2017

Design and Simulation of Device Failure Models for Electrostatic Discharge (ESD) Event

Meng Miao
University of Central Florida

DESIGN AND SIMULATION OF DEVICE FAILURE MODELS FOR ELECTROSTATIC DISCHARGE (ESD) EVENT

by

MENG MIAO
B.S. Zhejiang University, 2009
M.S. Zhejiang University, 2012

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2017

Major Professor: Kalpathy B. Sundaram

ABSTRACT

In this dissertation, the research mainly focused on discussing ESD failure event simulation and ESD modeling, seeking solutions for ESD issues by simulating ESD event and predict possible ESD reliability problem in IC design. The research involves failure phenomenon caused by ESD/ EOS stress, mainly on the thermal failure due to inevitable self-heating during an ESD stress. Standard Complementary Metal-Oxide-Semiconductor (CMOS) process and high voltage Doublediffusion Metal-Oxide-Semiconductor (DMOS) process are used for design of experiment. A multi-function test platform High Power Pulse Instrument (HPPI) is used for ESD event evaluation and device characterization. SPICE-like software ADICE is for back-end simulation.

Electrostatic Discharges (ESD) is one of the hazard that may affect IC circuit function and cause serious damage to the chip. The importance of ESD protection has been raised since the CMOS technology advanced and the dimension of transistors scales down. On the other hand, the variety of applications of chips is also making corresponding ESD protection difficult to meet different design requirement. Aside from typical requirements such as core circuit operation voltage, maximum accepted leakage current, breakdown conditions for the process and overall device sizes, special applications like radio frequency and power electronic requires ESD to be low parasitic capacitance and can sustain high level energy. In that case, a proper ESD protection design demands not only a robust ESD protection scheme, but co-design with the inner circuit. For that purpose, it is necessary to simulate the results of ESD impact on IC and find out possible weak point of the circuit and improve it. The first step of the simulation is to have corresponding

models available. Unfortunately, ESD models, especially there are lack of circuit-level ESD models that provide quick and accurate prediction of ESD event.

In this dissertation paper, ESD models, especially ESD failure models for device thermal failure are introduced, with modeling methodology accordingly. First, an introduction for ESD event and typical ESD protection schemes are introduced. Its purpose is to give basic concept of ESD. For ESD failure models, two typical types can be categorized depends on the physical mechanisms that cause the ESD damage. One is the gate oxide breakdown, which is electric field related. The other is the thermal-related failure, which stems from the self-heating effect associated with the large current passing through the ESD protection structure. The first one has become increasingly challenging with the aggressive scaling of the gate dielectric in advanced processes and ESD protection for that need to be carefully designed. The second one, thermal failure widely exists in semiconductor devices as long as there is ESD current flow through the device and accumulate heat at junctions. Considering the universality of thermal failure in ESD device, it is imperative to establish a model to simulate ESD caused thermal failure.

Several works related to ESD model can be done. One crucial part for a failure model is to define the failure criterion. As common solution for ESD simulation and failure prediction. The maximum current level or breakdown voltage is used to judge whether a device fails under ESD stresses. Such failure criteria based on measurable voltage or current values are straightforward and can be easy to implemented in simulation tools. However, the shortcoming of these failure criteria is each failure criterion is specifically designed for certain ESD stress condition. For example, the failure voltage level for Human Body Model and Charged Device Model are quite different, and it is hard to judge a device's ESD capability under standard test

conditions based on its transmission line pulse test result. So it is necessary to look deeper into the physical mechanism of device failure under ESD and find a more universal failure criterion for various stress conditions.

As one of the major failure mechanisms, thermal failure evaluated by temperature is a more universal failure criterion for device failure under ESD stress. Whatever the stress model is, the device will fail if a critical temperature is reached at certain part inside the device and cause structural damage. Then finding out that critical temperature is crucial to define the failure point for device thermal failure. One chapter of this dissertation will focus on discussing this issue and propose a simple method to give close estimation of the real failure temperature for typical ESD devices.

Combined these related works, a comprehensive diode model for ESD simulation is proposed. Using existing ESD models, diode I-V characteristic from low current turn-on to high current saturation can be simulated. By using temperature as the failure criterion, the last point of diode operation, or the second breakdown point, can be accurately predicted.

Additional investigation of ESD capability of devices for special case like vertical GaN diode is discussed in Chapter IV. Due to the distinct material property of GaN, the vertical GaN diode exhibits unique and interesting quasi-static I-V curves quite different from conventional silicon semiconductor devices. And that I-V curve varies with different pulse width, indicating strong conductivity modulation of diode neutral region that will delay the complete turn-on of the vertical GaN diode.

To my parents Zhenqing Miao and Hongli Wang

ACKNOWLEDGMENTS

It is a great experience for me as a graduate student in UCF. It won't be possible for me to accomplish this and finish this work without the support from people around me. It is their selfless help and encouragement that keeps me push forward.

First, I'd like to thank Prof. Juin J. Liou for his guidance during my Ph.D study. His expert advice always help me a lot when I encounter obstacle in my research and keep me working on the right track. Prof. Liou is my academic advisor as well as my friend. We can discuss freely on research problems, which is a great help on my projects. Prof. Liou put his trust on me and let me participate in the funded research projects with Analog Devices Inc. (ADI), in which, I learned a lot and accomplished several publication that enables me to graduate as a Ph.D student.

I want to thank Dr. Kalpathy B. Sundaram who helped me a lot during my Ph.D program. He is helpful and give support to students. Also, his advice is always inspiring with his expertise in device physics and fabrication.

I want to show my great gratitude to Dr. Yuanzhong Zhou, who mentors me on my research projects. It was a great experience to work with him on research projects funded by ADI. His ideas and advice are invaluable and always help me to find the right solution in my research. He is helpful and resourceful, revising papers for me with great patience.

I want to also thank Dr. Jean-Jacques Hajjar and members in ESD group, ADI. Dr. Hajjar is an experienced manager, and his team always solve problems and guarantee the reliability of

the product. I also would like to thank Javier Salcedo for his support on my research. And Srivatsan Parthasarathy offering valuable suggestions and discussions to this research work.

I would like to thank my dissertation committee members, Dr. Jiann S. Yuan, Dr. Xun Gong, Dr. Javier Salcedo and Dr. Yier Jin, for their invaluable suggestions and comments on improving this research work. I want to thank Analog Device Inc. for giving the opportunity to me and let me participate in the joint research project between ADI and UCF, from which I gained real world industrial experience.

I acknowledge and thank many valuable UCF professors who shared with me their wisdom and invaluable time, and the talented students from our research group. Special thanks to Dr. Wen Liu (GF), David Elis, Dr. Qiang Cui (Qorvo), Shurong Dong (ZJU), Zhixin Wang (UCF), Yunfeng Xi (UCF), Wei Liang (UCF), Aihua Dong (UCF), Linfeng He (UCF), Hang Li (UCF), for the valuable time we spent together in UCF.

I would like to express my gratitude to the financial, academic and technical support from University of Central Florida and its staff. And all the instructors at UCF, Zhejiang University, Zhoushan Secondary School, who guided my learning and provide me exemplary academic and moral teachings.

Last but not least, I am deeply indebted to my parents, Zhenqing Miao and Hongli Wang, Their love and support is my greatest motivation to accomplish this research work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

BJT	Bipolar Junction Transistor
CDM	Charged Device Model
CMOS	Complementary Metal-Oxide-Semiconductor
DMOS	Doublediffusion Metal-Oxide-Semiconductor
DUT	Device Under Test
EOS	Electrical Overstress
ESD	Electrostatic Discharge
GGNMOS	Gate-Grounded N-type MOSFET
GaN	Gallium Nitride
HBM	Human Body Model
HMM	Human Metal Model
HPPI	High Power Pulse Instrument
IC	Integrated Circuits
IEC	International Electrotechnical Commission
MM	Machine Model
MEMS	Micro Electro Mechanical System
SCR	Silicon Controlled Rectifier
SOG	Spin on Glass
TCAD	Technology Computer Aided Design

TDR	Time Domain Reflection
TDT	Time Domain Transmission
TDRT	Time Domain Reflection and Transmission
TLP	Transmission Line Pulsing
VFTLP	Very Fast Transmission Line Pulsing

CHAPTER 1 INTRODUCTION

1.1 ESD Event and ESD Failure

Since the first Integrated Circuit (IC) was invented in 1950s, the semiconductor IC has evolved dramatically in both dimension and scale. Following the "Moore's Law" for decades, the performance of IC has improved a lot with more transistors integrated on a single small chip [1]. On the other hand, with the shrinking dimension of transistors, problems in manufacture and reliability have emerged increasingly. To meet various requirement of applications (radio frequency [2], power electronics [3], MEMS [4], magnetic recording [5]), the fabrication and function of ICs has been greatly specified.

