(a). This model is based on earlier works [9]–[13],[23],[56]. It consists of a "core" MOS transistor model that describes the device behavior under normal operating conditions, plus a "wrapper" model that extends the model validity to high current levels typical of ESD. To avoid "double modeling" of the same effects by both the core and wrapper models, some of the parameters for the core model must be zeroed-out, specifically, those related to the source and drain resistances, body resistance, and impact ionization current. An important component of the wrapper model is the lateral NPN device, LNPN. One of the LNPN model parameters is the forward transit time τ_F, as indicated in Figure 2.8(b). This parameter cannot be obtained with any accuracy from TLP or VFTLP measurement data such as those shown in Figure 2.1 and Figure 2.2.

GGNMOS devices were characterized using a variety of DC sweeps and pulse testing, and parameter extraction was performed based on these results [9]–[13]. However, as noted previously, the value of τ_F cannot be obtained for these GGNMOS devices using the usual measurement techniques. The DC and pulse I-V characteristics were simulated, with two different values assumed for τ_F; the simulation results are shown in Figure 2.9. The simulated I-V curves show an excellent agreement with the measured I-V

curves of Figure 2.1, regardless of the τ_F value. It might seem odd that the rise-time dependent V_{t1}, i.e. the dV/dt triggering effect, can be predicted without including C_{diff} in the model or, equivalently, without knowing the value of τ_F. However, the V_{t1} value shown on a pulsed I-V curve is the voltage across a device in which the LNPN is off. Thus, at V_{t1}, C_{diff} is negligibly small and the trigger voltage is a function primarily of C_{jDB}, C_{jSB}, and R_{WELL}.

Next, the response of the GGNMOS to EETLP is simulated; results are shown in Figure 2.10 (τ_F=400ps) and Figure 2.11 (τ_F=0). In this case, the value of the transit time parameter does affect the simulation results. When τ_F is set to 0, a snap-up is predicted to occur regardless of the decay time constant, and the very fast snap-up causes the device to retrigger due to the dV/dt triggering effect. When τ_F is set to 400ps, the voltage snap-up is much smaller and the magnitude is dependent on the decay time constant, similar to the measurement results. These results show that the value of the transit time affects the device's transient behavior in the negative differential resistance region (NDR) of operation. That is, it affects the snap-back and snap-up dynamics. In actual device measurements, the effect of τ_F on snap-up will be more pronounced than its effect on snap-back, due to the smaller current available from the tester at the time of snap-up. Setting τ_F to 400ps provides a good match to the measurement data of Figure 2.4. Had the devices been characterized using only TLP or VFTLP, a model with τ_F=0 would have been erroneously judged to be correct. In the next section, it will be demonstrated that the device-tester interactions observed in practice can be simulated if the device dynamic operation in the NDR is properly modeled. Of course, proper modeling of the NDR also requires that the holding point be extracted with precision.

2.3.2 The Holding Point

The holding voltage V_{hold} can be defined formally in (2.1):

$$\left.\frac{\partial I_{DUT}}{\partial V_{DUT}}\right|_{V_{hold}} = 0 \tag{2.1}$$

The corresponding current is I_{hold}. An analytic solution to (2.2) is presented in [57]. In that chapter, the impact ionization multiplication factor M was modeled using the Miller expression [58],

$$M(V_{DUT}) = \frac{1}{1 - \left(\frac{V_{DUT}}{BV_{DB}}\right)^n} \tag{2.2}$$

where BV_{DB} is the drain-body junction breakdown voltage and n is an empirical fitting parameter. The expression obtained for V_{hold} is given in (2.3).

$$V_{hold} = BV_{DB} \times \sqrt[n]{1 - \left(\frac{\beta}{\beta + 1}\right)\left(\frac{1}{1 + \frac{kT}{q} \cdot \frac{1}{R_{BODY} I_e}}\right)} \tag{2.3}$$

Eq. (2.3) indicates that V_{hold} is a function of several important model parameters: the LNPN current gain β, the body resistance R_{BODY}, and the multiplication parameter n. A comparison between the simulated V_{hold} and the measured V_{hold} can serve as a rigorous check on the model parameters, provided that V_{hold} has been measured with precision.

Figure 2.12 shows EETLP measurement results for three GGNMOS with different channel lengths (L_{gate}). Based on the measurement data, it can be concluded that channel length affects both V_{hold} and $V_{snap\text{-}up}$. A device with a shorter L_{gate} has a lower V_{hold} because its β is higher [50], with (2.3) showing the exact functional dependency. A device with a shorter L_{gate} also has a lower $V_{snap\text{-}up}$. This is simply the result of the holding voltage being lower, as can be demonstrated using a simple charge sharing analysis [46]. This analysis results in the formula for $V_{snap\text{-}up}$ that is given in (2.4).

$$V_{snap-up} = V_{hold} + \frac{I_{hold} \cdot R_S}{1 + \frac{C_{TB}}{C_{tester}}} \tag{2.4}$$

where R_S is the tester output resistance. C_{TB} represents the parasitic capacitance adjacent to the DUT. If (2.4) is applied in the context of wafer-level testing, C_{TB} includes the probe and probe pad capacitances; for HBM testing, it represents the test board capacitance. C_{tester} is the source of any residual charge

dumped into the DUT after snap-up; for an HBM tester this is C_{HBM} (see Figure 2.13(a)), and for EETLP it is C_{term}. Equation (2.4) is consistent with the data shown in Figure 2.12. The three GGNMOS are identical except for their L_{gate}, and C_{term} was held constant in this experiment. The data in Figure 2.12 show that all three devices have the same I_{hold}. Thus, according to (2.4), $V_{snap\text{-}up}$ should be a monotonically increasing function of V_{hold} (and L_{gate}), consistent with the measurements. The devices with shorter L_{gate} also seem to have a longer on-time; this is because it takes more time for C_{term} to discharge through R_S to reach the lower V_{hold}.

2.4 Investigation of Device-Tester Interactions Using Circuit Simulation

Circuit simulation, enabled by well verified compact models, can be used to investigate potentially harmful device-tester interactions. Device-tester interactions are first analyzed and then simulated in the following sections.

2.4.1 Overvoltage Stress due to Snap-up

Voltage snap-up may occur during the decaying pulse tail of an HBM event or an EETLP stress [9],[10],[46],[59]. Eq. (2.4) shows that the larger the tester source capacitance C_{tester}, the larger the snap-up. Also, the larger the tester source resistance (R_S), the larger the snap-up. This second observation explains why a much smaller $V_{snap\text{-}up}$ is observed during EETLP testing (R_S=50Ω) than during HBM testing (R_S=1.5kΩ). During HBM testing, upon snap-up, V_{DUT} can become much larger than V_{hold}. The consequent, long-lasting overvoltage stress may significantly degrade IC reliability. This is one of the causes of miscorrelation between HBM and TLP results [43],[46]–[61].

However, the validity of (2.4) is limited because it does not address the case when C_{TB} is charged to V_{t1} before the charge sharing process is completed; in this case, the snapback device may be retriggered into its low impedance state one or more times. This phenomenon has been observed during HBM ESD testing [59]. The equivalent source resistances for HBM and system-level ESD testers are 1.5kΩ and 330Ω, respectively. Both are much larger than the R_S=50Ω of EETLP, and thus these testers produce

$V_{snap\text{-}up}$ that can easily exceed V_{t1}. The discharge of an HBM tester into a GGNMOS protection device was simulated; the results are shown in Figure 2.13, and retriggering is clearly visible. Even though this phenomenon takes place when I_{HBM} has decayed to a low level [59], it is hazardous since there is minimal resistance to limit the discharge current from C_{TB} into the DUT upon retriggering. C_{TB} is charged up to V_{t1}, and the burst of discharge current can have a peak value large enough to cause damage [61]. Pre-Si circuit simulation may be used to identify whether testing induced overstress will be problematic for a given design.

2.4.2 Relaxation Oscillator

Oscillation may occur during HBM testing and this has been reported to result in unexpected early failure [43],[46]. These oscillations are the result of the interaction between the tester impedance and the snapback device's S-shape I-V characteristic [9],[10] and may be understood by realizing that the tester and snapback device together form a relaxation oscillator [62]. As shown in Figure 2.14, the load portion of a relaxation oscillator consists of a device with an S-shape I-V characteristic in parallel with a capacitor, identified here as the test-board capacitance C_{TB}. The driving source consists of a power supply V_{SUPPLY} and a source impedance R_S. Note the similarity between the HBM tester model (Figure 2.13(a)) and the relaxation oscillator. Also note that V_{SUPPLY} can be more generally a time-varying supply, e.g., an HBM tester; however, here a constant supply is assumed to simplify the analysis.

The basic operation of the relaxation oscillator may be understood by making the following, simplifying assumptions.

1) $R_S \gg R_{ON} \approx 0$

2) The oscillation frequency is sufficiently low so that the snapback device will follow its static I-V curve.

3) An initial perturbation, such as a power-up ramp or noise glitch, is present to kick-start the oscillation.

The operation of a relaxation oscillator can be divided into two states: Charging and Discharging.

a) Charging:

I_{SUPPLY} charges up C_{TB} from V_{hold} to V_{t1} while the snapback device remains off. The time required to complete the charging is defined as T_{charge}, expressed by (2.5).

$$I_{SUPPLY} = \frac{V_{SUPPLY} - V_{DUT}}{R_S} = C_{TB}\frac{\partial V_{DUT}}{\partial t} \tag{2.5}$$

$$T_{charge} = R_S C_{TB} \cdot \ln\left(\frac{V_{SUPPLY} - V_{hold}}{V_{SUPPLY} - V_{t1}}\right) \tag{2.6}$$

b) Discharging

C_{TB} is discharged from V_{t1} to V_{hold} through the on-state resistance (R_{ON}) of the triggered-on snapback device. The time required to complete the discharging is defined as $T_{discharge}$, expressed by (2.8).

$$-\frac{V_{DUT}}{R_{ON}} = C_{TB}\frac{\partial V_{DUT}}{\partial t} \tag{2.7}$$

$$T_{discharge} \approx R_{ON} C_{TB} \cdot \ln\left(\frac{V_{t1}}{V_{hold}}\right) \tag{2.8}$$

The oscillation period (T) can be defined as

$$T = T_{charge} + T_{discharge} \tag{2.9}$$

According to this analysis, the snapback device turns on when V_{DUT} reaches its static V_{t1} and turns off when V_{DUT} falls below its static V_{hold}. Therefore, V_{DUT} swings in a periodic steady-state between V_{t1} and V_{hold}.

The assumed, very small value of R_{ON} implies that $T_{discharge}$ is much shorter than T_{charge}. One may also reasonably assume that V_{SUPPLY} is much larger than V_{t1} and V_{hold}, since a typical HBM pre-charge voltage will be one or more orders of magnitude larger than V_{t1} and V_{hold}. Making use of these observations and also the mathematical relation: $\ln(1-x) \approx -x \text{ if } x \to 0$, the period T can be approximated by the formulation in (2.10).

$$T \approx \frac{C_{TB} \cdot (V_{t1} - V_{hold})}{I_{SUPPLY}} \tag{2.10}$$

According to (2.10), the oscillation frequency (1/T) is proportional to the supply current I_{SUPPLY}. If I_{SUPPLY} is large, the oscillation frequency would be too fast for an actual snapback device to behave statically. Minority charge storage, such as the base charge in a GGNMOS, prevents it from turning off instantly even if the externally supplied current is removed. Therefore, the device may not turn off when V_{DUT} reaches the static V_{hold}, suggesting that the oscillating V_{DUT} can dip below the static V_{hold}. Moreover, if I_{SUPPLY} is sufficiently large, the device will not be completely turned off before being driven back above V_{hold}. Consequently, oscillation ceases and the device stably operates in the on-state.

The analysis indicates that oscillation will occur only at low values of I_{DUT}, consistent with the measurement data in [43],[46],[59]. Simulation is used to verify the analysis. Figure 2.15(a) shows the simulation set-up, which is consistent with the circuit model used to derive (2.10). Simulation shows that the system exhibits sustained oscillation for sufficiently low I_{SUPPLY}, Figure 2.15(b). For a higher I_{SUPPLY}, Figure 2.15(c), the oscillations are damped and the system self-stabilizes. Ringing disappears entirely when I_{SUPPLY} is raised further. The oscillation/ringing frequency as a function of I_{SUPPLY} is plotted in Figure 2.15(d) and matches well to the analysis presented above.

The simulation results of Figure 2.13 show that ringing occurs in the early part of the HBM discharge; experimental data show a similar behavior when the tester pre-charge voltage is low, as in this simulation [59]. These oscillations are not due to the parasitic inductance of the tester. This assertion is confirmed by the simulation results shown in Figure 2.16. Two HBM testers were simulated, one with parasitic inductance (L_S=12μH) and the other without, while all the other tester components and the DUT are kept the same. Simulation shows that V_{DUT}(t) oscillates in both cases, with virtually identical swing magnitude and ringing frequency. There is a time shift between the two waveforms because C_{HBM} must discharge through L_S which introduces a delay in the current waveform.

2.4.3 Instability during Turn-off in HBM Testing

Multiple on-off switching events have been observed during HBM testing [43],[46]. These, too, may be understood as a result of the interaction between the tester and the ESD protection snapback device. Simulations indicate that this oscillatory behavior is triggered by noise, which will always be present in an electronic system. In the simulation setup of Figure 2.17, a pulse current source is used to model noise. The initial pre-charge voltage on C_{HBM}, 300V, produces I_{HBM} values that are too high during the first 200ns of the HBM event for even a strong perturbation to initiate oscillation. However, if a noise pulse is injected when I_{HBM} is small, oscillation may occur, depending on the value of the DUT holding current.

To illustrate this point, two GGNMOS devices with the same (V_{t1}, I_{t1}, V_{hold}) but different I_{hold} were simulated using the HBM simulation setup of Figure 2.17. The HBM tester pre-charge voltage is the same in both simulations, as is the timing and the amplitude of the noise pulse. The simulation results are shown in Figure 2.18. The device with the lower I_{hold} remains stable, while oscillations are triggered in the one with the higher I_{hold}. The amplitude of the V_{DUT} oscillations grows in a regenerative fashion until the device undergoes repetitive, complete switching from on to off; similar behavior is displayed in measurement data presented in Figure 11 of [43]. In the simulations of Figure 2.18, V_{DUT} is seen to swing outside the limits [V_{hold}, V_{t1}]; the preceding analysis established that this is due to minority charge storage in the device, further highlighting the need to extract the device transient parameters (τ_F for a GGNMOS).

To initiate oscillation, the noise pulse must occur within a certain time window; simulation shows that this time window is wider for a device with high I_{hold}. This is a consequence of the high holding current device being able to oscillate over a wider range of I_{HBM} values. If high and low holding current devices are simulated using the relaxation oscillator model of Figure 2.14, a similar conclusion is reached, namely, that the high holding current device will oscillate over a wider range of I_{SUPPLY} values. Thus it is concluded that high holding current devices are more likely to experience multiple retriggering events during ESD testing than are devices with low holding current.

2.5 Summary

TLP and VFTLP measurement data may not be sufficient to fully evaluate the correctness of an ESD device compact model. This is especially true for snapback-type devices that have a very fast turn-on—too fast to fully resolve even using VFTLP. Turn-off transients occur at low current levels and are easier to resolve in measurements; these transients reveal themselves during HBM and EETLP testing. Furthermore, transient I-V data obtained from HBM or EETLP testing provide a far more accurate measure of the device holding point than may be read from the pulse I-V curve obtained using a 50-Ω TLP tester. For all of these reasons, it is recommended that model development and verification be performed using transient I-V data in addition to TLP/VFTLP data.

GGNMOS compact models that are tuned to provide the correct holding current and transient response were shown to enable predictive circuit simulation of the interaction between the protection circuitry and the ESD test environment. Measured HBM waveforms have been presented in the open literature; in some of these waveforms, large voltage oscillations are observed. This phenomenon can be explained by modeling the HBM tester and snapback-type ESD clamp as a relaxation oscillator. This physical picture reveals that the ESD clamp's holding current and its internal charge storage determine its susceptibility to test-induced oscillations, and further emphasizes the need to characterize ESD protection devices with techniques other than just square pulse testing.

2.6 Figures

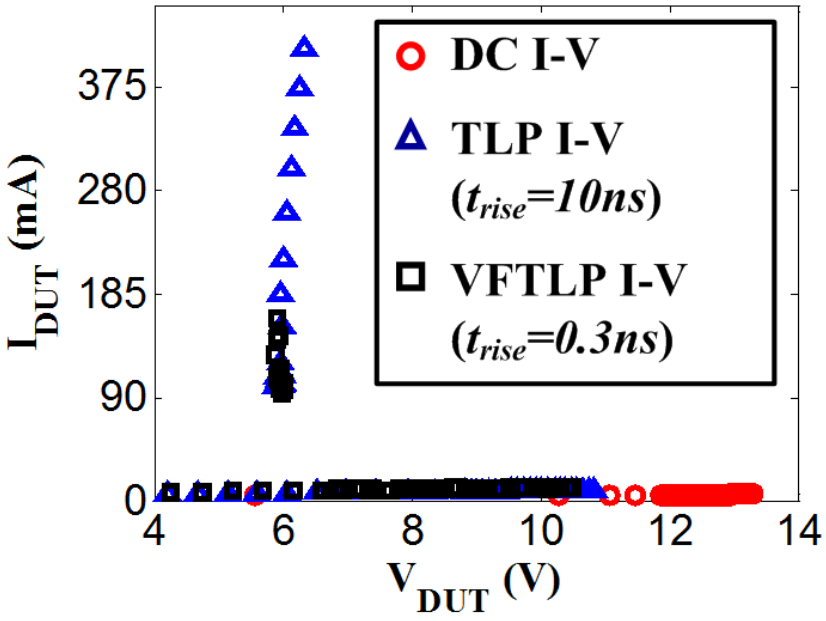


Figure 2.1. DC, TLP and VFTLP I-V curves. TLP pulse-width is 100ns; VFTLP is 25ns. The device-under-test is a 5V GGNMOS in a 90nm CMOS process; W/L = 100µm/0.72µm. The DUT failed upon snapback (13.3V) during DC IV sweep; It2 measured during TLP was 410 mA. The VFTLP IV sweep was not taken up to the failure point.

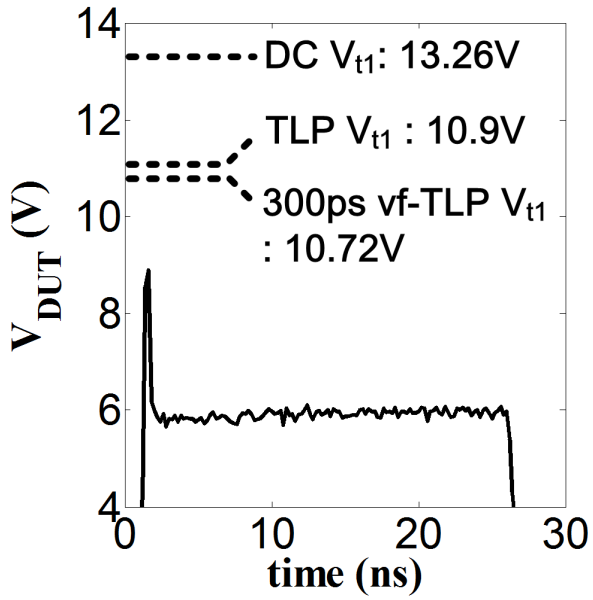


Figure 2.2 $V_{DUT}(t)$ from VFTLP at a current level I_{DUT}=125mA for the same device as in Figure 2.1. Data are obtained using a 12GHz digital sampling oscilloscope.

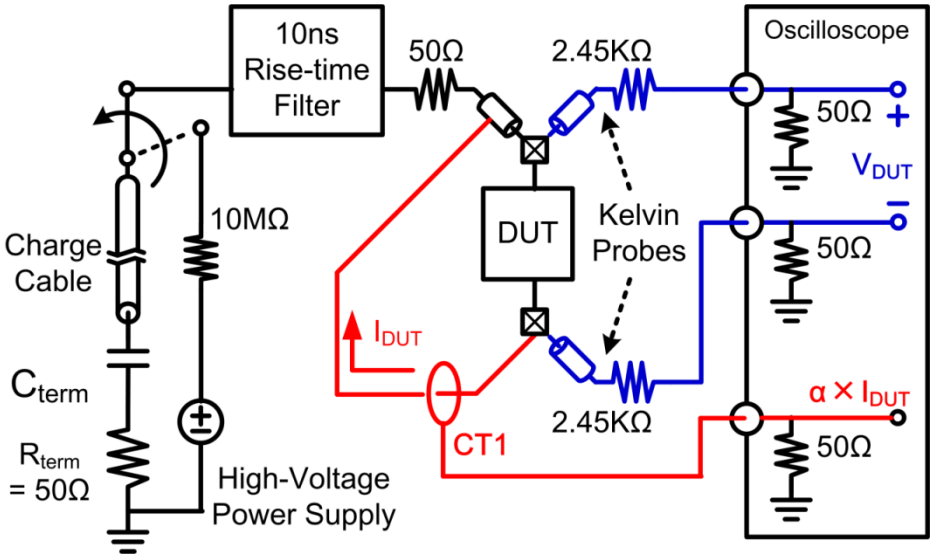


Figure 2.3 EETLP tester schematic [45]. The charge cable length defines the duration of the flat part of the pulse where $I_{DUT}(t)$ is at steady-state. The pulse falling edge decay time constant is equal to $C_{term}\cdot(R_{term}+Z_o) = C_{term}\cdot 100\Omega$.

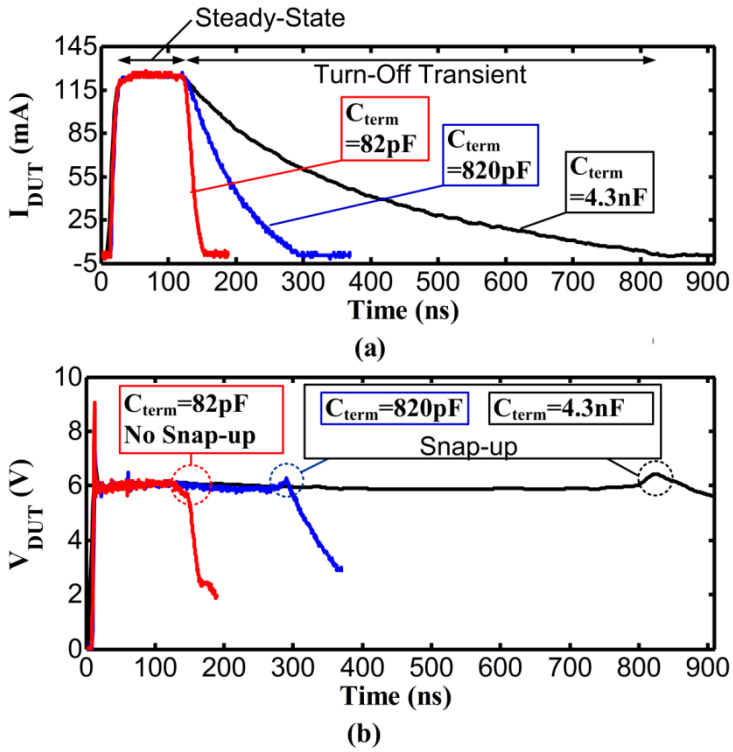


Figure 2.4 EETLP measurement results. (a) $I_{DUT}(t)$ and (b) $V_{DUT}(t)$. The device-under-test (DUT) is a 5V GGNMOS in a 90nm CMOS process; W/L = 100μm/0.72μm. The snap-up voltage $V_{snap\text{-}up}$ is observed to be an increasing function of C_{term}.

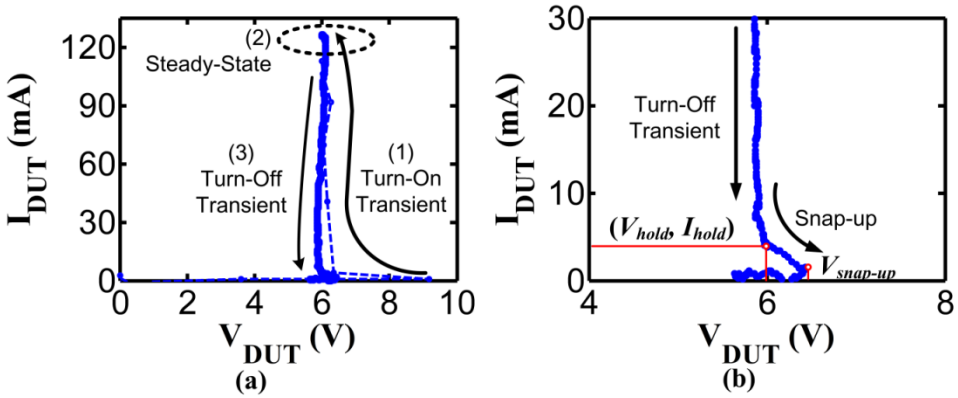


Figure 2.5 Transient I-V curve extracted from the dataset labeled "C_{term}=4.3nF" in Figure 2.4. (a) The entire I-V curve. (b) Zoomed-in view near the holding point (V_{hold}, I_{hold}).

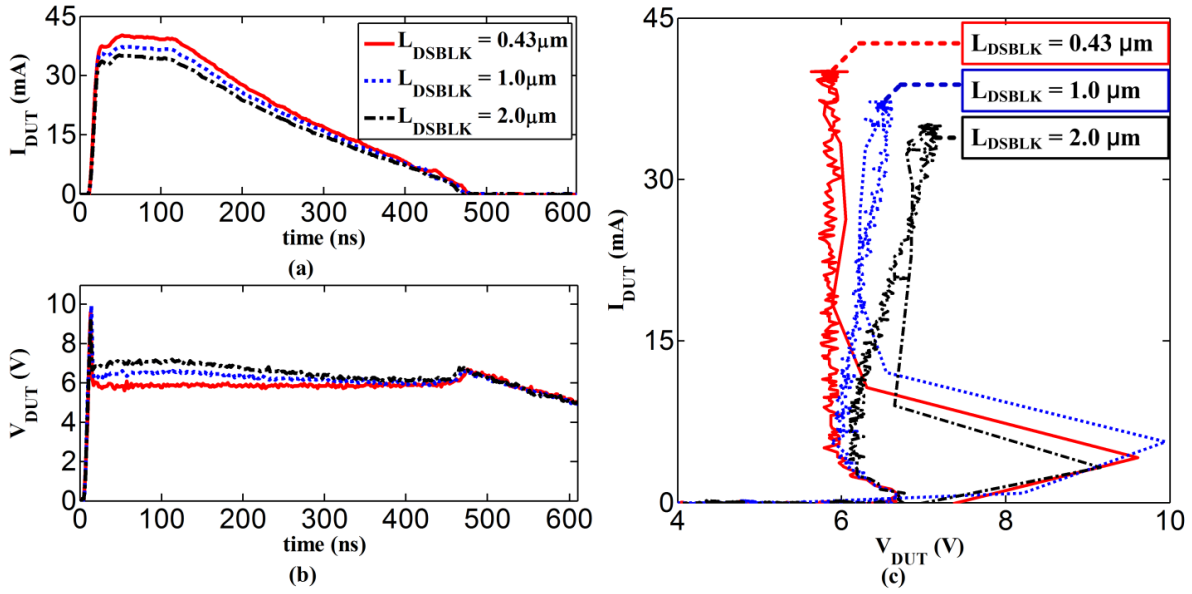


Figure 2.6 EETLP measurement results for 3 GGNMOS, each with a different drain silicide-block length (L_{DSBLK}). Each device has 10 fingers; each finger has W/L of 10µm/0.72µm. C_{term}=4.3nF. (a) $I_{DUT}(t)$. (b) $V_{DUT}(t)$. (c) Transient I-V curves.

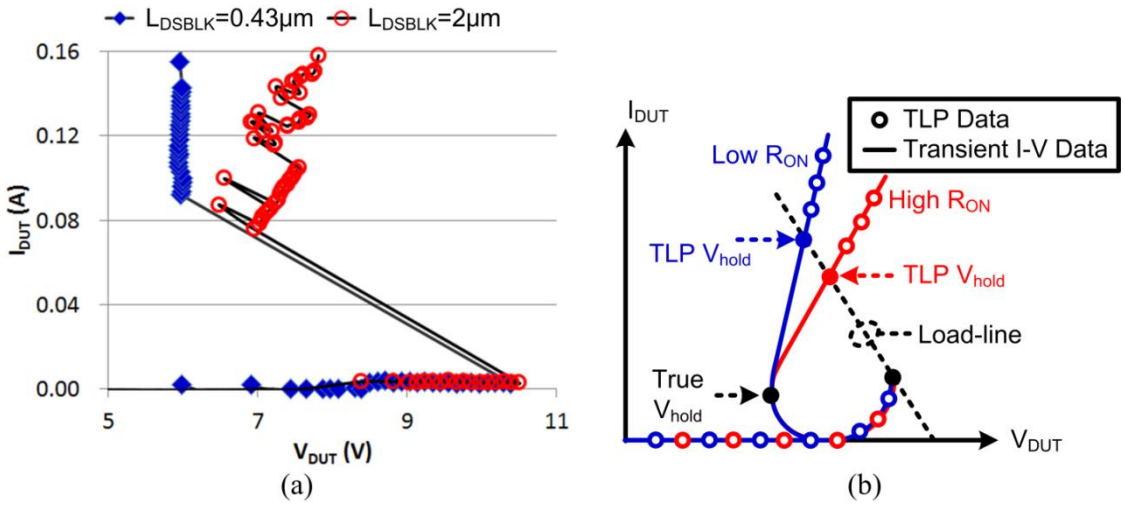


Figure 2.7 (a) Measured TLP I-V for the same GGNMOSs used in Figure 2.6. Both devices have 10 fingers, and the zig-zag I-V curve results from non-uniform turn-on among the fingers. (b) TLP system load-line prevents accurate extraction of V_{hold} from a quasi-static I-V curve; here, the device with higher R_{ON} falsely appears to have higher V_{hold}.

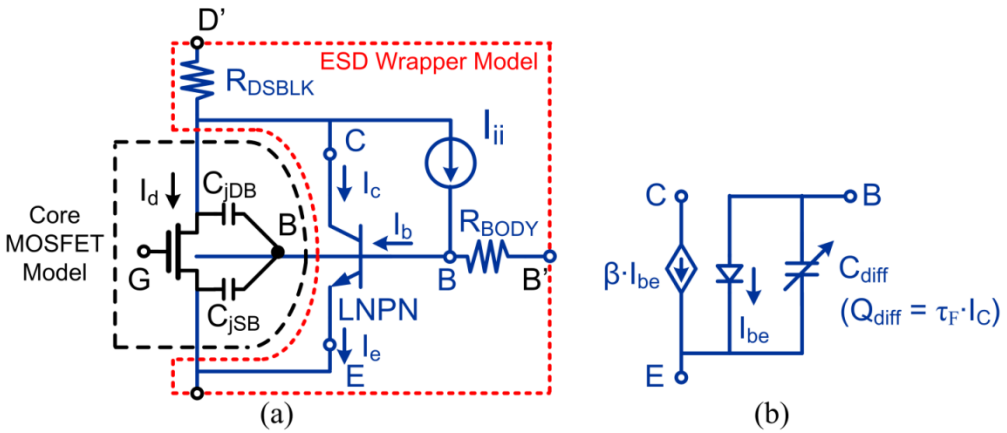


Figure 2.8 (a) MOSFET ESD model. As the devices used in this work contain a silicide-blocked drain ballasting resistance, the device series resistance is lumped into a single term on the drain-side R_{DSBLK}. (b) Implementation of the LNPN in the wrapper model. In this work, a uni-directional representation of the LNPN is used, since the GGNMOS were operated in just one polarity.

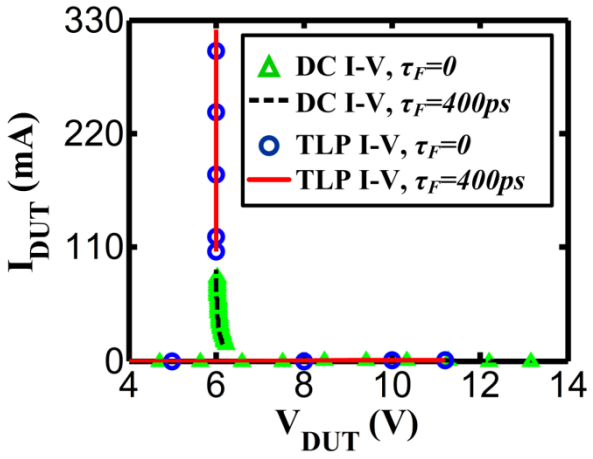


Figure 2.9 Simulated GGNMOS I-V. Both the DC I-V and the pulsed I-V match well to the measurements results of Figure 2.1, regardless of the τ_F value. The TLP I-V curves are constructed by simulating a pulse voltage source (10ns rise-time and 100ns pulse-width) connected to the DUT through a 50Ω series resistance. The DC I-V is simulated using a DC voltage sweep with high source impedance (500Ω).

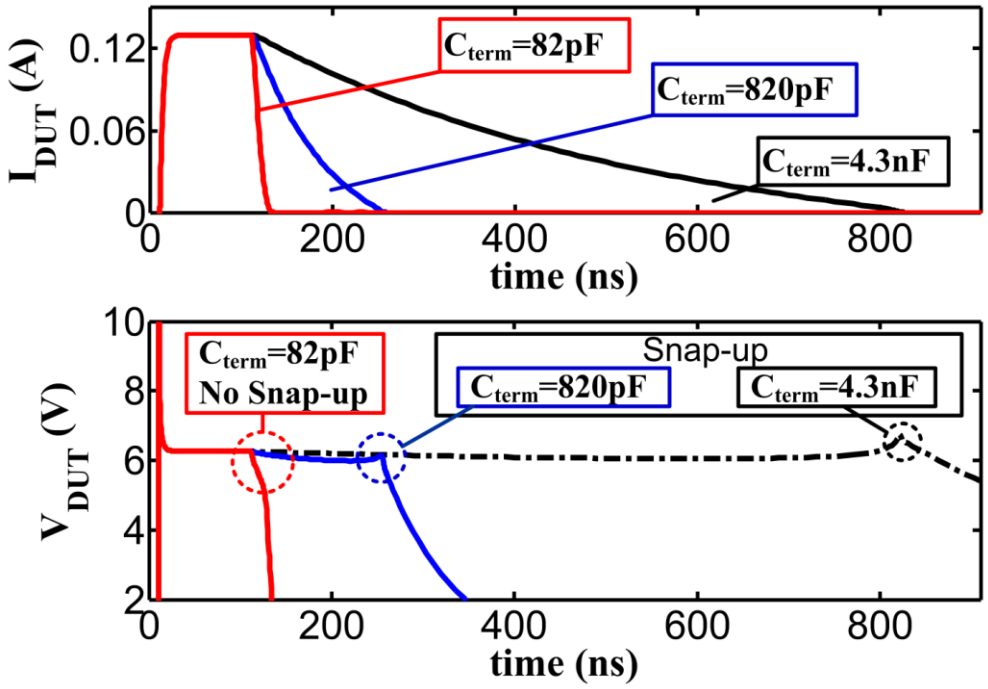


Figure 2.10 Simulated response of the GGNMOS to EETLP. The transit time parameter is to set τ_F=400ps. The device snap-up behavior well matches the measurement results of Figure 2.4. The transient V_{t1} from this simulation is a bit lower than that shown on the device pulsed I-V curve (Figure 2.1 and Figure 2.9); this is because the tester pre-charge voltage is larger in this case and thus so is the dV/dt triggering effect. At the same time, the transient V_{t1} seen in this simulation is larger than that obtained from a single pulse measurement (e.g. Figure 2.2), because GGNMOS snapback is too fast to be fully resolved by the 12 GHz oscilloscope.

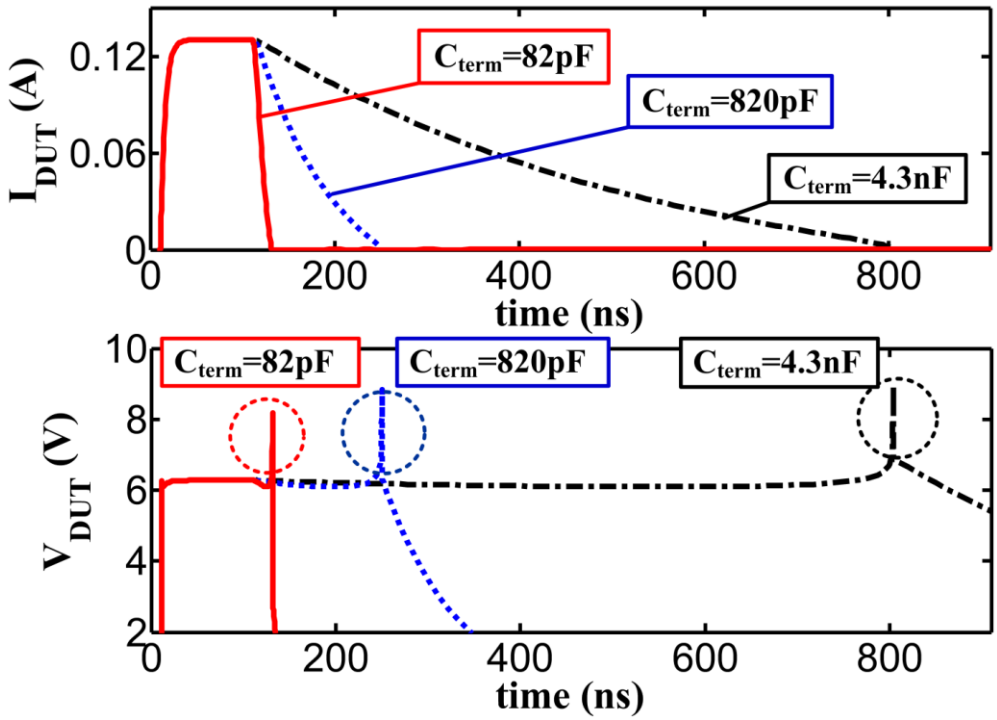


Figure 2.11 Simulated response of the GGNMOS to EETLP. The transit time parameter is set as τ_F=0. Voltage snap-up always occurs and is exaggerated, regardless of the value of C_{term}. This is contrary to the measurement results. Also note that the device is incorrectly predicted to turn on without any snapback because the exclusion of C_{diff} from the model allows the DUT to change between OFF- and ON-states instantaneously.