Long before the IC manufacture entered the era of sub-micrometer. Electrostatic Discharge (ESD) event is one of the major effect that influence the IC product reliability [6-8]. It is an inevitable event during the whole cycle of IC fabrication, packaging, transportation and application. Once there is electric charge generated and accumulated either by ordinary movement, like rubbing a doorknob or walking on a carpet, or transfer charge from one object to another, electrostatic discharge may occur, causing short-time but high voltage spark. Such discharge flowing through electric device may damage the vulnerable part of the IC, especially at the input/output port.

A whole protection design for ESD event is a systematic work, which involves both IC design and manufacture. Protection must include on-chip and off-chip design. Thus, it is a difficult work requires case-by-case solutions. It is reported that during the whole cycle of IC product, 37% failure is caused by ESD/EOS event [7]. The solution for ESD/EOS problem must be taken seriously.

Before IC manufacture entered sub-micrometer, ESD was not a big problem in IC design. The transistor was large with thick gate oxide that can sustain significant ESD stress without extra protection. IC was able to pass ESD test with only a few protection at the I/O port. The major attention is paid to other reliability issue, like circuit soft failure, threshold voltage shift and latch-up [9, 10]. While with advanced IC technology, semiconductor transistor has become smaller but it has to face the same level of ESD stress and it makes IC more vulnerable, hard failure (gate oxide breakdown, P-N junction secondary breakdown, metal interconnection melting) may occur.

On the other hand, different IC products are fabricated for different application. For example, the ESD protection design for radio frequency circuit require not only a robust ESD protection, but an inserted ESD unit with very low parasitic effect[11, 12]. Special fabrication process such as Micro-Electro-Mechanical Systems (MEMS) and carbon nano-tube may need extra consideration of ESD [13, 14]. Even ordinary IC may encounter the dilemma between ESD robustness and compact circuit area. Such special requirement for different circuit application means there is no universal ESD protection for every situation.

To better evaluate ESD stress under various situations, people have standardize ESD into several ESD models:

(1) HBM (Human-Body-Model) [15];

(2) CDM (Charged- Device-Model) [16];

(3) IEC (International-Electrotechnical-Commission);

(4) HMM (Human-Metal-Model) [17, 18];

(5) MM (Machine-Model) [19];

These are the main ESD models that IC designer must concern. They cover major ESD event a chip may encounter and require corresponding ESD protection schemes. The following section will introduce these ESD models briefly.

1.1.1 HBM Model

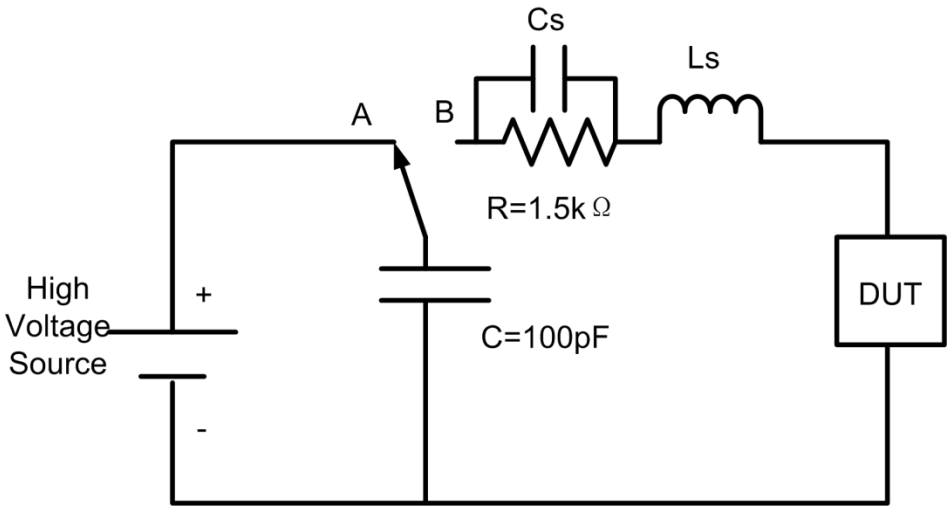


Figure 1-1 A simplified equivalent circuit for HBM ESD model.

HBM represents a charged human body contacts one pin of a grounded semiconductor device or an IC directly, transferring the charge into the device. Figure 1-1 shows its equivalent circuit. The C=100pF represents the equivalent capacitance of human body, the charge generated by rubbing or transferred from other objects is stored on this capacitance temporarily. Once the switch is closed with node B, namely the human body touches a grounded semiconductor device, electrostatic energy will discharge through a simple circuit and release on the device.

In the HBM equivalent circuit, R_{HBM} = 1.5kΩ and L_{HBM} = 7500nH are used to replicate a charged human body. Those values may vary a bit according to different standards include MIL-STD-883E [20], ESDA/JEDEC JS-001-2014 [21].

1.1.2 CDM Model

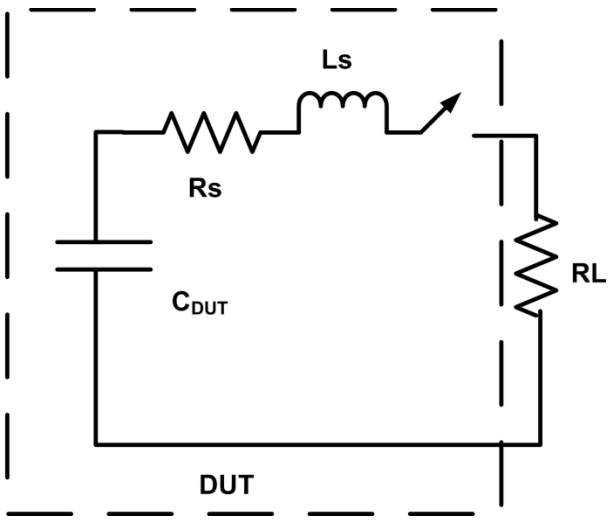


Figure 1-2 An equivalent circuit for CDM ESD stress model.

The charge device model (CDM) was aiming to model a package IC with charge already on it, and discharge the ESD energy through its pin when a pin contacts a grounded object or conductive surface. Such situation is highly possible during the manufacturing and assembly of ICs. Figure 1-2 gives the equivalent circuit of CDM. It should be noted that the parasitic resistance is smaller than the resistance in HBM, causing a much fast discharge event with higher overshoot peak at the very beginning of CDM event. This may be more deadly event for some applications, especially for the thin gate oxide of MOS transistors in more advanced process technology. Also due to the fast and short overshoot, damage may still occur even the circuit if

protected by ESD unit if the ESD unit doesn't respond fast enough. The commonly used standard for CDM are known as ESDA/JEDEC JS-002-2014 [22].

1.1.3 MM Model

Similar to HBM, MM is to model charged machines (metal tools, mechanical arm) contacts with one pin of a grounded chip. But MM has some parasitic inductance and lower parasitic capacitance than HBM, which will cause ESD impact in reverse direction, leading to more severe damage to the chip. Figure 1-3 is the typical equivalent circuit for MM. The commonly used standard for CDM are known as STM5.2-2012 – Electronic [23].

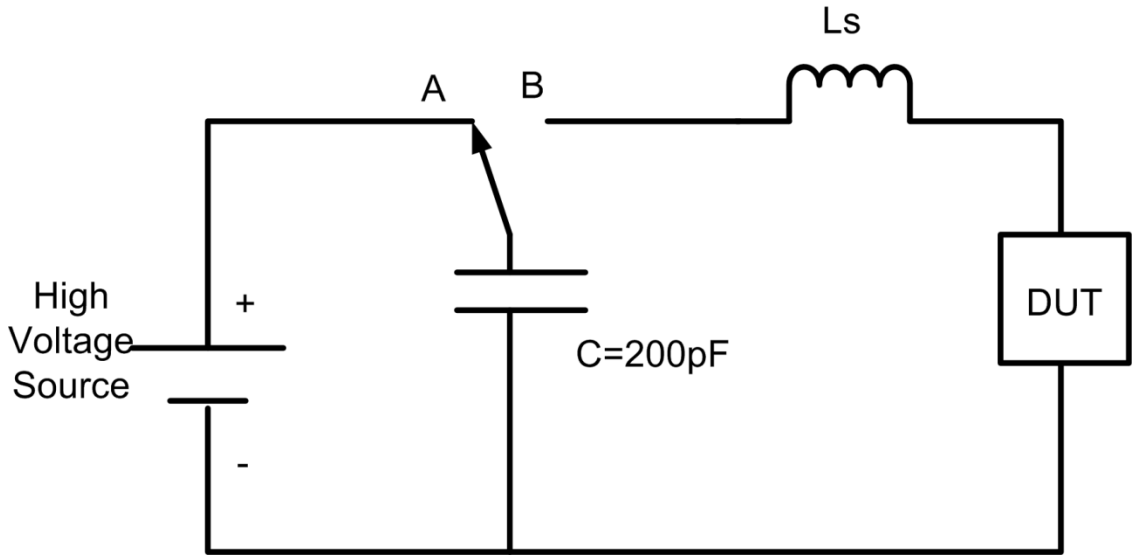


Figure 1-3 An equivalent circuit for MM ESD stress model.