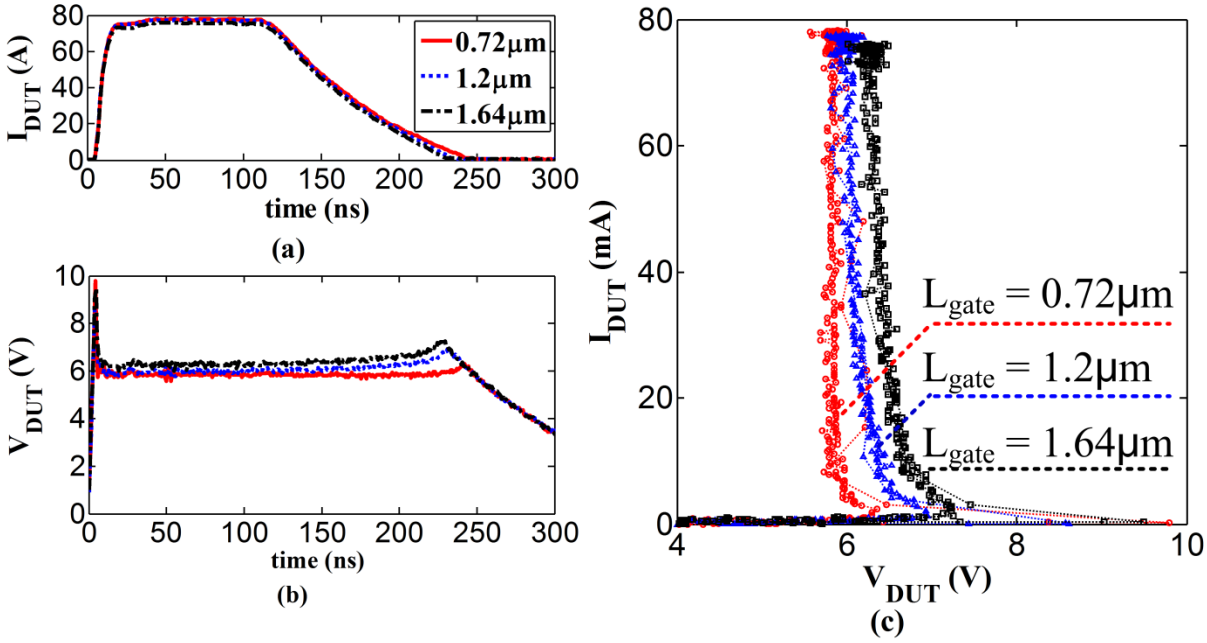


Figure 2.12 Three 5V GGNMOS of 100µm channel width in 90nm CMOS process were characterized using EETLP with C_{term}=820pF. Three devices with different channel length (L_{gate}) are tested: L_{gate}=0.72µm, 1.2µm, and 1.64µm, respectively. (a) $I_{DUT}(t)$. (b) $V_{DUT}(t)$. (c) Transient I-V curves.

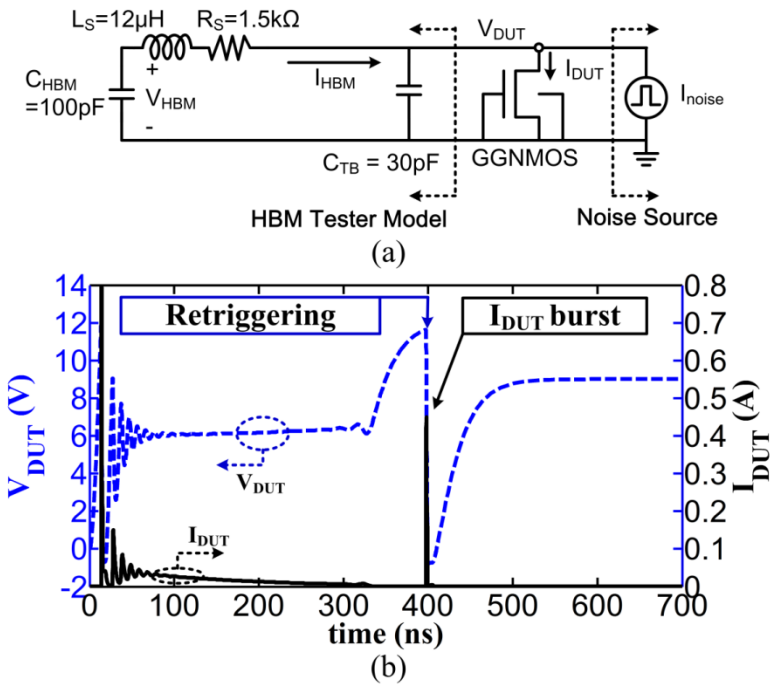


Figure 2.13 (a) Simulation set-up for an HBM tester with a GGNMOS DUT. The initial pre-charge voltage on CHBM in this simulation is 80V. (b) Simulated $V_{DUT}(t)$ and $I_{DUT}(t)$. After the GGNMOS DUT first turns off, C_{TB} gets charged up to V_{t1} and the GGNMOS is triggered a second time. C_{TB} then dumps its charge into the GGNMOS, resulting in a sudden burst of current. Very similar behavior is seen in the measurement data presented in Figure 2.14 of [59]. Notice that ringing occurs in the early part of the HBM discharge, ~10−50ns. This damped oscillatory behavior results from a device-tester interaction described in subsection 2.4.2.

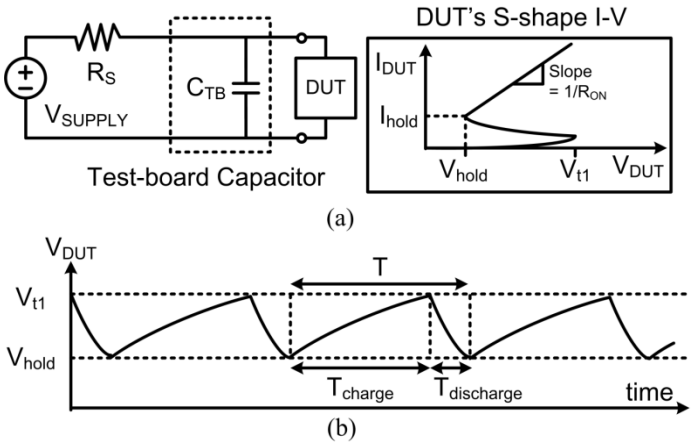


Figure 2.14 (a) Schematic representation of a relaxation oscillator [34]. (b) Corresponding $V_{DUT}(t)$. Notice the resemblance of (a) to an HBM tester. For oscillation to occur, C_{TB} must be present and the tester output resistance R_S must be large. Note that the inductor L_S in the HBM tester model (Figure 2.13 and Figure 2.17) is not a necessary component of a relaxation oscillator.

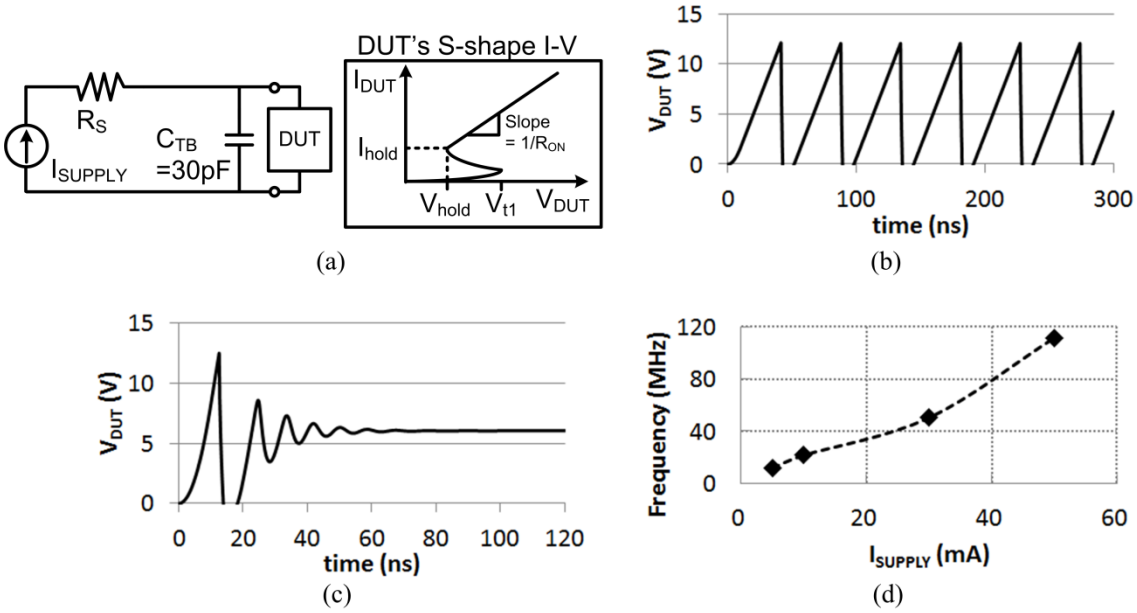


Figure 2.15 (a) Simulation model of a relaxation oscillator. (b) Simulated $V_{DUT}(t)$, I_{SUPPLY}=10mA. (c) Simulated $V_{DUT}(t)$, I_{SUPPLY}=50mA. (d) The oscillation frequency as a function of I_{SUPPLY}. The simulation results are consistent with eq. (2.10) until I_{SUPPLY} reaches 42mA. Above that current level, the system is stable.

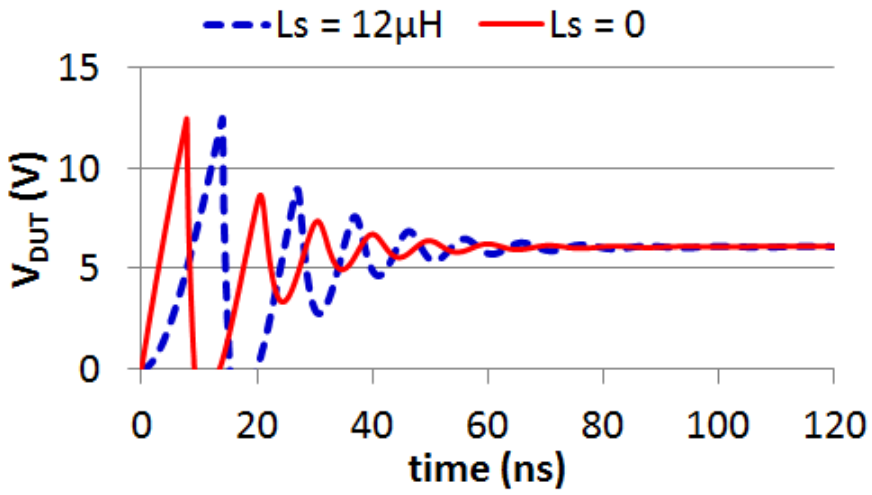


Figure 2.16 Simulated $V_{DUT}(t)$ under HBM testing with different tester parasitic inductance L_S. There is no artificial noise source in the simulation, so the oscillatory response is inherent to the system. Aside from the time shift, the two simulated $V_{DUT}(t)$ are almost identical in terms of swing and frequency for their oscillatory response at the beginning. This result slows that the initial instability is governed by the theory of relaxation oscillation. The time shift merely comes from the fact that the rate of C_{HBM} discharges through L_S is different because L_S is different.

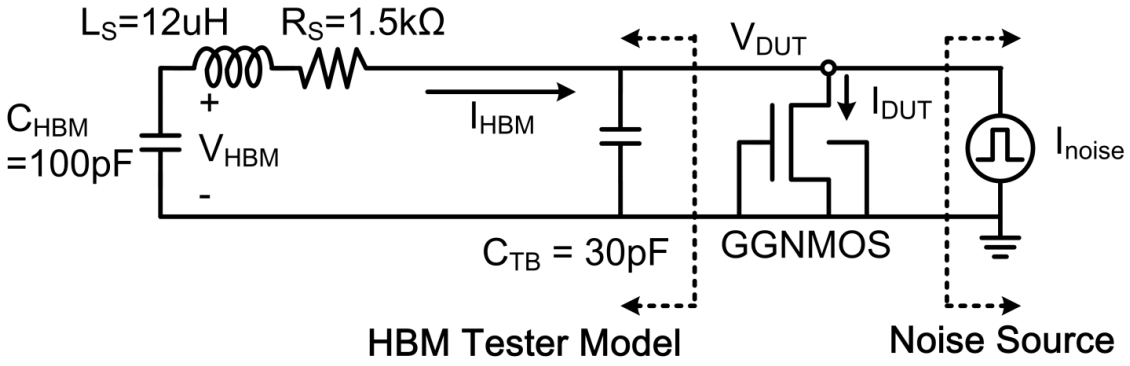


Figure 2.17 Simulation set-up used to illustrate instability during HBM testing. The initial pre-charge voltage on C_{HBM} is 300V. The current source I_{noise} represents noise in the system. Two GGNMOS devices with same (V_{t1}, I_{t1}, V_{hold}) but different I_{hold} (5mA and 10mA) are simulated.

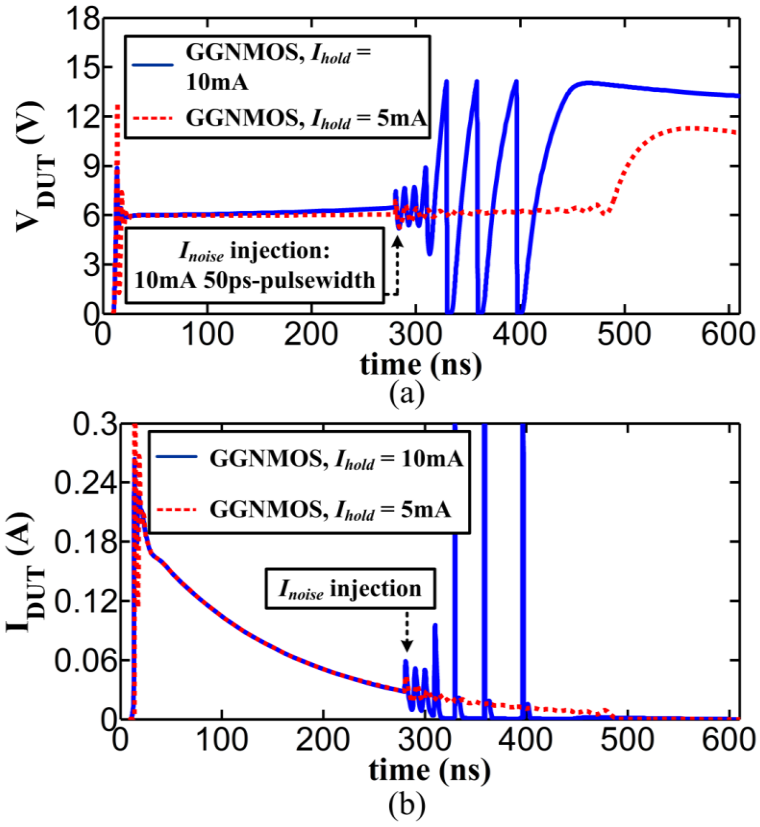


Figure 2.18 (a) Simulated $V_{DUT}(t)$ and (b) $I_{DUT}(t)$ for the set-up of Figure 2.17. In (b), a zoomed-in view is provided. The noise pulse induces oscillations only in the device with the higher holding current. Relative to Figure 2.13, there is very little ringing at the beginning of the HBM event, due to the higher pre-charge voltage and thus higher initial I_{HBM} in this case.

Chapter 3. Distributed Multi-finger MOSFET ESD Models

This chapter presents a model for multi-finger MOSFETs operating under ESD conditions. It is a distributed model that can reproduce the effect of layout geometry on trigger voltage, on-state resistance, and non-uniform turn-on of device fingers [11]. A three-terminal transmission line pulsing technique enables model parameter extraction. Analysis of measurement data and TCAD simulation reveals that self-heating is not uniform across the device, and this affects the relation between on-state resistance and the number of fingers [12]. With self-heating incorporated, the model correctly reproduces the device I-V curve up to high current levels.

3.1 Motivation

It is customary to characterize the snapback operation of a MOS device utilized as a snapback clamp; however, all MOSFETs that may be exposed to ESD should be characterized to prevent unexpected breakdown [63]. Transistors exposed to ESD include those used in I/O circuits and active rail clamp circuits. A multi-finger MOSFET layout style is usually employed both for protection devices and for transistors in output drivers. The device fingers generally do not get triggered into snapback all at the same current level, and the non-uniform triggering can result in ESD performance degradation [52],[53],[64],[65].

EDA tools can assist the design of on-chip ESD protection networks, but this requires the availability of device compact models that extend to the ESD operating region. Many ESD-relevant models for MOSFETs may be found in the literature, e.g., [13]–[24]. However, the majority of these works treat the body resistance as a single element, in effect representing the device as a single-finger MOSFET, and the resultant model does not capture non-uniform triggering among the fingers. Only a limited effort has been made to model non-uniform triggering by constructing a distributed model of the body resistance [65], and without providing a robust methodology for parameter extraction or model verification.

This chapter demonstrates a methodology to characterize and model multi-finger ESD-MOSFETs. The resulting compact model is scalable with respect to device layout geometry and accurately reproduces non-uniform triggering of the device fingers under ESD conditions. This chapter presents the model for the distributed body resistance network, and additionally focuses on the modeling of self-heating in multi-finger devices.

3.2 Model

Figure 3.1 illustrates the usual layout of a multi-finger MOSFET. Based on the layout geometry, a distributed model is proposed, similar to that in [65]. An illustration of the distributed model is provided in Figure 3.2, in which the parts of the model representing the well resistances are shown separately from the transistor part of the model for improved visibility. The resistor mesh connecting the body terminals of all fingers is defined as the R_{body} network.

Each finger of the transistor is represented using the ESD MOSFET model shown in Figure 3.3 [9],[10],[13],[23]; the model represented by Figure 3.3 will henceforth be referred to as the "finger model". The channel current (I_{ds}) is provided by the MOSFET model from the process design kit (PDK) but with certain terms zeroed out, such as those related to impact ionization [66],[67], because the PDK model is inaccurate at high current levels. An ESD "wrapper" model is used to extend the model to ESD current levels by the addition of a parasitic lateral bipolar transistor (LNPN) and an impact ionization current source (I_{ii}). The ESD wrapper model also includes the drain and source series resistances (R_D and R_S) which may be large due to silicide-blocking. The parameters for the ESD wrapper model are extracted at current levels typical of those during ESD.

The R_{body} network of well resistance in Figure 3.2 retains the spatial dependency of each finger's local body potential. Each of the four unique resistors in Figure 3.2 is assumed to have the geometry illustrated in Figure 3.4. The resistance of each element, R_{eq}, is modeled as

$$R_{eq} = R_{\square} \cdot \frac{L_{eff}}{W_{eff}} \tag{3.1}$$

where $R_{\square}$ denotes the sheet resistance (of the P-well, in this case). L_{eff} and W_{eff} are the effective length and width of the resistor, respectively.

A distributed model is used for the well resistance to retain the spatial dependency of each finger's local body potential. Each of the four unique resistors in Figure 3.2 is assumed to have the geometry illustrated in Figure 3.4. Empirical, layout-dependent expressions for the four unique elements shown in Figure 3.2 are given in (3.2)–(3.5). The functions used to model the R_{body} network resistors are inspired by those used previously to model trapezoid-shaped resistors [22] and spreading resistors [23].

$$R_{side} = R_{\square} \cdot \frac{s_S + d_{D,edge}}{W_{eff,side}} \tag{3.2}$$

where $W_{eff,side} = w_f + 2s_d$.

$$R_{finger} = R_{\square} \cdot \frac{s_D}{W_{eff,finger}} \tag{3.3}$$

where $W_{eff,finger} = L_f \left[1 + \ln\left(\frac{w_f}{2 \cdot w_{f0}}\right)^{\gamma}\right] + d_D + d_S$.

$$R_{base,S} = R_{\square} \cdot \frac{2 \times d_S}{w_f} \tag{3.4}$$

$$R_{base,D} = R_{\square} \cdot \frac{2 \times d_D}{w_f} \tag{3.5}$$

These equations provide a good global fitting of the measurement data to the simulation data for a wide range of layout geometries, especially in regards to the trigger voltage (V_{t1}) and on-state resistance (R_{ON}). Among all the layout geometric variables, N_f (number of fingers) and w_f (width of each finger) deserve the most attention. The well tap spacings, s_D and s_S, are often fixed at their minimum allowed values to maximize device density. The drain/source silicide-blocking lengths, d_D and d_S, are generally set

to the pre-determined values that maximize I_{t2} per layout area. Therefore, this chapter focuses on N_f and w_f, the primary design variables.

Due to limited test-chip area, only a small number of test structures are dedicated to explore the dependence of R_{on}, V_{t1}, and V_{hold} on s_D and s_S. With fixed nominal w_f=20μm, measurement shows that the I-V characteristic has no quantifiable dependence on typical s_D and s_S ranging from 1 to 5μm; device-to-device variation masks any observable pattern on V_{t1} or V_{hold}. Also, there is no test structure to quantify the effect of d_D and d_S because the given (d_D, d_S) is suggested to be optimal by the PDK for ESD MOSFET. However, compared to (s_D, s_S), the effect of (d_D, d_S) may deserve further study, as they are also the design variables for on-resistance and turn-on uniformity.

3.3 Model Parameter Extraction

3.3.1 Measurement Setup

Previous works have addressed parameter extraction for the finger model [9],[10],[13]–[23]. This chapter focuses on the extraction of the model parameters for the body resistance network. The test structures used are multi-finger NMOS with various layout geometries, fabricated in a 130nm CMOS process. All test devices are 1.2V thin-oxide NFET with 120nm gate length. Model parameters are extracted from pulsed I-V data. Pulse measurements allow the DUT to reach ESD current levels; under DC test conditions, thermal or gate dielectric breakdown would occur at an insufficient current level. A transmission line pulsing (TLP) system [32] is used to produce pulses with 100ns width and 10ns rise-time. The transmission line that delivers the pulse is connected between the drain (D) and source (S) terminals. To extract ESD characteristics as functions of gate bias, 3-terminal test structures with probe pads at the D, S, and gate (G) terminals are used, and a gate biasing probe is connected, as shown in Figure 3.5(a). To extract parameters for the R_{body} network, test structures equipped with probe pads at their D, S and body (B) terminals are characterized using the three terminal TLP set-up (3T-TLP) shown in Figure 3.5(b). The 3T-TLP system allows for body current (I_B) measurement. A current probe samples

I_{DUT}—the total current in the device-under-test ("DUT")—as it enters terminal D. The probe contacting terminal B is connected to the transmission line shield via a short jumper wire. The jumper is a 3-cm litz wire which passes through a second current probe that measures I_B. Kelvin probing is used to measure the potential difference (V_{DUT}) between nodes D and S.

3.3.2 R_{body} Network Component Extraction

Figure 3.6 shows the quasi-static I_{DUT}-V_{DUT} curve obtained from the 3T-TLP measurement on a 6-finger GGNMOS. The iconic zigzag shape of the curve results from a varying number of fingers triggering at different current levels. There is some overlap between the branches of the zigzag I-V curve because the number of fingers that are triggered by a single pulse is somewhat non-deterministic, presumably due to the carrier multiplication process. Figure 3.7 shows the corresponding I_B-I_{DUT} curve, which similarly contains multiple branches.

Most of the model parameters are obtained from the device layout; however, the effective sheet resistance ($R_{\square}$) must be obtained from measurement. The data of I_B-I_{DUT} curves are used to extract $R_{\square}$. First, I_B is simulated, assuming that $R_{\square}$ is equal to the P-well sheet resistance value given in the PDK documentation. Then, the value of $R_{\square}$ is fine-tuned by fitting the simulated I_B-I_{DUT} curve to the measurement result. The simulated I-V curves match the measured ones only if $R_{\square}$ is modeled as a function of the current through the corresponding resistor, using the empirical expression given in (3.6).

$$R_{\square}(I_R) = R_{S0} + R_{SM} \Big/ \left(1 + \frac{I_R}{J_0 \cdot W_{eff}}\right) \tag{3.6}$$

The hole current due to impact ionization flows primarily in a thin layer close to the surface prior to snapback; it spreads deeper into the P-well afterward, thereby widening the cross-sectional area of the body resistor [19],[20],[50],[68]. This is conjectured to be the reason that $R_{\square}$ decreases when the current I_R reaches a sufficiently high level, the effect modeled by (3.6). The model parameters R_{S0}, R_{SM}, and J_0 are obtained by fitting the simulated I_B to the measured values (e.g., Figure 3.7). At the start of the

parameter optimization procedure, R_{S0} is set equal to the provided value of the P-well sheet resistance and R_{SM} is set to be one order of magnitude larger than R_{S0}.

3.4 Model Verification

In this section, it will be demonstrated that the proposed compact model well represents:

1) $V_{t1}(V_{GS})$
2) $V_{t1}(N_f, w_f)$
3) *Non-uniform turn-on of the MOSFET fingers*

3.4.1 Gate Bias Effect

Figure 3.8(a) shows the measured and simulated TLP I-V curves measured at different values of V_{GS}. Figure 3.8(b) shows the extracted V_{t1} as a function of V_{GS}. Simulation accurately represents the I-V curves and V_{t1} as functions of gate bias, except in the case that only one finger of the device is triggered on, which is elaborated on in the following sections. V_{GS} modulates the impact ionization generated body current which, in turn, provides the base current for the parasitic LNPN [13],[57]; snapback occurs when the LNPN turns on. Accurate modeling of $V_{GS}(V_{t1})$ is critical for designing gate-coupled MOSFET protection circuits and active rail clamp protection circuits [63].

3.4.2 Trigger Voltage Vt1

The trigger voltage, V_{t1}, depends on the body resistance seen by the finger that first turns on. This first finger is most likely to reside at the center of the multi-finger structure because it is farthest away from the body pick-up ring at the periphery. Figure 3.9 shows the V_{t1} values extracted from both measurement and simulation for GGNMOS with a varying numbers of fingers (N_f) and finger width (w_f). Although the measured V_{t1} varies by about 0.1V from pulse to pulse or sample to sample, the simulated V_{t1} values match the measurement results reasonably well, indicating that the distributed MOSFET model captures the dependency of V_{t1} on layout parameters.

3.4.3 Non-uniform Turn-on of Fingers

Figure 3.10 compares the measured I_{DUT}-V_{DUT} curve of Figure 3.6 with that from simulation. Simulation correctly predicts non-uniform triggering of the gate fingers and confirms that it is the central two fingers which turn on at the lowest current, because the central fingers see a larger effective body resistance. The on-state resistance (R_{ON}) of each branch of the I_{DUT}-V_{DUT} curve is accurately represented in simulation.

Because the distributed model in Figure 3.2 is constructed based on the symmetric layout in Figure 3.1, the model is also symmetric and thus cannot represent conduction by an odd number of fingers. Therefore, it will not reproduce the I_{DUT}-V_{DUT} branch labeled as "1-finger" in Figure 3.6. However, this branch can be forced to appear in simulation by disabling all fingers except a central one or by introducing a small asymmetry into the model. The dashed line in Figure 3.10 shows the 1-finger simulation result; it matches up well against the measurement data, confirming that this branch of the curve is indeed due to conduction by just 1 finger.

Figure 3.11 shows the simulated I_B-I_{DUT} curve. The discontinuities and the slope changes evident in the measurement data are replicated in the simulation; these occur each time a new finger or pair of fingers is triggered on, as will be demonstrated next. At any given point on the curve, the total body current I_B is the sum of the individual fingers' body currents as indicated in (3.7).

$$I_B = \sum_{on} I_B[i], \qquad i \in [1, N] \tag{3.7}$$

where $I_B[i]$ is the body current contributed by the ith finger of a N-finger device; note that the sum in (3.7) is taken over only the fingers that are triggered on. I_B jumps, i.e. $I_B(I_{DUT})$ is discontinuous, whenever a new finger is triggered on and the number of current sources in the sum changes. This will be illustrated for the specific case of the 6-finger GGNMOS shown in Figure 3.2. Define $\hat{I}_{DUT}$ as the minimum current for which 4 fingers are triggered on; for $I_{DUT} < \hat{I}_{DUT}$, only the center 2 fingers will be on. The body potential for each of the triggered on fingers is expressed in (3.8).

$$V_B[i] = V_{BE,on} + I_S[i] \cdot R_S \tag{3.8}$$

where $I_S[i]$ is the current flowing in the source of the ith finger and R_S is its source-side series resistance; R_S is the same for each finger. Once a finger is triggered, the potential drop across its LNPN's base-emitter junction is roughly pinned at $V_{BE,on}$. In the case that $I_{DUT} < \hat{I}_{DUT}$ ("case (a)"), the current divides equally between the center two fingers, i.e., $I_S[1] = I_S[2] = \frac{1}{2}(I_{DUT} - I_B)$. It follows that $I_B[1] = I_B[2] = \frac{1}{2}I_B$. The total body current for this case, $I_B^{(a)}$, is thus given by Eq. (3.9).

$$I_B^{(a)} = \frac{2V_{BE,on}}{R_B^{(a)} + R_S} + \frac{R_S}{R_B^{(a)} + R_S} I_{DUT} \tag{3.9}$$

where

$$R_B^{(a)} = \left(\frac{R_{finger}}{2}\right) \Big\| \left\{ R_{base,D} + \left(\frac{R_{finger}}{2}\right) \Big\| \left(R_{base,S} + R_{side}\right) \right\} \tag{3.10}$$

Next consider the case that $I_{DUT} \geq \hat{I}_{DUT}$ ("case (b)"). Due to the layout symmetry, $I_S[1] = I_S[2]$ and $I_S[3] = I_S[4]$; also, $I_B[1] = I_B[2]$ and $I_B[3] = I_B[4]$. It also follows that $I_S[1] + I_S[3] = \frac{1}{2}(I_{DUT} - I_B)$. The total body current, $I_B^{(b)}$, in this case is given by (3.11).

$$I_B^{(b)} = \frac{4V_{BE,on}}{2R_B^{(b)} + R_S} + \frac{R_S}{\left(2R_B^{(b)} + R_S\right)} I_{DUT} - \frac{\left(R_S \| 2R_B^{(b)}\right)}{R_{base,S} + R_{side}} (I_S[1] - I_S[3]) \tag{3.11}$$

where

$$R_B^{(b)} = \left(\frac{R_{finger}}{2}\right) \Big\| \left(\frac{R_{finger}}{2}\right) \Big\| \left(R_{base,S} + R_{side}\right) \tag{3.12}$$

At the current level $I_{DUT} = \hat{I}_{DUT}$, (3.9) and (3.11) yield different values of I_B, indicating that the function is discontinuous. This becomes especially evident if one neglects the right-most term in (3.11), which is smaller than the others. This results in a linear relation between I_B and I_{DUT}, with the slope and y-intercept being different for cases (a) and (b). Eq. (3.8) also explains why in Figure 3.11 I_B is observed to

be an increasing function of I_{DUT}, even after all the fingers are operating in snapback; the increasing voltage drop across the source resistor R_S causes $V_B[i]$ to rise.

The lowest current branch in Figure 3.11 is the result of conduction in just a single finger and will not appear in the simulation results unless it is forced to do so. The simulation results predict that the GGNMOS will have negligibly small I_{DUT} until I_B is large enough to turn on the first finger, but in measurement I_{DUT} appears at lower I_B levels than predicted. This discrepancy is attributed to non-uniform conduction across the finger width, which occurs at very low current levels [69],[70],[71].

3.5 On-state Resistance and Self-heating

The on-state resistance (R_{ON}) of a multi-finger MOSFET is a function of the total device width as well as the series resistance of each finger. Since each finger carries a significant current during ESD, its series resistance is affected by self-heating. However, each finger may not carry the same amount of current, so the amount of self-heating is not the same. Consequently, each finger contributes differently to the overall R_{on}. The distributed compact model should reproduce this effect.

3.5.1 Single-finger On-state Resistance, R_{on1}

In [69], it was proposed that the differential on-state resistance of a single-finger NMOS operating in snapback, R_{on1}, can be expressed by the analytical formula shown in (3.13).

$$R_{on1} = \frac{\partial V_D}{\partial I_D} = R_D + R_B \cdot \frac{\beta(M-1)-1}{\beta M} \tag{3.13}$$

However, Eq. (3.13) does not include the effect of the source-side series resistance R_S. The following analysis provides a more detailed formulation for R_{on1}. R_{on1} at a DC bias point (V_{DUT}, I_{DUT}) is obtained from a linear, small-signal model of the device. In snapback mode, a single-finger GGNMOS can be represented by the forward Ebers-Moll model, as shown in Figure 3.12(a). Note that the channel current of the GGNMOS is not included in the model; under snapback operation, its contribution to R_{on1} is negligible relative to that of the LNPN. All elements in the schematic shown in Figure 3.12(a) are

linearized to obtain the small-signal model shown in Figure 3.12(b). The detailed steps of this exercise are presented in subsection 3A.1. Using the small-signal model in Figure 3.12(b), the differential on-state resistance R_{on1} can be derived, with the result shown in (3.14).

$$R_{on1} = R_D + \frac{R_B(r_e + R_S)}{R_B + r_e + R_S} + r_\mu \cdot \left(1 - \frac{M\alpha R_B}{R_B + r_e + R_S}\right) \tag{3.14}$$

Next, Eq. (3.14) is evaluated in the limiting case that I_D is very large ($I_D \to \infty$). Under this condition, $r_e \ll R_S$, $r_\mu \approx 0$, and $M \cdot \alpha$ approaches 1 [57]. The resultant R_{on1} is given by the expression in (3.15).

$$\lim_{I_D \to \infty} R_{on1} = R_D + \frac{R_B R_S}{R_B + R_S} \tag{3.15}$$

In the usual case that $R_B \gg R_S$, R_{on1} at high current levels is well approximated by (3.16).

$$R_{on1} \cong R_D + R_S \tag{3.16}$$

In a multi-finger device that is fully triggered on, each finger will contribute to the overall on-state resistance. Using (3.16) to represent the on-state resistance of a single finger, the overall on-state resistance of a multi-finger MOSFET is expected to have the value given by (3.17).

$$R_{on} = \frac{\partial V_D}{\partial I_D} = \left[\sum_{i=1}^{N} \frac{1}{R_{on1}[i]}\right]^{-1} = \left[\sum_{i=1}^{N} \frac{1}{(R_D + R_S)}\right]^{-1} = \left(\frac{d_D + d_S}{w_f}\right) \cdot \frac{R_{diff}}{N_f} \tag{3.17}$$

where d_D and d_S are the lengths of the silicide-blocked regions on the drain and source sides, respectively, and R_{diff} is the sheet resistance of the silicide-blocked drain/source diffusion. The measurement and simulation results shown in Figure 3.13 both confirm that the overall R_{on} varies as $(N_f \times w_f)^{-1}$, as predicted by (3.17).

3.5.2 Modeling Non-uniform Self-heating

Although the measurement and simulation results in Figure 3.13 appear to be in good agreement, further investigation reveals that a systematic error in the simulation results appears if the normalized on-state resistance ($R_{on,norm}$) is plotted as a function of N_f, as shown in Figure 3.14. $R_{on,norm}$ is obtained by

multiplying the overall R_{ON} by the total channel width ($N_f \times w_f$). Based on (3.17), $R_{on,norm}$ is expected to be a constant, independent of N_f, which is indeed seen in the simulation results of Figure 3.14 (data labeled "Sim. w/o Self-Heating"). In contrast, the measurement results in Figure 3.14 show that the normalized on-state resistance is an increasing function of N_f. This phenomenon is attributed to non-uniform self-heating, which has two root causes.

a) Heat generation differs among the fingers.

Even after all the fingers of a GGNMOS have been triggered on, they do not each conduct the same amount of current. As demonstrated in (3.20), the center fingers carry more current than the outer fingers. The higher current density for the center fingers causes them to experience more Joule heating.

b) Less heat is transported away from the central fingers than the outer fingers.

A finger located in the middle of a multi-finger structure is surrounded by other fingers which act as heat sources. Consequently, there is a reduced temperature gradient; the gradient is what drives heat transport. The fingers located at the outer edges of the multi-finger structure are adjacent to "cold" silicon on one side, which promotes heat dissipation.

The net effect is that the temperature near the central fingers is higher than that near the outer fingers. The source and drain resistances of an individual finger are increasing functions of the local temperature because of reduced carrier mobility [72]. The GGNMOS model was reformulated to include self-heating, following the approach described in [13]. The drain and source series resistors, R_D and R_S, of a finger are modeled as functions of the local temperature rise, ΔT, by (3.18).

$$\begin{cases} R_D = R_{D,T0} \cdot \left(1 + \dfrac{\Delta T}{T_0}\right)^{\delta} \\ R_S = R_{S,T0} \cdot \left(1 + \dfrac{\Delta T}{T_0}\right)^{\delta} \end{cases} \tag{3.18}$$

where T_0 is the ambient temperature. $R_{D,T0}$ and $R_{S,T0}$ are the nominal drain and source series resistances at the ambient temperature. The self-heating coefficient, δ, is a fitting parameter. ΔT of each finger is found during circuit simulation using the thermal equivalent circuit shown in Figure 3.15. The thermal circuit elements R_{TH} and C_{TH} are extracted from measurement of a 2-finger GGNMOS test structure. Pulsed I-V simulations were performed using the enhanced compact model; in the simulations, each finger of the N-finger GGNMOS has an identical thermal equivalent circuit applied. The normalized on-state resistance was extracted from the new simulation results. As seen in Figure 3.14, the model that includes self-heating reproduces the measurement results.

The R_{ON} values plotted in Figure 3.13 and Figure 3.14 are extracted from the I-V curves at moderate current levels ($I_{DUT} \approx \frac{1}{2} I_{t2}$), before the I-V curves show significant bending due to severe self-heating. The device on-state resistance can be found by summing the on-state resistance of each finger in an inverse fashion, e.g. as done in (3.17), only if the I-V characteristic of each finger has the same x-intercept, V_{hold}, as illustrated in Figure 3.16(a). However, circuit simulation can be used to obtain the device I-V curve all the way up to its second breakdown point, as demonstrated in Figure 3.16(b).

3.5.3 TCAD Simulation of Self-heating

Non-uniform self-heating was attributed, in part, to the fact that each finger does not carry equal current, as demonstrated in (3.21). After triggering, the current flowing through the ith finger, $I_D[i]$, is given by (3.19).

$$I_D[i] = \frac{V_{DUT} - V_{hold}[i]}{R_D[i] + R_S[i]} \tag{3.19}$$

where $V_{hold}[i]$, $R_D[i]$, and $R_S[i]$ are the holding voltage, drain-side and source-side series resistances of the ith finger, respectively. Once the device starts to heat up, the lattice temperature is higher for the center fingers, causing them to have larger R_D, R_S and V_{hold}; the latter is a result of the LNPN's temperature-dependent gain [72]. This increases the input resistance of the central fingers, which steers current toward the colder, outer fingers in a sort of ballasting effect. This analysis is justified based on TCAD transient

simulation of multi-finger GGNMOS response to a pulse current input. The TCAD results shown in Figure 3.17 confirm that the lattice temperature is highest near the central fingers, and the results shown in Figure 3.18 confirm that ballasting becomes significant at high current levels. As shown in Figure 3.18(a), when the device is operating on the lower part of the snapback I-V curve ($I_{DUT}=0.16I_{t2}$), the current per finger consistently decreases as one moves from the center fingers to the outer fingers. However, at a higher total current level (e.g., $I_{DUT}=0.81I_{t2}$), Figure 3.18(b) shows that the current per finger is virtually a constant for all fingers except the two on the outside edges, confirming the current steering effect.

The detailed dynamics of the time-dependent current distribution inside the MOSFET could be modeled by making all the device parameters—including R_D, R_S, M, and β—be functions of temperature, and assigning position dependent thermal parameters to the fingers. The resulting complexity of both the model and the parameter extraction procedure is not compatible with the aims of compact modeling, especially considering that the simpler model already provides a good representation of the device terminal characteristics. TCAD is recommended for the microscopic analysis of device behavior.

3.6 Summary

A distributed, compact model of a MOSFET operating under ESD conditions is shown to produce pulsed I-V characteristics that scale with respect to the device layout geometry. The distributed model accurately represents the device I-V from low to high current levels, reproducing non-uniform turn-on among the device fingers and non-uniform self-heating. The distributed model provides a pulsed I-V response that is accurate across a wider range of ESD conditions than that of a single-finger model.