1.1.4 IEC and HMM model

HBM, CDM and MM are ESD models for single component, they can be tested on wafer level or packaged chips. When the ICs are integrated with other components and assembled in a

system, the performance of component level ESD may degrade. The overall performance of a system must be evaluated with new ESD models. To ensure the system-level ESD robustness, IEC standard was developed for system level ESD testing [24-26]. An ESD gun, specialized tester, is usually used to produce such ESD stress. The HMM model is used to predict system level ESD robustness prior to assembly [27, 28]. It reproduce the situation that ESD discharge caused by a human touching a pin of a grounded component with a metal tool.

1.1.5 ESD Testing Method

Most ESD models can only evaluated the ESD level of a device under test (DUT), but cannot characterize it. It requires a method to reveal how the DUT behave under ESD stresses. The transmission-line pulse (TLP) system has been used to characterize on-chip ESD protection structure in the integrated circuit industry [29, 30]. This technique has been introduced by T.J.Maloney and N.Khurana in 1985 [31]. A schematic of a TLP system is shown in Figure 1-4. It is composed of a pulse generator, the transmission line and the DUT, together they form a closed transmission line. The incident and reflection waveforms on the transmission line are captured and recorded. Unlike a DC measurement using a step in test waveform, TLP generates a sequence of discrete waveforms but with increasing magnitude. Each waveform and its results are recorded and a I-V characterization curve is plotted.

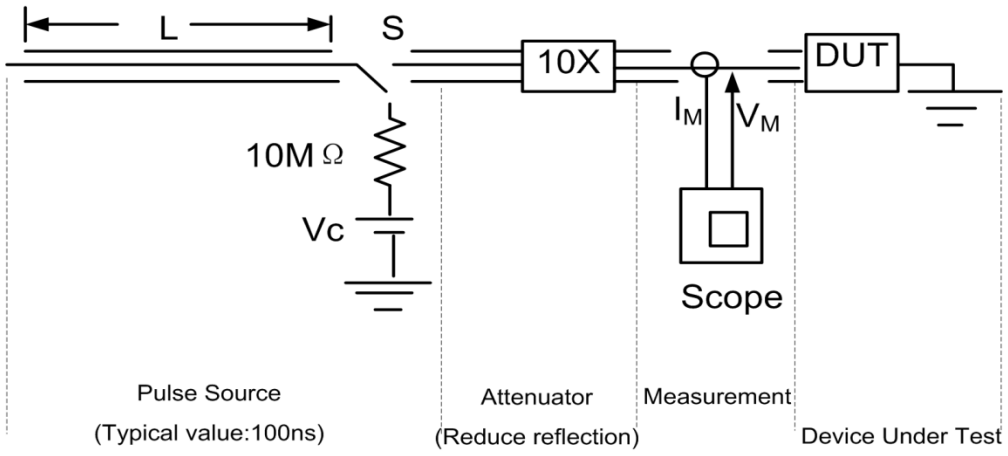


Figure 1-4 A TLP equivalent diagram.

Typical TLP system generated pulses with a rise time of 0.5 ns ~ 10ns, pulse width of 75 ns ~ 150 ns. There are several types of TLP system depending on how the pulse is generated and measured, such as Time Domain Reflection (TDR), Time Domain Transmission (TDT), Time Domain Reflection and Transmission (TDRT) and Current source (with 500 ohm load resistance). For the TDR, it can be categorized into TDR-O and TDR-S. Figure 1-5 gives one example of how a TLP I-V curve is plotted. As pulse is generated and flow through the transmission line, the incident and reflected voltage/current waveforms are captured by a high sampling rate oscilloscope. Using the stabilized portion (70% ~ 90%) of the voltage/current waveforms as one pair of voltage/current values, a sequence of voltage/current data points can be obtained with the increasing magnitude of the TLP pulses. Connecting those data dots, a TLP I-V curved can be drawn as represented in Figure 1-5. It must be noticed that even the other portion of the TLP waveforms contains valuable information. It can be seen that at the very beginning of the transient voltage waveform, there is a sharp peak, which is called overshoot [32]. The overshoot

with its peak value and the duration is the key criterion that reflects the response time of the ESD protection unit. With a smaller overshoot, the ESD device can turn-on much faster to handle fast ESD impact with large energy [33].

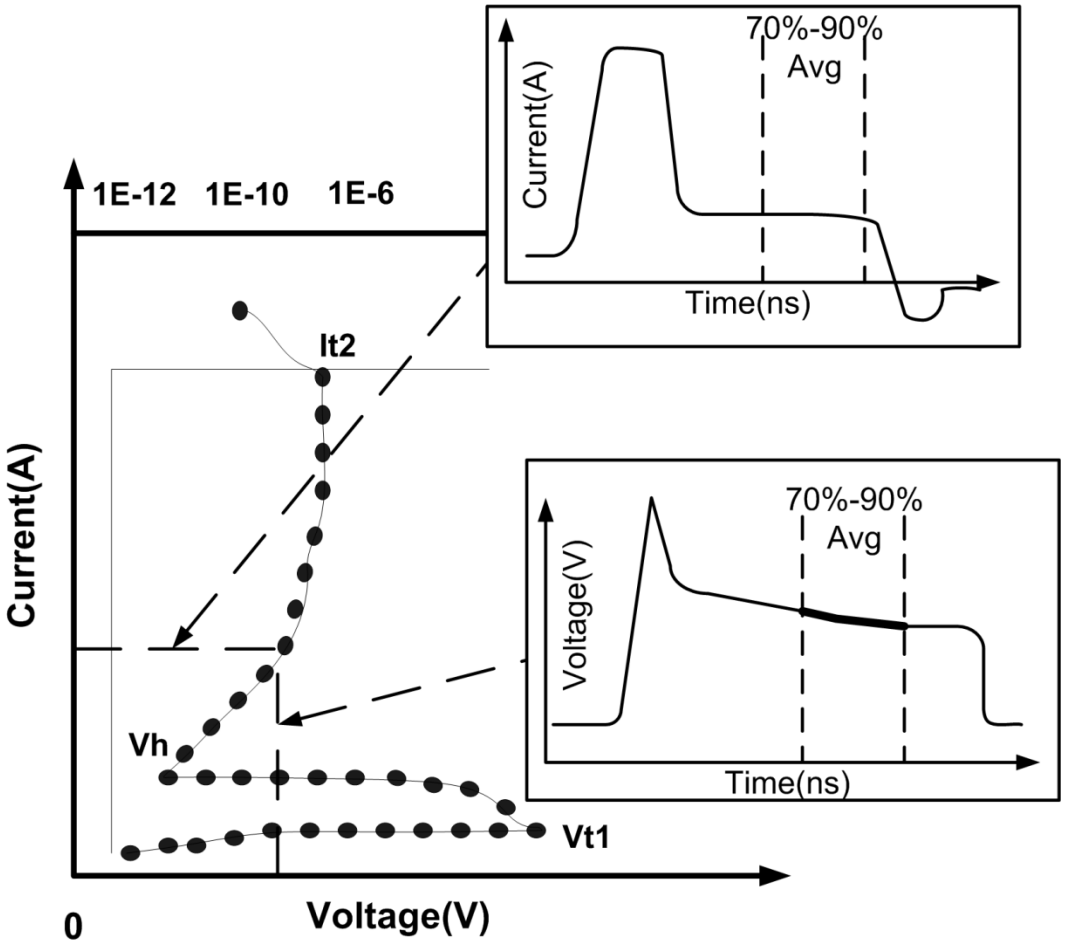


Figure 1-5 A diagram shows how a TLP I-V curve is plotted.

1.1.6 Multi-pin Test

For packaged chip, it usually has multi-pin configuration, depending on the chip and package type. To evaluate the ESD robustness of the product comprehensively, it requires the

chip being tested under different ESD models, and pass certain level for each model. In general, an industrial IC product will require HBM ≥ 2KV, CDM ≥ 1KV, MM ≥ 200V. For applications that the I/O contacts outer interface frequently, like USB, HDMI and RFIC, higher ESD levels are mandatory. Besides ESD models, most IC product have to pass the latch-up test to ensure the chip functions normally under ESD/EOS situations. As aforementioned, the product must pass several IEC model tests to ensure the product at system level ESD and its Electro Magnetic Compatibility (EMC).

Since a chip usually has multi-pin configuration but conventional ESD test is a pin-to-pin test, a test pattern is needed when performed a full chip test. The below Figure 1-6 depicts eight test configurations for full chip test. It covers most situation that ESD current flow through the chip.

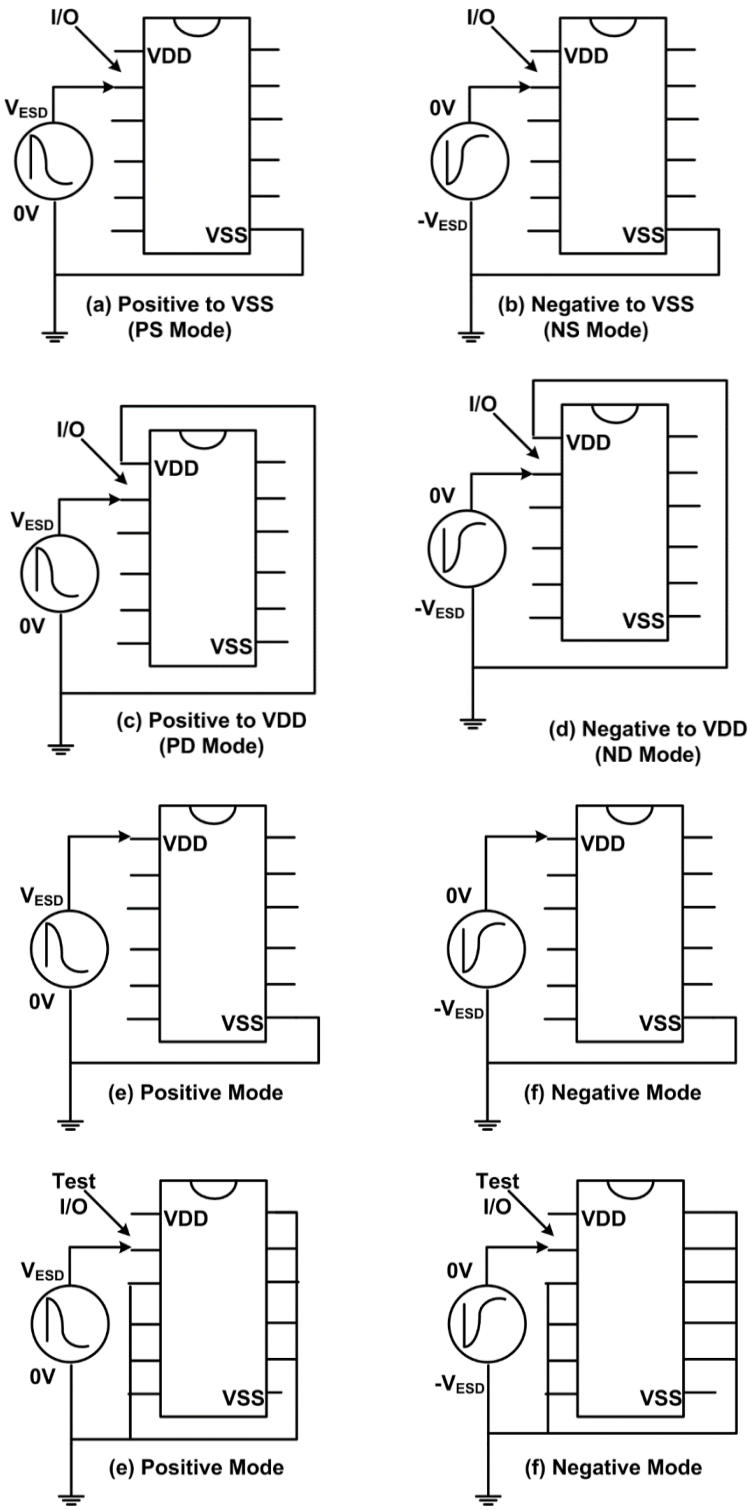


Figure 1-6 Diagrams show 8 test pattern for ESD test for multi-pin chips.

1.1.7 Latch-up Test

Although it's not an ESD damage, latch-up event is usually boned with ESD test. By draining voltage/current through the chip, it can check latch-up immunity of the chip and prevent possible hard/soft failure due to the defects of the design. Below Figure 1-7 is a standard procedure of latch-up test:

1) All inputs tie to high;
2) Outputs are floating;
3) At time T2, the DUT is tested at its normal operation condition (from VDD to VSS);
4) Adding drain voltage/current test signal during time T3 to T4 (current at I/O pins, voltage on VDD and VSS);
5) Measure the current between VDD and VSS at time T5, compare it with recorded current value at time T2;
6) If $I_{norm} < 25mA$, then $I_{T5} > I_{norm} + 10mA$ is a latch-up. If $I_{norm} > 25mA$, then $I_{T5} > 1.5*I_{norm}$ is a latch-up.

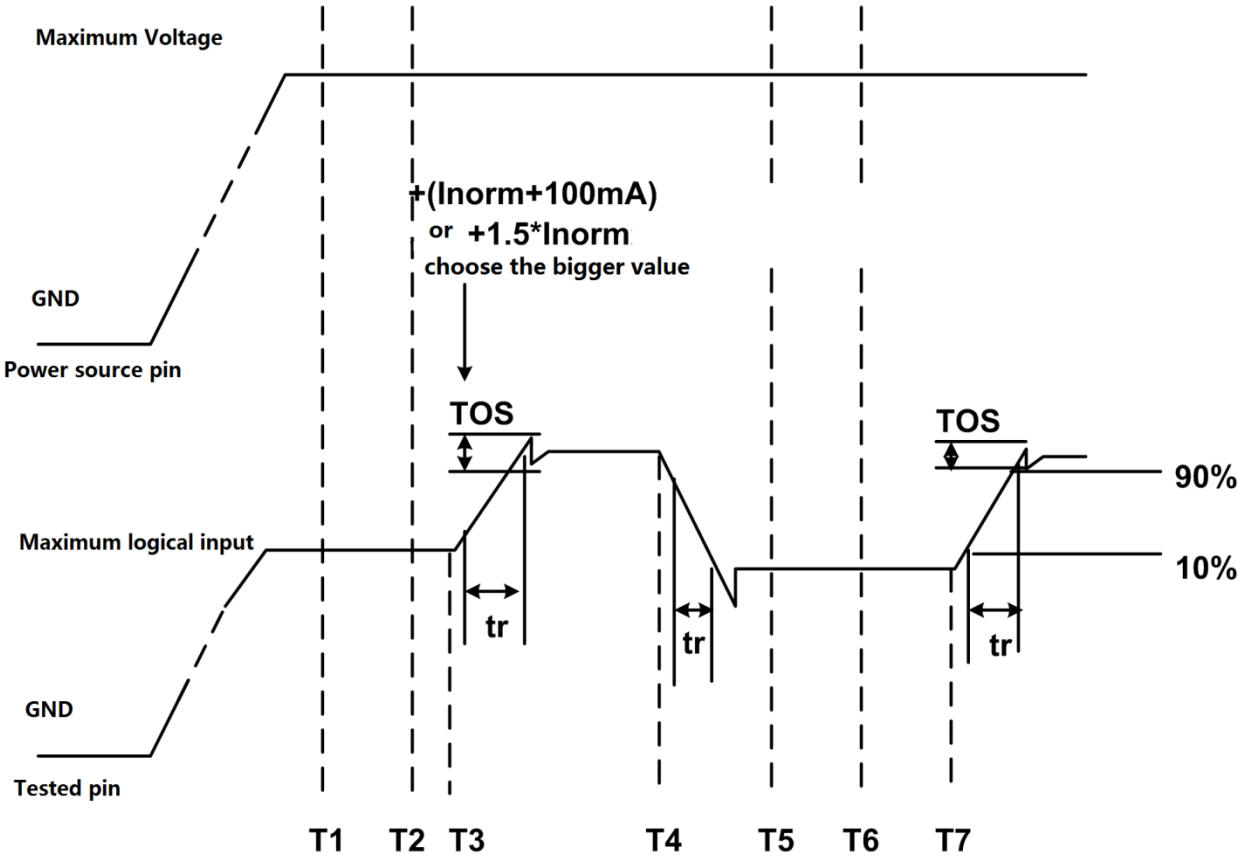


Figure 1-7 Latch-up test and the input test signal.

1.2 Basic ESD protection devices

1.2.1 Introduction

Before we discuss about different device used for ESD protection, it is necessary to introduce the basic concept of ESD design window and ESD protection schemes.