3.7 Supplement: Derivation of Single-finger R_{on1} Model

The expression for R_{on1} is derived using the model of Figure 3.12(a). The DC bias current, I_D, can be expressed by (3.20).

$$I_D = \alpha I_E + (M-1)\cdot \alpha I_E = M(V_{CB})\cdot \alpha I_E(V_{BE}) \tag{3.20}$$

The forward, common-base current gain parameter, α, is a function of V_{CB}, but this bias dependency is weak relative to those of M and I_E. Thus, it is not included in the analysis. In the vicinity of the DC bias point, the drain current may be expressed by (3.21).

$$I_D(V_{BE}+v_{BE}, V_{CB}+v_{CB}) = I_D(V_{BE}, V_{CB}) + i_D(v_{BE}, v_{CB}) \tag{3.21}$$

A linearized modeling approach is used for the small signal and thus i_D is represented by (3.22).

$$i_D = v_{BE}\cdot \left.\frac{\partial I_D}{\partial v_{BE}}\right|_{V_{BE}} + v_{CB}\cdot \left.\frac{\partial I_D}{\partial v_{CB}}\right|_{V_{CB}} \tag{3.22}$$

Expressions for the derivatives are derived in (3.23) and (3.24).

$$\frac{\partial I_D}{\partial v_{BE}} = M\alpha \frac{\partial I_E}{\partial v_{BE}} = M\alpha \frac{I_E}{\frac{kT}{q}} = M\alpha r_e = M g_m \tag{3.23}$$

$$\frac{\partial I_D}{\partial v_{CB}} = \frac{\partial M}{\partial v_{CB}}\cdot \alpha I_E \equiv \frac{1}{r_\mu} \tag{3.24}$$

The small-signal model derivation is completed by finding an expression for i_E. Using Figure 3.12(a), this is found to be

$$i_E = \frac{v_{BE}}{r_e} \tag{3.25}$$

where r_e was defined in (3.23). The third terminal current, i_B, is found using

$$i_B = i_D - i_E \tag{3.26}$$

Eqs (3.22)–(3.26) constitute the small-signal model. A schematic representation of the model is shown in Figure 3.12(b). In contrast to the modeling approach used in [69], here it was not assumed that $\partial I_D/\partial V_{CB}$ and $\partial M/\partial V_{CB}$ are constant.

Given the small-signal single-finger model, R_{on1} is obtained using (3.27).

$$R_{on1} = \frac{v_D}{i_D} \tag{3.27}$$

Based on the schematic of Figure 3.12(b), $v_D(i_D)$ is obtained as follows:

$$v_D = i_D R_D + v_C \tag{3.28}$$

$$v_C = v_B + r_\mu \cdot \left(i_D - M g_m \cdot (v_B - v_E)\right) \tag{3.29}$$

$$v_B = i_D \cdot R_Z \tag{3.30}$$

where $R_Z = \frac{R_B \cdot (r_e + R_S)}{R_B + r_e + R_S}$.

$$v_E = \frac{R_S}{r_e + R_S} \cdot v_B \tag{3.31}$$

Plugging (3.29)–(3.31) into (3.28), one obtains

$$v_D = i_D \cdot \left(R_D + R_Z + r_\mu \cdot \left[1 - \frac{\alpha M R_B}{R_B + r_e + R_S}\right]\right) \tag{3.32}$$

Solving for R_{on1}, one obtains

$$R_{on1} = R_D + \frac{R_B(r_e + R_S)}{R_B + r_e + R_S} + r_\mu \cdot \left(1 - \frac{M\alpha R_B}{R_B + r_e + R_S}\right) \tag{3.33}$$

(3.33) can be further reduced if I_D is sufficiently large. This is the condition of $I_D \to \infty$ applied in (3.15). This condition can be reasonably satisfied once the device snapback because it only requires (i) $r_e \ll R_S$ and (ii) $r_\mu \approx 0$.

To satisfy (i), (3.34) is required.

$$r_e \ll R_S \;\rightarrow\; I_E \gg \frac{kT/q}{R_S} \rightarrow\; I_D > I_E \gg \frac{kT/q}{R_S} \tag{3.34}$$

Because $kT/q \approx 27\text{mV}$ is small, the condition of $I_D \gg kT/qR_S$ can be easily satisfied even at low I_{DUT} after snapback.

To satisfy (ii), (3.35) is required.

$$r_\mu = \frac{\partial v_{CB}}{\partial I_D} \approx 0 \rightarrow V_{CB} \approx \text{constant} \tag{3.35}$$

Since V_{CB} is the potential drop across the drain-body junction where impact ionization occurs, it is relatively constant after snapback even at low current level, and that $r_\mu \approx 0$ is easily satisfied.

3.8 Supplement: Analysis of Finger Current Distribution

It is claimed that the center fingers conduct more current than the outer fingers. This statement is verified using proof by initially assuming the opposite, i.e., that each finger conducts an equal current. If the various $I_D[i]$ are all equal, and similarly the $I_S[i]$, it follows that the body currents $I_B[i]$ must all be equal. With this assumption, the body potential of each finger, $V_B[i]$, can be found analytically. This is done for the case of a 6-finger device, with the aid of the circuit model shown in Figure 3.19. The results are given in Eqs. (3.36)–(3.38).

$$V_B[1] = I_B \cdot \frac{\frac{R_{finger}}{2}\left\{R_{base,D}\left(\frac{R_{finger}}{2} + R_{base,S} + R_{side}\right) + \frac{R_{finger}}{2}\left(R_{side} + R_{base,S} + R_{side}\right) + \frac{R_{finger}}{2}\left(R_{side} + R_{base,S}\right)\right\}}{\frac{R_{finger}}{2}\left(\frac{R_{finger}}{2} + R_{base,S} + R_{side}\right) + R_{base,D}\left(\frac{R_{finger}}{2} + R_{base,S} + R_{side}\right) + \frac{R_{finger}}{2}\left(R_{base,S} + R_{side}\right)} \tag{3.36}$$

$$V_B[3] = I_B \cdot \left\{\frac{R_{side}\frac{R_{finger}}{2}\left(R_{base,D} + \frac{R_{finger}}{2}\right) + \left(R_{base,S} + R_{side}\right)\frac{R_{finger}}{2}\left(R_{base,D} + \frac{R_{finger}}{2}\right) + \left(R_{base,S} + R_{side}\right)\frac{R_{finger}}{2}\frac{R_{finger}}{2}}{\frac{R_{finger}}{2}\left(R_{base,D} + \frac{R_{finger}}{2}\right) + \left(R_{base,S} + R_{side}\right)\left(R_{base,D} + \frac{R_{finger}}{2}\right) + \left(R_{base,S} + R_{side}\right)\frac{R_{finger}}{2}}\right\} \tag{3.37}$$

$$V_B[5] = I_B \cdot \frac{R_{side}\left\{R_{base,S}\left(\frac{R_{finger}}{2} + R_{base,D} + \frac{R_{finger}}{2}\right) + \frac{R_{finger}}{2}\left(\frac{R_{finger}}{2} + R_{base,D} + \frac{R_{finger}}{2}\right) + \frac{R_{finger}}{2}\left(\frac{R_{finger}}{2} + R_{base,D}\right)\right\}}{R_{side}\left(\frac{R_{finger}}{2} + R_{base,D} + \frac{R_{finger}}{2}\right) + R_{base,S}\left(\frac{R_{finger}}{2} + R_{base,D} + \frac{R_{finger}}{2}\right) + \frac{R_{finger}}{2}\left(R_{base,D} + \frac{R_{finger}}{2}\right)} \tag{3.38}$$

If the condition shown in (3.39) holds, further manipulation of the above expressions reveals the inequality shown in (3.40).

$$R_{side} < \frac{R_{finger}}{2} \tag{3.39}$$

$$V_B[5] < V_B[3] < V_B[1] \tag{3.40}$$

Indeed, (3.39) is an accurate statement given the typical multi-finger MOSFET layout; refer to Figure 3.4 and observe that the outermost fingers have the widest cross-section for body current flow. Because the body potential decreases from the center to the outer fingers (i.e., $V_B[5]<V_B[3]<V_B[1]$), it

follows that $I_B[5]<I_B[3]<I_B[1]$, in contradiction to the assumption of constant $I_B[i]$. Because the center fingers have higher $I_B[i]$ and $V_B[i]$, they must also have higher I_S and I_D. Therefore, it is verified that the center fingers conduct more current than the outer fingers, consistent with the simulation results.

3.9 Figures

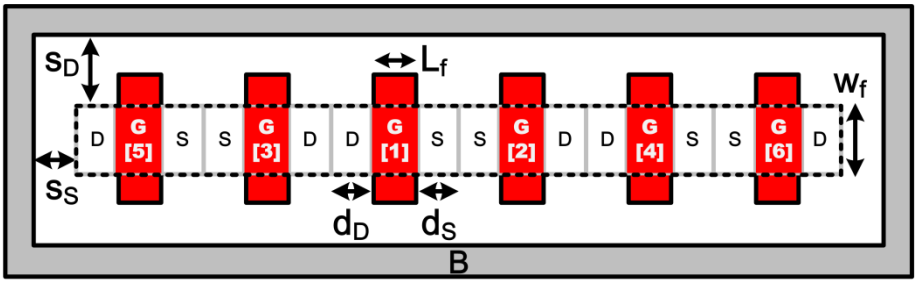


Figure 3.1 Typical layout of a multi-finger MOSFET. The regions marked D, G and S are connected together by on-chip metal, forming the transistor drain, gate, and source terminals. The region marked B provides the body contact. The finger index ([i]) matches that in Figure 3.2 and serves an illustrative purpose only. Layout geometric parameters include the number of fingers (N_f; here $N_f=6$), finger width (w_f) and length (L_f), spacing to body pick-up ring (s_D and s_S), lengths of drain and source diffusions (d_D and d_S).

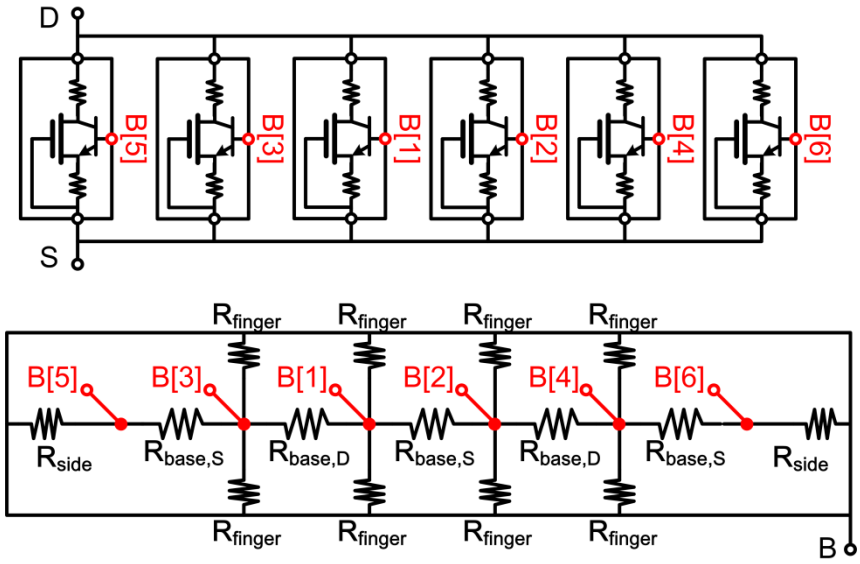


Figure 3.2 Distributed model of the multi-finger NMOS shown in Figure 3.1. The internal body terminal of the ith finger taps into the body resistance network at the same node B[i]. There are 4 uniquely-valued resistors in the body network—R_{side}, R_{finger}, $R_{base,S}$ and $R_{base,S}$; these may be mapped to the device layout as shown in Figure 3.4. In this figure, it is assumed that the gate fingers are tied to the ground terminal ("GGNMOS").

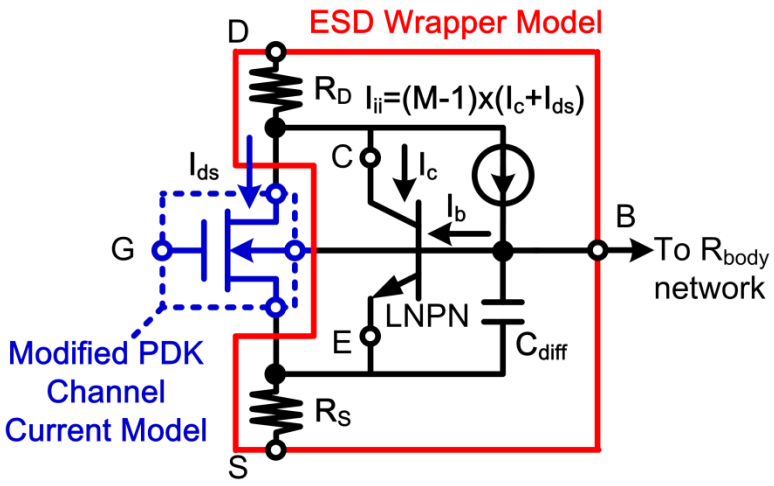


Figure 3.3 Model of one MOSFET finger includes a channel current model and an ESD wrapper model. The channel current model is obtained from the PDK but with certain features disabled, including impact ionization, substrate current induced body effect (SCBE), and drain/source resistances [66],[67]. Note that the PDK MOSFET model already includes parasitic junction capacitors from drain and source to body. The ESD wrapper model includes collector and base currents (I_c and I_b) of the parasitic lateral bipolar transistor (LNPN), base-emitter diffusion capacitance (C_{diff}), impact ionization current source (I_{ii}), and drain and source resistances (R_D and R_S). The components in the ESD wrapper model are extracted under high current conditions to ensure validity.

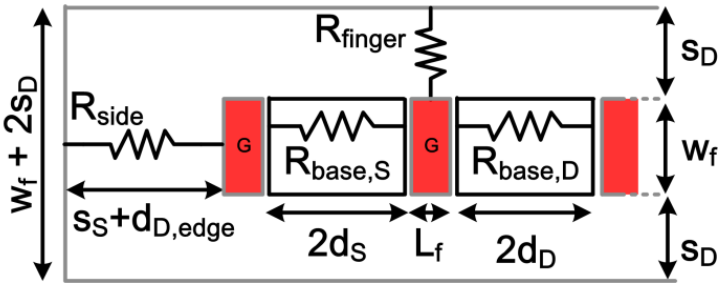


Figure 3.4 The values of R_{side}, R_{finger}, $R_{base,D}$, and $R_{base,S}$ are functions of their physical dimensions, obtained from the layout.

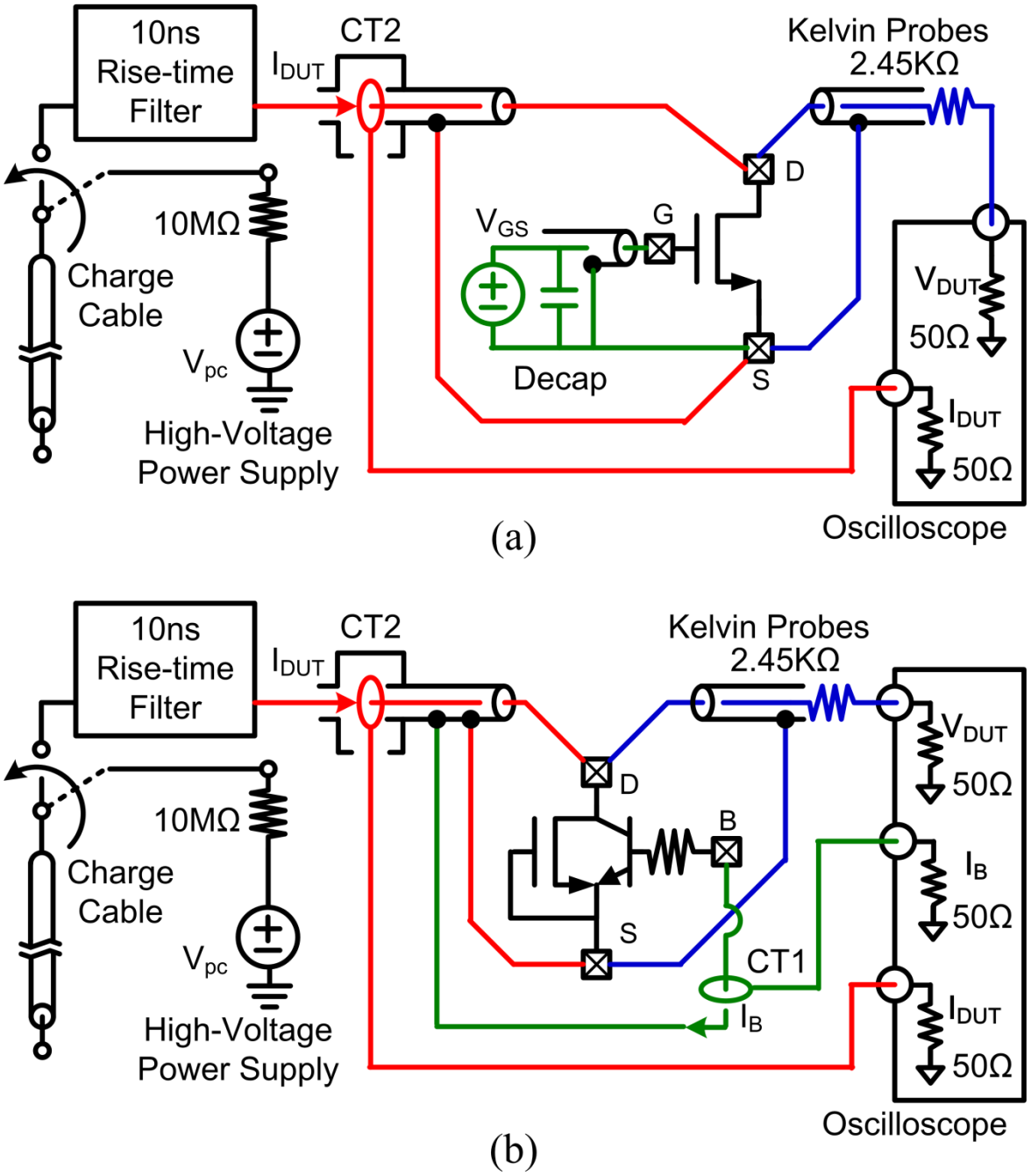


Figure 3.5 (a) Schematic of the TLP system with gate-biasing probe which has an embedded decoupling capacitor (Decap) to stabilize the gate-to-source potential during pulsing. (b) Schematic of the 3T-TLP system. CT1 and CT2 are current probes.

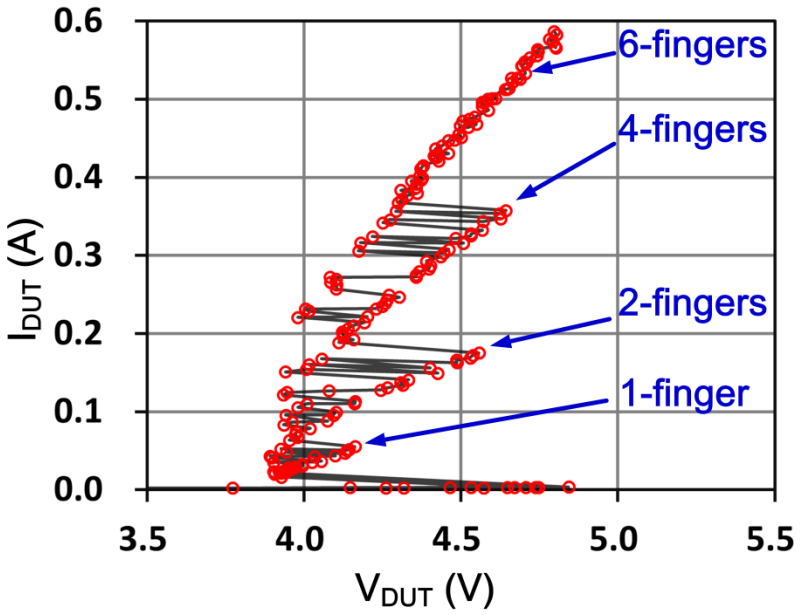


Figure 3.6 Quasi-static TLP I-V curve of a 6-finger GGNMOS, with w_f = 20μm. Several distinct on-state branches result due to non-uniform turn-on among fingers. The solid lines connecting discrete data points indicate the time evolution of the pulse measurements. The number of fingers triggered on can be inferred from the R_{ON} value for each branch.

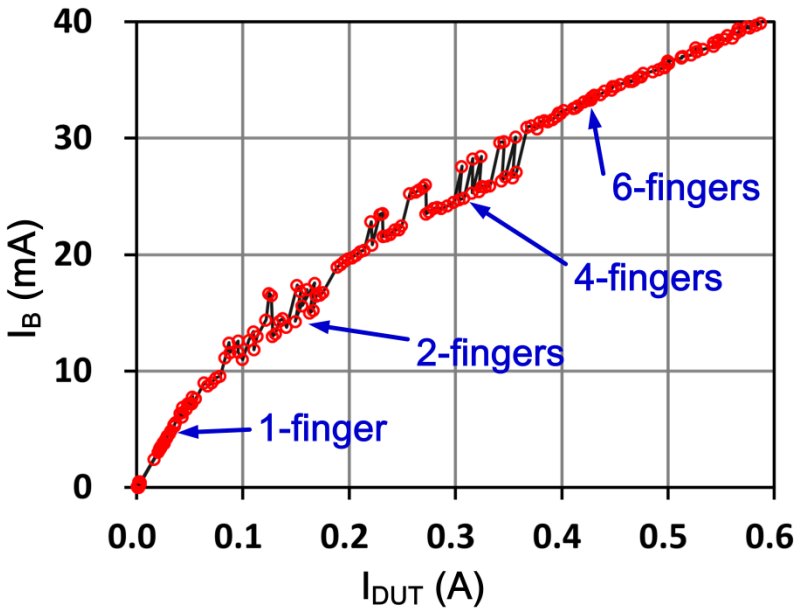


Figure 3.7 Quasi-static I_B vs. I_{DUT} for the same GGNMOS as in Figure 3.6.

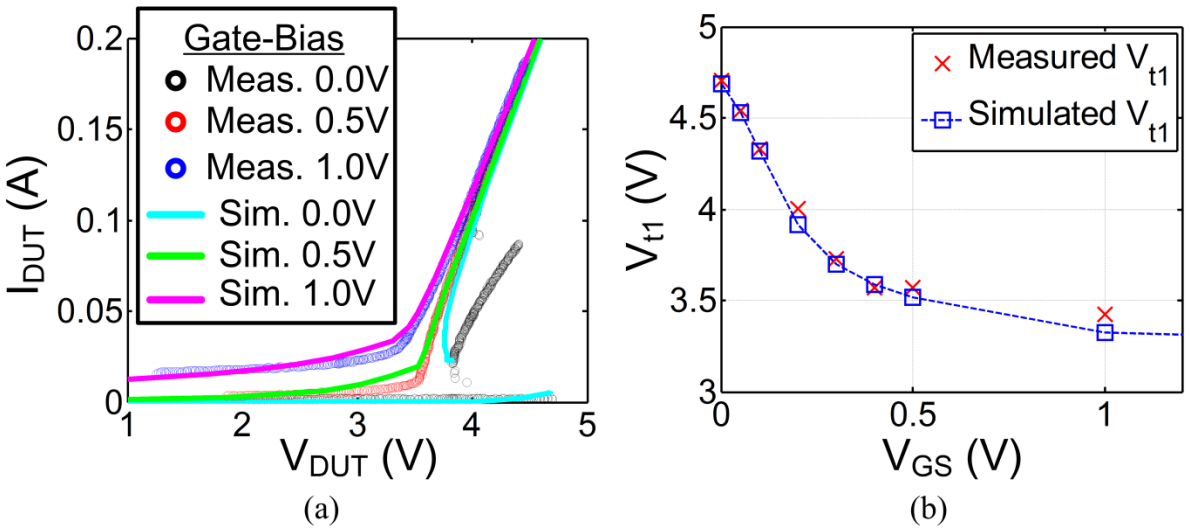


Figure 3.8 Gate bias effect on trigger voltage. (a) Measured and simulated TLP I-V curves as functions of V_{GS}. (b) Measured and simulated V_{t1} as extracted from the data shown in part (a). The device is a 2-finger MOSFET with w_f =20μm. The probe used to apply the gate bias contains a decoupling capacitor to ground which stabilizes the applied DC bias when pulses are applied at the drain.

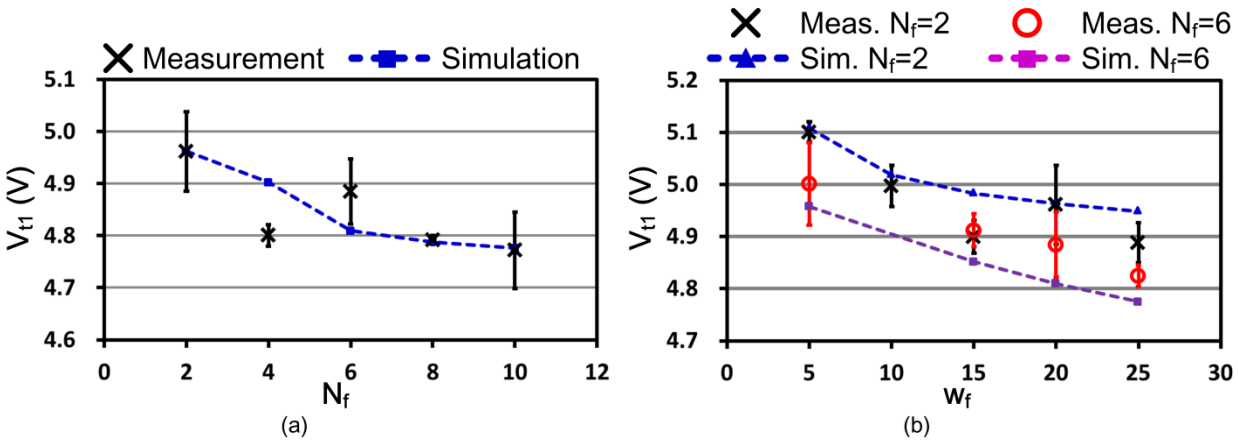


Figure 3.9 Measured and simulated V_{t1}. (a) w_f =20μm and N_f is variable. (b) w_f is variable; N_f=2 and N_f=6. The measured devices are GGNMOSs with s_D=s_S=2μm. Error bars show the variability of the measured V_{t1}.

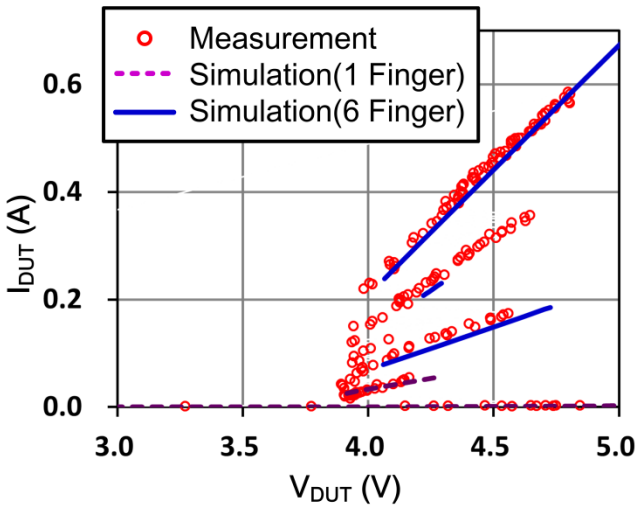


Figure 3.10 Measured (red circles) and simulated (blue lines) $I_{DUT}(V_{DUT})$ for a GGNMOS with w_f=20µm and N_f=6. The 4-finger branch of the measured curve is longer than in simulation due to the stochastic nature of snapback; the number of fingers triggered on fluctuates between pulses. The lowest on-branch appearing on the measured I-V curve is replicated by simulations in which just one finger is activated.

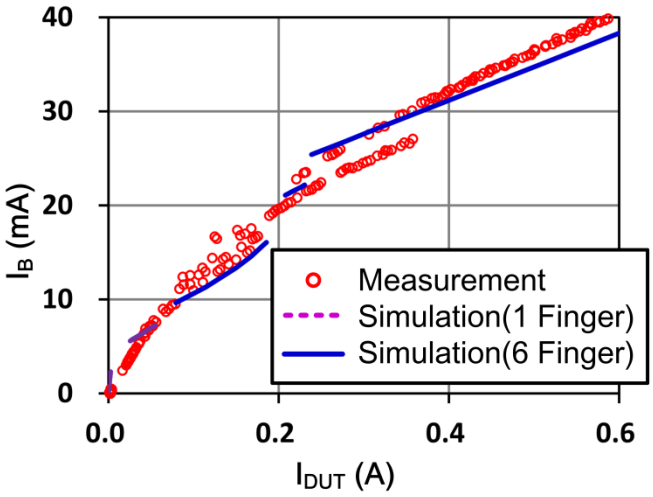


Figure 3.11 Measured and simulated I_B vs. I_{DUT} for the GGNMOS of Figure 3.10. The region below I_{DUT}=0.07A can be replicated in simulation only if single-finger conduction is forced. This mismatch that appears at very low current levels may be due to non-uniform conduction across the width of the first finger.

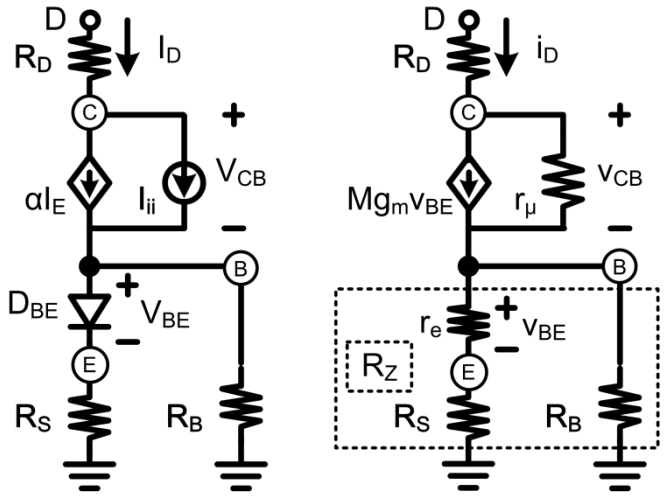


Figure 3.12 (a) DC model of a single-finger GGNMOS operating in snapback mode. The impact ionization current source is modeled as $I_{ii} = (M-1)\cdot \alpha I_E$. (b) Linearized small-signal model of (a).

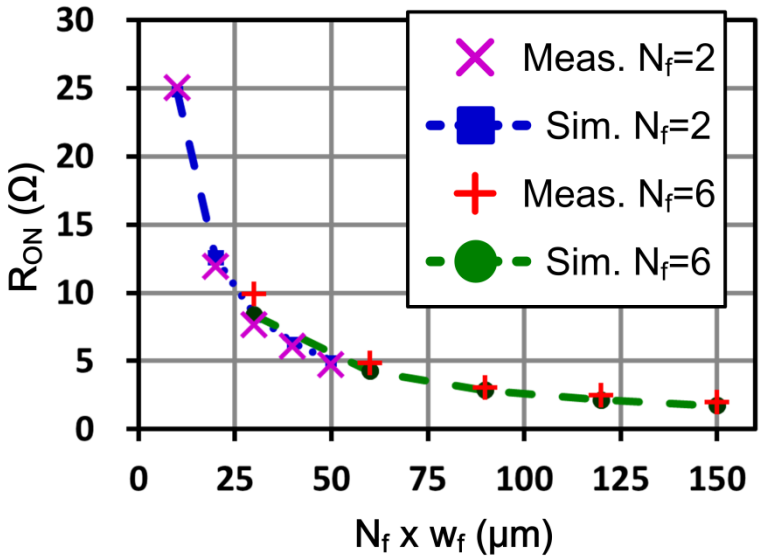


Figure 3.13 Measured and simulated R_{ON} as a function of the total channel width. Both N_f and w_f are variables. The data are obtained at sufficiently high I_{DUT} to ensure all the fingers are turned on. This result suggests that R_{ON} is inversely proportional to the total channel width.

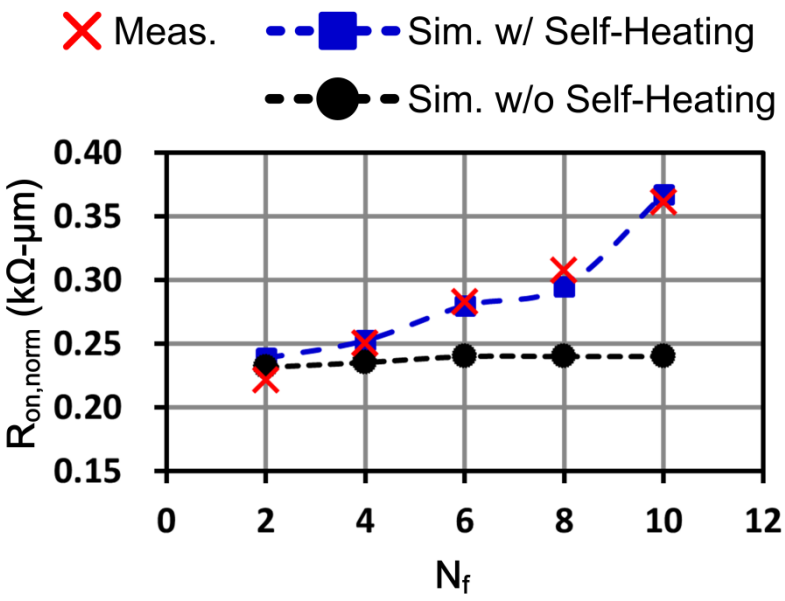


Figure 3.14 Measured and simulated values of the normalized on-state resistance. Test structures are GGNMOS with w_f=20μm. The normalized on-state resistance is a function of N_f due to the non-uniform current distribution, and non-uniform self-heating among the device fingers.

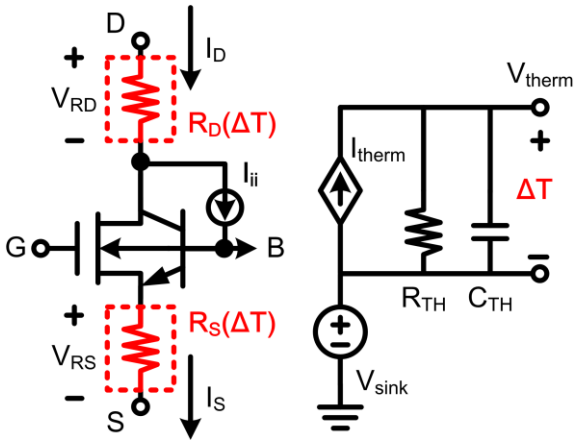


Figure 3.15 In the compact model, the drain and source series resistances are temperature-dependent. A thermal equivalent circuit is used to simulate the temperature increase $\Delta T = V_{therm} - V_{sink}$ due to Joule heating in each finger. The current source represents power dissipation; $I_{therm} = V_{RD} \cdot I_D + V_{RS} \cdot I_S$ [13]. This numerically efficient approach is justified by the fact that simulation can reproduce the pulsed I-V characteristic.

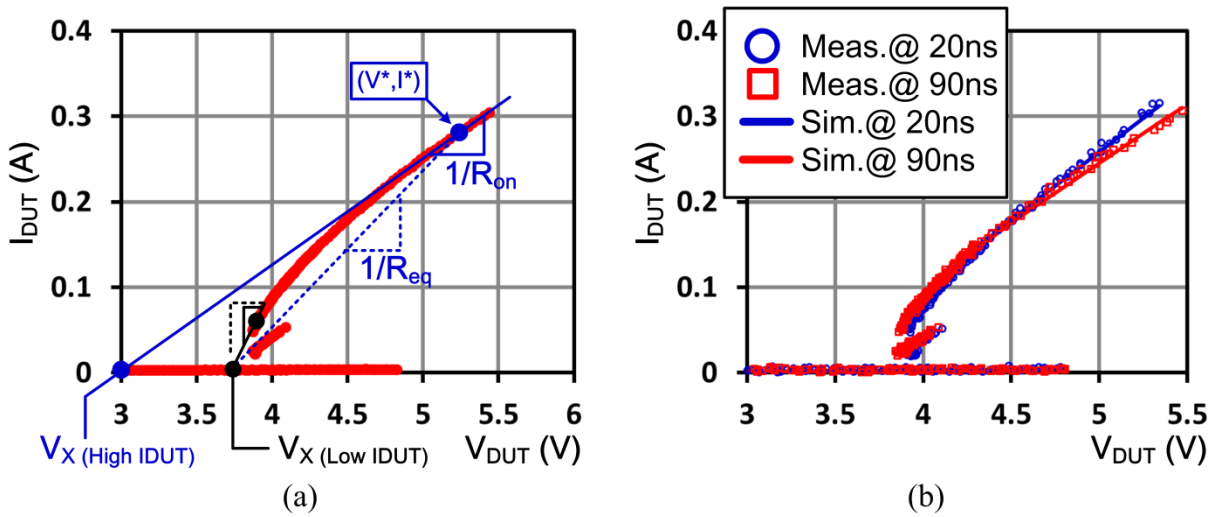


Figure 3.16 (a) Illustration of on-state resistance. R_{ON}, the differential on-state resistance, satisfies the equation $V_{DUT} = R_{on} \cdot I_{DUT} + V_X$, where V_X is the extrapolated offset voltage measured at the specific I_{DUT}. The equivalent resistance R_{eq} satisfies $V_{DUT} = R_{eq} \cdot I_{DUT} + V_{hold}$, where V_{hold}, the extrapolated offset voltage measured at low current levels, is a constant. The device's net R_{ON} may be used to infer information about the per finger on-state resistance only if each finger has the same V_X, which is guaranteed only at relatively low currents where all fingers V_X=V_{hold}. (b) Measured and simulated I-V curves sampled at 20ns and at 90ns after the current pulse is initiated. DUT: GGNMOS, N_f=2, w_f=20μm, I_{t2}=0.39A. Self-heating is a transient phenomenon; on-state resistance increases with I_{DUT} and with sampling time. The model with empirical self-heating incorporated correctly reproduces these I-V curves.