As aforementioned that ESD protection is a case-by-case design, for different protection target, proper ESD unit must be chosen carefully. There is a concept of ESD design window must be met for each design. Figure 1-8 shows a typical ESD design window and I-V characteristics of a basic ESD protection device. The ESD design window has two boundaries.

The lower limit is defined by the normal operation voltage of the circuit being protected. The holding voltage, or the lowest voltage that ESD unit can reach must higher than this lower limit, or the circuit may in risk of latup-up during ESD/EOS event. The upper limit of the ESD design window is defined by the hard failure voltage of the protected circuit, usually it is the gate oxide breakdown voltage of an input MOSFET or the drain-source breakdown voltage of the output drive transistor [34, 35]. The ESD unit cannot work beyond the upper limit, otherwise the inner circuit may be damaged by the ESD/EOS stress. A 10% safety margin is usually needed to make ESD protection designs more reliable. An ideal ESD protection unit has a fixed trigger voltage and very low on-resistance, however, which is contradicted with reality. A conventional ESD device will turn-on with a certain on-resistance and this on-resistance will gradually increase at high current region due to self-heating effect. Moreover, with more advanced IC process technology, the ESD design window is shrinking with upper and lower boundary closing in. Thus, it's getting difficult to design proper ESD protection scheme for advanced CMOS technology.

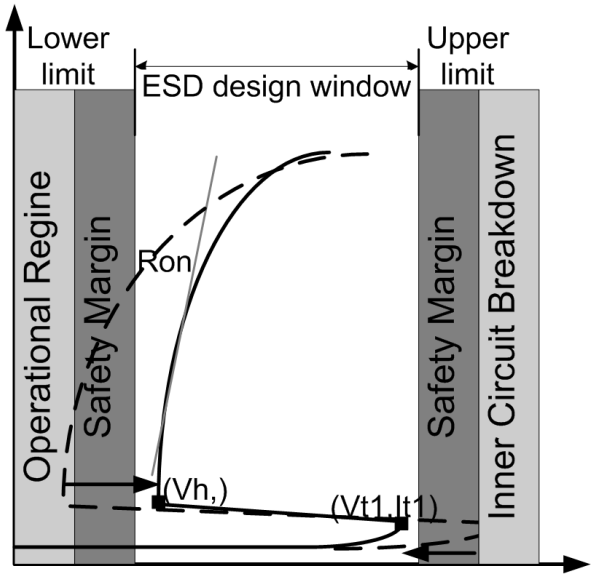


Figure 1-8 ESD design window.

There are several typical devices that can be used for ESD protection, all are evolved from conventional device but due to the purpose for ESD protection, modifications are made.

1.2.2 Diode

P-N junction diode is the simplest structure but a very effective ESD protection device. It has very low parasitic effect and easy to do metal interconnection [36, 37]. However, as most semiconductor devices designed for ESD protection, it comes with pros and cons.

1) Diode has very low turn-on voltage (0.6V ~ 0.8V) at forward direction, which could be either an advantage or disadvantage. For low voltage protection, we cannot find any other device that provides such low turn-on voltage. However, for operation voltage a bit higher, like over 2V, it requires three or four diodes in series, that series

connection will probably introduce a parasitic effect called Darlington effect and increase the leakage of diodes in series.

2) Although the turn-on voltage of diode's reverse direction is pretty high, the current capability of reverse conduction is very low. To achieve acceptable ESD current capability by using a reverse-biased diode, that diode must be very large, which will occupy a significant layout area.
3) Nonetheless, the parasitic capacitance per unit area for diode is very low, which makes diode an ideal candidate for RF ESD protection. Moreover, due to the simple structure of diode, its SPICE model is more accurate than other devices with complicated structures, thus circuit design could simulate and predict circuit performance with diode ESD protection in-situ.
4) Fast turn-on speed is another merit comes with diode protection. Short term ESD event like CDM requires ESD protection unit to respond within 1ns. To meet such rigid requirement, diodes are usually used as a secondary protection for CDM event.

There are several variations of diodes, below Figure 1-9 gives some typical examples.

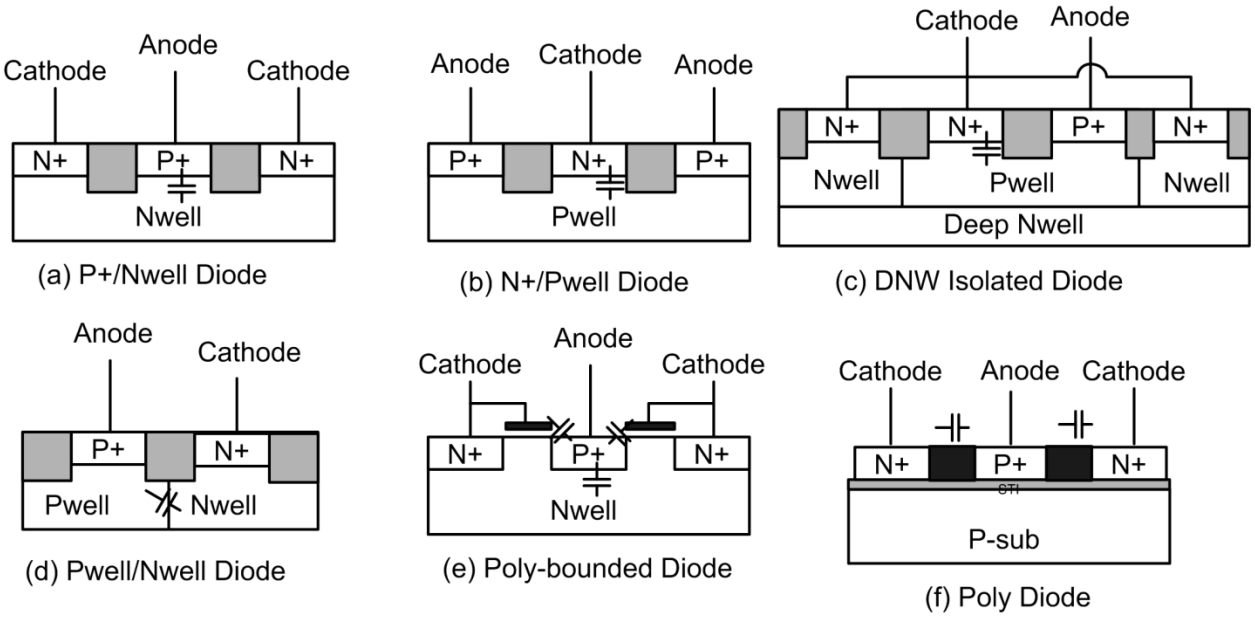


Figure 1-9 Cross-sections of typical diode structures for ESD application in CMOS process.

As we may expect, diodes with poly gate structure has fast turn-on speed but will introduce extra parasitic capacitance. To prevent this, metal interconnection and layout of these diodes is a delicate work.

As aforementioned, diodes in series may increase the leakage current due to the Darlington effect. It is a negative side-effect we must diminish when designing diode string protection scheme. As Figure 1-10 shows there is always a substrate to support the P-N junction diodes, thus for P+/Nwell diode, there is a parasitic P+/Nwell/Psub BJT as T1, T2 and T3 in the figure. The emitter of T2 is connected with the base of T1, The emitter of T3 is connected with base of T2. In that case, three BJTs form a cascade configuration and amplify the leakage current. As current operation equation of BJT states

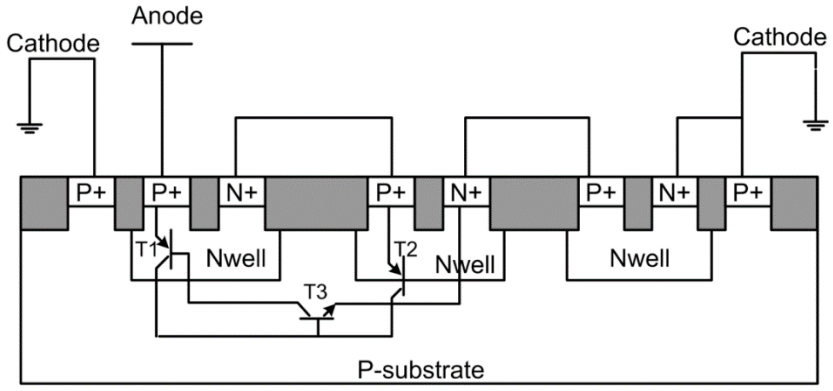


Figure 1-10 Cross-sections of P+/Nwell diodes with parasitic BJT paths embedded.