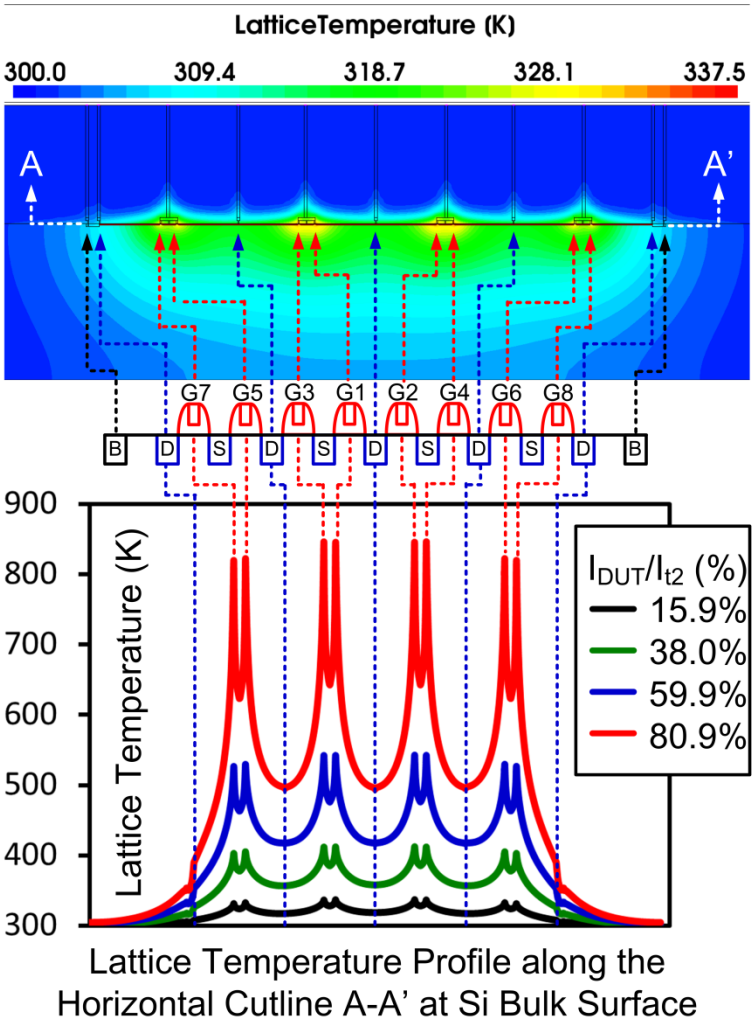


Figure 3.17 The top plot shows the 2D temperature distribution of an 8-finger GGNMOS obtained from transient TCAD simulation using Synopsys Sentaurus®. It is sampled 90ns after a current step appears at the DUT; I_{DUT}/I_{t2}=15.9%. The lattice temperature profile shown at the bottom is taken along a cutline near the surface of the silicon. The average temperature in the source diffusions is higher than that in the drain diffusions because the latter are significantly longer and extend farther away from the hot spot at the drain junction right under the gate edge. These simulation results should be considered as providing only a qualitative picture of what is happening inside the device since the doping profile was not obtained from process simulation, and the drift-diffusion simulator uses an empirical impact ionization model.

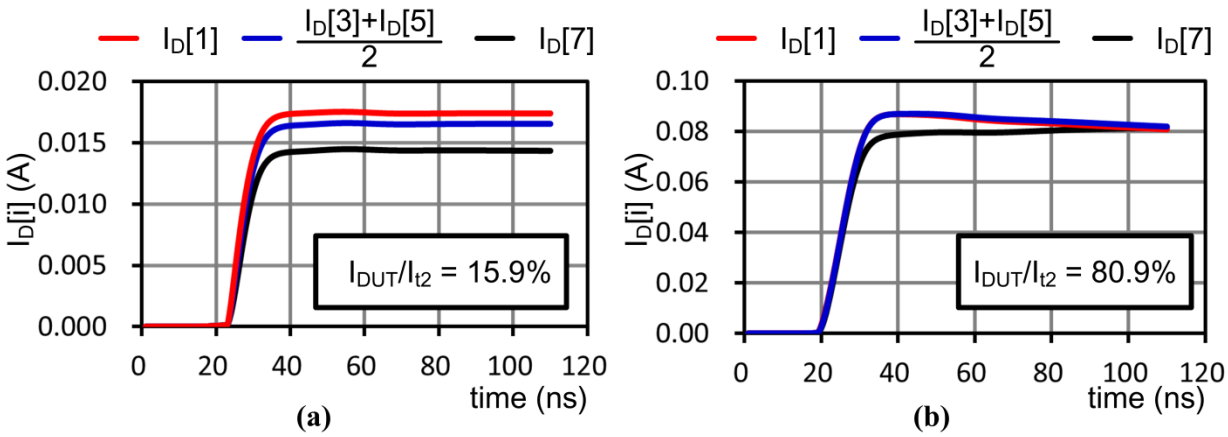


Figure 3.18 TCAD simulated current distribution in an 8-finger GGNMOS subjected to TLP. $I_D[i]$ is defined as the current through the ith finger, similar to the numbering system shown in Figure 3.1. Since the 5th and 3rd fingers share a drain contact, it is difficult to separate their currents, so the average $(I_D[3] + I_D[5])/2$ is plotted instead. Nevertheless, a monotonic trend, $I_D[1]>I_D[3]>I_D[5]>I_D[7]$, can be inferred based on the temperature profile shown in Figure 3.17. (a) At a relatively low I_{DUT} level, current through the central fingers is higher than that through the outer fingers because of higher local body potential. (b) At a high I_{DUT} level, current through each finger becomes equalized at the end of the pulse due to the ballasting effect described in the text. That $I_D[1]$ is higher than $I_D[7]$ at the beginning of the pulse indicates the current through centermost fingers (1st and 2nd) is still higher than that through the outermost fingers (7th and 8th). At the same time, however, current through the 3rd to 6th fingers is almost indistinguishable from that through the 1st and 2nd fingers. Because the final I_{DUT} is so high, the device conducts significant current even during the rising edge of the pulse, and the thermal ballasting effect quickly equalizes the current distribution.

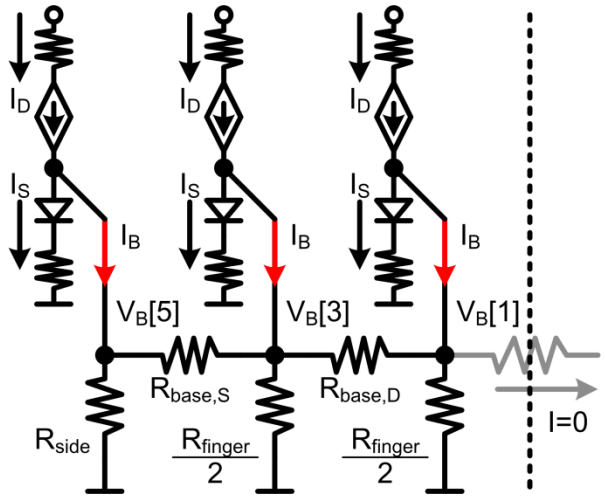


Figure 3.19 Circuit model for the analysis of body potential in a 6-finger GGNMOS. Because of the symmetry in the model, there is no current flow between the 1st and 2nd fingers, so the analysis can be performed on only half of the model.

Chapter 4. Piecewise-Linear Model with Transient Relaxation for Circuit-level ESD Simulation

This work presents a new type of model, piecewise-linear with transient relaxation (PWL-TR), to describe the nonlinear transient characteristics of ESD protection devices and circuits. The PWL-TR model represents the device as a finite state machine, so no proprietary information is disclosed. The PWL-TR model offers accuracy comparable to a compact model but enables more computationally efficient simulation.

4.1 Motivation

Circuit simulation of an IC's response to ESD requires models of the on-chip ESD protection devices and circuits [73]–[76]. Compact models can reproduce the electrical response of ESD devices [11],[13],[28],[29],[77]. However, the compact model formulation may reveal details about the fabrication process and internal circuitry; furthermore, its nonlinearity increases the simulation run-time. Therefore, behavioral models, i.e. "black box" models, are preferred. It has been proposed that the IBIS syntax [78] be used to describe the behavior of an IO pin during ESD; toward this end, ESD protection circuits are represented by a piecewise-linear (PWL) model implemented as a finite state machine (FSM) [79]. However, a PWL model cannot reproduce the transient voltage overshoot displayed by many ESD protection devices [26],[28],[29],[77]. The transient response impacts the efficacy of the ESD protection device/circuit, especially in regards to fast ESD events, e.g. CDM or HMM. Thus, the piecewise-linear model with transient relaxation (PWL-TR) is proposed. The ESD device or circuit is abstracted as a PWL model at each simulated time-point; the PWL representation of the device evolves during simulation by a process named transient relaxation.

4.2 Piecewise-Linear Model with Transient Relaxation

4.2.1 Piecewise-linear I-V Branches

Many ESD protection devices undergo snapback; Figure 4.1(a) illustrates the response of such a device to current pulses. V_{t1} denotes the trigger voltage measured under quasi-static conditions; when a fast transient is applied, the voltage across the device-under-test (DUT) can exceed V_{t1} before it snaps-back toward the holding voltage. Figure 4.1(b) shows the I-V characteristics obtained by sampling the pulse response at different time-points. At each simulation time-point, the device can be described as a resistor R_{EQ} in series with a voltage source V_{EQ}. Note that there are multiple I-V branches, each with its own values of (R_{EQ}, V_{EQ}). The most important information in Figure 4.1(b) is captured by three of the I-V branches:

1) Off-branch (R_{off}, 0)
2) On-branch (R_{on}, V_{on})
3) Trigger branch (R_{tr}, V_{tr})

4.2.2 Transient Relaxation

Figure 4.2 illustrates the framework of a PWL-TR model for any two-terminal device or circuit. The transient behavior is governed by a FSM. At the nth simulated time-point (t_n), the state of the device is denoted as $S(t_n)$. Transitions between states are based on the bias condition (V_{DUT}, I_{DUT}). The rules governing the assignment of values to $R_{EQ}(t_n)$ and $V_{EQ}(t_n)$ are different for the three states, and are described below.

1) Hi-Z State ($S_{Hi\text{-}Z}$)

$S_{Hi\text{-}Z}$ is the initial state before the ESD stimulus is applied. Eq. (4.1) describes the I-V behavior in this state.

If $S(t_n) = S_{Hi-Z}$,

then $R_{EQ}(t_n) = R_{off}$ and $V_{EQ}(t_n) = 0$. (4.1)

2) Turn-on State ($S_{T\text{-}On}$)

When V_{DUT} exceeds V_{t1}, the device is in state $S_{T\text{-}On}$. The trigger time $t_{0,on}$ is defined as follows: if $V_{DUT}(t_{n-1}) < V_{t1}$ and $V_{DUT}(t_n) \geq V_{t1}$, then $t_{0,on} = t_n$. The device cannot instantaneously transition to its steady-state on-condition at time $t_{0,on}$ and may exhibit transient voltage overshoot. Eq. (4.2) gives (R_{EQ}, V_{EQ}) in $S_{T\text{-}On}$.

If $S(t_n) = S_{T-On}$,

then $R_{EQ}(t_n) = \alpha \cdot R_{on} + (1-\alpha) \cdot R_{t0}$ and $V_{EQ}(t_n) = \alpha \cdot V_{on} + (1-\alpha) \cdot V_{t0}$. (4.2)

$$\alpha(t_n) = 1 - \exp\left\{-\frac{t_n - t_{0,on}}{\tau_{on}}\right\} \tag{4.3}$$

As time progresses from $t_{0,on}$, the turn-on relaxation coefficient α (4.3) increases from 0 to 1, and the I-V characteristic gradually "relaxes" from a higher impedance mode (R_{t0}, V_{t0}) to a lower impedance mode (R_{on}, V_{on}). The speed of relaxation is controlled by the time constant τ_{on}. τ_{on} may be modeled as a constant or as a bias-dependent variable; section 4.3.2 addresses the trade-offs associated with these two options.

In (4.2), (R_{t0}, V_{t0}) depend on $S_{0,on}$, the previous state from which the device transitioned to $S_{T\text{-}On}$.

a) If $S_{0,on} = S_{Hi\text{-}Z}$, then ($R_{t0}$, V_{t0})=(R_{tr} , V_{tr}).

b) If $S_{0,on} = S_{T\text{-}Off}$, then ($R_{t0}$, V_{t0})=($R_{EQ}(t_{0,on})$, $V_{EQ}(t_{0,on})$).

Case (a) occurs when the device is first triggered on. Case (b) applies to a device that has already been turned on and off, and is triggered on again at t_n.

3) Turn-off State ($S_{T\text{-}Off}$)

The model transitions from $S_{T\text{-}On}$ to $S_{T\text{-}Off}$ when I_{DUT} drops below the holding current (I_{hold}); when this occurs, the time-point is recorded as $t_{0,off}$. Eq. (4.4) gives (R_{EQ} , V_{EQ}) in $S_{T\text{-}Off}$.

If $S(t_n) = S_{T-Off}$, (4.4)

then $R_{EQ}\,(t_n) = (1-\beta)\cdot R_{EQ}\left(t_{0,off}\right) + \beta \cdot R_{off}$ and $V_{EQ}(t_n) = (1-\beta)\cdot V_{EQ}\left(t_{0,off}\right)$.

$$\beta(t_n) = 1 - \exp\left\{-\frac{t_n - t_{0,off}}{\tau_{off}}\right\} \tag{4.5}$$

After $t_{0,off}$, the turn-off relaxation coefficient β (4.5) increases from 0 to 1, and (R_{EQ},V_{EQ}) relaxes from ($R_{EQ}(t_{0,off})$,$V_{EQ}(t_{0,off})$) to (R_{off}, 0V) at a rate determined by the time constant τ_{off}. Previous works have shown that the turn-off process is not instantaneous [9],[10],[45],[46]. Regardless of whether the device stores significant charge when I_{DUT} drops below I_{hold}, it is beneficial to set τ_{off} equal to a few simulation time-increments to avoid a rapid change in (V_{DUT}, I_{DUT}), as this may cause simulation convergence problems. After the device has been in state $S_{T\text{-}Off}$ sufficiently long, as defined in (4.6), the device is considered fully turned off, and $S(t_n)$ returns to $S_{Hi\text{-}Z}$.

$$t_n - t_{0,off} > 6 \cdot \tau_{off} \tag{4.6}$$

The condition in (4.6) ensures β>0.998, so (R_{EQ},V_{EQ}) are very close to (R_{off}, 0V).

4.2.3 Application to Non-snapback Devices

The PWL-TR modeling technique may also be applied to ESD devices and circuits that do not undergo snapback, such as a diode. In this case, the three primary I-V branches will intersect at a single I-V point (V_{t1}, I_{hold}), as indicated in (4.7).

$$V_{t1} = I_{hold} \cdot R_{off} = V_{on} + I_{hold} \cdot R_{on} = V_{tr} + I_{hold} \cdot R_{tr} \tag{4.7}$$

4.3 Simulation Results and Discussion

4.3.1 Diode

Figure 4.3(a) shows the measured and simulated responses of a P+/N-well diode to current pulses with different rise-times. Figure 4.3(b) plots the peak (V_{peak}) and the steady-state (V_{SS}) values of V_{DUT} against the corresponding steady-state current level. Here, a PWL-TR model with constant τ_{on} is clearly sufficient to reproduce V_{peak} as a function of pulse rise-time.

4.3.2 Silicon-controlled Rectifier (SCR)

Figure 4.4 shows the pulsed I-V characteristic and a sample transient waveform for a grounded-gate NMOS-triggered SCR (GGSCR) [80]. Two PWL-TR model implementations are compared: (i) τ_{on} is a constant, and (ii) $\tau_{on} = f(I_{DUT})$, with the function f implemented as a lookup table. For this device, a PWL-TR model with bias-dependent τ_{on} provides a better fit over the entire range of I_{DUT}. Two physical phenomena underlie the bias-dependent τ_{on}. First, the N-well and P-well of the SCR need to be fully charged before the device can achieve its minimum on-resistance, a process that proceeds more rapidly at high current levels [81],[82]. Second, at high current levels, V_{DUT} can be so high that the N-well/P-well junction begins to avalanche [28],[29]. The function $\tau_{on} = f(I_{DUT})$ is extracted from transient waveforms such as that in Figure 4.4(a). For each pulse current amplitude, τ_{on} is adjusted by matching the simulated waveform to the measured one.

A PWL-TR model with bias-dependent τ_{on} is inherently nonlinear because (R_{EQ}, V_{EQ}) in $S_{T\text{-}On}$ depend on the instantaneous current $I_{DUT}(t_n)$. In contrast, a PWL-TR model with constant τ_{on} always acts as a linear model because of the bias-independent (R_{EQ}, V_{EQ}) in each state. Regardless, both PWL-TR model realizations are less complicated than conventional compact models of ESD devices, resulting in a reduced simulation time. To achieve the largest reduction in simulation time, one should use PWL-TR models with constant τ_{on}.

4.4 Summary

Any two-terminal ESD protection device (e.g. diode or GGSCR) or circuit (e.g. ESD rail clamp) can be described as a PWL-TR model for circuit-level ESD simulation. Piecewise-linear models enable speed-up of circuit simulation and obscure the circuit's internal structure. Transient relaxation allows the model to reproduce a device's transient response, which is needed to accurately simulate overvoltage stress caused by fast ESD transients. Setting the turn-on time constant (τ_{on}) to a bias-independent value yields a model that is linear at each time point, thus minimizing the simulation time and without

compromising accuracy in many cases. A bias-dependent τ_{on} may be used to more accurately model the voltage overshoot of complicated protection circuits, while still obscuring the circuit details and providing simulation speed-up relative to compact models.

4.5 Figures

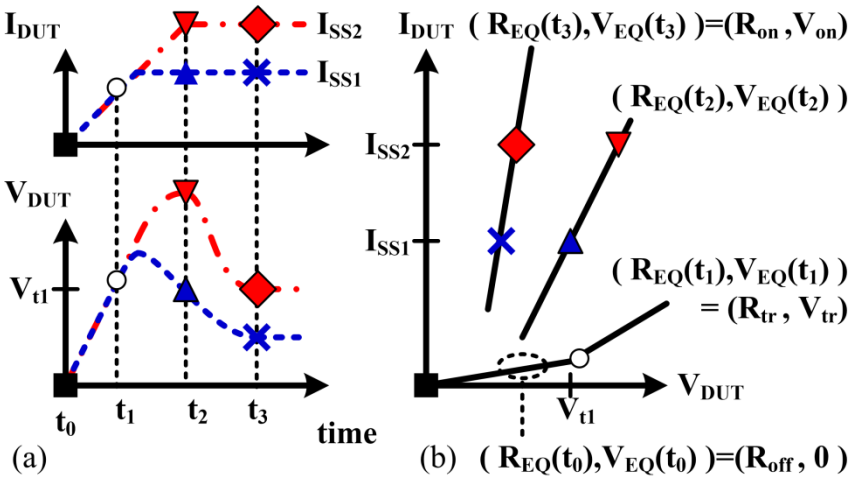


Figure 4.1 (a) Snapback device's terminal voltage $V_{DUT}(t)$ in response to two current pulses $I_{DUT}(t)$ of different amplitudes. (b) Multiple I-V branches are obtained by sampling $V_{DUT}(t)$ and $I_{DUT}(t)$ at different time-points. Each branch may be represented by a Thévenin equivalent circuit (R_{EQ}, V_{EQ}).

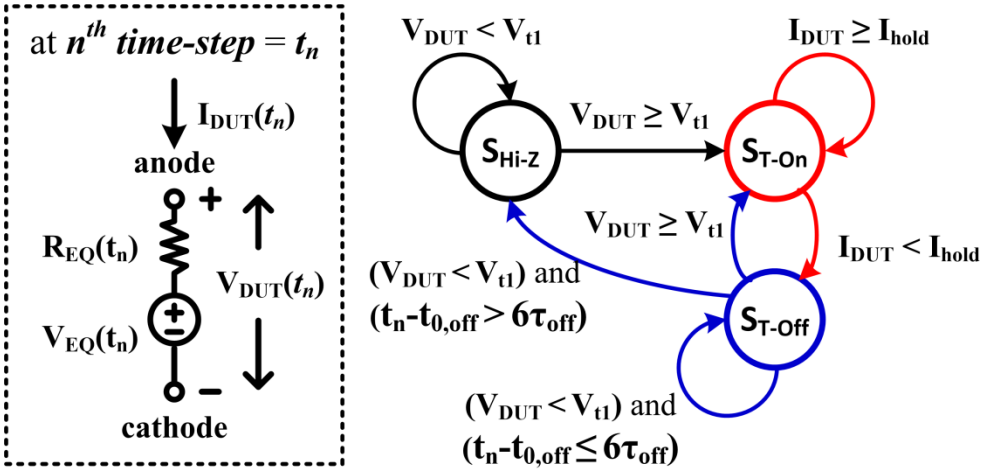


Figure 4.2 Piecewise-linear model with transient relaxation (PWL-TR), implemented as a finite state machine (FSM).

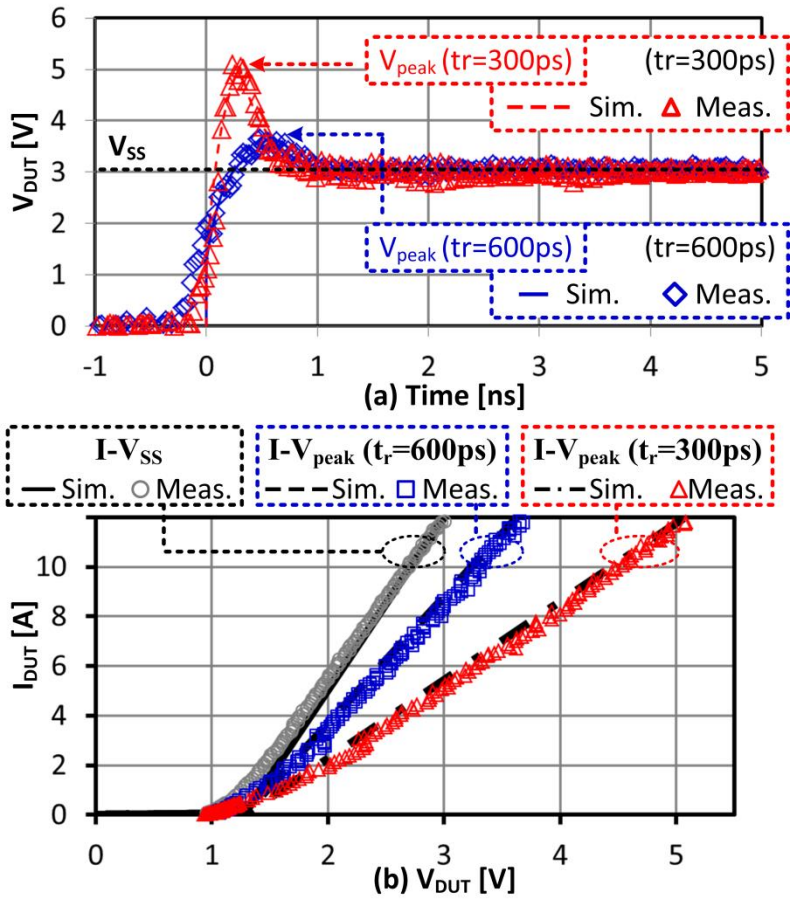


Figure 4.3 (a) $V_{DUT}(t)$ for a 130nm CMOS P+/N-well diode that is subjected to 12A current pulses with different rise-times (t_r). Total junction perimeter is 450µm. (b) Pulsed I-V curves. I-V_{SS} is the I_{DUT} and V_{DUT} at steady-state. I-V_{peak}(t_r) shows the peak voltage for a given steady-state current level and t_r.

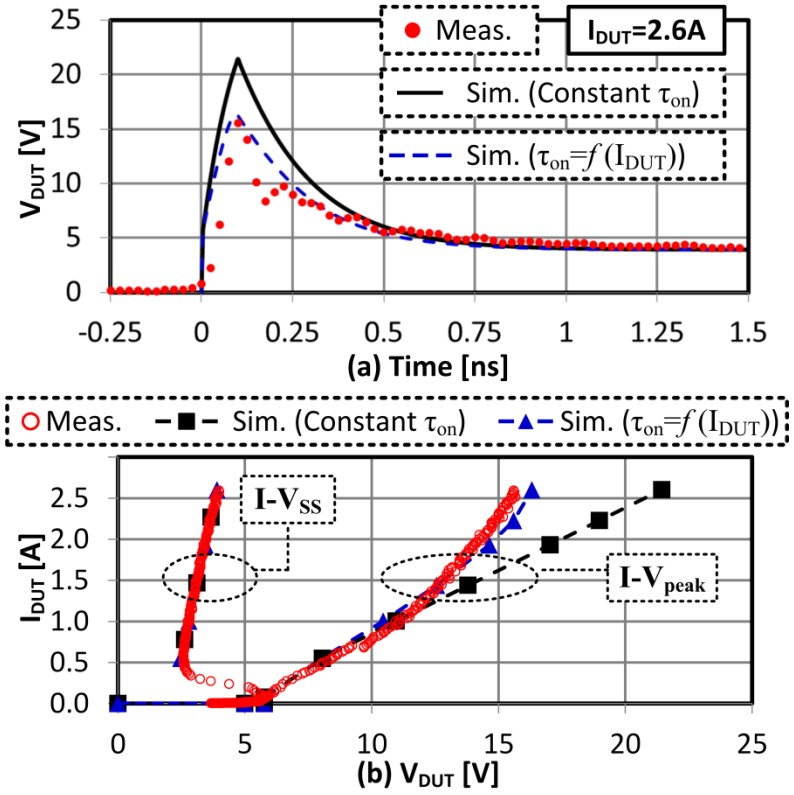


Figure 4.4 (a) Response of a 65nm CMOS GGSCR [83] to 2.6A current pulse with t_r = 100ps. The SCR is 40μm wide. (b) Pulsed I-V characteristics. t_r=100ps.

Chapter 5. Full-Component Modeling and Simulation of Charged Device Model ESD

This chapter presents a methodology to construct an equivalent circuit model of a packaged integrated circuit mounted on a field-induced charged device model electrostatic discharge tester. Circuit simulation is used to obtain the full-component current and voltage distributions. This work focuses on predicting overvoltage stress at power domain crossing circuits.

5.1 Motivation

The charged device model (CDM) emulates electrostatic discharges (ESD) associated with the automated manufacturing and handling of integrated circuits. Product return rates correlate closely with CDM qualification results [84]; failure signature can be replicated by the field-induced CDM (FICDM) test [85],[86]. FICDM tester may be calibrated to either JEDEC [5] or ESDA standard [4]. Figure 5.1 depicts a packaged IC mounted on a FICDM tester. Each pin is tested ("zapped") at a progressively higher pre-charge voltage (V_{CDM}) until failure occurs; both positive and negative V_{CDM} are used. The peak discharge current measured at the zapped pin (zap-pin) is on the order of several amperes. Due to the high current level, circuits on the die may get damaged [87],[88].

Circuit simulation can predict if the component will have adequate CDM robustness. Previous works on circuit-level modeling and simulation of CDM-ESD [7],[89]–[97] have been successful in optimizing and verifying pad-ring I/O and ESD circuitry design [98]. Circuit simulation has also been used to predict the CDM susceptibility of domain-crossing circuits between different core power domains [91],[99],[100]. These works demonstrate that the "cross-domain stress" is affected by the package interconnections, die substrate, pad-ring ESD circuitry, and the on-die power distribution network (PDN), all of which should be represented in the simulation model. To this end, this work presents a methodology to construct a full-component CDM model. The objective is to limit the size of the netlist without unduly

compromising the simulation accuracy; practical guidelines for modeling the core PDN, N-well/P-well junctions, and substrate taps are provided.

5.2 Analysis of Cross-Domain Stress

During FICDM testing, the field-induced charge is stored on each unshielded conductive segment of the package, on-die metal, and the die substrate, exposed to the field charge plate (FP) or the ground plate (GP). During a CDM event, current travels between the zap-pin and the many charge storage capacitors; the portion of this current that flows on-die generates overvoltage to the devices near the zap-pin, elsewhere in the pad-ring, or even deep in the core circuitry, due to the full-component nature of the discharge.

Within the circuit core, it is the domain-crossing circuits that are likely to be stressed most heavily [88]. Figure 5.2 illustrates a zap to a V_{DD} pin in domain-1 of a chip with two independent core power domains and a separate IO supply. Prior to the CDM event, the chip is unpowered, and the logic gates in the domain-crossing circuit are not in any defined state. Thus, it is difficult to predict the potential on the signal line (V_X) during the discharge. However, considering the given setup, V_X is bounded by (5.1).

$$V_{SS,TX} \leq V_X \leq V_{DD,TX} \tag{5.1}$$

$V_{SS,TX}$ and $V_{DD,TX}$ denote the potentials on the $V_{SS,1}$ and $V_{DD,1}$ buses in the vicinity of TX, respectively. The cross-domain voltage stresses on the gate dielectric of RX NMOS and PMOS are $|V_{GS,N}|$ (5.2) and $|V_{GS,P}|$ (5.3), respectively.

$$\begin{aligned} |V_{GS,N}| \equiv V_X - V_{SS,RX} &\leq V_{DD,TX} - V_{SS,RX} \\ &= [V_{DD,TX} - V_{SS,TX}] + [V_{APD,1} + V_{Bus,IO} + V_{APD,2}] + [V_{SS,TX} - V_{SS,1,Edge}] \\ &+ [V_{SS,2,Edge} - V_{SS,RX}] \end{aligned} \tag{5.2}$$

$$\begin{aligned}\left|V_{GS,P}\right| \equiv V_X - V_{DD,RX} \le V_{DD,TX} - V_{DD,RX} &= \left\{V_{DD,TX} - V_{SS,RX}\right\} + \left[V_{SS,RX} - V_{DD,RX}\right] \\ &= \left|V_{GS,N}\right| + \left|V_{DD,RX} - V_{SS,RX}\right|\end{aligned} \tag{5.3}$$

In (5.2) and (5.3), $V_{SS,RX}$ and $V_{DD,RX}$ denote the potentials on the $V_{SS,2}$ and $V_{DD,2}$ buses in the vicinity of RX, respectively. Each component in the cross-domain stress can be placed into one of three categories:

1) V_{DD}-to-V_{SS} potential difference across power rails: $\left[V_{DD,TX} - V_{SS,TX}\right]$ in $\left|V_{GS,N}\right|$ and $\left[\left|V_{DD,TX} - V_{SS,TX}\right| + \left|V_{DD,RX} - V_{SS,RX}\right|\right]$ in $\left|V_{GS,P}\right|$.
2) Voltage drop in the pad-ring V_{SS}/V_{SSIO} network, $\left[V_{APD,1} + V_{Bus,IO} + V_{ADP,2}\right]$, appears in both equations. Due to the high CDM current level, the voltage drop across an APD is mainly contributed by the ohmic regions rather than the junction. For example, at 10A current, the diode series resistance of 0.5Ω contributes 5V, much larger than the junction voltage drop.
3) Core-to-ring voltage drop. For a given net, say $V_{SS,1}$, this is the voltage drop from the chip periphery to its core. The relevant term is $\left[V_{SS,TX} - V_{SS,1,Edge} + V_{SS,2,Edge} - V_{SS,RX}\right]$.

$V_{GS,N}$ and $V_{GS,P}$ may be either positive or negative, depending on the polarity of VCDM and which pin is zapped. The domain-crossing circuit is at risk if the simulated $V_{GS,N}$ or $V_{GS,P}$ exceed the gate dielectric breakdown voltage (BV_{OX}) of the corresponding devices. BV_{OX} differs between PMOS and NMOS and between bias in accumulation and inversion [101]. However, the process design kit (PDK) often only provides the lowest value for all possible stress situations. While the magnitude of BV_{OX} for thin-oxide MOSFETs at ESD time-scale is about 4 to 6V, it can vary by about 1V [102],[103],[49], so both $V_{GS,N}$ and $V_{GS,P}$ should be monitored for CDM-risk assessment.

To predict whether the voltage stress during a given pin-zap is excessive and to identify the worst-case zap-pin requires full component transient simulation. Details on how to construct the simulation netlist are provided in the following sections.

5.3 Full-Component Model for CDM Simulation

5.3.1 FICDM Tester Model

As shown in Figure 5.1, the FICDM tester model represents the pogo-pin, spark, FP-to-GP capacitor (C_{FG}), and the pre-charge voltage supply. The pogo-pin is modeled as a 1Ω resistor and an inductor (L_{pogo}, typically 4 to 6nH [90]). The spark is modeled as a resistor (R_{spark}), typically 20Ω to 25Ω [90]. Values of C_{FG}, R_{spark} and L_{pogo} can be extracted from the measured current waveform with a calibration target [95].

5.3.2 Package Interconnect Model

Each pad-to-pin interconnection may be represented by a multi-stage RLC circuit. The number of stages is set such that the delay of each stage is much smaller than the rise-time of the ESD current. For instance, in [96], each trace in a 40x40mm2 BGA package is represented by a 20-stage RLC circuit because the interconnect delay is comparable to the CDM current pulse rise-time [97]. In contrast, 1-stage [89],[75],[92] or 2-stage [94] models are sufficient for small packages. In Figure 5.3, a 3-stage model is used for a TQFP package [93]. Many of the electrical parameters in the model are given in the package date-sheet, but a few adjustments may be required. In particular, the inductance must represent the partial self-inductances, rather than the loop self-inductances which are based on an assumed ground plane in a printed circuit board. If the partial self-inductances are not provided in the date-sheet, they can be estimated analytically [93].

Two or more on-chip bond pads may be connected together by package-level metal which provides an off-die discharge path to reduce the overvoltage generated on chip. The stress reduction is especially pronounced if package metal is used to connect V_{SS} busses that are isolated on die [99]. Therefore, it is critically important that the package model reflects any connections between the package-level traces.

5.3.3 Pad-ring Model

To describe the discharge path in the pad-ring, the model should reflect the ESD protection implemented in each pad-ring cell and the topology of the pad-ring structure [89],[76],[98]. The connectivity between cells must also be preserved. Each bus segment between adjacent pad-ring cells is modeled as a resistor. In the example in Figure 5.4, the I/O power supply is formed by a continuous ring of V_{DDIO}/V_{SSIO} buses. In contrast, D0 and D1 are isolated core power domains, so their pad-ring supply buses (V_{DD0}/V_{SS0} and V_{DD1}/V_{SS1}, respectively) do not extend around the entire chip periphery. The pad-ring model also reflects the placement of anti-parallel diodes (APD) that connect the V_{SS} buses to the V_{SSIO} bus.

5.3.4 Core Power Distribution Network Model

Figure 5.5 illustrates a V_{DD} network of a single core domain using a mesh-style PDN planning. The V_{DD} networks of the core PDN are represented by meshes of resistors (R_{PDN}'s). The same modeling principle applies to the V_{SS} network as well. The core PDN model must capture the floorplan and connectivity of the V_{DD}/V_{SS} network of each core power domain to the pad-ring buses and to other domains. If isolated core domains are adjacent to each other, the corresponding V_{DD}/V_{SS} meshes should not be interconnected.

The following guidelines help limit the number of nodes in the model without unduly compromising accuracy.

1) Only the upper-level metal portion of the power distribution network is represented by the model. This constitutes the main, low-impedance discharge path. Relatively little current flows through the high-resistivity thin-metal layers used for local power distribution.
2) The connectivity between the core PDN and the pad-ring buses must be preserved. Each tap from the pad-ring to core PDN is represented by a resistor.

3) Each intersection in the top-level PDN metalization corresponds to one PDN node in the core PDN model.
4) A long stripe of PDN interconnect with no intersections can be represented by only one R_{PDN}, provided that its maximum I-R drop during CDM is small (less than $\frac{1}{10}BV_{OX}$ for reasonable spatial resolution). Otherwise, it should be further partitioned accordingly.
5) A special case of 3). If there are only a few domain-crossing circuits supplied by the long stripe, and their locations are known, the placement of the PDN nodes can be based on the physical locations of these domain-crossing circuits.

The nonlinear devices that comprise the domain-crossing circuitry need not be included in the core PDN model. These devices conduct a negligible fraction of the CDM current, so excluding them speeds up simulation without significantly affecting the full-chip current distribution. The resulting setup provides a conservative estimate of the nodal potentials in the PDN, because it excludes the (relatively small) V_{DD} to V_{SS} leakage current. Using the notation in Figure 5.2, the voltage stresses on a domain-crossing circuit can be estimated using (5.4) and (5.5) at a given simulated time-point (t_n).

$$\left|V_{GS,P}\right| \leq \max\left\{\left|V_{DD,TX} - V_{DD,RX}\right|, \left|V_{SS,TX} - V_{DD,RX}\right|\right\} \tag{5.4}$$

$$\left|V_{GS,N}\right| \leq \max\left\{\left|V_{DD,TX} - V_{SS,RX}\right|, \left|V_{SS,TX} - V_{SS,RX}\right|\right\} \tag{5.5}$$

$V_{DD,TX}$, $V_{SS,TX}$, $V_{DD,RX}$, and $V_{SS,RX}$ are simulated nodal potentials at t_n on the PDN nodes supplying the domain-crossing circuit. A post-processing script automatically monitors both $\left|V_{GS,P}\right|$ and $\left|V_{GS,N}\right|$ of each domain-crossing circuit at each t_n, and records their highest magnitudes and polarities throughout the simulated time-span. Should any recorded voltage stress exceed the corresponding BV_{OX}, the domain-crossing circuit is at risk. A subsequent "localized simulation" can be performed to design the domain-crossing circuit and local ESD protection using the full-component CDM simulation results. Details are given in section 5.5.

5.3.5 Die Substrate Model

The p-type substrate is modeled as a 3D resistive mesh, as shown in Figure 5.6. The semiconductor substrate can be represented as a RC model. However, the typical substrate resistivity (ρ_{sub}) for a CMOS process is in the range of 0.01 to 10 Ω-cm, and the dielectric relaxation time spans from 0.01ps to 10ps. In comparison, the rising and falling edges of a CDM pulse are several hundreds of picoseconds. Therefore, a model containing only resistors is reasonably accurate as long as the simulation time-step is longer than the dielectric relaxation time. The value of each resistor in the 3D mesh ($R_{bulk\text{-}x}$, $_{-y}$, and $_{-z}$) is calculated based on the resistivity of the substrate, the geometry of the die, and mesh sizes in each direction [7],[73],[92].

The mesh density of the 3D mesh in the x- and y-directions, i.e., parallel to the core PDN, is the same as that of the core PDN model, so the first layer of the substrate mesh (n_Z=1) can be tapped into the core V_{SS} network. Each V_{SS}-tap merges the contacts, vias, and low-level metal stacks into one equivalent resistor ($R_{VSS,Tap}$) to capture the connectivity from the p-substrate (and p-well) to the core V_{SS} mesh. Note that even if adjacent core V_{SS} meshes are not connected with PDN metalization, they still share a common p-type substrate, which reduces the cross-domain stress during CDM [73].

In the z-directional, the p-bulk mesh is partitioned into N_Z layers; a heuristic method to determine N_Z is provided in section 5.3.8. If the p-substrate is conductively bonded to a metallic die-attach plate (DAP), the DAP can be treated as a short, and all the nodes in the last p-bulk layer (n_Z=N_Z) are merged into a single node [73]. If the DAP is made of insulating material, all the nodes in the last layer remain separated [92].

5.3.6 On-Die Decoupling Capacitance Model

The dominant capacitances between the V_{DD} and V_{SS} nets are those associated with the decoupling capacitors (“de-caps”) and the well junctions; they help limit the magnitude of the CDM-induced voltage stress [73],[99]. De-caps are usually made of arrays of MOS capacitors, designed to

provide V_{DD}-to-V_{SS} noise suppression, and explicitly placed throughout the chip. In the pad-ring model, de-cap is represented as a capacitor with series resistor, and integrated into the pad-ring cell model. In the core PDN model, for each pair of PDN nodes in the V_{DD}/V_{SS} meshes, the associated de-caps are lumped as a capacitor with series resistor between the corresponding PDN nodes.

A complete model for the parasitic N-well/P-well junction is illustrated in Figure 5.7. These junction elements are distributed between the V_{DD} and V_{SS} nets. Thus, the granularity of the model of the N-well/P-well junctions should match that of the core PDN model and the pad-ring model. The parameters for the junction element can be estimated from the layout and given in the PDK.

Explicit de-caps made of MOS capacitors are bias-dependent. The N-well junction capacitors and diodes (C_{JSW}/D_{JSW} and C_{JBTM}/D_{JBTM}) are also bias-dependent. In section 5.4.2, it will be shown that these nonlinear capacitors may be approximated as linear capacitors without significantly affecting the simulation results, and thus simpler de-cap models can be used to speed up the simulation.

5.3.7 Tester-Die Coupling Model

Charges stored on the die are modeled by arrays of capacitors: the $C_{Die\text{-}FP}$'s and the $C_{Die\text{-}GP}$'s, representing coupling from the die to FP and GP, respectively, as illustrated in Figure 5.8. Any package can be placed into one of two rough categories, reflecting the types of tester-die coupling:

1) The p-type substrate (attached to the die attach plate) faces FP, as illustrated in Figure 5.8. $C_{Die\text{-}FP}$ is coupled only to the common p-bulk, and $C_{Die\text{-}GP}$ to the nets in the top metal layers.
2) The top metal layers face FP, as in the case in Figure 5.1. $C_{Die\text{-}FP}$ is coupled to the top metal layers which are usually reserved for the PDN, and $C_{Die\text{-}GP}$ to the common p-bulk.

The methodology to estimate the total values of $C_{Die\text{-}FP}$ and $C_{Die\text{-}GP}$ is provided in [93]. The total capacitances are then partitioned into arrays, connected to the nodes in PDN and the p-bulk model. In the example of Figure 5.8, GP is coupled to the die top-level metal layers, so the $C_{Die\text{-}GP}$ array is connected to the PDN with the same mesh density as the core PDN model. Also, since FP is coupled to the p-substrate,

the $C_{Die\text{-}FP}$ array has the same mesh density as the x-y plane of the die substrate model. However, if the last layer of the p-bulk mesh can be merged into a single node, $C_{Die\text{-}FP}$ may be modeled as a single lumped capacitor.

5.3.8 P-bulk Network Z-directional Meshing

A heuristic is devised to determine optimal N_Z, given ρ_{sub} and the x- and y-directional meshes, using the simplified representation of an IC in Figure 5.9. The setup assumes single power domain. It only includes models for the FICDM tester, p-bulk, tester-die coupling, and the V_{SS} network in the pad-ring and core PDN. CDM simulation of the simplified component is performed with increasing N_Z from 2. Consecutive simulation results are compared using two metrics: error (5.6) and cost (5.7).

$$e[N_Z] = \frac{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(V_{[N_Z]}(t_n) - V_{[N_Z-1]}(t_n)\right)^2}}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(V_{[N_Z-1]}(t_n)\right)^2}} \tag{5.6}$$

where $t_{n\in[0,N-1]}$ is the nth simulation time-point.

$$c[N_Z] = \frac{simulation\ time\ [N_Z]}{simulation\ time\ [N_Z = 2]} \tag{5.7}$$

$e[N_Z]$ is the root-mean-square (RMS) difference between the nodal voltage with N_Z and that with $N_Z - 1$, normalized to the RMS value of the nodal voltage obtained at $N_Z - 1$. All simulations in the heuristic method use the same constant time-step, so for each n, t_n is the same in all simulations. $e[N_Z]$ is calculated for all the nodes in the netlist; the maximum is used as the measure of convergence rate against N_Z. As N_Z increases, $e[N_Z]$ decreases and asymptotically approaches a lower bound. $c[N_Z]$ represents the overhead in the simulation run-time due to not using the minimum N_Z=2. Given the purely linear netlist, the simulation result is only affected by the z-directional meshing of the p-bulk. Figure 5.10(a) shows an example of $e[N_Z]$ and $c[N_Z]$. Observe that $e[N_Z]$ become worse (higher) when the substrate resistivity is lowered. Because a larger fraction of the CDM current will flow through the substrate with lower ρ_{sub}, it requires a finer mesh.

To optimize N_Z, a figure of merit (FOM) is defined in (5.8).

$$FOM[N_Z] = \frac{1}{e[N_Z] \cdot c[N_Z]} \tag{5.8}$$

The optimal z-directional mesh density, defined as N_Z*, is what maximizes the FOM. It aims to provide the best trade-off to simultaneously minimize the error e[N_Z] and the run-time overhead c[N_Z]. For the case study of Figure 5.10(b), N_Z*=6.

5.4 Simulation Results and Discussion

The importance of including models representing the FICDM tester, package, pad-ring, and substrate has been illustrated in [73],[74],[7],[93],[94],[76]. For the rest of the model components, a series of test-cases are analyzed in the following subsections. These test-cases share the same pad-ring cell sequence and core power domain floorplan, as illustrated in Figure 5.11. It has three core power domains. Domain D0 resembles the application-specific logic domain, composed of standard-cell blocks or "sea of gates". D1 and D2 are noise-sensitive domains, such as analog-to-digital converter or clock generating circuit. Being noise sensitive, they require their own V_{DD}/V_{SS} supplies (V_{DD1}/V_{SS1} and V_{DD2}/V_{SS2}), isolated from the supplies of D0 (V_{DD0}/V_{SS0}).

There is one continuous ring of I/O V_{DD}/V_{SS} buses (V_{DDIO}/V_{SSIO}) in the pad-ring. Core domains are also supplied by the associated V_{DDk}/V_{SSk} (k=0,1,2) buses in the pad-ring. Wherever the pad-ring V_{SSk} bus is broken, APDs are placed to connect the broken V_{SSk} to nearby V_{SSIO}. The test-cases assume a 65nm mixed-signal CMOS technology, with substrate resistivity of 2Ω-cm, in a lead-frame TQFP package of 96 pins. The size of the package is 14x14 mm^2; the die size is 6.8x6.8 mm^2, attached to the die attach plate of 8x8mm^2.

5.4.1 Current Bypass Effect by Core Power Distribution Network

The CDM current flows both in the pad-ring buses and in the core PDN, as suggested by the simulation result in Figure 5.12. If the core V_{DD}/V_{SS} meshes are not included in the simulation, all the

current is forced to flow along the pad-ring buses. As a result, the I-R drops along the pad-ring buses and APDs are exaggerated, causing overestimation of cross-domain overvoltage. Without the core PDN model, V_{SS0}–V_{SS2} becomes 7.00V instead of 3.83V, and V_{SS0}–V_{SS1} becomes 3.65V instead of 1.14V. Figure 5.12 proves that the core PDN model must be included in the full-component CDM simulation, so the on-die CDM current flow and cross-domain stresses can be properly simulated.

5.4.2 Effect of Decoupling Capacitors

The second test case is devised to evaluate if decoupling capacitors, or "de-caps," reduce the cross-domain voltage stress. In this test case, de-caps are distributed uniformly throughout the core PDN and are modeled as ideal, i.e., they have zero series resistance and a bias-independent capacitance. Figure 5.13 shows the simulated worst-case cross-domain stress as a function of total decoupling capacitance. Even a small amount of de-cap can drastically reduce the stress. Simulation shows that it is the V_{DD}-to-V_{SS} term in the cross-domain stress that gets reduced by the de-caps. This example demonstrates that, in addition to the rail-clamps, de-caps must be included in the full-component model to correctly predict the V_{DD}-to-V_{SS} potential differences during simulated CDM events.

CMOS ICs contain a large amount of parasitic N-well/P-well junction capacitance. For example, in a standard-cell design flow, N-wells take up roughly half of the chip area. Both the area capacitance and the side-wall capacitance contributed by the N-well/P-well junctions are significant. A test-case is devised to investigate whether these parasitic capacitances augment the intentional de-caps in a significant way, thereby necessitating their inclusion in the CDM model. The full-component model is modified to include the N-well/P-well junction elements (Figure 5.7), assuming half of the chip area is occupied by N-wells. All of the "intentional" de-caps are removed from the chip model. For the test-case used in Figure 5.14, simulation indicates that without the N-well/P-well junctions, the maximal cross-domain stress would be 8.72V, but in the presence of these N-well/P-well junctions, the cross-domain stress is reduced to 7.19V.

The bias-dependency of the well capacitance makes the junctions non-linear elements; it is worthwhile to consider if significant error would be introduced by instead modeling these as linear capacitors. Two constant capacitance values are used: one is the 0V-bias value and the other is the value obtained for a reverse bias of 4.5V. As summarized in Figure 5.14, the simulation results obtained with the linear capacitor models do not differ much from those obtained with the nonlinear capacitor model.

Based on Figure 5.14, the junction element model can be simplified. Simulation also confirms that eliminating the diodes from model of the N-well/P-well junction element does not cause an observable difference in in the simulation results. Thus, the junction element only needs to include linear elements: the parasitic resistances ($R_{VSS,Tap}$, $R_{VDD,Tap}$, R_{PW}, R_{NW}, and R_{BTM}) and linear capacitors (C_{JSW} and C_{JBTM}) with well capacitances calculated at a fixed bias. Reducing the model complexity also results in a reduced simulation run-time and improved convergence.

Figure 5.13 implies that, once the total impedance of all de-cap models in parallel is low enough, the V_{DD}-to-V_{SS} stress is insensitive to the exact capacitance. For the test-case used in Figure 5.14, this is valid even at a reverse bias of 4.5V with parasitic well resistances in series. This situation may not be true for any test-case; however, the nonlinear capacitor model can still simplified as linear capacitor to boost simulation efficiency, as long as the reverse bias V_R is conservatively estimated. Larger non-zero reverse bias V_R lowers the bias-dependent capacitance, thereby resulting in a more conservative value of the CDM-induced voltage stress (see Figure 5.14). A conservative estimate of V_R is the highest V_{DD}-to-V_{SS} stress for the test-case to withstand. It is suggested that V_R be chosen as follows. The full-chip ESD network design is usually undertaken with a V_{DD}-to-V_{SS} design target (e.g., a fraction of BV_{OX} of core NFET and PFET). V_R can be set to the V_{DD}-to-V_{SS} design target. If the V_{DD}-to-V_{SS} target is yet to be determined, one can perform full-component simulation without the de-cap model. Once V_R is acquired, both bias-dependent MOS capacitor model and well junction capacitor model can be replaced by linear capacitor, using information provided by the PDK and the acquired conservative V_R.

5.4.3. Effect of Package-level Ground Network

To demonstrate the power of package-level common grounding, test-cases TC-C0 and TC-C1 are constructed. TC-C0 is the baseline setup that includes all the basic model components. TC-C1 augments TC-C0 with an additional package-level ground ring that connects all the V_{SS} and V_{SSIO} pads. Figure 5.15 compares the simulated stresses in the pad-ring V_{SS} buses from the two test-cases. The package-level ground ring serves as an additional conductance in parallel with the APDs in the pad-ring, and prevents CDM current from entering the die. As a result, the I-R drops along the pad-ring buses and APDs can be drastically reduced. Without ground-ring, TC-C0 has maximal $|V_{SS1} - V_{SS0}| = 6.65V$; with ground-ring, TC-C1 has reduced $|V_{SS1} - V_{SS0}| = 2.87V$ at the same location. This example demonstrates that a CDM-robust design should be comprehensive, considering not only the on-die ESD protection network in the pad-ring and the core PDN floor-planning, but also the interconnect design at the package level.

5.4.4 Effect of Tester-Die Coupling

Table 5.1 compares the simulation results of two test-cases:

1) TC-D1 emulates the package whose $C_{Die\text{-}FP}$ is coupled to the common p-substrate.
2) TC-D2 is a hypothetical test-case in which there is no tester-die coupling.

Both test-cases assume the same design (Figure 5.11), same geometry for the die, and die attach plate. Simulation shows that the amount of charges stored directly on the die contributes substantially to the total charge of the DUT (23% in the given test-case). Without the tester-die coupling model, TC-D2 predicts lower pogo-pin current and cross-domain stress, especially the I-R drop along the VSS nets. This result shows that the tester-die coupling model must always be included; otherwise, the total field-induced charge storage cannot be properly modeled, and the CDM current flow and on-die overvoltage will be underestimated.

5.5 Localized Simulation of Domain-Crossing Circuit

The models proposed above for full-component CDM simulation do not include any of the nonlinear devices in the core circuitry. As argued in 5.3.4, the corresponding simulation will provide a fast, numerically stable, conservative yet reasonable estimate of overvoltage stress (5.4) and (5.5) at all the domain-crossing circuits. These results can be used to ascertain the worst-case pin-zap and which domain-crossing circuits are at risk of CDM-induced damage.

Once the at-risk domain-crossing circuits are identified, one may wish to modify the circuit design to remove the CDM hazard, e.g. by deploying local ESD protection [91]. It is proposed that a "localized" simulation be performed to facilitate this process. The schematic for a localized simulation is illustrated in Figure 5.16; its input sources — the voltage waveforms on the supply nets of the TX and RX — are extracted from the earlier, full-component CDM simulation. The stimuli, $V_{DD,TX}(t)$, $V_{SS,TX}(t)$, $V_{DD,RX}(t)$, and $V_{SS,RX}(t)$, should be extracted at PDN nodes in the vicinity of the domain-crossing circuit being simulated.

Figure 5.17 compares two simulation setups:

a) This setup includes the domain-crossing circuit of interest with various local ESD protection schemes, placed at the edge of two domains in the core, inside a full-component CDM model.
b) This setup corresponds to the localized simulation. It includes the same domain-crossing circuit and local ESD protections as in setup (a). The stimuli are extracted at the same location where the core devices in setup (a) are placed, from the full-component simulation without core devices.

The results of the two setups are very similar. Since the localized simulation uses only a small netlist, the run-time is fast, even if the included devices are highly nonlinear.

Although there may be a large number of domain-crossing circuits in the core, only a few localized simulations are necessary. The following guidelines should be applied.

1) Only if the estimated cross-domain stresses ((5.4) and/or (5.5)) from full-component simulation exceed safe limits, should localized simulation of a domain-crossing circuit be performed.
2) If a domain crossing-circuit needs to undergo localized simulation, it should be simulated only one time, corresponding to the worst-case pin-zap. The input stimuli should be extracted from the worst-case full-component pin-zap simulation.
3) If a cluster of domain-crossing circuits are in close proximity to one another, and if they share a similar design, e.g. cross-domain communication on an M-bit bus, then only one localized simulation will be sufficient to represent all.

5.6 Summary

The CDM ESD robustness of a packaged IC can be evaluated by circuit simulation with a full-component netlist. Only components that constitute the main charge storage and discharge current paths are included in the netlist; unnecessary details such as low-level metallization or nonlinear core devices are excluded to speed up the simulation. The CDM-induced voltage stress is determined by the on-chip ESD protection, the on-die power distribution networks, the package-level inter-pin connections, and the on-die decoupling capacitors; all of these affect the discharge current path. From the full-component CDM simulation, the worst-case pin-zap and the power domain-crossing circuits at high risk can be identified. Subsequently, the at-risk circuits may be redesigned to include local ESD protection; a localized simulation is performed to assess the design modifications. The stimuli for the localized simulation are extracted from the full-component simulation and are implemented as independent excitations.

5.7 Supplement: Full-component Netlist Extraction

The procedure to extract the full-component netlist from a given design is outlined as follows.

For the upper-level metal layers that are laid out in a regular pattern to form the pad-ring buses and core PDN meshes, the corresponding models can either be extracted from the actual layout file

(GDSII) or calculated analytically. To maintain the simplicity of the model, only the equivalent resistance of a **repeating unit** is necessary. For the pad-ring, it is the bus resistance (R_{BUS}) between two adjacent cells. For the core PDN, it is the resistance (R_{PDN}) between two PDN nodes. After R_{BUS}'s and R_{PDN}'s are obtained, a script can be used to **automatically** generate the entire pad-ring and core PDN models, given an outline of the pad-cell sequence and core PDN floorplan. It is found that using a commercial extractor (e.g., Calibre PEX) is less efficient because it produces a redundant, complicated resistor **network**, even for a simple stripe of metal line. On the other hand, the resistance of a stripe of metal stack can be easily calculated accurately, using the sheet resistances provided by the PDK. It has been verified that the terminal resistance of a stipe of metal stack extracted by the commercial tool is the same as that by calculation. Manually building the model guarantees the result to be minimal but sufficient, without redundant circuit elements or nodes.

For irregular PDN, it may be acceptable to use commercial a layout extractor. However, an irregular PDN usually corresponds to a customized circuit block with limited layout footprint. Therefore, it remains possible to manually construct the model. After all, the customized PDN should simply be a small collection of vertical and horizontal metal stripes whose resistances can easily be calculated.

Extracting the p-type substrate is similar to extracting a regular core PDN mesh. Given the mesh density in each direction, the dimension of each unit of $R_{bulk\text{-}x,\ -y,\ -z}$ can be determined, and the corresponding resistance can be calculated simply as a three-dimensional semiconductor.

To extract the model of each pad-cell, any circuit block that is not relevant to ESD current path is first removed. The remaining layout thus contains only the ESD devices (including de-cap). The path resistance between each ESD device to the relevant I/O pad or pad-ring buses is extracted and accounted for in the model. Here, it is recommended to use a commercial tool to extract the path resistances because the local routing within each pad-cell tends to be highly congested and irregular. However, it is possible to replace each extracted path resistance network with a single equivalent resistor.

To extract the on-die decoupling capacitor model, the designer should devise a repeating unit which represents a circuit macro that is powered by one set of V_{DD}/V_{SS} PDN. The repeating unit assumes a collection of core circuits which are most likely to be re-instantiated frequently. Then, the explicit de-cap and the parasitic well decap per PDN node can be estimated, and the on-die decoupling capacitor model can be built. Although this heuristic unrealistically assumes that the decoupling capacitance is uniform among the core PDN nodes, the total amount of de-cap is usually high, so the cross-domain stress is not sensitive to the exact value of de-cap, as illustrated in section 5.4.2.

The procedure to obtain the rest of the model component has been outlined in section 5.3.

5.8 Table and Figures

Table 5.1 Effect of tester-die coupling. Both test-cases (TC-D1 and -D2), described in section 5.4.4, undergo zaps to D1 V_{DD}-pin (No.76) at V_{CDM}= –500V.

Simulation Results	Test Cases	
	TC-D1	TC-D2
Field Charge Plate (FP) to Die Coupling	to P-substrate (reference)	N/A
Total field-induced charge [nC]	4.90	3.79 (-23%)
Peak pogo-pin current [A]	10.56	9.89 (-6.4%)
Cross-Domain Stress V_{DD1}-V_{DD0} [V]	8.21	7.38 (-10%)
I-R drop on V_{SS} Nets V_{SS1}-V_{SS0} [V]	-4.80	-4.05 (-16%)

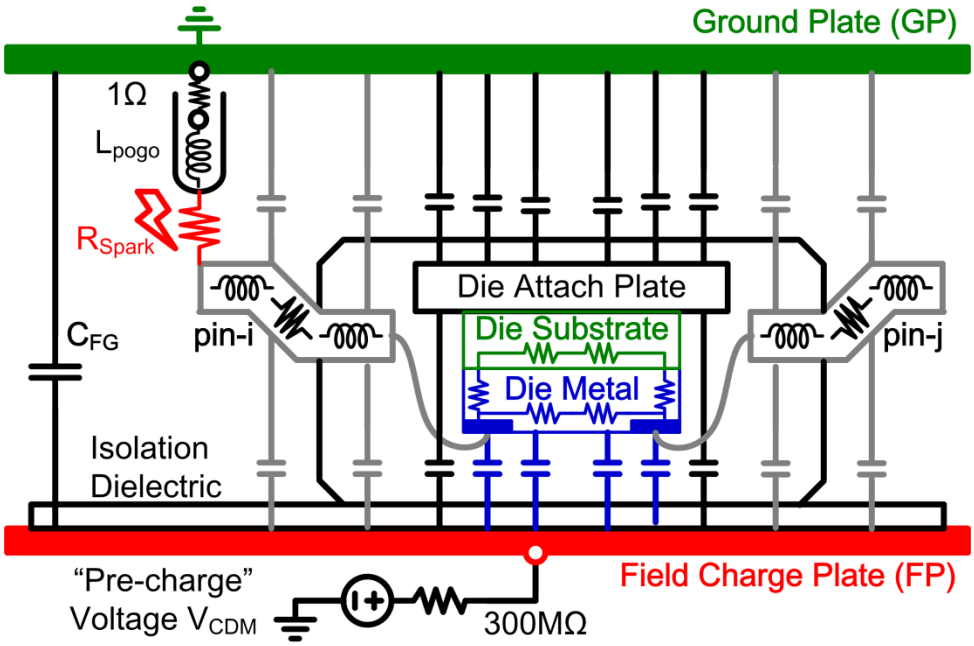


Figure 5.1 A packaged IC as the device under test (DUT) placed on a FICDM tester. The DUT is of a "cavity-up" wire-bonded lead-frame package.

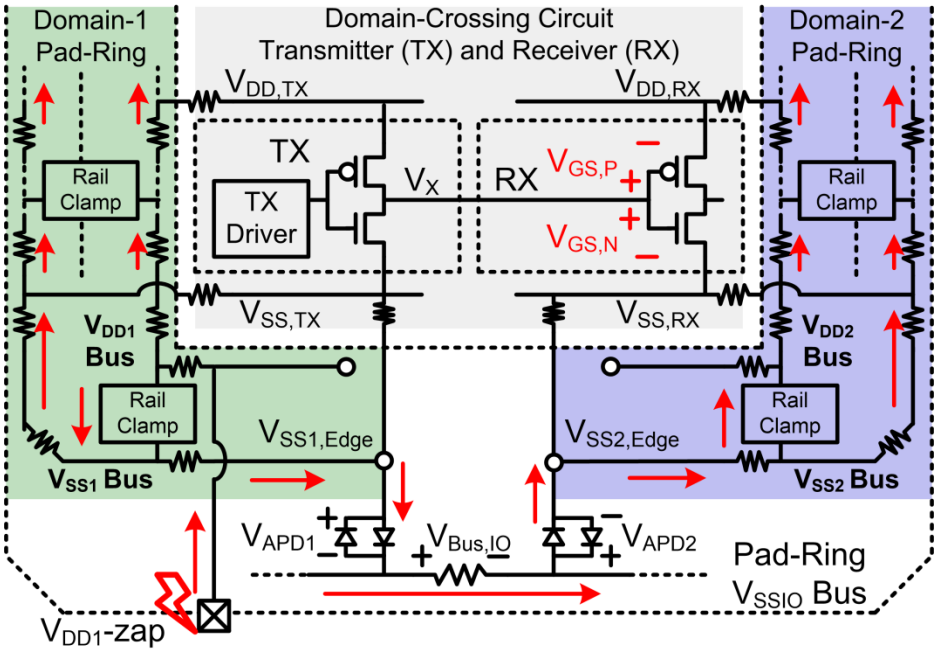


Figure 5.2 A domain-crossing circuit consisting of two digital buffers. The arrows indicate the direction of discharge current, assuming a V_{DD1} pin is zapped under negative V_{CDM}; i.e., CDM current is forced into the V_{DD1} pin.

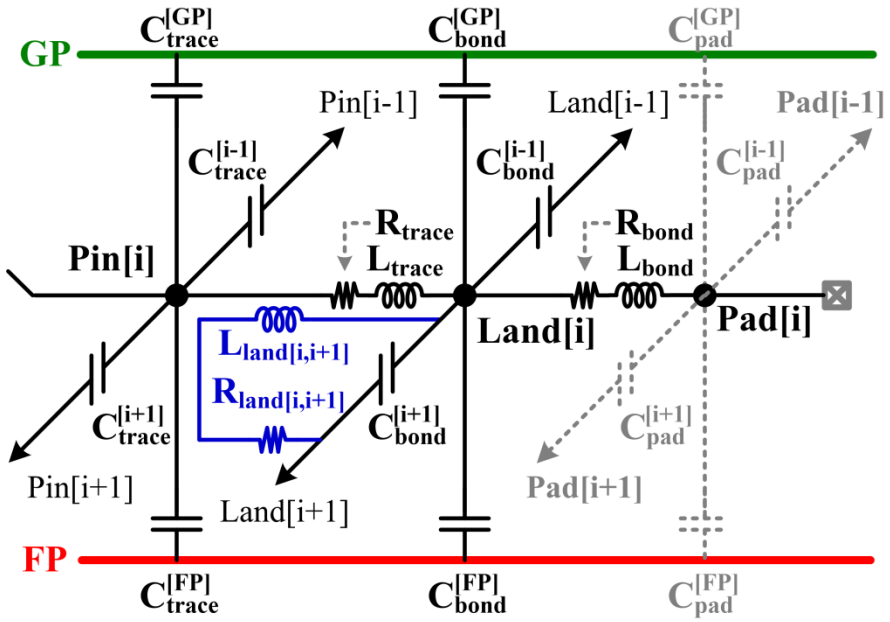


Figure 5.3 Each TQFP pin-to-pad interconnection is modeled as a 3-stage RLC circuit. For the ith interconnect, Pin[i] is connected to Pad[i] through (R_{trace}, L_{trace}) and (R_{bond}, L_{bond}) which join at Land[i]. $C_{pad}^{[FP]}$, $C_{bond}^{[FP]}$, and $C_{trace}^{[FP]}$ represent coupling to FP, and $C_{pad}^{[GP]}$, $C_{bond}^{[GP]}$, and $C_{trace}^{[GP]}$ to GP. Mutual capacitors from interconnect [i] to its nearest neighbors [i±1] include $C_{pad}^{[i\pm1]}$, $C_{bond}^{[i\pm1]}$, and $C_{trace}^{[i\pm1]}$. If Pad[i] and Pad[i+1] are connected together at package-level, the connection between Land[i] and Land[i+1] is represented by $(R_{land[i,i+1]}, L_{land[i,i+1]})$. The capacitors associated with Pad[i] (i.e., $C_{pad}^{[i\pm1]}$, $C_{pad}^{[FP]}$, $C_{pad}^{[GP]}$) will be included in the ith pad-cell in the pad-ring model (Figure 5.4); the rest of the RLC components are in the package model.

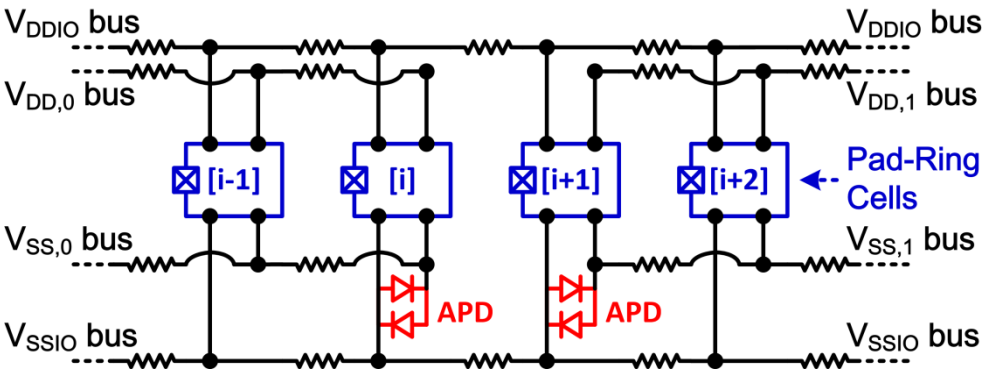


Figure 5.4 The pad-ring model includes a series of pad-ring buses and pad-ring cells that reflects actual ESD protection design.

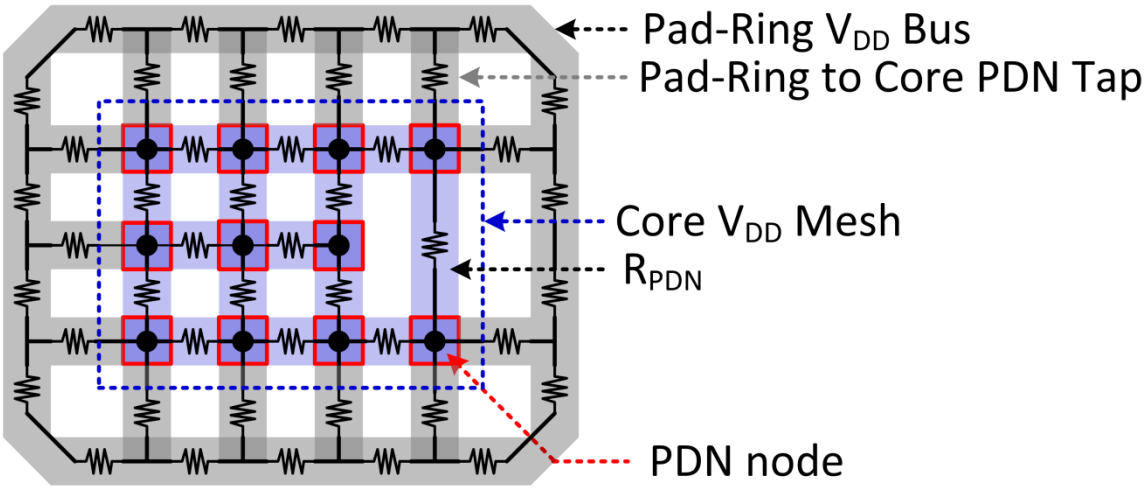


Figure 5.5 Illustration of the V_{DD} network in the pad-ring and in the core. The V_{DD} network of the core PDN is modeled as 2D meshes of resistors.

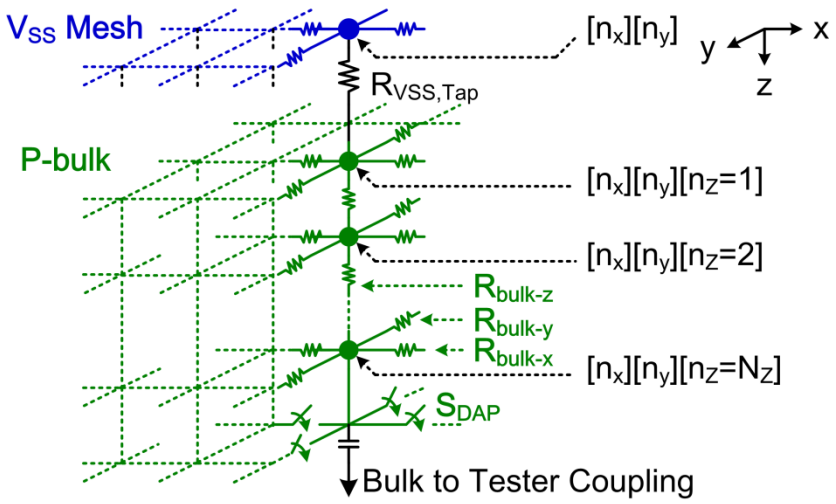


Figure 5.6 A 3D resistive network represents the die substrate. The first layer (n_Z=1) is resistively tapped to the V_{SS} mesh in the core PDN model. For the last layer (n_Z=N_Z), the switches (S_{DAP}) are closed if a metallic DAP is conductively bonded to the p-substrate; all S_{DAP} are open if the DAP is of insulating material. These S_{DAP} switches are not implemented in the netlist, but merely to show the connectivity of the last layer.

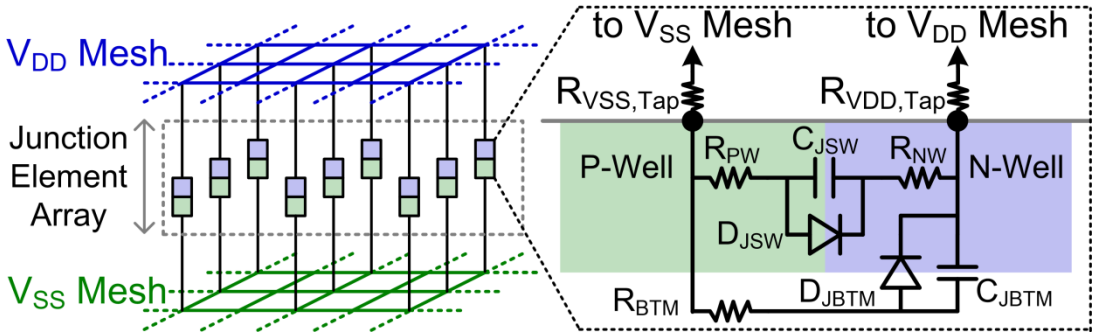


Figure 5.7 The many N-well/P-well junctions are represented as an array of junction elements between the nodes in the V_{SS} and V_{DD} networks. Each junction element includes a side-wall capacitor/diode (C_{JSW}/D_{JSW}) and a bottom capacitor/diode (C_{JBTM}/D_{JBTM}). R_{PW} and R_{NW} are parasitic resistances of the side-wall junction; R_{BTM} is the parasitic resistance of the bottom junction. $R_{VDD,Tap}$ and $R_{VSS,Tap}$ are the resistances associated with the contacts, vias, and low-level metal stacks that connect the wells to the V_{DD}/V_{SS} networks.

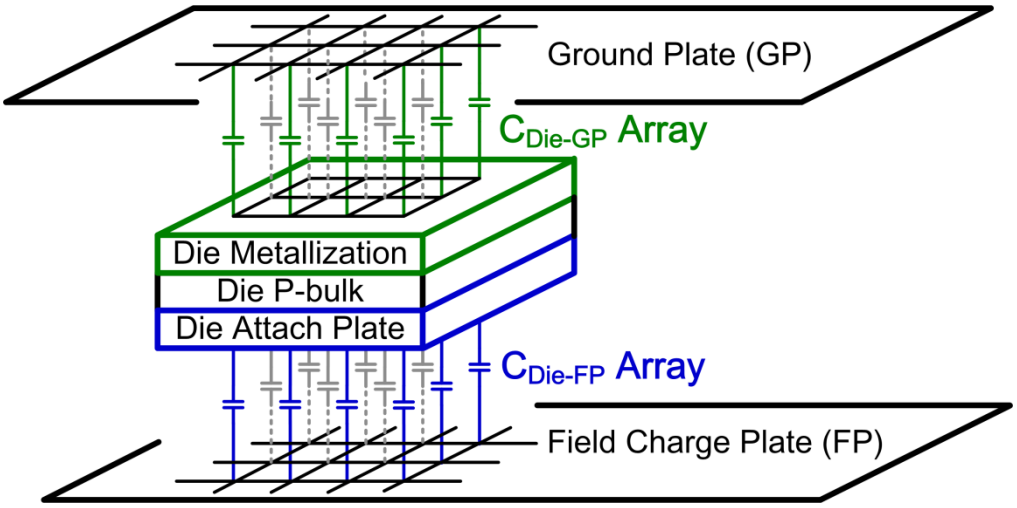


Figure 5.8 Tester-die coupling capacitors. For the illustrated orientation of the die, the C_{Die-GP} array is connected from GP to each node in the PDN in the die top metal. The C_{Die-FP} array is connected from FP to the last layer of nodes (n_Z=N_Z) in the p-bulk 3D mesh.

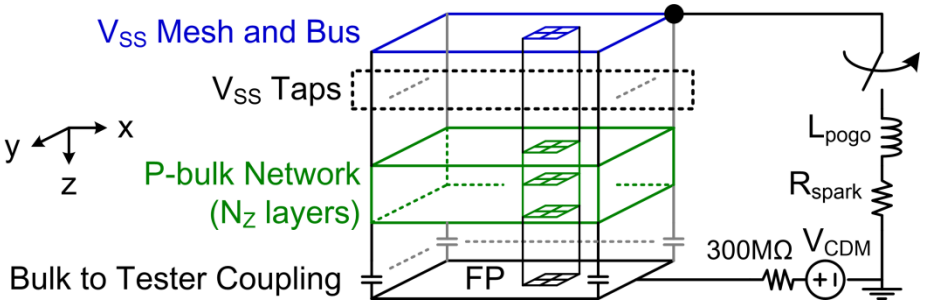
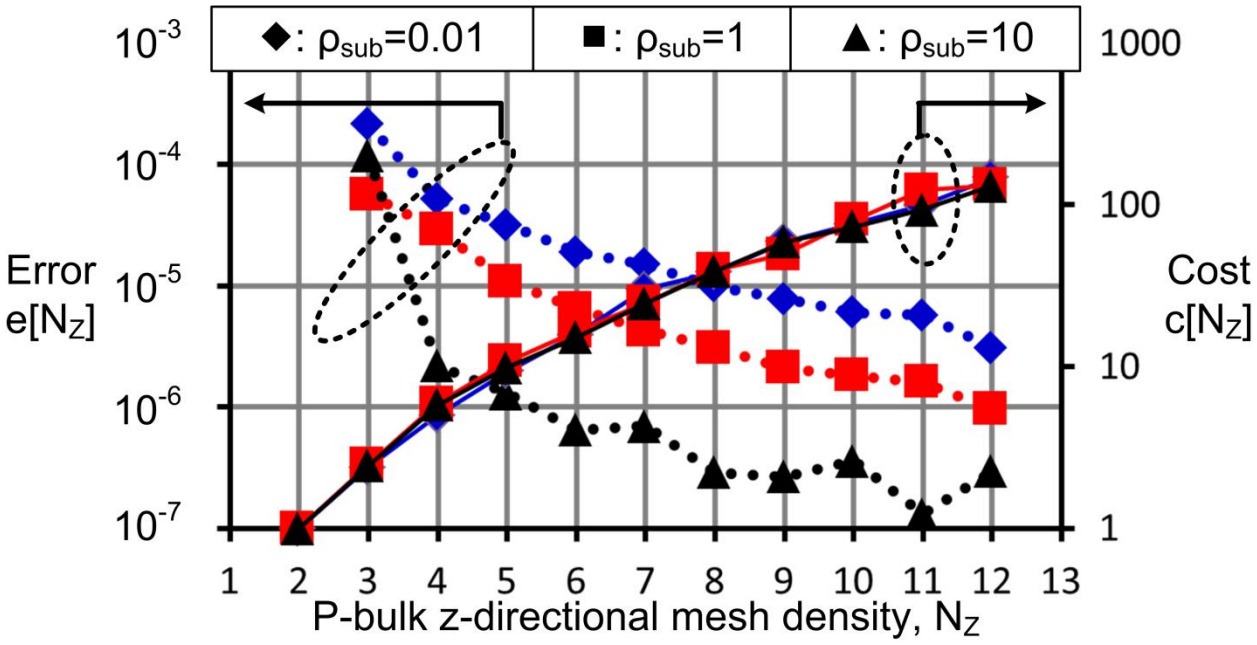
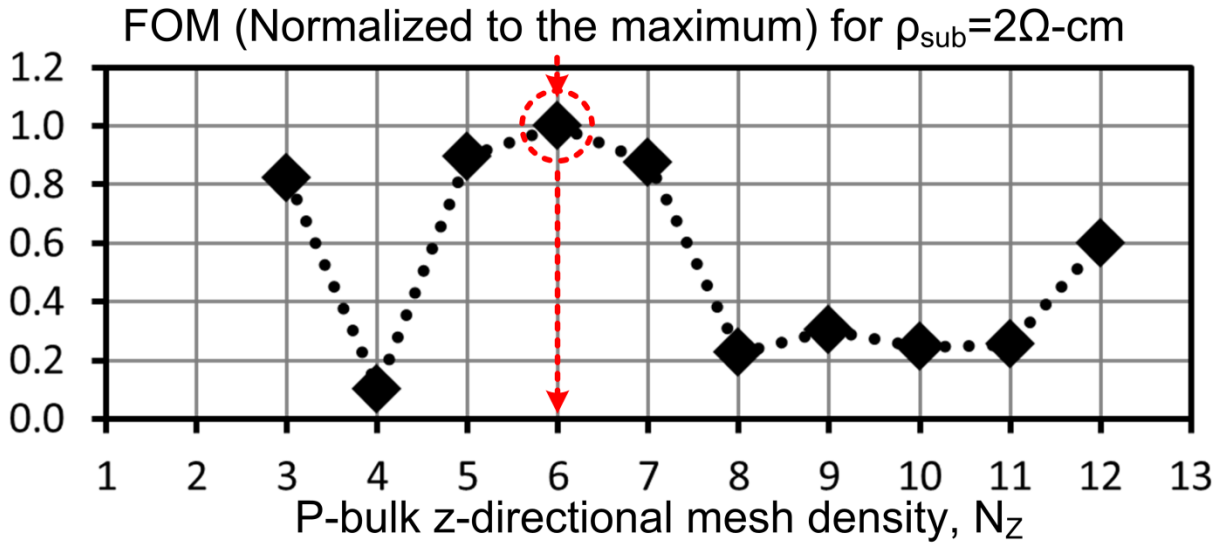


Figure 5.9 Test-substrate model for the heuristic for determining N_Z*, the optimal p-bulk z-directional mesh density.