$$I_E = (1+\beta)I_B \tag{1.1}$$

and

$$I_D = I_s \cdot \exp\left(\frac{V_D}{V_T}\right) \tag{1.2}$$

we got

$$V_1 = V_T ln\frac{I_{E1}}{I_S} = V_T ln\frac{(\beta+1)I_{E2}}{I_S} = \cdots V_T ln\frac{(\beta+1)^{n-1}I_{En}}{I_S} = V_n + (n-1)V_T \ln(\beta+1) \tag{1.3}$$

suppose the gain of each diode is β, then the total voltage across the diode string is

$$V_{sum} = \sum_{i=1}^{n} V_n = nV_1 - \frac{n(n-1)}{2} V_T \ln(\beta+1) = nV_T[ln\frac{I_{E1}}{I_S(\beta+1)^{\frac{n-1}{2}}}] \tag{1.4}$$

finally, the leakage current caused by Darlington effect could be

$$I_{Leakage} = I_{E1} = I_s \cdot \exp\left(\frac{V_{sum}}{nV_T}\right) \cdot (1+\beta)^{\frac{n-1}{2}} \tag{1.5}$$

compared with the static current of an ideal diode string, which is

$$I_{Leakage} = I_s \cdot \exp\left(\frac{V_{sum}}{nV_T}\right) \tag{1.6}$$

the leakage is increased by a factor of $(1+\beta)^{\frac{n-1}{2}}$, and $I_{Bn} = I_{E1} \cdot \frac{1}{(1+\beta)^n}$

As the above analysis, due to the Darlington effect, the static leakage is a large portion of the total current. Due to the Webster Effect [38], such high current injection will lower the resistance of the base region and reduce efficiency of the emitter. Thus the ESD current capability of the diode string is reduce equivalently.

This negative effect is not very obvious when diodes are place far away from each other, however, as a compact layout is required in advanced CMOS process, devices are putting closely to each other to save layout area. In that case, the gain of the paracitc BJT will increase significantly, causing the inevitable Darlington Effect. Technology Computer Aided Design (TCAD) simulation gives an example in detail (Figure 1-11). When current increases, the high current region in red color has moved from diode string to the substrate and continuously reduce current flowing through the diode string. Such phenomenon has been reported in [39].

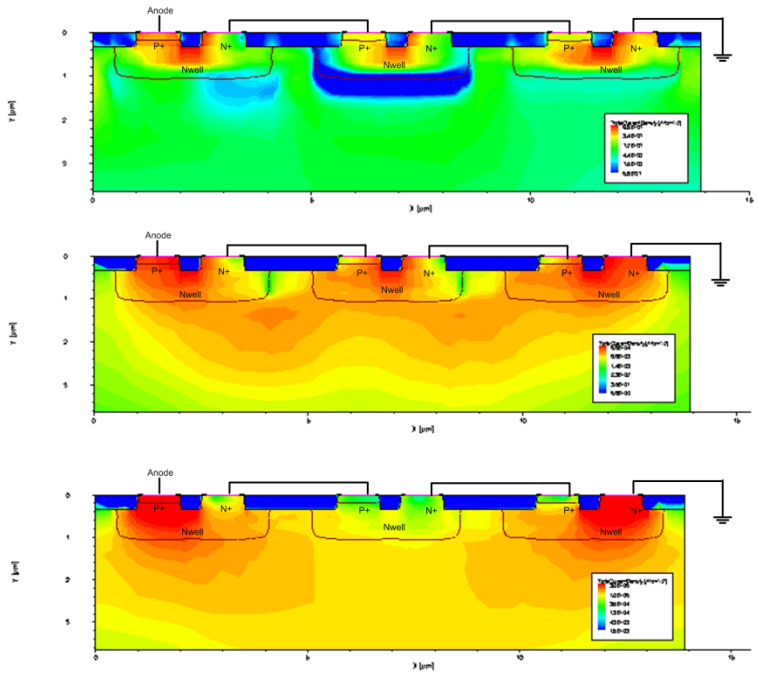


Figure 1-11 TCAD simulation of the total current density charts in a diode string.

1.2.3 MOSFET

Compared with diodes and SCRs, grounded-gate NMOS (GGNMOS) is more like a compromise option. It has better ESD area efficiency but still lower than SCR. The snapback is not as deep as SCR's, but enough for GGNMOS provide proper current capability. Because of these merit of GGNMOS, it can easily meet corresponding ESD window and widely used in standard I/O cell.

Unlike a NMOS for circuit application using the channel region to conduct current, rather, GGNMOS works as a BJT using the body region to conduct large portion of the ESD current. When a GGNMOS triggers under ESD stress, the channel will conduct the early portion of the ESD current. The gate and channel structure ensure the GGNMOS will turn-on soon enough

before the ESD energy cause damage to the inner circuit. As current increases, current flows through the channel will assist the embedded BJT to turn-on. Although the single GGNMOS structure is simple, it is more concerned about GGNMOS in multi-finger configuration when used in real situations.

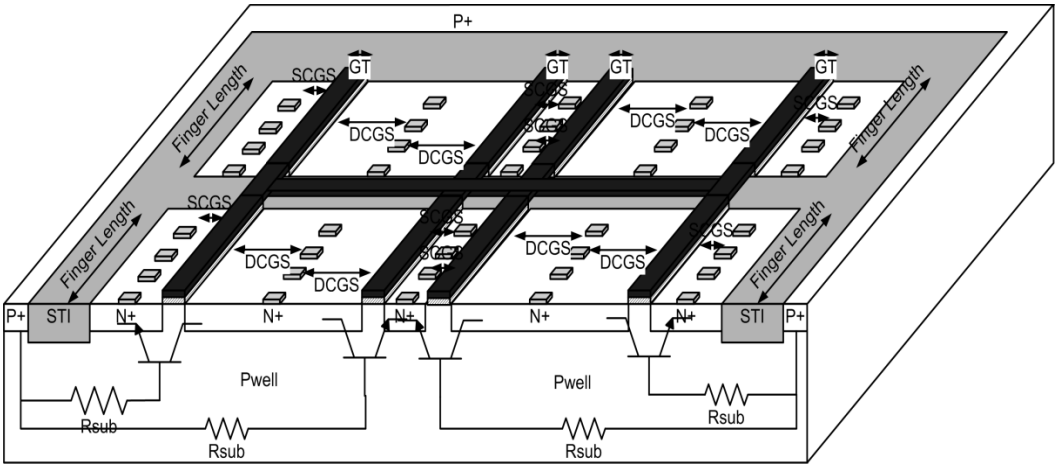


Figure 1-12 Layout and cross-section of a multi-finger GGNMOS configuration.

Figure 1-12 is an overview of a multi-finger GGNMOS. It can be seen that for multi-finger configuration, except the gate length, finger width and drain contact to gate space these single finger parameters, another major change introduced is the body tie. Since the P+ body tie is only around the multi-finger GGNMOS array, each finger has different distance to the body tie. This may cause non-uniformly conducting when triggering the multi-finger GGNMOS. To mitigate this, several technology is introduced. The most commonly used one is the silicide block and ballast resistor. The silicide block layer will prevent the generation of silicide at drain side,

adding a small resistance at drain side. Although this will slightly degrade the performance of each GGNMOS, the overall uniformly conducting is improved.

Another advantage of using GGNMOS as I/O ESD protection is saving area. Albeit the GGNMOS itself is not the ESD device with the most area efficiency, its value in I/O protection is intrinsic. As we know, for port with input and output function both, input MOSFET transistors are in parallel with output MOSFET transistors. In that case, those MOSFET can not only be drive transistors, but provide ESD protection capability in off-state.

1.2.4 SCR

Another type of commonly used snapback ESD protection device is silicon controlled rectifier (SCR) [40, 41]. It is a commonly used power electronic device, but its high power, high current capability is very valuable for ESD protection.

The cross-section of a standard SCR structure is shown in below Figure 1-13. By adding P+ and N+ in the adjacent Pwell and Nwell, a PNP BJT and a NPN BJT are interconnected. Figure 1-13 gives the equivalent circuit of SCR. The collector of PNP is connected with the base of NPN, and NPN is draining current from the base of PNP. Thus, a positive feedback is formed. Once ESD current flow through the well resistor of both BJTs and forms a voltage bias, BJTs are turned on and finally SCR start to conduct [42].