(a)



(b)

Figure 5.10 (a) Error ($e[N_Z]$) and cost ($c[N_Z]$) as functions of z-directional mesh density (N_Z). (b) Figure of merit (FOM) as a function of N_Z. The fluctuations are due to simulator artifacts, such as tolerance and truncation in parameter values. The simulated die-size is 6.8x6.8mm^2, with die thickness of 230μm.

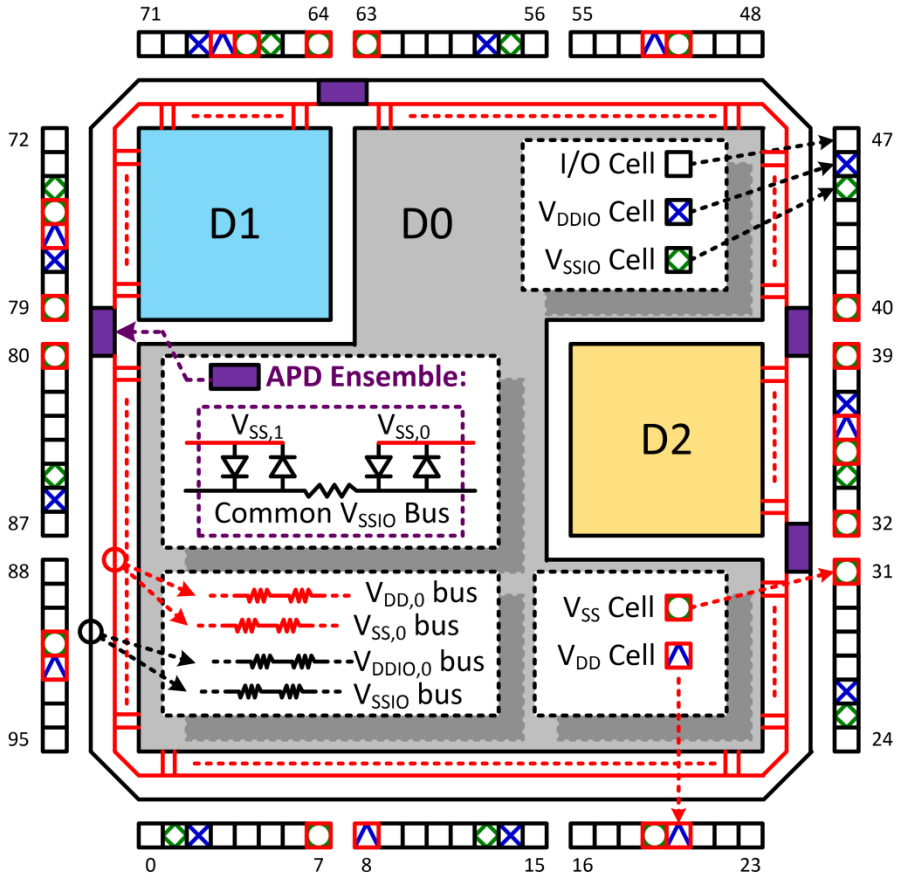


Figure 5.11 Pad-ring cell sequence and core power domain floorplan used in all test-cases. PDNs for different core domains (D0, D1, and D2) are isolated. V_{SS} nets are connected to the common V_{SSIO} bus in the pad-ring through the APD ensembles. Each pad-ring cell includes a selection of ESD protection devices suitable for the associated functionality [76],[98].

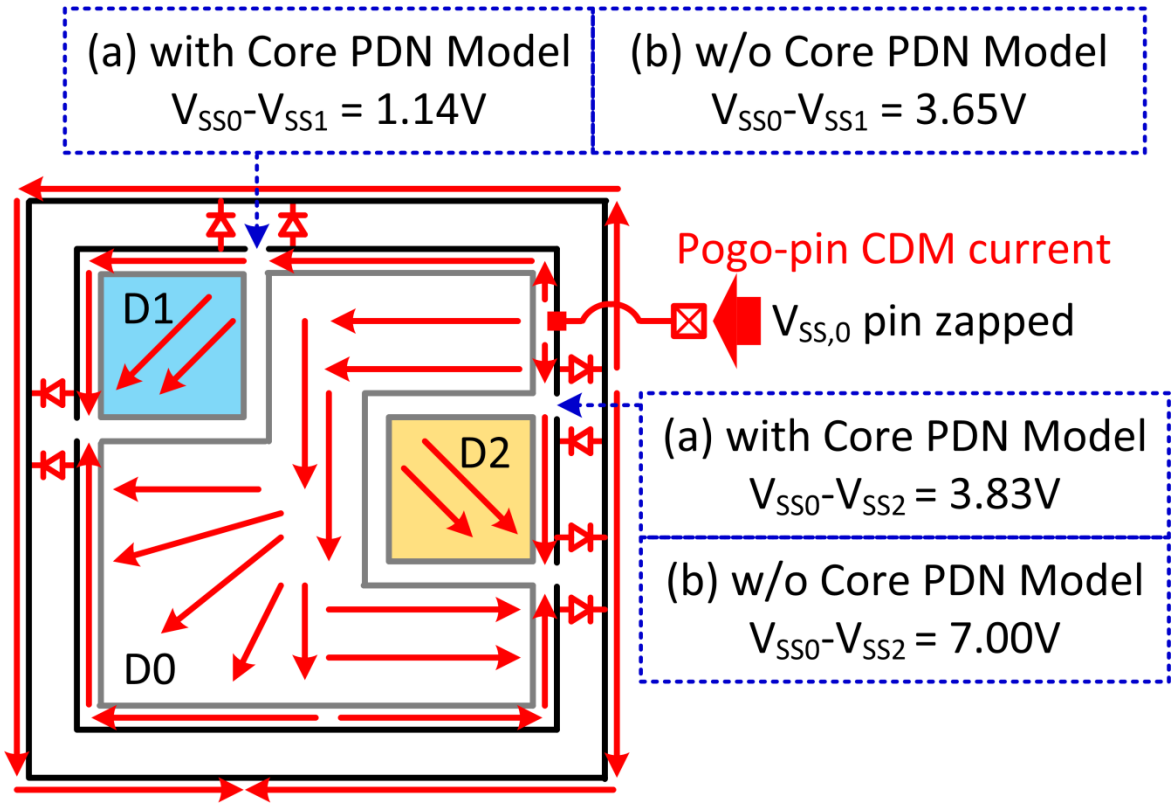


Figure 5.12 The figure summarizes the simulated potential differences along the pad-ring V_{SS} buses in two test-cases: (a) with vs. (b) without core PDN model, during a zap to a D0 V_{SS}-pin at V_{CDM}= –500V. If the core PDN model is included, a significant portion of the CDM current (indicated by the arrows) flows directly into the core PDN. De-caps in both test-cases are ideal capacitors and exist only in the pad-ring cells.

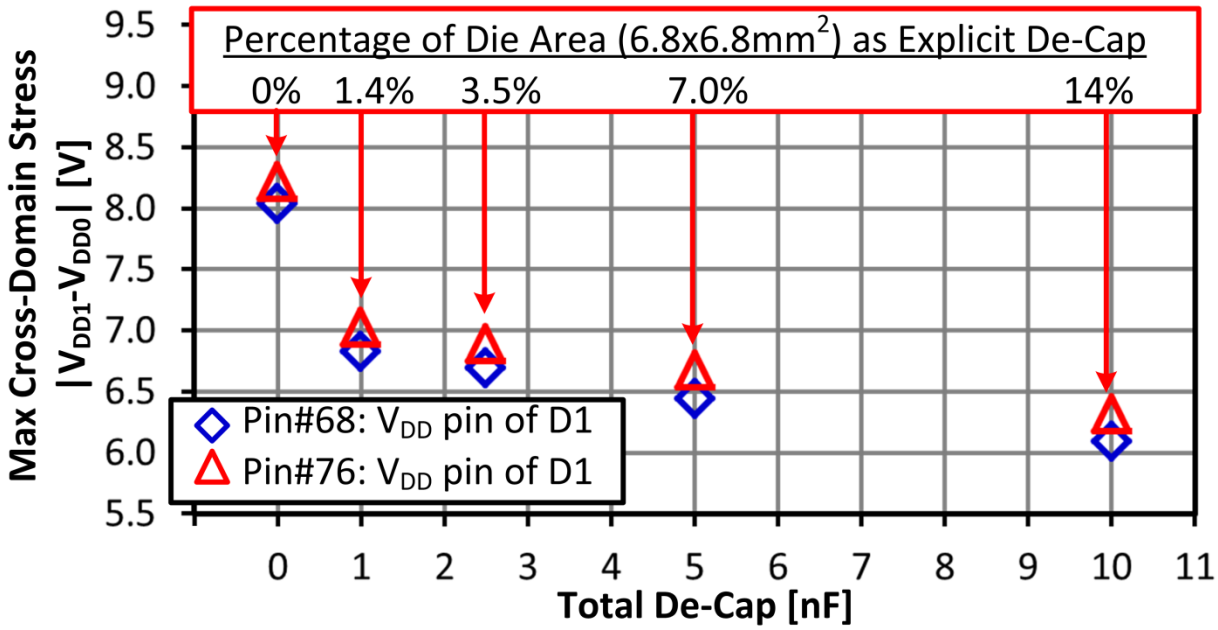


Figure 5.13 Effect of de-cap on the cross-domain stress in two different pin-zap configurations (V_{DD}-pins No. 68 and 76 in of D1 in Figure 5.11) under V_{CDM}= –500V. The de-cap assumes the capacitance of thick-oxide NMOSCAP in N-well whose capacitance per unit area is 15fF/μm2 if biased at 2.5V in depletion mode. The series resistance is assumed zero.

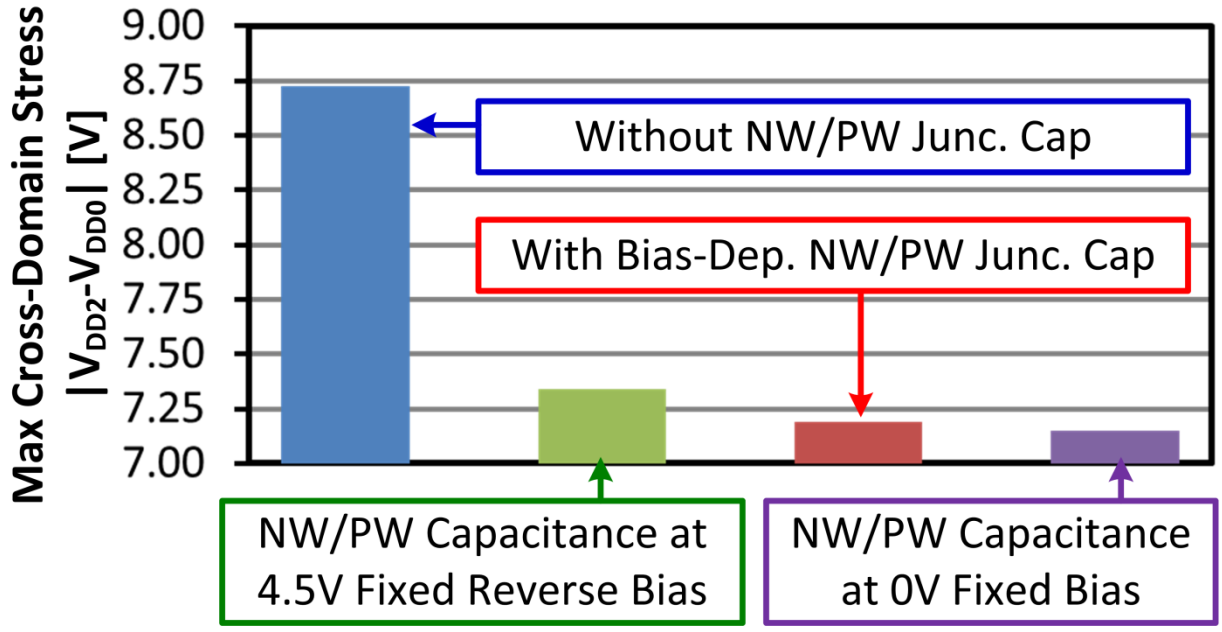


Figure 5.14 Effect of N-well/P-well junction capacitors. The figure shows the cross-domain stress when the V_{DD} pin (No. 36) of D2 in Figure 5.11 is zapped under V_{CDM}= –500V. Only the junction elements contribute to the capacitances between the V_{DD} and V_{SS} nets in the test-case.

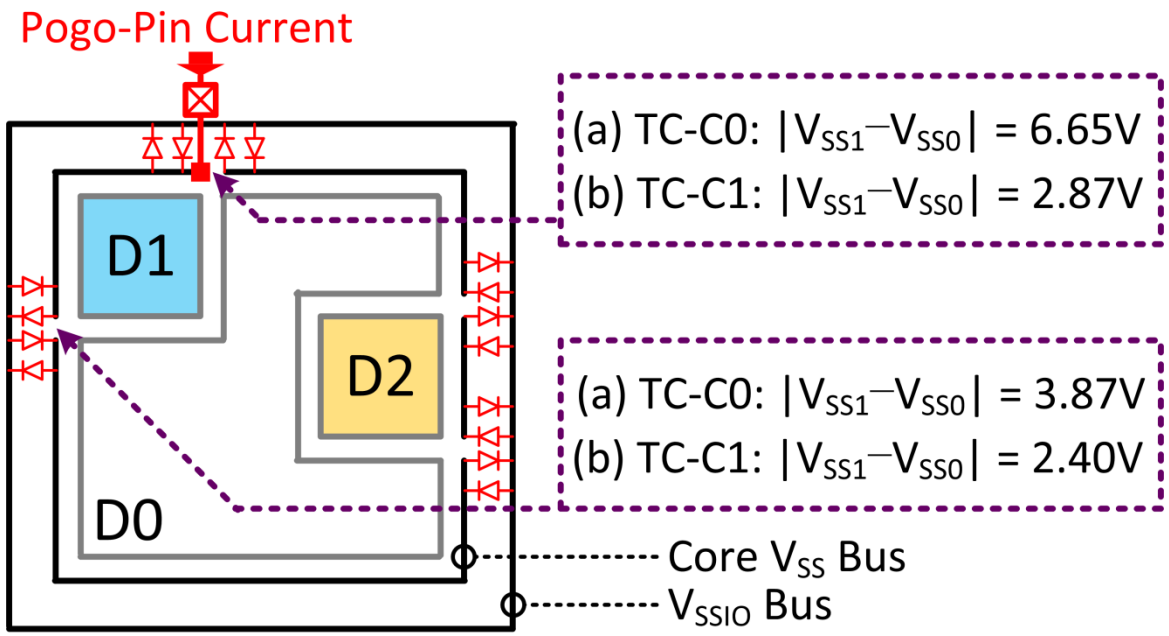


Figure 5.15 Comparison of simulated potential differences along the pad-ring V_{SS} buses of two test-cases: (a) TC-C0 vs. (b) TC-C1, during a FICDM zap to a D1 core V_{SS}-pin at V_{CDM}= –500V. Both share the same design (Figure 5.11), but TC-C1 has an additional package-level ground ring that connects all the V_{SS} and V_{SSIO} pads. The associated model parameters defined in Figure 5.3 are L_{land}=0.6nH and R_{land}=1mΩ.

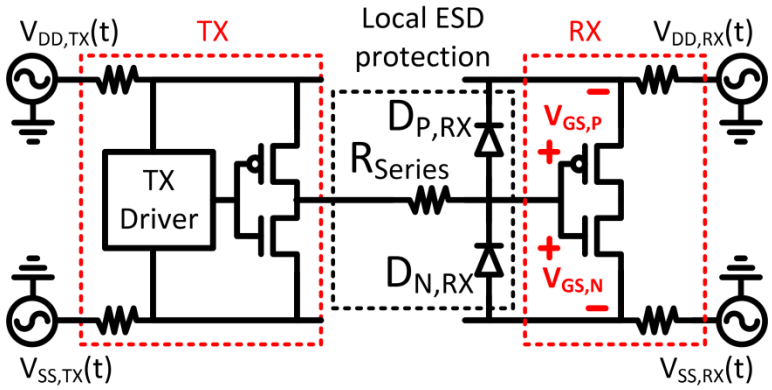


Figure 5.16 Setup for localized simulation of a domain-crossing circuit. The stimuli are the voltage waveforms on the V_{DD} and V_{SS} nets supplying the domain-crossing circuit: $V_{DD,TX}(t)$, $V_{SS,TX}(t)$, $V_{DD,RX}(t)$, and $V_{SS,RX}(t)$. These are obtained from the full-component simulation without the domain-crossing circuit.

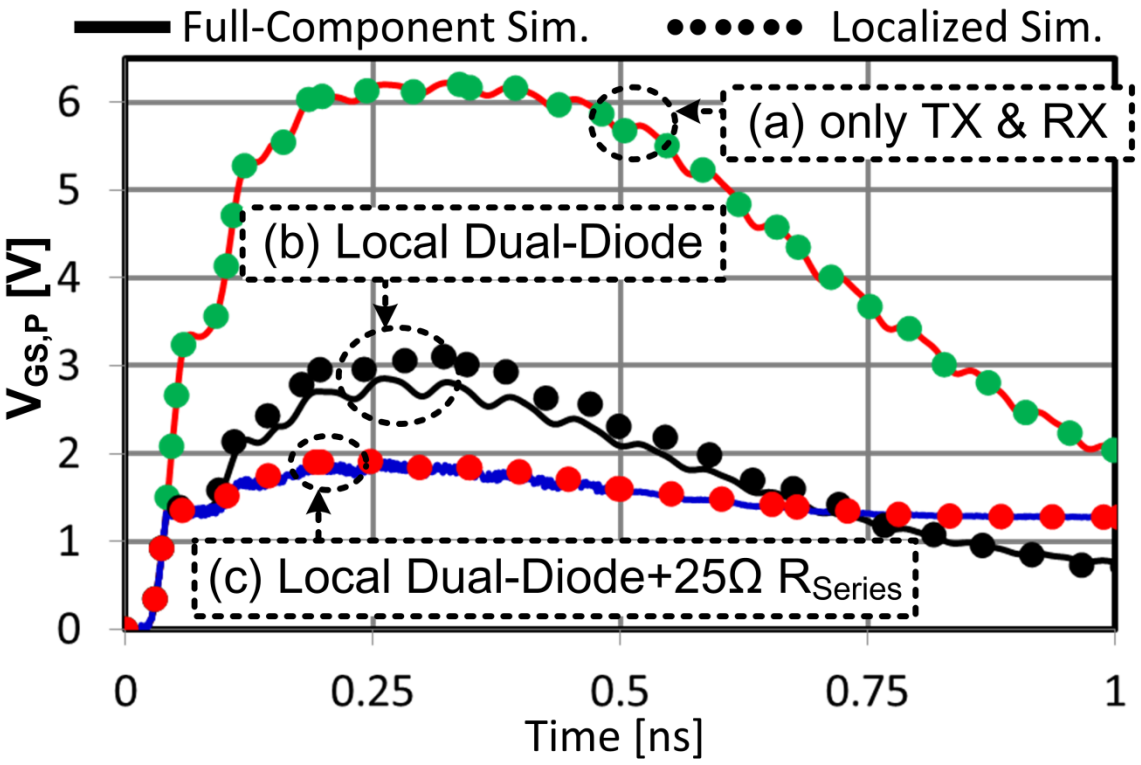


Figure 5.17 Simulated voltage stresses on the RX PMOS gate using localized simulation and full-component simulation. Both setups include the same domain-crossing circuit, along with three different configurations of local ESD protection defined in Figure 5.16. (a) No local ESD protection. (b) Only local dual-diode ($D_{P,RX}$ and $D_{N,RX}$) at the RX input. R_{Series}=0. (c) In addition to the local dual-diode, R_{Series}=25Ω.

Chapter 6. Fast Circuit Simulator for Transient Analysis of CDM ESD

Run-time for circuit-level CDM ESD simulation can be prohibitively long. By leveraging the specific properties of the problem formulation, run-time can be reduced without compromising accuracy. A reduced run-time is obtained by using cluster computing, and additional speed-up is achieved using specialized device models and a customized simulator engine.

6.1 Motivation

Predictive circuit-level simulation of CDM ESD [104] is by necessity a transient simulation because the circuit netlist contains devices that cannot be accurately described by static models. In general, the netlist is composed of capacitors, inductors and ESD devices; the latter may display non-quasi-static behavior on the CDM time scale. In order to identify the worst case pin-zap (i.e., pin number, discharge polarity), as well as the magnitude and location of the stress generated on chip, one must simulate both positive and negative discharges to every pin [105]. If the simulations are performed using a SPICE-like simulator, the run-time may be prohibitive—on the order of days—especially if the simulations are run sequentially. This work proposes a specialized simulator which aims to speed up CDM simulation without sacrificing accuracy by exploiting the unique properties of the circuit to be analyzed.

First, since the pin-zaps are mutually independent, cluster computing can be used to execute multiple pin-zap simulations in parallel. Next, the run-time for each pin-zap simulation must be reduced. Toward that goal, this work introduces ICE-TEA, the Illinois Circuit simulator for Efficient Transient ESD Analysis.

General purpose circuit simulators that are based on SPICE perform transient analysis use the Newton-Raphson method and modified nodal analysis (MNA) [106],[107],[108]. At each simulation time-point, the netlist with nonlinear devices is iteratively formulated as a system of linearized KCL and

KVL equations (i.e., the MNA system [106]) through the Newton-Raphson method. The solution to the MNA system is found using a numerical linear system solver. The solver must use a (slow, inefficient) direct method because the MNA matrix for a generic circuit netlist may not be symmetric and positive definite (SPD).

This work will show that the netlist representing an IC undergoing CDM ESD can be formulated using only resistors, inductors, capacitors, current sources, and piecewise linear devices. Specifically, the non-linear ESD protection devices are modeled using the piecewise-linear with transient relaxation (PWL-TR) modeling technique [109]. Given that the netlist contains only a restricted set of circuit elements, the circuit can be formulated as a set of KCL equations only and solved by nodal analysis (NA). Furthermore, the corresponding linear system (i.e., the NA system) is guaranteed to be SPD and solvable by a computationally efficient iterative method, the preconditioned conjugate gradient method. A similar approach has been successfully used to speed up circuit simulation to analyze I-R drops in on-chip power distribution networks [110],[111].

This chapter is organized as follows. In Section 6.2, full-component IC modeling for CDM ESD analysis is briefly reviewed; the simulation run-time is estimated as a function of the package pin count. Section 6.3 details how to construct circuit element models to maintain a SPD NA system; the algorithmic architecture of ICE-TEA is also outlined. The preconditioned conjugate gradient method (PCG) [112],[113], which is ICE-TEA's iterative linear system solver, is described in Section 6.4. Section 6.5 presents the techniques to implement an efficient simulator engine. An example full-component CDM simulation is presented in Section 6.6.

6.2 Computational Complexity of CDM ESD Simulation

Given an IC whose package has Q pins, the total simulation run-time for CDM analysis can be expressed by (6.1).

$$\text{Total Run-Time} = \left(\frac{2Q}{\Sigma}\right) \times \left(\frac{T}{\Delta t}\right) \times \mathcal{T}(N) \tag{6.1}$$

The variables in (6.1) are defined below.

- Σ: The number of pin-zap simulations executed simultaneously using cluster computing.
- T: Time duration of simulated ESD event, typically a few nanoseconds.
- Δt: Simulation time-step, typically a few picoseconds. T/Δt is the total number of simulation time-points.
- $\mathcal{T}(N)$: Average run-time of transient simulation per time-point. $\mathcal{T}$ is an increasing function of N, the number of nodes in the netlist.

The total number of transient simulations is equal to 2Q, since each pin must be tested at both polarities of the pre-charge voltage, according to the FICDM test standard [104].

The simulation netlist must represent the FICDM tester, the IC package, the pad-ring, and power distribution network on the die. Ref. [114] describes how to construct a full-component CDM simulation model, such as that illustrated in Figure 6.1 and Figure 6.2. N, the number of nodes in the netlist, is correlated with the package pin count Q, as demonstrated in (6.2).

$$N(Q) = c_{FICDM} + c_{PIN}Q + (2 + c_Z)\left(c_{CORE}\frac{Q}{4}\right)^2 \tag{6.2}$$

The parameters in (6.2) are defined below.

- c_{FICDM}: Number of nodes in the FICDM tester model. Typically, $c_{FICDM} = 9$.
- c_{PIN}: Number of nodes per pin required to model each set of package interconnect, pad-ring buses, and ESD protection devices. Typically, $c_{PIN} = 10$.
- c_{CORE}: Ratio of core V_{DD}/V_{SS} mesh pitch to the pad pitch. Typically, $c_{CORE} = 1$.
- c_Z: Number of layers in the die substrate model. $c_Z = 6$ is optimal for a substrate resistivity of 2 Ω-cm [114].

As illustrated in Figure 6.2, $(2 + c_Z)$ is the number of nodes in the z-direction at any point (x, y); arranged along the z-axis are one V_{DD} mesh, one V_{SS} mesh, and c_Z layers of p-bulk meshes. Each mesh extends along the x-y plane and consists of a 2D network of resistors; there are $\left(c_{CORE}\frac{Q}{4}\right)^2$ nodes per 2D mesh. The parameters c_{FICDM}, c_{PIN}, c_{CORE}, and c_Z in (6.2) are relatively constant. Therefore, N increases asymptotically with the square of Q. As an example, the simulation netlist for a particular component in a 96-pin TQFP package contained 5086 nodes.

The run-time for a direct linear system solver grows by $O(N^x)$, where x is in the range of 2 to 3. In contrast, for an iterative solver, x typically ranges from 1 to 2. Therefore, the run-time per time-point grows as $O(Q^{\mu})$, where

- $4 \leq \mu \leq 6$ if using a direct solver
- $2 \leq \mu \leq 4$ if using an iterative solver

Cluster computing can boost the simulation speed by a factor of Σ, but Σ is constant, regardless of the problem size (i.e. with Q or N), and dependent on available hardware resources, which are often shared among multiple users. The only speed-up technique that scales with the problem size is to reduce the complexity order (x or μ) by formulating the circuit netlist such that a faster simulator engine can be used.

6.3 Problem Formulation

The architecture of ICE-TEA is shown in Figure 6.3. Table 6.1 provides a summary comparison between ICE-TEA and a commercial SPICE-like simulator.

6.3.1 Transient Nodal Analysis

At each simulation time-point t_n, each circuit element (device) is formulated as a Norton equivalent circuit, so the entire circuit can be modeled as a linear system of KCL equations for nodal analysis (NA),

$$G\mathrm{v} = I \tag{6.3}$$

where

- $\mathrm{v} = [\mathrm{v}_i]_{i=1\ldots N}$ is the length-N vector of nodal voltages. It is the solution that will be found by NA.
- $G = \left[g_{ij}\right]_{i,j=1\ldots N}$ is the N-by-N conductance matrix of the NA system.
- $I = [i_i]_{i=1\ldots N}$ is the length-N vector of independent current sources.

6.3.2 Linear Elements

1. Resistor and Independent Current Source

Figure 6.4 shows the "stamps" [106] of a resistor and a current source. When parsing the netlist, ICE-TEA consecutively constructs G and I of the NA system. Whenever ICE-TEA encounters a resistor in the netlist, the stamp of the resistor is added into the conductance matrix G. Or, if an independent current source is encountered, its stamp is added into the vector I.

2. Capacitor and Inductor

The Norton equivalent models for capacitors and inductors are formulated using numerical integration. The trapezoidal integration method is chosen because it offers the best accuracy and simulation efficiency. The I-V characteristic of a capacitor is given by

$$\mathrm{v}_C(t_n) = \mathrm{v}_C(t_n - \Delta t) + \frac{1}{C}\int_{t_n-\Delta t}^{t_n} i_C(\tau)d\tau \approx \mathrm{v}_C(t_n - \Delta t) + \frac{\Delta t}{2C}[i_C(t_n) + i_C(t_n - \Delta t)] \tag{6.4}$$

Thus, $i_C(t_n) = g_{C,EQ} \cdot \mathrm{v}_C(t_n) - i_{C,EQ}$ (6.5)

where $g_{C,EQ} = \frac{2C}{\Delta t}$ and $i_{C,EQ} = \frac{2C}{\Delta t}\mathrm{v}_C(t_n - \Delta t) + i_C(t_n - \Delta t)$.

The I-V characteristic of an inductor is given by

$$i_L(t_n) = i_L(t_n - \Delta t) + \frac{1}{L}\int_{t_n-\Delta t}^{t_n} \mathrm{v}_L(\tau)d\tau \approx i_L(t_n - \Delta t) + \frac{\Delta t}{2L}[\mathrm{v}_L(t_n) + \mathrm{v}_L(t_n - \Delta t)] \tag{6.6}$$

Thus, $i_L(t_n) = g_{L,EQ} \cdot \mathrm{v}_L(t_n) + i_{L,EQ}$ (6.7)

where $g_{L,EQ} = \frac{\Delta t}{2L}$ and $i_{L,EQ} = \frac{\Delta t}{2L}\mathrm{v}_L(t_n - \Delta t) + i_L(t_n - \Delta t)$.

Once the equivalent conductance, $g_{C,EQ}$ or $g_{L,EQ}$, and current source, $i_{C,EQ}$ or $i_{L,EQ}$, are obtained, they can be stamped into G and I as a resistor and a current source.

3. Voltage Source and Dependent Sources

ICE-TEA can only solve for nodal voltages of a circuit that can be formulated as a SPD system, so the preconditioned conjugate gradient method can be used [112],[113]. Therefore, the following linear circuit elements cannot be supported by ICE-TEA, although they would be available in a general purpose circuit simulator

- Voltage source (VS)
- Voltage-controlled current source (VCCS)
- Voltage-controlled voltage source (VCVS)
- Current-controlled current source (CCCS)
- Current-controlled voltage source (CCVS)

A voltage source (V_S) is disallowed because its stamp introduces a zero entry on the diagonal of G, so the resultant linear system may not be positive definite. Likewise, VCVS and CCVS must be excluded because they both contain output voltage sources. Since CCCS requires a voltage source to measure the input current, it is also disallowed. The stamp of a VCCS is asymmetric and makes the overall system non-SPD.

The exclusion of voltage sources and all dependent sources (XCXS) does not diminish the utility of ICE-TEA because full-component CDM simulation does not require any of those elements. In a realistic circuit model, there are no ideal voltage sources; rather, all real sources must have finite source impedance. Using Thévenin-to-Norton transformation, any independent voltage source in series with

source impedance can be described as an independent current source in parallel with the source impedance, and thus made compatible with ICE-TEA. While dependent sources may be useful for describing devices with complex behavior, such as for an ESD protection device, accurate ESD device models can be constructed without dependent sources by using the PWL-TR modeling technique. This is addressed in the next section.

6.3.3 PWL-TR Models

The piecewise-linear with transient relaxation (PWL-TR) modeling technique [109] can reproduce the quasi-static I-V characteristic as well as transient voltage overshoot of any ESD protection device. At each simulation time-point t_n, the device is described by the parallel combination of a linear resistor $R_{EQ}(t_n)$ and a current source $i_{EQ}(t_n)$, which are then stamped into the NA system as shown in Figure 6.5. The PWL-TR method can even be used to represent an entire nonlinear protection circuit, such as a rail-clamp circuit, thereby simplifying the simulation netlist both in terms of node count and model complexity.

6.3.4 State-based Newton-like Iteration

Conventional circuit simulators use the Newton-Raphson (N-R) method to iteratively solve the system of nonlinear equations until the solution converges [106],[107],[108]. If nonlinear device models are employed, the circuit netlist must be re-linearized for each N-R iteration, so several N-R iterations ($\mathrm{itr}_{\mathrm{N-R}} > 1$) are required for each time-point. If PWL-TR device models are used instead, a less computationally intensive "state-based Newton-like" method suffices, as described in Figure 6.5. Since all devices are either linear or piecewise-linear, no linearization is required. In addition, instead of checking the solution against a set of convergence criteria as in the conventional N-R method, the simulator only needs to check if any device has changed state. Only when one or more devices change state are the corresponding PWL-TR models updated, and the new NA system is re-formulated and re-solved. Simulation shows that, during the course of a CDM event, the relevant protection devices turn on or off only within a few intervals of time-points. For the rest of the simulation time-points, no device

changes state, so $\mathrm{itr}_{\mathrm{N-R}} = 1$ for most of the time-points. For example, in one instance of full-component CDM simulation, only one iteration was needed for 83% of 1626 time-steps. By reducing the total number of iterations, the run-time is greatly reduced.

6.4 Preconditioned Conjugate Gradient Method

The preconditioned conjugate gradient (PCG) method [113] is used to find the solution v of (6.3). Its pseudo-code is given in Figure 6.6. The solution converges quickly if the initial guess is close to the solution to be found, making PCG very well suited to transient analysis, because the solution from previous time-point can be used as an estimate for the current time-point. This is especially true when a small time-step is used, as is normally the case for CDM simulation.

6.4.1 Choice of Preconditioner

Preconditioning is used to improve the condition number of the admittance matrix in order to speed up convergence. ICE-TEA uses the Jacobi preconditioner:

$$P = \left[p_{ij}\right]_{i,j=1\ldots N} = \begin{cases} p_{ii} = \dfrac{1}{g_{ii}} \\ p_{ij} = 0 \quad \forall\, i \neq j \end{cases} \tag{6.8}$$

The term p_{ii} can always be computed because all the diagonal terms (g_{ii}) in the conductance matrix G are guaranteed to be non-zero. Since each node [i] will be connected to at least one device, it contributes a positive stamp (Figure 6.4) to g_{ii}. One might consider the case in which node [i] is only connected to a current source; however, any current source must be accompanied by source impedance which then contributes to g_{ii}. Therefore, g_{ii} must always be a non-zero positive quantity. This guarantees the Jacobi preconditioner can always be computed.

Since the conductance matrix G may be different at each time-point, the preconditioner must be recalculated accordingly. As there may be thousands of time-points in the transient simulation, the accumulated run-time spent in re-computing the preconditioner can be substantial. Therefore, the Jacobi preconditioner is chosen for its low computational cost.

6.4.2 Convergence Criteria

ICE-TEA uses the same convergence criteria in PCG as commercial simulators do to achieve the same accuracy. The two convergence criteria are listed below.

- Update Convergence Criterion

For the ith entry in v

$$\left|\mathrm{v}_i^{(k)} - \mathrm{v}_i^{(k-1)}\right| < \mathrm{reltol} \cdot \mathrm{v}_{i,max} + \mathrm{v}_{\mathrm{i,tol}} \tag{6.9}$$

where reltol is the relative tolerance and $\mathrm{v}_{\mathrm{i,tol}}$ is the absolute tolerance on the solution v. The update convergence criterion ensures the iterative solver eventually arrives at a stable, convergent solution v.

- Residue Convergence Criterion

For the ith entry in the residue $r = (I - G\mathrm{v})$

$$\left|r_i^{(k)}\right| < \mathrm{r}_{\mathrm{i,tol}} \tag{6.10}$$

where $\mathrm{r}_{\mathrm{i,tol}}$ is the absolute tolerance on the residue r. The residue convergence criterion ensures that all KCL equations in the nodal analysis are satisfied.

6.5 Speed-Up Techniques

The efficiency of ICE-TEA is contingent on the careful implementation of the PCG solver engine. This section provides guidelines on the implementation of the simulator engine. For the rest of the chapter, unless indicated otherwise, simulations are performed on a desktop computer with a quad-core 2.8GHz processor and 8GB memory.

6.5.1 Parallel Programming

PCG is well suited to be implemented using parallel programming because it is composed of a collection of vector-vector operations and matrix-vector multiplications. In this work, C++ along with

OpenMP API is used for parallel programming because it is supported on multiple hardware and OS platforms.

1. Linear Vector-Vector Operations

Any linear vector-vector operation can be described and parallelized by (6.11).

For each $z_{i=1\ldots N}$ in $z = x + \alpha \cdot y$,

$$\mathrm{z}_i = \mathrm{x}_i + \alpha \cdot y_i \tag{6.11}$$

The known variables are vectors $x = [x_i]_{i=1\ldots N}$ and $y = [y_i]_{i=1\ldots N}$, and a scalar quantity α. Ideally, all unknown entries z_i's can be computed simultaneously since their computations are independent from one another.

2. Sparse Matrix-Vector Multiplication

A sparse matrix-vector multiplication can be computed by (6.12).

For each $z_{i=1\ldots N}$ in $z = Gx$,

$$\mathrm{z}_i = \sum_{\substack{k=1 \\ g_{ik}\neq 0}}^{N} g_{ik} x_k \tag{6.12}$$

The known variables are the sparse matrix $G = [g_{ik}]_{i,k=1\ldots N}$ and the vector $x = [x_i]_{i=1\ldots N}$. The unknown entries $z_{i=1\ldots N}$ can be computed simultaneously because each $\mathrm{z_i}$ depends only on the entries in the $\mathrm{i^{th}}$ row of G and the input x, not on the other entries.

The matrix G is represented by a special data structure (i.e., the sparse matrix representation) that stores only the non-zero entries of G. As argued in [106],[108], a circuit netlist usually corresponds to a sparse matrix. For each row in G, say the $\mathrm{i^{th}}$ row, the number of non-zero terms is limited and constant, contributed by the devices connected to node [i]. For example, each node in the 3D p-bulk mesh (Figure 6.2) is only connected to 6 other nodes. While N of a full-component CDM netlist can be large, the

number of floating point operations in (6.12) is not O(N) but is much smaller, determined by the average number of non-zero entries in each row. By using the sparse matrix data structure, unnecessary computations can be avoided. It is important to leverage the sparsity of the netlist to speed up the matrix-vector multiplication because it dominates the run-time of PCG.

3. Preconditioner-Vector Multiplication

The matrix-vector multiplications in steps 4 and 12 in the PCG algorithm (Figure 6.6) actually behave like a vector-vector operation because the Jacobi preconditioner is purely diagonal. The preconditioner-vector multiplication can be computed by (6.13).

For each $z_{i=1\ldots N}$ in $z = Px$,

$$z_i = p_{ii} x_i \tag{6.13}$$

The known variables are the Jacobi preconditioner *P* (6.8) and the vector $x = [x_i]_{i=1\ldots N}$. The computation of each entry z_i is very efficient; only one floating-point multiplication is required, because all off-diagonal entries in *P* must be zero. In addition to the low computational cost of recalculating the Jacobi preconditioner for each time-point, the cost for carrying out the preconditioner-vector multiplication is also low, which further justifies the choice of using the Jacobi preconditioner.

6.5.2 Speed-up from Customized Linear System Solver

To demonstrate how ICE-TEA benefits from the customized simulation engine, the test-case in Figure 6.7 is devised. It is derived from the netlist of a full-component CDM model. The test-case only represents the FICDM tester, p-bulk, tester-die coupling, and the V_{SS} networks on the die and in the package. The on-die V_{SS} nets include the pad-ring V_{SS} bus and the core V_{SS} mesh. The V_{SS} nets in the package include a series of package interconnects connected to the pad-ring V_{SS} bus at intervals. There are no V_{DD} networks or I/O pad-cells. As a result, the netlist does not include any ESD protection devices, and the simulation netlist is composed only of built-in linear circuit elements. Therefore, the run-time is

only affected by the simulator's linear system solver, without the effect of any nonlinear ESD device models.

Figure 6.8 compares the run-times of ICE-TEA and Spectre® (i.e., a general purpose circuit simulator). For the test-case illustrated in Figure 6.7, the relevant curves are labeled 1a (Spectre®) and 2a (ICE-TEA). It is evident that ICE-TEA is more efficient because it takes advantage of the numerical properties of the specialized problem formulation to enable the use of PCG. Regression on the data shows that the run-time T_1 of Spectre® exhibits a cubic growth rate ($T_1 = c_1 N^{3.02}$) which is characteristic for a direct linear system solver used in a general purpose simulator. In contrast, the run-time T_{2a} of ICE-TEA is almost linear ($T_{2a} = c_{2a} N^{0.99}$) for an efficient iterative method. This result indicates that the speed-up from ICE-TEA can become greater for larger problems because the complexity order of PCG (an iterative solver) is lower than that of the direct solver in Spectre®. Since the complete netlist for a full-component CDM simulation may contain several thousand nodes, it may be imperative to use a specialized simulator such as ICE-TEA.

The efficiency of the PCG algorithm is further explored with the aid of a modified test-case, in which the package-level V_{SS} interconnects are removed from the original test-case (Figure 6.7); the relevant curves in Figure 6.8 are labeled 1b (Spectre®) and 2b (ICE-TEA). The run-times of ICE-TEA on the two test-cases are fit to $T_{2a,b} = c_{2a,b} N^{x_{2a,b}}$:

- Original (2a): $x_{2a} = 0.99$ and $c_{2a} = 1.74 \times 10^{-2}$
- Modified (2b): $x_{2b} = 1.04$ and $c_{2b} = 4.3 \times 10^{-3}$

Both results exhibit linear growth rate ($x_{2a} \approx x_{2b} \approx 1$) because this is determined by the (same) algorithmic complexity per PCG iteration. However, the constant coefficients are different (i.e., $c_{2a} > c_{2b}$); the coefficient is proportional to the average number of PCG iterations needed to find the solution. Recall that the PCG converges faster if the initial guess is close. In the original test case, the package interconnects are modeled as a series of RLC networks. The solution (nodal voltages) would exhibit

significant ringing, so it may not serve well as a good initial guess for the next time-point. In the modified test-case, without the RLC networks of the package interconnects, the system is more damped. The nodal voltages are more stable and closer to those of the next time-points, so PCG converges quickly.

The same comparison is also made with Spectre®; see curves 1a and 1b in Figure 6.8. The two sets of Spectre® run-times are virtually identical because the efficiency of a direct solver is not aided by the solution from the previous time-step. This example shows that the run-time of ICE-TEA not only is a function of the size of the netlist, but also depends on the nature of the netlist. If the netlist is composed of many resonant structures (i.e., LC tanks), there may be ringing on internal nodes that slows down PCG. Nevertheless, ICE-TEA's run-time is always a linear function of the node count, and will always outperform Spectre®, because ICE-TEA has lower run-time growth rate, and the effort to find the solution of the current time-point is not wasted but is used as the initial guess to help find the solution for the next time-point.

6.5.3 Speed-up from PWL-TR Modeling

To evaluate the benefit of using PWL-TR models instead of nonlinear models, a pad-ring test-case (Figure 6.9) is devised. To emphasize the impact of ESD device models on run-time, the test-case consists only of one ring of pad-cells with just one I/O power/ground network (V_{DDIO}/V_{SSIO}). Therefore, the simulation netlist is heavily populated with ESD devices. Two sets of pad-ring simulations are performed. One uses PWL-TR models to represent the ESD diodes and rail-clamps; the other uses the nonlinear model [77] instead. The PWL-TR and nonlinear models are calibrated to exhibit virtually identical voltage overshoot and pulsed I-V curves. Both sets of simulations use PCG as the linear system solver. The one with PWL-TR models uses state-based Newton-like iterations to resolve the piecewise-linearity. The other with nonlinear compact models uses the "true" Newton-Raphson method to resolve the nonlinearity. At each N-R iteration, the nonlinear compact model is linearized and stamped as a two-terminal Norton equivalent circuit. The convergence criterion of (6.9) is used to determine if the solutions

of two consecutive N-R iterations converge. If so, the simulation moves on to the next time-point; otherwise, an additional N-R iteration is required.

Run-times are summarized in Figure 6.10. The run-time obtained with nonlinear compact models is longer because the stamp of the nonlinear model is more difficult to compute, and more N-R iterations are required. Note that, thanks to the iterative nature of PCG, the run-time complexity is linear, regardless of the type of models being used. Also, as the solution from the current N-R iteration serves as a good initial guess for the next, successive N-R iterations converge quickly. This explains why there is only a modest speed-up obtained from PWL-TR modeling, ranging from about 1.3 to 1.6 times. Greater speed-up can be expected if a direct method solver is used, as in the case of general circuit simulator.

The speed-up factor is not only sensitive to the size and complexity of the netlist, but also to the convergence criteria. While the speed-up factor varies for the reasons detailed above, simulation with PWL-TR models is always faster. In addition, simulation with nonlinear device models sometimes does not converge, requiring the subsequent use of looser convergence criteria and/or smaller time-step. This effect results in reduced accuracy and/or unnecessarily longer simulation time. With PWL-TR models, simulation converges over a wider range of time-steps and tolerances, without sacrificing accuracy [109].

6.5.4 Speed-up from Built-in Models

Previous studies [115],[116] have demonstrated that simulation with Verilog-A models is slower than simulation which uses models built in to the simulator source code. However, these comparisons focus on nonlinear compact models, and have not demonstrated how the slow-down is correlated with the problem size. Since this work focuses on linear and piecewise-linear (PWL) models, the following two model implementation options are considered.

1) Models are constructed using only built-in piecewise-linear elements provided by the simulator
2) PWL models are implemented in Verilog-A (external to the simulator)

Although ICE-TEA has a built-in PWL-TR model and such models can be coded in Verilog-A, it is not readily implemented using only the built-in piecewise-linear elements provided by Cadence Spectre®. Therefore, to enable the comparison of modeling options 1 and 2 described above, a simplified PWL model is devised, as shown in Figure 6.11. Two sets of pad-rings (Figure 6.9) are simulated using Spectre®: one with ESD devices implemented by option 1, and the other by option 2. Run-times are compared in Figure 6.12. It is evident that simulation with Verilog-A models is slower, even though these models reduce the total number of nodes in the simulation netlist by obscuring the internal nodes of the device. Thus, although it is convenient to prototype a device model in Verilog-A, the finalized ESD device model should be integrated into the simulator to improve simulation efficiency, as is done in ICE-TEA.

6.6 Full-chip Simulation Example

In this section, ICE-TEA is benchmarked against Spectre® for the case of full-component CDM simulation. The component to be simulated assumes a 6.8x6.8mm^2 die fabricated in a 65nm CMOS process, placed in a 14x14mm^2 TQFP package. Figure 6.13 illustrates the pad-cell sequence and power domain floorplan on the die. The rail-clamp between the V_{DDIO} and V_{SSIO} buses is distributively placed in all pad-cells. Each I/O pad is protected by dual diodes. Rail-clamps between the V_{DD} and V_{SS} buses are placed in all V_{DD}/V_{SS} cells. The methodology to construct the full-component CDM simulation netlist is detailed in [114].

6.6.1 Cluster Computing

All pins in the given IC design must be simulated to verify the design's ESD robustness. As each pin-zap is independent from the others, multiple pin-zaps can be simulated in parallel using cluster computing. The Illinois Campus Cluster [117] was used to evaluate the benefits of cluster computing for CDM simulation. A request to simulate 96 independent pin-zaps was issued using batch commands. Ideally, all 96 simulation jobs would be carried out simultaneously, and the total run-time would be that of a single pin-zap. However, according to maximum available hardware resources per-user account, only

8.19 jobs on average could be executed at the same time. Therefore in this case, the speed-up factor, Σ in (6.1), is equal to 8.19.

6.6.2 Simulator-dependent Run-times

Run-times for full-component CDM simulation are summarized in Table 6.2. Using ICE-TEA, the run-time for different pin-zap simulations varied less than 5% around the average. The two simulators with PWL-TR models produce virtually identical waveforms (Figure 6.14); both identify the same pin-zap as the worst case and the same location of power domain crossing circuits receiving the largest voltage stress. However, ICE-TEA provides more than 4x speed-up (#4 vs. #5 in Table 6.2). In the given test-case, only 404 devices are represented by PWL-TR models. Based on the results shown in Figure 6.12, the slower run-time of Spectre® can still be attributed to its linear system solver, rather than to the use of Verilog-A for PWL-TR model implementation. This confirms that full-chip CDM simulation benefits from ICE-TEA's customized simulation engine.

The 4x speed-up factor is relatively modest because the test-case assumes a TQFP package which includes highly inductive lead-frame traces and strong capacitive coupling to the FICDM tester. As explained in Section 6.5.2 (Figure 6.8), such a "high-Q" setup causes ringing on the nodes near the package interconnects, and degrades the efficiency of ICE-TEA. When simulating packages intended for high-speed applications, e.g. those with low L/C package interconnects, more speed-up can be expected.

It is also worthwhile to compare case #1 to cases #2 through #4 in Table 6.2. These CDM simulations were all performed using Spectre®: case #1 with nonlinear (compact) models, the others with PWL-TR models. With the same set of convergence criteria, the simulation with the nonlinear device models cannot converge. (The ESD diode compact model is implemented in Verilog-A; the rail-clamp circuits are described using built-in compact models). To obtain a convergent result, the convergence criteria must be loosened 1000 times, and the less efficient but more stable Gear2 integration method is required. Even so, the run-time is over 3 hours for one pin-zap simulation. This example highlights why

nonlinear models should be avoided in full-component CDM simulation: the combination of nonlinearity and large problem size results in inefficient, inaccurate, or non-convergent simulation.

Spectre® simulations produce results that are virtually independent of the selected method of integration, and Table 6.2 shows that the trapezoidal integration method yields the fastest run-time. This observation explains why ICE-TEA formulates the companion models for capacitors and inductors using the trapezoidal integration method.

6.7 Summary

Full-chip CDM simulation often is computationally expensive. Nonlinearity in the models of the on-chip ESD protection devices requires the simulator to spend more Newton-Raphson iterations to find the solution, and may even prevent the simulation from converging. Alternatively, accuracy, efficiency, and numerical stability of the CDM simulation can all be achieved by representing the on-chip ESD protection devices using PWL-TR models rather than nonlinear compact models.

A full-component CDM simulation netlist can be represented using only linear circuit elements (R, L, C, I) and piecewise-linear elements (PWL-TR). The selected circuit elements can all be represented as Norton equivalent circuits, and the corresponding stamps guarantee the system of nodal analysis to be symmetric positive definite at each simulation time-point. This enables the use of a fast, iterative linear system solver. Simulation results confirm that the run-time of the specialized simulator engine exhibits almost linear growth rate with respect to the netlist size, while that of a general purpose simulator is cubic. ICE-TEA's linear growth rate keeps the simulation run-time manageable even for large netlist, but the runtime of a general purpose simulator will explode due to its cubic growth rate.

While a general purpose circuit simulator is versatile by accommodating the full set of linear and nonlinear model and Verilog-A model prototyping, such flexibility is not needed for full-component CDM simulation. For comprehensive analysis of a CDM simulation netlist with thousands of nodes, it is optimal to use a specialized circuit simulator.

6.8 Tables and Figures

Table 6.1 Comparison of ICE-TEA and Spectre®, which is the general-purpose circuit simulator used for benchmarking. XCXS denotes all four types of dependent sources discussed in 6.3.2.3.

Simulator	ICE-TEA (Specialized)	Cadence Spectre® (General Purpose)
Circuit Elements (O = supported; × = not supported)		
R, L, C, I	O (Built-in)	O (Built-in)
V, XCXS	×	O (Built-in)
PWL-TR	O (Built-in)	O (Verilog-A)
Nonlinear Devices	×	O (Built-in or Verilog-A)
Circuit Analysis	Nodal Analysis (NA) (KCL only)	Modified Nodal Analysis (MNA) (KCL and KVL)
Simulator Engine		
Linear System Solver	Preconditioned Conjugate Gradient (Iterative Method)	Proprietary Sparse Matrix Solver (Direct Method)
Nonlinear System Solving Method	State-based Newton-like Method	Newton-Raphson Method

Table 6.2 Run-time of the worst-case pin-zap (Figure 6.14). The convergence criteria are held constant: reltol $= 10^{-6}$, $v_{i,tol}$ = vabstol $= 5 \times 10^{-6}$, and $r_{i,tol}$ = iabstol $= 10^{-7}$. N = 5086. With this set of convergence criteria, simulation with nonlinear models cannot converge (*).

No.	Simulator	Integration Method	Device Model	Run-Time (sec)
#1	Spectre®	Any	Nonlinear	∞*
#2	Spectre®	Backward Euler	PWL-TR	1977
#3	Spectre®	Gear2	PWL-TR	520
#4	Spectre®	Trapezoidal	PWL-TR	441
#5	ICE-TEA	Trapezoidal	PWL-TR	107

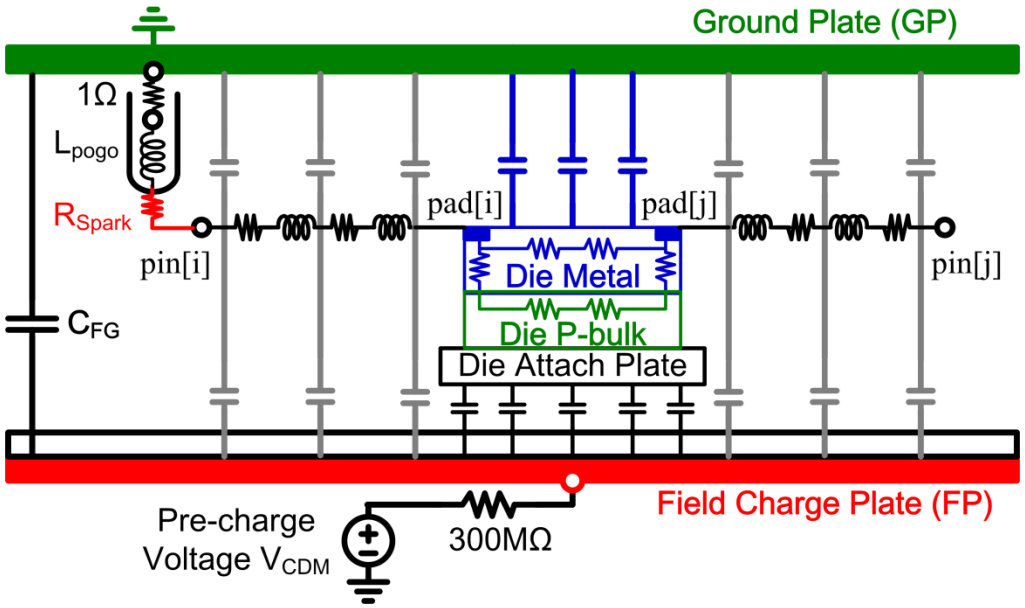


Figure 6.1 Model of a packaged IC mounted on a FICDM tester.

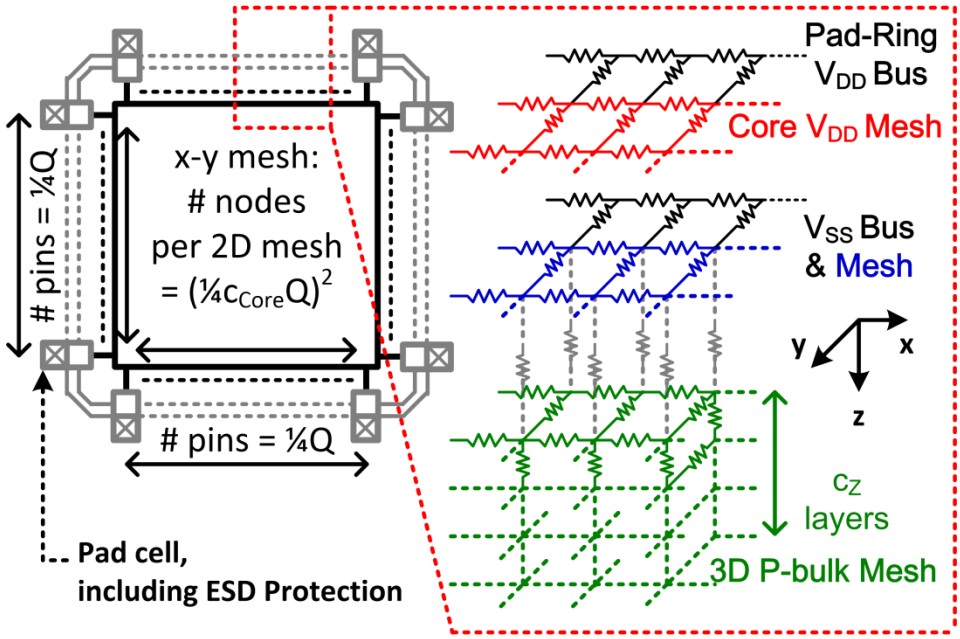


Figure 6.2 Die-level portion of the full-component CDM model in Figure 6.1. It includes the on-die power distribution network (V_{DD}/V_{SS} buses in the pad-ring and V_{DD}/V_{SS} meshes in the core) and the chip substrate (p-bulk). The on-chip protection devices (not pictured) are placed in the pad-ring cells.

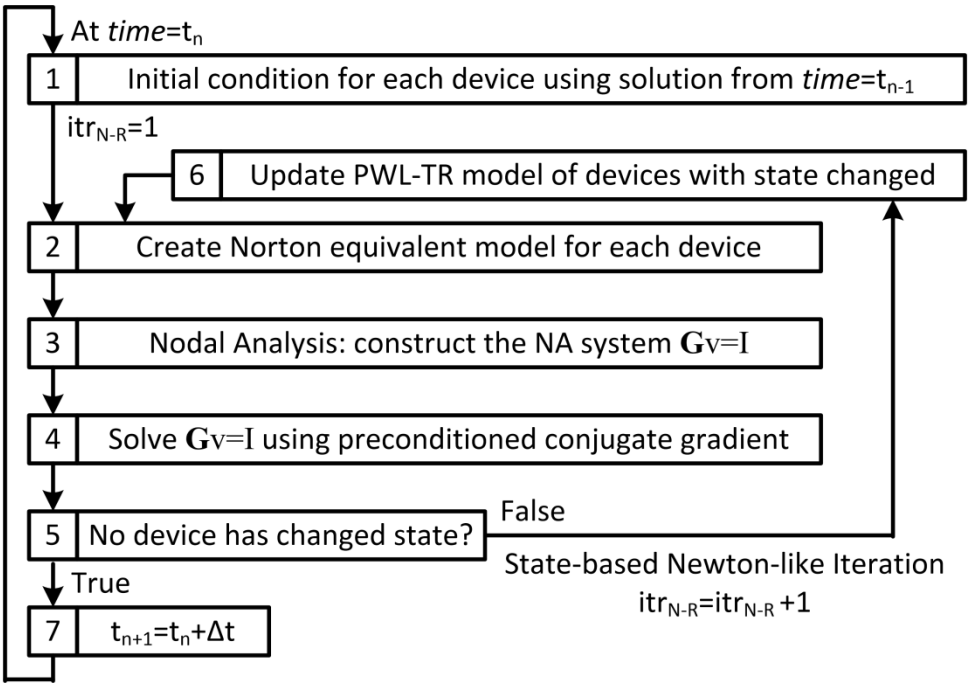


Figure 6.3 ICE-TEA flowchart. ICE-TEA performs transient nodal analysis (NA) by formulating the circuit netlist as a system of KCL equations at each time-point t_n.

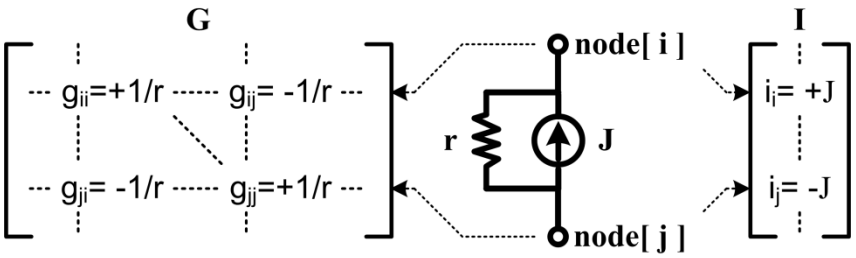


Figure 6.4 Stamps for resistor and current source in nodal analysis.

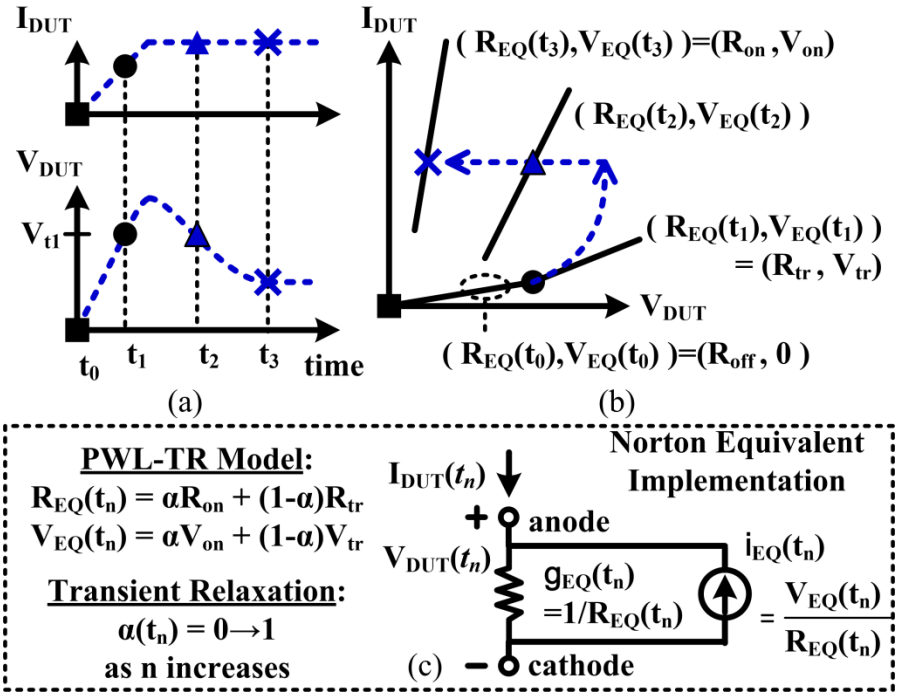


Figure 6.5 (a) A snapback device's terminal voltage $V_{DUT}(t)$ in response to a current pulse $I_{DUT}(t)$. (b) Each branch represents the Thévenin equivalent circuit (R_{EQ}, V_{EQ}) of the device at time=t_n. (c) The PWL-TR model [109] can capture the evolution of (R_{EQ}, V_{EQ}). In ICE-TEA, the PWL-TR model is implemented as a Norton equivalent circuit (g_{EQ}, i_{EQ}).

1.	Given an initial guess on the solution v	
2.	$k = 0$	# k: PCG iteration counter
3.	$r = I - G$v	# r: residue
4.	$d = Pr$	# d: conjugate search direction
5.	$\delta_{new} = r^T d$	
6.	While the convergence criteria (6.9) and (6.10) not met	
7.	$q = Gd$	
8.	$\alpha = \frac{\delta_{new}}{d^T q}$	# α: search step-size
9.	v = v + αd	# new estimate of the solution v
10.	If k is divisible by $\sqrt{N}$:	$r = I - G$v
11.	else:	$r = r - \alpha q$
12.	$s = Pr$	
13.	$\delta_{old} = \delta_{new}$	
14.	$\delta_{new} = r^T s$	
15.	$\beta = \frac{\delta_{new}}{\delta_{old}}$	# β: Gram-Schmidt coefficient
16.	$d = s + \beta d$	# update search direction d
17.	$k = k + 1$ and go to step 5	

Figure 6.6 Preconditioned conjugate gradient method (PCG). Steps 10 and 11 are used alternatively to minimize round-off error accumulation without overusing the required but expensive matrix-vector multiplication. Details can be found in [113].

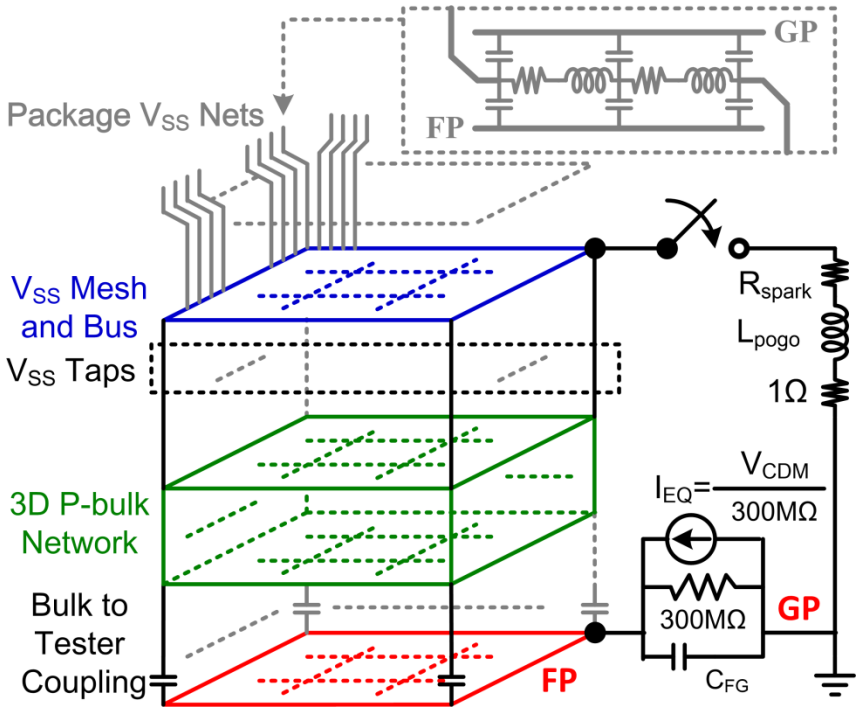


Figure 6.7 Simplified netlist to facilitate the comparison of the solver engines of ICE-TEA and Spectre®. The netlist contains only built-in linear elements to represent the on-die V_{SS} network, substrate, package interconnects, and the FICDM tester. Note that the pre-charge supply (V_{CDM} + 300MΩ) is replaced by its Norton equivalent circuit.

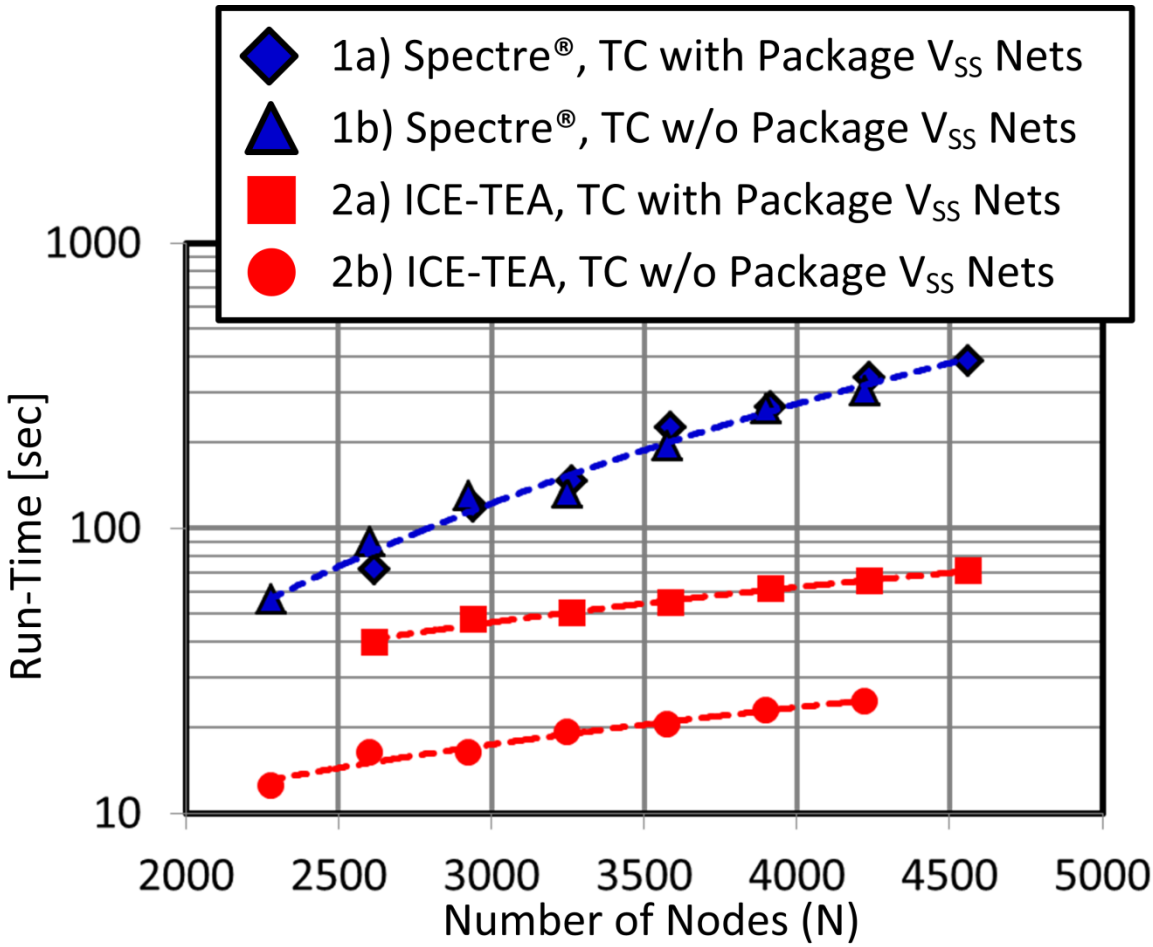


Figure 6.8 Run-time of Spectre® and ICE-TEA. Both simulators perform transient analysis on two kinds of test-cases: a) the complete netlist in Figure 6.7, and b) the same netlist in Figure 6.7 but with the package V_{SS} nets removed. Number of nodes is increased by making the z-directional mesh representing the 3D p-bulk model finer [114]. Both simulators use the same numerical setup: reltol= 10^{-6}, $v_{i,tol}$=vabstol=5×10^{-6}, and $r_{i,tol}$=iabstol=10^{-7}, with trapezoidal integration method.

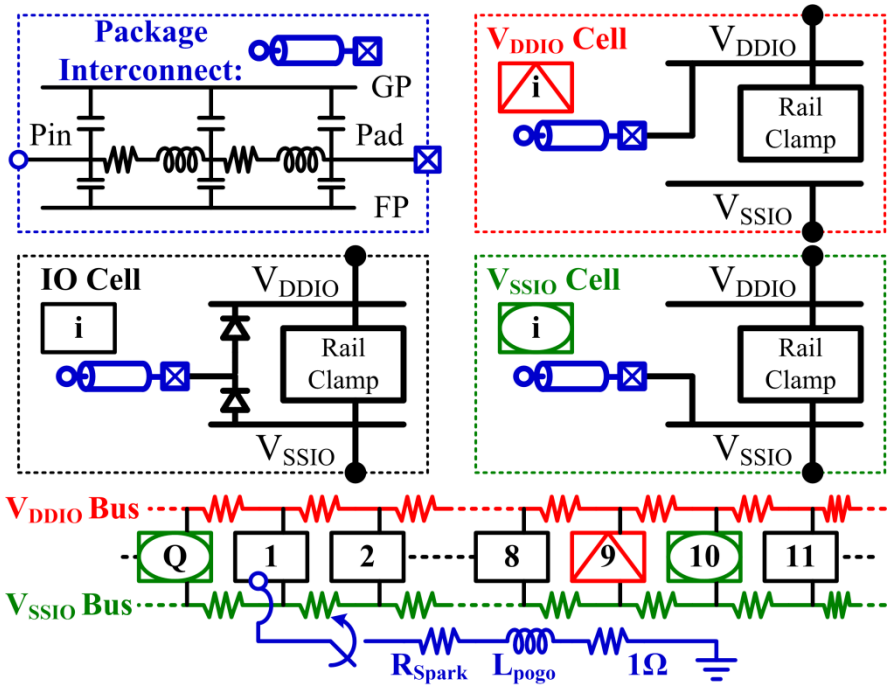


Figure 6.9 Example of pad-ring model. In a Q-pin setup, the V_{DDIO} and V_{SSIO} ports of the ith pad-cell are connected to its neighbors by the resistors comprising the V_{DDIO} and V_{SSIO} buses. One V_{DDIO} cell and one V_{SSIO} cell are placed for every 8 IO cells. The Qth pad-cell is connected back to the 1st pad-cell to complete the pad-ring.

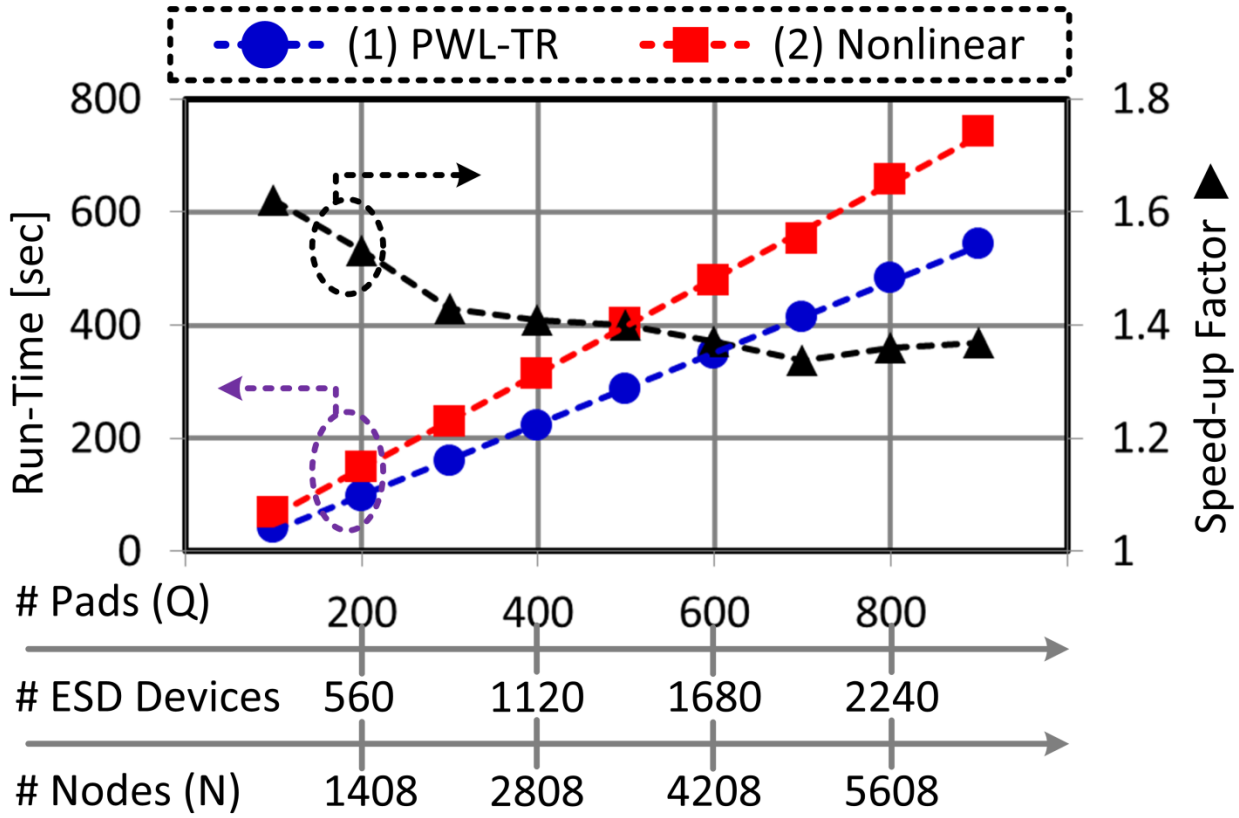


Figure 6.10 Pad-ring of Figure 6.9 is simulated using (1) PWL-TR models and (2) nonlinear compact models [77]. Linear regression on run-time with PWL-TR model yields $T_1 = m_1 Q + b_1$ with m_1=0.632; the run-time with nonlinear models is given by $T_2 = m_2 Q + b_2$ with m_2=0.839. Speed-up factor (=T_2/T_1) asymptotically approaches m_2/m_1= 1.328.

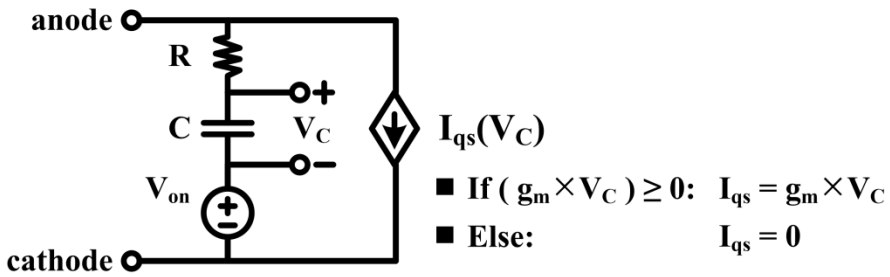


Figure 6.11 Simple model using only built-in elements. The piecewise-linear VCCS $I_{qs}(V_C)$ captures the quasi-static I-V behavior. Gradual turn-on is approximated by R, C and the offset voltage V_{on}. The delay for the device to evolve into its steady-state is set by the RC time-constant.