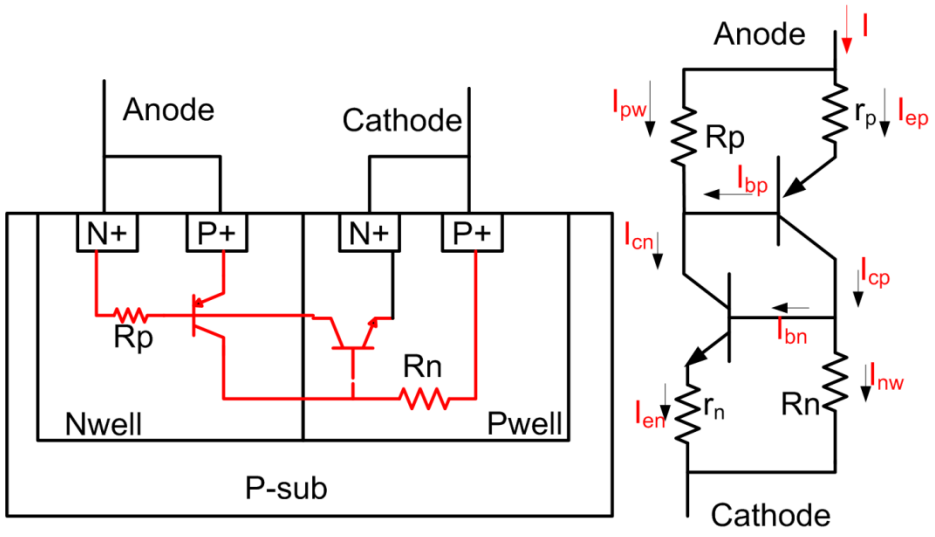


Figure 1-13 Cross-section and the equivalent circuit of the SCR structure.

The emitter current flowing through both BJTs occupies the majority of the ESD current, as $I_{en} = I_{ep} \gg I_{pw}, I_{nw}$. For both PNP and NPN, we have

$$I_{cn} \cong I_{bp} \tag{1.7}$$

$$I_{cp} \cong I_{bn} \tag{1.8}$$

in that case we have

$$I_{bp} \cong I_{cn} = \beta_n I_{bn} \leq \beta_p I_{cp} = \beta_n \beta_p I_{bn} \tag{1.9}$$

or, as SCR in snapback condition

$$\beta_n \beta_p \geq 1 \tag{1.10}$$

when including well current I_{pw} and I_{nw}, it will be

$$I = I_{ep} + I_{nw} \tag{1.11}$$

$$I = I_{en} + I_{pw} \tag{1.12}$$

thus, the total current should be

$$I = I_{cp} + I_{cn} = \alpha_p I_{ep} + I_{cp0} + \alpha_n I_{en} + I_{cn0} \tag{1.13}$$

$$I = \alpha_p (I_{ep} + I_{nw}) = \alpha_p I_{nw} + I_{cp0} + \alpha_n (I_{en} + I_{pw}) = \alpha_n I_{pw} + I_{cn0} \tag{1.14}$$

where α_p, $\alpha_n$are the common base gain of PNP and NPN BJT, I_{cp0}, I_{cn0} are saturation current of PNP and NPN in reverse direction. Finally, we can get

$$I = \frac{\beta_p(\beta_n+1)I_{nw}+\beta_n(\beta_p+1)I_{pw}}{\beta_p\beta_p-1} = I_H \tag{1.15}$$

where I_H is the holding current when SCR enters snapback state.

for the holding voltage, the emitter resistors r_{ep} and r_{en} must be considered.

$$V_H = V_{pBE} + V_{nBE} + I \cdot r_{ep} \cdot V_{nBE} + I \cdot r_{en} + V_{potential} \tag{1.16}$$

As we can see, with increasing I, the voltage across $I \cdot r_{ep} \cdot V_{nBE} + I \cdot r_{en}$ is increasing, which represents as the on-resistance of the SCR. The $V_{potential}$ is barrier across the reversed Nwell/Pwell. It is influenced by the junction shape, doping concentration of the Nwell/Pwell junction. In that case, one way to increase the holding voltage is to stretch the base width of two BJTs or lower the doping concentration, however, this will reduce the current gains of PNP and increase the junction reverse breakdown voltage, or the trigger voltage. If the gains are lowered too much, the condition $\beta_n\beta_p \geq 1$ may not meet, the SCR will not turn-on into a positive feedback.

Due to the positive feedback and deep snapback feature, SCRs usually possess the feature of high trigger voltage and very low holding voltage. Although this will provide very high ESD current capability and very low on-resistance, the I-V curve of SCR may not meet the ESD design window. To improve this, many modifications are made to SCRs.

1.3 Dissertation outline

Different technologies and applications require customized ESD consideration at the early stage of design and development. The goal of this research is to develop compact models for circuit level ESD simulation and investigate ESD failure related issues. This dissertation is summarized as following:

Chapter 1 presents some of related the background information, from the ESD event, ESD test models and test method. Since this dissertation focuses on ESD models, ESD failure and failure model related topics are. Chapter 2 gives one approach of ESD failure model and model parameter extraction method. This modeling methodology uses temperature instead of voltage or current values to predict device failure due to thermal energy. Chapter 3 takes a deep look into the physical mechanism behind device thermal failure, and proposed a new method to estimate the failure temperature for typical ESD devices. Chapter 4 reports a comprehensive diode model for circuit level ESD simulation. ESD related issues like overshoot, on-resistance variation and in-device temperature increment during stress, are included in this model. Chapter 5 investigate the ESD performance of vertical Gallium nitride (GaN) diode in forward conducting. The experiment results show interesting turn-on delay phenomenon for vertical GaN diode, especially under short duration pulses.

CHAPTER 2 JUNCTION THERMAL FAILURE MODEL FOR DEVICES SUBJECT TO ELECTROSTATIC DISCHARGE STRESSES

2.1 Introduction

As the ICs is highly integrated with shrinking transistor size, the silicon ICs are becoming more vulnerable to ESD stresses, thus, simulating ESD event and finding potential reliability defects prior to fabrication is highly valuable for developing and improving IC designs. Since ESD stresses tend to push devices to operate beyond their normal region of operation, conventional models are incompetent to support simulations in the ESD current-voltage regime. Moreover, as ESD stress may push devices to their final failure point, failure prediction models are also required when simulating for ESD Classification test levels.

Compact models are established based on physical mechanism, and there are two major ESD-induced damages responsible for the catastrophic failure[43, 44]; dielectric breakdown, which is electric field-related [45], and thermal failure, which is energy dissipation/temperature-related [46-48]. Dielectric breakdown typically occurs at the gate of MOS devices, and is getting increasingly important due to the aggressive scaling of MOS-gate dimensions as well as dielectric thicknesses in advanced CMOS processes [49]. On the other hand, thermal failure is more common as current flows through any layer or junction in the IC and continuously heats the structure.

Thermal failures have been reported in the literature and can be simulated in many ways [46-52]. It is a delicate way using TCAD software to solve the differential heat conduction equations numerically and predict thermal conduction in devices. Such method, however, are

limited by the complexity when simulating a small circuit performed in mixed-mode. They are incompatible with parameter-based compact circuit models.

The thermal failure of a semiconductor component is evaluated by the power-to-failure (P_f) within a stress pulse time (t_f), assuming the stress power is constant. Reports in the literature [46, 47] established physical electro-thermal models to describe the P_f as a function of pulse duration. Wunsch-Bell first described the thermal failure process as linear curves with time dependency of $t^{-1/2}$ in a log timescale plot using one dimensional analysis [46]. Later, a wider pulse time spectrum was expanded and divided into different time ranges, describing the observed dependence of P_f with respect to t_f in experiments [47, 48]. Dwyer's work [48] extended the Wunsch-Bell model into a larger time-scale range in which the three dimensional thermal diffusion is divided into several time regions using the time scales t_a, t_b, and t_c. As the curve depicted in Figure 2-1, the first portion of the curve is the adiabatic region, typically within short time of 10ns. It has a time dependence of $1/t$. From tens of nanoseconds to microseconds, it is the Wunsch-Bell region where the plot fits a $1/t^{1/2}$ slope. For a very long stress time, typically beyond 100 μs, the power enters a steady state as the thermal equilibrium has been established.

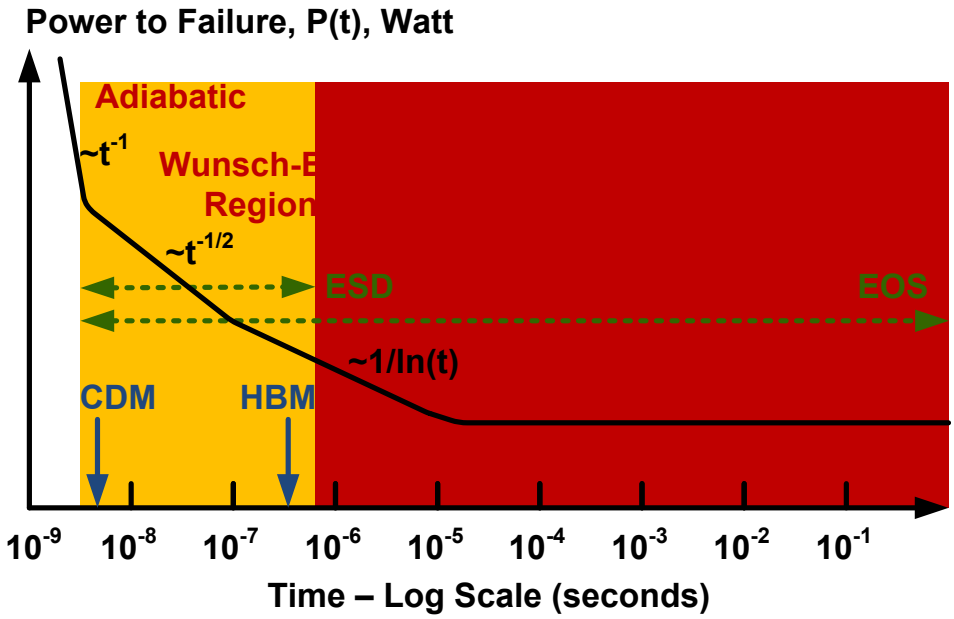


Figure 2-1 Input power-to-failure as a function of electrical pulse width, for a constant ΔT, across a reverse-biased junction.