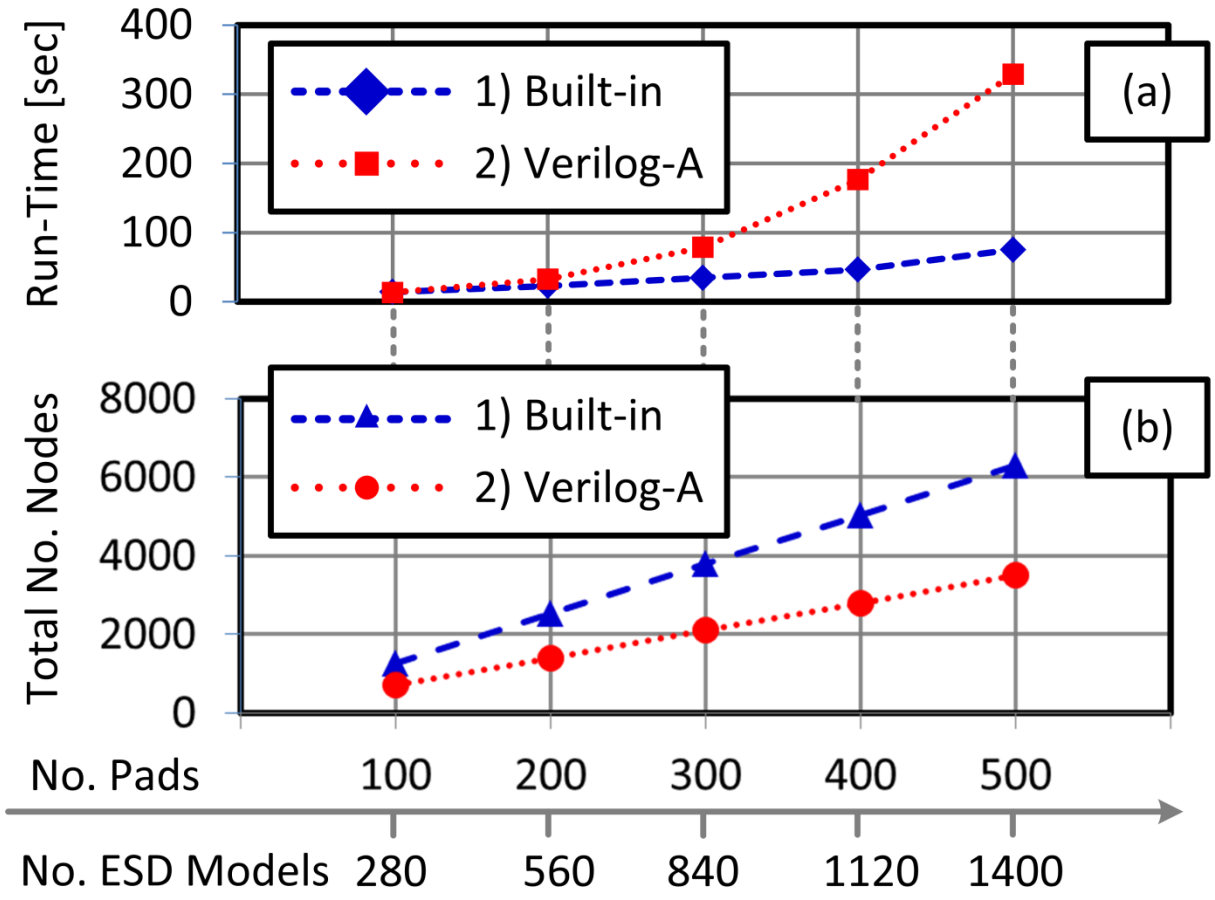


Figure 6.12 Spectre® run-time for the two implementations: 1) the ESD device models are implemented using only built-in linear elements, and 2) the ESD models are implemented in Verilog-A.

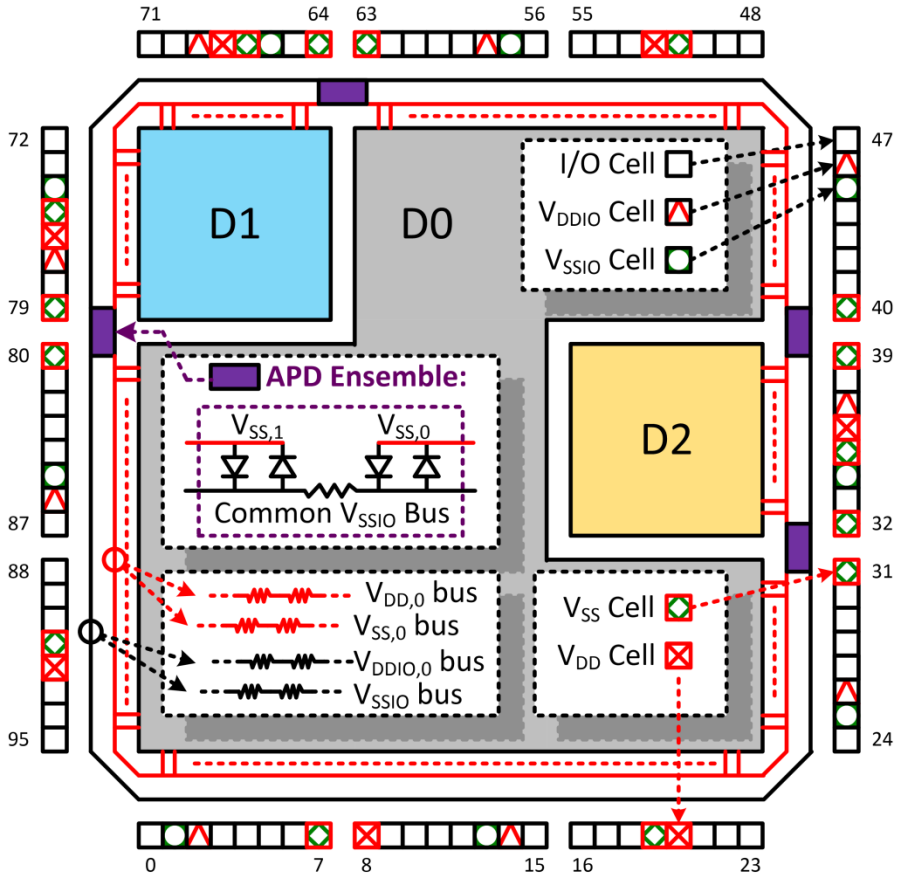


Figure 6.13 Pad-cell and power domain floorplan for the test-case. There are three “core” power domains, D0, D1 and D2, as well as an IO power domain.

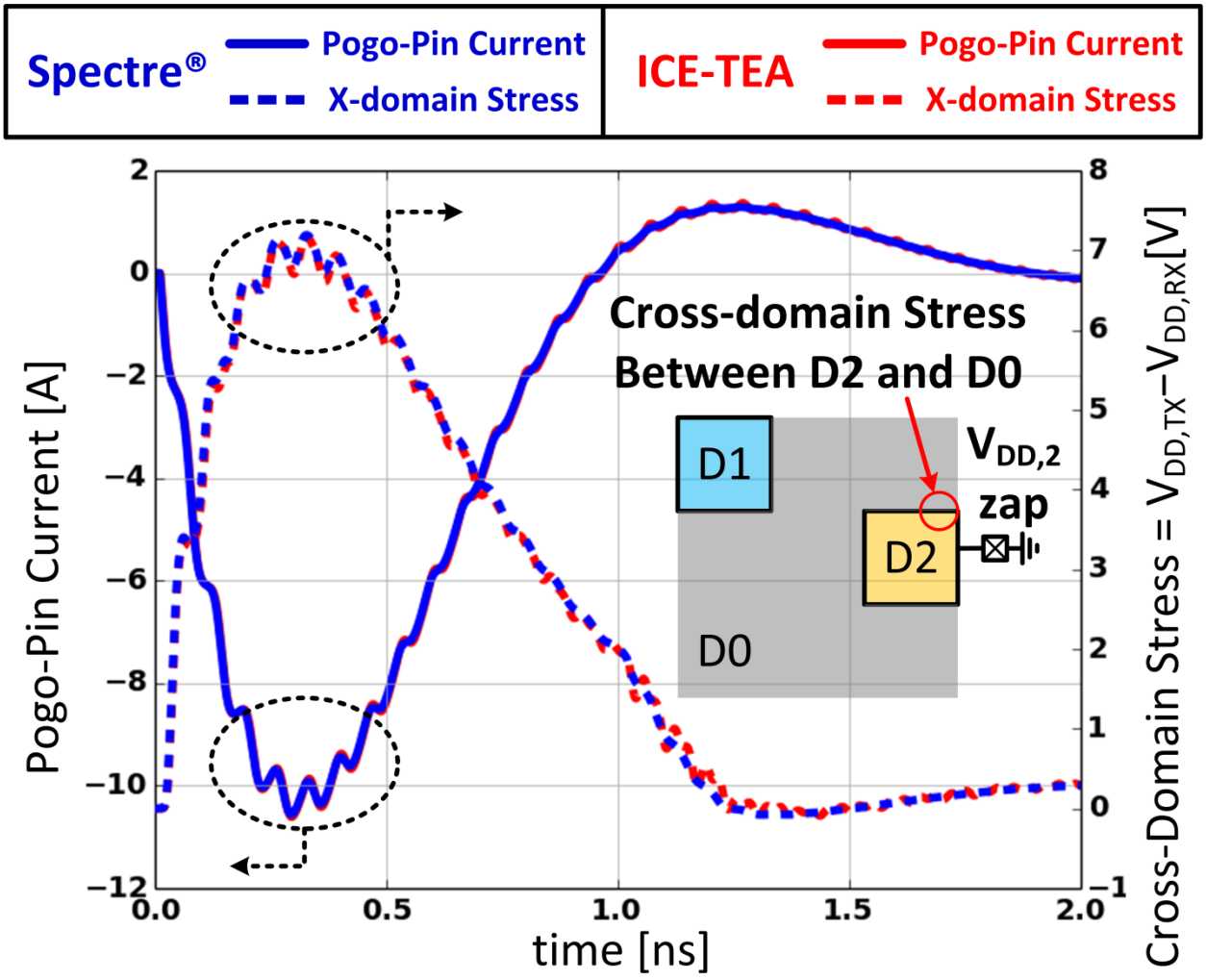


Figure 6.14 Simulated waveforms for FICDM testing of the design shown in Figure 13. Both simulators predict the same worst-case pin-zap (pin No. 36 — a V_{DD} pin in D2 — at V_{CDM} = –500V), the same pogo-pin current waveform, and the same cross-domain voltage stress. Analysis of the cross-domain stress follows the procedure in [114].

Chapter 7. Conclusions and Future Works

MOSFET compact model allows design and verification of MOSFET-based ESD protection circuits through circuit simulation. The semi-physical MOSFET ESD model can reproduce the quasi-static I-V behavior and transient response to arbitrary ESD-like stimuli, so accuracy of the simulation is ensured.

Full-component IC model includes tester environment, package, pad-ring (including models for ESD devices/circuits), core power distribution networks, substrate, and N-well/P-well parasitics. Through transient CDM ESD simulation, the worst-case pin-zap can be identified, along with the location, stressing pattern, and magnitude of the cross-domain stress on the domain-crossing circuit in the core.

Due to the sheer size of the full-component models and the nonlinear nature of the ESD protection elements, simulation run-time can become prohibitively long. Therefore, speed-up techniques must be implemented, including the piecewise-linear modeling technique with transient relaxation (PWL-TR), specialized problem formulation, and customized simulator engine. By representing ESD devices/circuits with PWL-TR models instead of nonlinear compact models, speed-up can be achieved without compromising accuracy. Also, parallelism in FICDM test standard can be exploited: all pin-zap simulations are mutually independent, and can be performed simultaneously using cluster computing. Finally, the full-component model can be described using only a restricted set of linear circuit elements and the PWL-TR models. Such specialization guarantees that the model netlist can be formulated as a system of nodal analysis equations, and solvable by a customized simulator engine. The customized simulator engine uses pre-conditioned conjugate gradient method as the linear system solver and state-based Newton-like iteration to resolve the responses of the built-in PWL-TR models. The efficiency of the customized simulator engine greatly outperforms a general-purpose circuit simulator. Speed-up factor increases with larger problem size because the complexity order of the customized simulator engine is much lower than that of a general-purpose simulator.

Further extensions of this work are suggested as follows. A control port can be augmented to the 2-terminal PWL-TR model, so it becomes a 4-terminal piecewise-linear **voltage-controlled current source** (VCCS) with transient relaxation (4T PWL-TR). This allows more complicated behavior to be modeled, such as the I-V characteristic of a transistor and modulation of trigger voltage (V_{t1}) by gate bias. The 4T PWL-TR model still behaves as a linear circuit element, so the corresponding circuit simulation remains numerically stable, as compared to that with nonlinear models. The stamp of the 4T PWL-TR model — equivalent to a VCCS which can be represented as a Norton circuit— is still compatible with nodal analysis (NA). The resultant NA system will remain positive definite, but it is not symmetric. As a result, improvement on ICE-TEA must be made to accommodate the 4T PWL-TR model. Fortunately, there are iterative methods to solve an **asymmetric positive definite** system, such as the **bi-conjugate gradient stabilized method (BiCGSTAB)** and **generalized minimal residual method (GMRES**). Being iterative methods, they still exhibit linear growth rate against problem size, so simulation run-time remains manageable for large netlists.

With the capability to simulate transistor-like behavior using the 4T PWL-TR model, more complicated ESD protection architecture can be simulated, such as the distributed rail-clamp. Big-NFETs in each pad-cell can be represented by the 4T PWL-TR model, interconnected by pad-ring power/ground buses, and triggered by an additional bus driven by the distributed trigger circuits. In addition, if the core transistors can also be represented by the 4T PWL-TR model, all the domain-crossing circuits can then be included to form a more complete full-component model. The effect of local ESD protection and the sizing of domain-crossing circuits can then be evaluated **during** full-component simulation, instead of being postponed to a later phase of localized simulation.

A **multi-level** comprehensive FICDM analysis can be devised. Static I-R drop analysis can first be performed for each pin-zap. The results can be used to ascertain a group of "high-risk" pin-zaps — those that are likely to produce damage-inducing cross-domain stress. A second phase of full-component transient simulation can then be performed **only** on the high-risk group. Without simulating thousands of

time-points, static I-R drop analysis can be orders of magnitude faster than transient simulation. Also, with reduced number of necessary full-component transient simulations, overall run-time and demand on computer cluster is lessened. However, one must be cautious to ensure the validity of approximating the inherently transient event of FICDM as a static analysis. First, the validity of the static model representing all the charge storage elements and the package RLC model must be verified. Second, the static ESD device model must take into account transient voltage overshoot as a function of rise-time.

Finally, full-component FICDM simulation is more than a diagnosis tool for aftermath failure analysis; it also serves as a design tool to detect and prevent failure before tape-out. Because ESD current floods the entire chip, the pad-ring ESD architecture and core power domain floorplan should be co-designed to guarantee overall ESD resilience. The ability to simulate full-component ESD current distribution should be exploited to formulate a new design methodology which simultaneously optimizes the pad-ring and core PDN. The full-chip optimization process would therefore necessitate an automated netlist extraction tool which directly obtains the full-component FICDM model from the layout or an equivalent description of the design.

References

[1] Electrostatic Discharge Association, Rome, NY, "ESD Fundamentals," June 2001. [Online]. Available: http://www.esda.org/esd_fundamentals.html.

[2] A. Amarasekera and C. Duvvury, *ESD in Silicon Integrated Circuits*, 2nd ed. New York, NY: John Wiley and Sons, 2002.

[3] *Electrostatic Discharge Sensitivity Testing, Human Body Model, Component Level*, Standard Test Method ANSI/ESD STM 5.1-2007, 2007.

[4] *Electrostatic Discharge Sensitivity Testing, Charged Device Model, Component Level*, Standard Test Method ANSI/ESD STM 5.3.1-1999, 1999.

[5] *Field-Induced Charged-Device Model Test Method for Electrostatic Discharge Withstand Thresholds of Microelectronic Components*, JESD22-C101C, Dec. 2014.

[6] V. Shukla, "Predictive transient circuit simulation of charged device model ESD events in system in package chips," Ph.D. dissertation, University of Illinois at Urbana-Champaign, USA, 2013.

[7] M. Etherton, "Charged device model (CDM) ESD in ICs," Ph.D. dissertation, Eidgenössische Technische Hochschule, Zurich, 2006.

[8] S. Sowariraj, "Full chip modelling of ICs under CDM stress," Ph.D. dissertation, University of Twente, Netherlands, 2005.

[9] K.-H. Meng and E. Rosenbaum, "The need for transient I-V measurement of device ESD response," in *Proc. EOS/ESD Symp.*, 2012, pp. 1–8.

[10] K.-H. Meng and E. Rosenbaum, "Verification of snapback model by transient I-V measurement for circuit simulation of ESD response," in *IEEE Trans. Dev. Mat. Rel.*, vol. 13, no. 2, pp. 371–378, Jun. 2013.

[11] K.-H. Meng and E. Rosenbaum, "Layout-aware, distributed, compact model for multi-finger MOSFETs operating under ESD conditions," in *Proc. EOS/ESD Symp.*, 2013, pp. 1–8.

[12] K.-H. Meng, Z. Chen, and E. Rosenbaum, "Compact distributed multi-finger MOSFET model for circuit-level ESD simulation," submitted to *Microelectronics Reliability*, 2014.

[13] J. Li, S. Joshi, R. Barnes, and E. Rosenbaum, "Compact modeling of on-chip ESD protection devices using Verilog-A," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 6, pp. 1047–1063, Jun. 2006.

[14] A. Amerasekera, M.-C. Chang, C. Duvvury, and S. Ramaswamy, "Modeling MOS snapback and parasitic bipolar action for circuit-level ESD and high-current simulations," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 1996, pp. 318–326.

[15] S. Beebe, "Simulation of complete CMOS I/O circuit response to CDM stress," in *Proc. EOS/ESD Symp.*, 1998, pp. 259–270.

[16] H. Wolf, H. Gieser, and W. Stadler, "Bipolar model extension for MOS transistors considering gate coupling effects in the HBM ESD domain," in *Proc. EOS/ESD Symp.*, 1998, pp. 271–280.

[17] M. Mergens, W. Wilkening, S. Mettler, H. Wolf, and W. Fichtner, "Modular approach of a high current MOS compact model for circuit-level ESD simulation including transient gate coupling behavior," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 1999, pp. 167–178.

[18] S. Ramaswamy, E. Li, E. Rosenbaum, and S.-M. Kang, "Circuit-level simulation of CDM-ESD and EOS in submicron MOS devices," in *Proc. EOS/ESD Symp.*, 1996, pp. 316–321.

[19] S. Ramaswamy, A. Amerasekera, and M.-C. Chang, "A unified substrate current model for weak and strong impact ionization in sub-0.25μm NMOS devices," in *Tech. Dig. IEDM*, 1997, pp. 885–888.

[20] X. Y. Zhang, K. Banerjee, A. Amerasekera, V. Gupta, Z. Yu, and R. W. Dutton, "Process and layout dependent substrate resistance modeling for deep sub-micron ESD protection devices," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2000, pp. 295–303.

[21] M.-J. Chen, H.-S. Lee, and S.-T. Chen, "Extraction of eleven model parameters for consistent reproduction of lateral bipolar snapback high-current I-V characteristics in NMOS devices," *IEEE Trans. on Elec. Dev.*, vol. 48, no. 6, pp. 1237–1244, Jun. 2001.

[22] F.-C. Hsu, P.-K. Ko, S. Tam, C. Hu, and R. S. Muller, "An analytical breakdown model for short-channel MOSFET's," *IEEE Trans. on Elec. Dev.*, vol. 29, no. 11, pp. 1735–1740, Nov. 1982.
[23] J. Li, H. Li, R. Barnes, and E. Rosenbaum, "Comprehensive study of drain breakdown in MOSFETs," *IEEE Trans. on Elec. Dev.*, vol. 52, no. 6, pp. 1180–1186, Jun. 2005.
[24] M. Stockinger and J. W. Miller, "Characterization and modeling of three CMOS diode structures in the CDM to HBM timeframe," in *Proc. EOS/ESD Symp.*, 2006, pp. 46–53.
[25] J.-R. Manouvrier, P. Fonteneau, C.-A. Legrand, P. Nouet, and F. Azaïs, "Characterization of the transient behavior of gated/STI diodes and their associated BJT in the CDM time domain," *Microelectronics Reliability*, vol. 49, no. 12, pp. 1424–1432, Dec. 2009.
[26] J. P. Di Sarro and E. Rosenbaum, "A scalable SCR compact model for ESD circuit simulation," *IEEE Trans. on Elec. Dev.*, vol. 57, no. 12, pp. 3275–3286, Oct. 2010.
[27] A. Romanescu, H. Beckirch-Ros, P. Fonteneau, C.-A. Legrand, P. Ferrari, and J.-D. Arnould, "Scalable modeling studies on the SCR ESD protection device," in *Proc. EOS/ESD Symp.*, 2011, pp. 1–8.
[28] R. Mertens and E. Rosenbaum, "A physics-based compact model for SCR devices used in ESD protection circuits," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2013, pp. 2B.2.1–2B.2.7.
[29] 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.
[30] Y. Zhou, J.-J. Hajjar, A. W. Righter, and K. P. Lisiak, "Modeling snapback of LVTSCR devices for ESD circuit simulation using advanced BJT and MOS models," in *Proc. EOS/ESD Symp.*, pp. 175–184, 2007.
[31] L. Lou and J. Liou, "An improved compact model of silicon-controlled rectifier (SCR) for electrostatic discharge (ESD) applications," in *IEEE Trans. on Elec. Dev.*, vol. 55, no. 12, pp. 3517–3524, Dec. 2008.
[32] T. J. Maloney and N. Khurana, "Transmission line pulsing techniques for circuit modeling of ESD phenomena," in *Proc. EOS/ESD Symp.*, pp. 49–54, 1985.
[33] H. Gieser and M. Haunschild, "Very-fast transmission line pulsing of integrated structures and the charged device model," in *Proc. EOS/ESD Symp.*, pp. 85–94, 1996.
[34] M. Mergens et al., "ESD-level circuit simulation—impact of gate RC-delay on HBM and CDM behavior," in *Proc. EOS/ESD Symp.*, pp. 446–455, 2000.
[35] N. Jack, J. Davis, M. Chaine, and E. Rosenbaum, "HBM cross power domain failure due to secondary tester pulse," in *Proc. EOS/ESD Symp.*, pp. 1–7, 2009.
[36] S. Joshi and E. Rosenbaum, "Transmission line pulsed waveform shaping with microwave filters," in *Proc. EOS/ESD Symp.*, 2003, pp. 1–8.
[37] J. E. Barth, K. Verhaege, L. G. Henry, and J. Richner, "TLP calibration, correlation, standards, and new techniques," *IEEE Trans. on Electronics Packaging Manufacturing*, vol. 234, no. 2, pp. 99–108, 2001.
[38] E. Grund and R. Gauthier, "TLP systems with combined 50 and 500-ohm impedance probes and kelvin probes," in *Proc. EOS/ESD Symp.*, pp. 1–10, 2003.
[39] T. Daenen, S. Thijs, M. I. Natarajan, V. Vassilev, V. De Heyn, and G. Groeseneken, "Multilevel transmission line pulse (MTLP) tester," in *Proc. EOS/ESD Symp.*, pp. 1–6, 2004.
[40] E. Grund, "Snapback device studies using multilevel TLP and multi-impedance TLP testers," in *Proc. EOS/ESD Symp.*, pp. 1–9, 2005.
[41] D. Johnsson, D. Pogany, J. Willemen, E. Gornik, and M. Stecher, "Avalanche breakdown delay in ESD protection diodes," in *IEEE Trans. on Elec. Dev.*, vol. 57, no. 10, pp. 2470–2476, 2010.
[42] M. Scholz et al., "Calibrated wafer-level HBM measurements for quasi-static and transient device analysis," in *Proc. EOS/ESD Symp.*, pp. 89–94, 2007.
[43] E. Grund and M. Hernandez, "Obtaining TLP-like information from an HBM simulator," in *Proc. EOS/ESD Symp.*, pp. 95–101, 2007.
[44] M. Scholz, "ESD on-wafer characterization: is TLP still the right measurement tool?" *IEEE Trans. on Inst. and Meas.*, vol. 58, no. 10, pp. 3418–3246, 2009.

[45] N. Thomson, N. Jack, and E. Rosenbaum, "Exponential-edge transmission line pulsing for snap-back device characterization," in *Proc. IEEE Int. Reliab. Phys. Symp.*, pp. 3E.2.1–3E.2.6, 2012.
[46] V. A. Vashchenko et al., "Turn-off characteristics of the CMOS snapback ESD protection devices — new insights and its implications," in *Proc. EOS/ESD Symp.*, pp. 39–45, 2006.
[47] A. Salman et al., "Characterization and investigation of the interaction between hot electron and electrostatic discharge stresses using NMOS devices in 0.13μm CMOS technology," in *Proc. IEEE Int. Reliab. Phys. Symp.*, pp. 219–225, 2001.
[48] A. Ille et al., "Ultra-thin gate oxide reliability in the ESD time domain," in *Proc. EOS/ESD Symp.*, pp. 285–294, 2006.
[49] A.Ille et al., "Reliability aspects of gate oxide under ESD pulse stress," in *Proc. EOS/ESD Symp.*, pp. 328–337, 2007.
[50] K. Bock, B. Keppens, V. De Heyn, G. Groeseneken, L. Y. Ching, and A. Naem, "Influence of gate length on ESD-performance for deep submicron CMOS technology," in *Proc. EOS/ESD Symp.*, 1999, pp. 95–104.
[51] V. De Heyn, G. Groeseneken, B. Keppens, M. Natarajan, L. Vacaresse, and G. Gallopyn, "Design and analysis of new protection structures for smart power technology with controlled trigger and holding voltage," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2001, pp. 253–258.
[52] C. Duvvury, C. Diaz, and T. Haddock, "Achieving uniform nMOS device power distribution for sub-micron ESD reliability," in *Tech. Dig. IEDM*, 1992, pp. 131–134.
[53] C. Jiang, E. Nowak, and M. Manley, "Process and design for ESD robustness in deep submicron CMOS technology," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 1996, pp. 233–236.
[54] M.-D. Ker, C.-H. Chuang, and W.-Y. Lo, "Layout design on multi-finger MOSFET for on-chip ESD protection circuits in a 0.18-μm salicided CMOS process," in *Proc. IEEE Int. Symp. Electronics, Circuits and Systems (ICECS)*, 2001, pp. 361–364.
[55] J.-H. Lee, K.-M. Wu, S.-C. Huang, and C.-H. Tang, "The dynamic current distribution of a multi-fingered GGNMOS under high current stress and HBM ESD events," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2006, pp. 629–630.
[56] J. Li et al., "Predictive full circuit ESD simulation and analysis using extended ESD compact models methodology and tool implementation," in *Proc. EOS/ESD Symp.*, pp. 111–117, 2010.
[57] S. Joshi, R. Ida, P. Givelin, and E. Rosenbaum, "An analysis of bipolar breakdown and its application to the design of ESD protection circuits," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2001, pp. 240–245.
[58] S. L. Miller, "Ionization rates for holes and electrons in silicon," *Phys. Rev.*, vol. 105, no. 4, pp. 1246–1249, 1957.
[59] C. Russ, H. Gieser, and K. Verhaege, "ESD protection elements during hbm stress tests—further numerical and experimental results," in *Qual. Reliab. Engng. Int.*, Vol. 11: pp.285–294, 1995.
[60] A. Salman et al., "Characterization and investigation of the interaction between hot electron and electrostatic discharge stresses using NMOS devices in 0.13μm CMOS technology," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2001, pp. 219–225.
[61] J.-H. Lee, J. R. Shih, H. P. Kuan, and K. Wu, "The influence of decoupling capacitor on the discharge behavior of fully silcided power-clamped device under HBM ESD event," in *IEEE Symp. on the Physical and Failure Analysis of Integrated Circuits*, pp. 1–5, 2010.
[62] A. S. Elwakil, "Low-voltage relaxation oscillator," in *Electronics Letters*, vol. 35, no. 15, pp. 1256–1257, 2000.
[63] S. Ruth, M. Stockinger, J. W. Miller, V. Whitney, M. Kearney, and S. Ngo, "A CDM robust 5V distributed ESD clamp network leveraging both active MOS and lateral NPN conduction," in *Proc. EOS/ESD Symp.*, 2011, pp. 1–9.
[64] M.-D. Ker and H.-C. Hsu, "The Impact of inner pickup on ESD robustness of multi-finger NMOS in nanoscale CMOS technology," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2006, pp. 631–632.

[65] J.-H. Lee, K.-M. Wu, S.-C. Huang, and C.-H. Tang, "The dynamic current distribution of a multi-fingered GGNMOS under high current stress and HBMESD events," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2006, pp. 629–630.
[66] S. Beebe, "Simulation of complete CMOS I/0 circuit response to CDM stress," in *Proc. EOS/ESD Symp.*, 1998, pp. 259–270.
[67] M. V. Dunga, X. Xi, J. He, W. Liu, K. M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu, "BSIM4.6.0 MOSFET model," BSIMGroup, Univ. California at Berkeley, Berkeley, CA, USA, 2012. [On-line]. Available: http://www-device.eecs.berkeley.edu/~bsim3/BSIM4
[68] T. Skotnicki, G. Merckel, and A. Merrachi, "New physical model of multiplication-induced breakdown in MOSFETs," *Solid-State Electronics*, vol. 34, no. 11, pp. 1297–1307, Nov. 1991.
[69] G. Boselli, C. Duvvury, and V. Reddy, "Efficient pnp characteristics of pMOS transistors in sub-0.13µm ESD protection circuits," in *Proc. EOS/ESD Symp.*, 2002, pp. 260–269.
[70] M. Litzenberger, R. Pichler, D. Pogany, E. Gornik, K. Esmark, and H. Gossner, "Influence of layout parameters on triggering behavior in 0.35µm and 0.18µm process gg-nMOS ESD protection devices," in *Proc. ESSDERC*, 2001. pp. 335–338.
[71] M. Heer et al., "Analysis of the triggering behavior of low voltage BCD single and multi-finger gc-NMOS ESD protection devices," in *Proc. EOS/ESD Symp.*, 2006, pp. 275–284.
[72] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 3rd ed., New York: Wiley, 2002.
[73] V. Shukla, N. Jack and E. Rosenbaum, "Predictive simulation of CDM events to study effects of package, substrate resistivity and placement of ESD protection circuits on reliability of integrated circuits," in *Proc. IEEE Int. Rel. Phys. Symp.*, 2010, pp. 485–493.
[74] V. Shukla and E. Rosenbaum, "CDM simulation study of a system-in-package," in *Proc. EOS/ESD Symp.*, 2010, pp. 1–8.
[75] D. Abessolo-Bidzo, T. Smedes, and A. J. Huitsing, "CDM simulation based on tester, package and full integrated circuit modeling: case study," in *IEEE Trans. on Elec. Dev.*, vol. 59, no. 11, pp. 2869–2875, 2012.
[76] C. A. Torres, J. W. Miller, M. Stockinger, M. D. Akers, M. G. Khazhinsky, and J. C. Weldon, "Modular, portable, and easily simulated ESD protection networks for advanced CMOS technologies," in *Microelectronics Reliability*, vol. 42, no. 6, pp. 873–885, 2002.
[77] J.-R. Manouvrier et al., "A physics-based compact model for ESD protection diodes under very fast transients," in *Proc. EOS/ESD Symp.*, 2008, pp. 67–75.
[78] IBIS (Input Output Buffer Information Specification) ver. 6.0, Sept. 2013. [Online]. Available: http://www.eda.org/ibis/about/
[79] N. Monnereau, F. Caignet, N. Nolhier, D. Trémouilles, and M. Bafleur, "Behavioral-modeling methodology to predict electrostatic-discharge susceptibility failures at system level: An IBIS improvement," in *Proc. EMC Eur. Symp.*, Sep. 2011, pp. 457–463.
[80] C. C. Russ et al., "GGSCRs: GGNMOS triggered silicon controlled rectifiers for ESD protection in deep sub-micron CMOS processes," in *Proc. EOS/ESD Symp.*, 20001, pp. 22–31.
[81] J. Di Sarro, K. Chatty, R. Gauthier, and E. Rosenbaum, "Study of design factors affecting turn-on time of silicon controlled rectifiers (SCRS) in 90 and 65nm bulk CMOS technologies," in *Proc. IEEE Int. Rel. Phys. Symp.*, 2006, pp. 163–168.
[82] K. Esmark et al., "Transient behavior of SCRs during ESD pulses," in *Proc. IEEE Int. Rel. Phys. Symp.*, 2008, pp. 247–253.
[83] Z. Chen, R. Mertens, C. Reiman, and E. Rosenbaum, "Improved GGSCR layout for overshoot reduction," in *Proc. IEEE Int. Rel. Phys. Symp.*, 2015, pp. 3F.2.1–3F.2.8.
[84] T. Smedes and Y. Christoforou, "On the relevance of IC ESD performance to product quality," in *Proc. EOS/ESD Symp.*, 2008, pp. 14–20.
[85] ESD Association. (2002). White paper — ESD phenomena and the reliability for microelectronics. [Online]. Available: http://www.esda.org/documents/whitepaper_000.pdf

[86] Industry Council on ESD Target Levels. (Apr. 2010). White paper 2: A case for lowering component level CDM ESD specifications and requirements. (rev.2.0) [Online]. Available: http://www.esda.org/documents/IndustryCouncilWhitePaper2.pdf
[87] K. Watanabe, T. Hiraoka, K. Sato, T. Sei, and K. Numata., "New protection techniques and test chip design for achieving high CDM robustness," in *Proc. EOS/ESD Symp.*, 2008, pp. 332–338.
[88] E. R. Worley, "Distributed gate ESD network architecture for inter-power domain signals," in *Proc. EOS/ESD Symp.*, 2004, pp. 1–10.
[89] J. Lee, Y. Huh, J.-W. Chen, P. Bendix and S. M. Kang, "Chip level simulation for CDM failures in multi-power ICs," in *Proc. EOS/ESD Symp.*, 2000, pp. 456–464.
[90] M. Dissegna et al., "CDM circuit simulation of a HV operational amplifier realized in 0.35μm smart power technology," in *Proc. EOS/ESD Symp.*, 2007, pp. 1B.3-1–1B.3-10.
[91] N. Olson, V. Shukla and E. Rosenbaum, "Test chip design for study of CDM related failures in SoC designs," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2011, pp. 719–724.
[92] M. S. B Sowarariraj, T. Smedes, C. Salm, A. J. Mouthaan, and F. G. Kuper, "Role of package parasitics and substrate resistance on the CDM failure levels," in *Microelectronics Reliability*, vol. 43, pp. 1569–1575, 2003.
[93] V. Shukla, G. Boselli, M. Dissegna, C. Duvvury, R. Sankaralingam, and E. Rosenbaum, "Prediction of charged device model peak discharge current for microelectronic components," in *IEEE. Trans. Dev. Mat. Rel.*, vol. 14, no. 3, pp. 801–809, 2014.
[94] I. Doerr et al., "Electrical characterization of a PowerSO-package in the context of electrostatic discharge," in *Proc. IEEE Int. Symp. Electromagn. Compat.*, 2003, pp. 390–393.
[95] T. Maloney and N. Jack, "CDM tester properties as deduced from waveforms," in *Proc. EOS/ESD Symp.*, 2013, pp. 1–9.
[96] J. Di Sarro, B. Reynolds, and R. Gauthier, "Influence of package parasitic elements on CDM stress," in *Proc. EOS/ESD Symp.*, 2013, pp. 1–9.
[97] V. Kireev, J. Karp, M. Hart, S. Jeong, and P. Upadhyaya, "Effect of delay in package traces on CDM stress and peak current," in *Proc. EOS/ESD Symp.*, 2009, pp. 1–10.
[98] K.-H. Meng, A. Gerdemann, J. W. Miller, and E. Rosenbaum, "A co-optimization methodology on ESD robustness and functionality for pad-ring circuitry," in *Proc. EOS/ESD Symp.*, 2014, pp. 1–10.
[99] N. Olson, N. Jack, V. Shukla and E. Rosenbaum, "CDM-ESD induced damage in components using stacked-die packaging," in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 6-4.1–6-4.4, 2011.
[100] N. Jack, V. Shukla, and E. Rosenbaum, "Comparison of wafer-level with package-level CDM stress facilitated by real-time probing," in *IEEE. Trans. Dev. Mat. Rel.*, vol. 11, no. 4, pp. 522–530, 2011.
[101] E. Y. Wu and J. Suñé, "Power-law voltage acceleration: a key element for ultra-thin gate oxide reliability," in *Microelectronics Reliability*, vol. 45, no. 12, pp. 1809–1834, 2005.
[102] E. Wu et al., "Polarity-dependent oxide breakdown of NFET devices for ultra-thin gate oxide," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2002, pp. 60–72.
[103] B. E. Weir, C.-C. Leung, P. J. Silverman, and M. A. Alam, "Gate dielectric breakdown in the time-scale of ESD events," in *Microelectronics Reliability*, vol. 45, no. 3–4, pp. 427–436, 2005.
[104] ANSI/ESDA/JEDEC Joint Standard for Electrostatic Discharge Sensitivity Testing, Charged Device Model (CDM) - Device Level, JEDEC JS-002-2014, 2014
[105] E. Rosenbaum, K.-H. Meng, Y. Xiu, and N. Thomson, "Current challenges in component-level and system-level ESD simulation," in *Asian-Pacific EMC Symp.*, 2015.
[106] C.-W. Ho, A. E. Ruehli, and P. A. Brennan, "The modified nodal approach to network analysis," in *IEEE Trans. Circuits and Systems*, vol. 22, no. 6, pp. 504–509, June 1975.
[107] K. Kundert, *The Designer's Guide to Spice and Spectre®* , New York, Springer US, 1995.
[108] Virtuoso® Spectre® Circuit Simulator Reference (ver. 11.1), Sept. 2011. [Online] Available: https://projects.cecs.pdx.edu/attachments/download/3248/spectreref111.pdf
[109] K.-H. Meng, R. Mertens, and E. Rosenbaum, "Piecewise-linear model with transient relaxation for circuit-level ESD simulation," submitted to *IEEE Trans. Device Mater. Rel.*

[110]C.-H. Cho, N.-Y. Tsai, H. Yu, C.-R. Lee, Y. Shi, and S.-C. Chang, "On the preconditioner of conjugate gradient method — A power grid simulation perspective," in *Proc. Int. Conf. Comput.-Aided Des.*, pp. 494–497, 2011.
[111]S. S.-Y. Liu, C.-J. Lee, C.-C. Huang, H.-M. Chen, C.-T. Lin, and C.-H. Lee, "Effective power network prototyping via statistical-based clustering and sequential linear programming," in *Proc. Conf. Des., Automation and Test in Europe*, pp. 1701–1706, 2013.
[112]M. R. Hestenes and E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Systems," in *Journal of Research of the National Bureau of Standards*, vol. 49, no. 6, Dec. 1952.
[113]J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Aug. 1994. [Online] Available: http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf
[114]K.-H. Meng, V. Shukla, and E. Rosenbaum, "Full-component modeling and simulation of charged device model ESD," submitted to *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*
[115] G. Depeyrot, F. Poullet and B. Dumas "Verilog-A compact model coding whitepaper," in *Proc. Nanotechnology Conf.*, vol. 2, pp. 821–824, June, 2010.
[116]M.-A. Chalkiadaki, C. Valla, F. Poullet, and M. Bucher, "Why- and how- to integrate Verilog-A compact models in SPICE simulators," in *Int. J. Circuit Theory Appl.*, vol. 41, no. 11, pp. 1203–1211, Nov. 2013.
[117]https://campuscluster.illinois.edu/