Other Circuit-level simulation of thermal failure in ESD events has been proposed using different approaches. Some [51, 52] used the above described power-to-failure or Wunsch-Bell relations to predict failure in rectangle pulses with constant power. An approach has been using thermal networks with infinite stages, based on the differential heat diffusion equation [50]. This necessitated elaborated program coding to incorporate them in circuit simulators, which limited its adoption. These methods assumes thermal failure under constant stress, but the reality is that ESD stresses are characterized by electrical waveforms with irregular shapes. Others try to use the average power of an ESD pulse [53] or exponentially decaying current profile [54] to represent a given ESD stress, such as the HBM, however, such methods indicate we must establish different failure models according to different ESD stress model, such as Charged

CDM or HMM. Additionally, the input stress waveform may be distorted after passing through various circuitry interfaces. To address these shortcomings in the successful adoption of thermal failure models for large scale circuit simulation, defining a modeling framework that can be implemented in SPICE and applicable in simulations using arbitrary ESD stress conditions urgently needed [55].

2.2 Modeling and Characterization

A sub-circuit compact model for device failure under ESD stress is proposed. The model is designed to be fully compatible with SPICE circuit simulation tools. For a given ESD stress, it reproduces the thermal behavior in the device by simulating temperature variations in discrete stages. Model parameters are extracted using measurement data from a High Power Pulse Instrument (HPPI) tester which generates pulses with widths ranging from 5ns to 500 ns. The p-n junction structures are first picked up to establish the model, since they are the most fundamental building blocks in many devices. Once model parameters are extracted, temperature variation can be simulated within the sub-circuit model. The compact model is verified by comparing the measured and calculated failure power with respect to failure time results.

Most thermal failures occurring in semiconductors are related to current filament that pushes lattice temperature to a critical level. As a fundamental structure in the CMOS process, p-n junctions function as a trigger element of most passive ESD protection devices and across which, the majority of ESD stress drops on it. Thus, it is susceptible to ESD damage when the discharged energy flows through the device.

Figure 2-2 shows the TCAD simulation of temperature distribution in a reverse biased p-n diode at the end of an ESD stress. When the pulse voltage exceeds the junction reverse breakdown voltage, leakage current increases exponentially and heats the junction area continuously. The silicon region with the highest temperature is the hotspot and is located near the silicon surface [56]. Due to the competition between heat generation and heat dissipation, heat id generated at hotspot and diffuses into deep silicon. The temperature keeps rising until the end of the pulse stress or a thermal equilibrium state is reached. Whenever a critical temperature is reached at the hotspot during this period, the device fails.

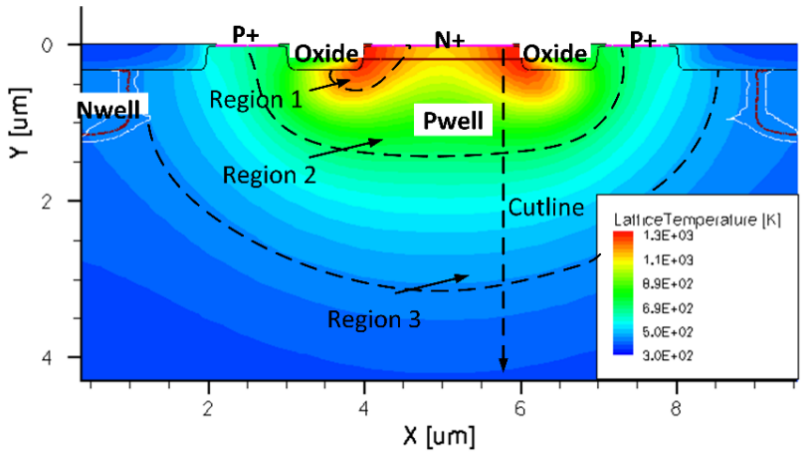


Figure 2-2 Simulated lattice temperature contours with three different thermal regions.

A cutline in Figure 2-2 is drawn from the hotspot near the surface of the device to deep silicon region, by combining temperature contours at different time frame, the time dependent temperature variation of a 100 ns pulse stress can be probed (Figure 2-3). The temperature at the hotspot region rises when the electrical pulse stress is applied. The region deep into the silicon

follows slowly due to the delay in the heat diffusion. After pulse ends in 100 ns, the highest temperature is recorded in the hotspot region. The temperature begins to drop thereafter.

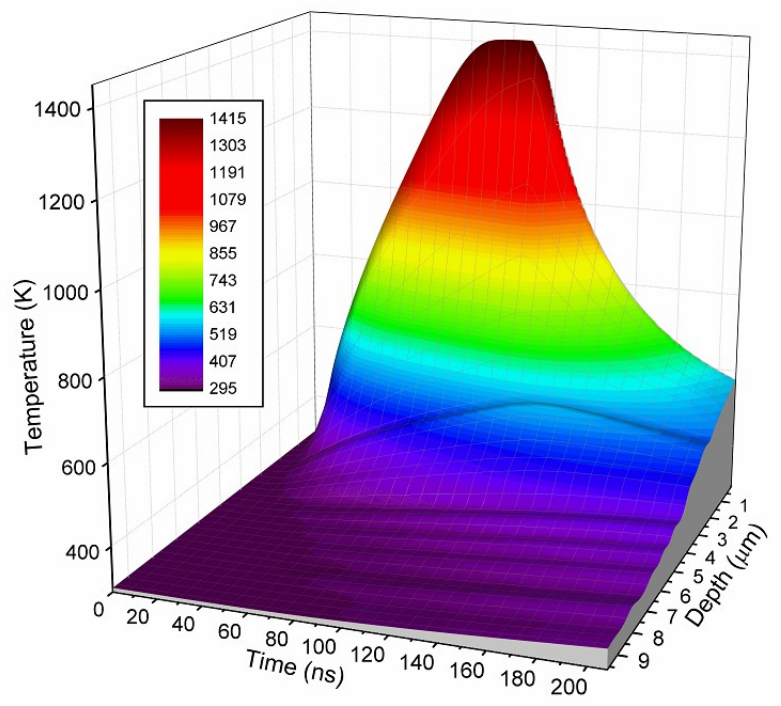


Figure 2-3 Time dependent temperature variation along the depth of junction structure under a transient pulse stress.

2.3 Model Parameter Extraction

According to time-dependent temperature variation depicted in Fig. 2-3, heat dissipates from the hotspot towards the silicon bulk, which contributes to a temperature gradient in device. Thus, the device may be partitioned into thermal regions with different temperature levels they represent. The junction region in Figure 2-2 is divided into three regions as the dashed lines indicate. The small one in red color is the hotspot region with the highest temperature,

surrounded by the Region-2 with a lower temperature. Between the Region-2 and the rest of the bulk material is Region-3, in which the temperature gradually degrades until the ambient temperature is reached.

A Cauer ladder thermal equivalent circuit has been used to model device self-heating [57]. Based on the analogy between electrical and thermal conductions [58, 59], the thermal behavior can be reproduced as an electrical circuit. The electrical equivalent of power and temperature are current and voltage, respectively. The thermal capacity is similar to a capacitor in an electrical circuit, and the efficiency of heat dissipation is measured by thermal resistors. Thus, the larger the thermal capacitor, the more thermal energy it can store, and the smaller thermal resistor, the faster heat transfers between adjacent regions.

The three stage circuit shown in Figure 2-4 has been used as the thermal failure model in this study. The current source P_0 represents the input ESD power that generates heat in unit of watt. The capacitor and resistor pairs represent the volume of thermal energy and heat dissipation efficiency in the thermal regions as shown in Figure 2-2. The hotspot region is equivalent to the C1 capacitor in the thermal circuit. Since the capacitors represent the region from the hotspot (small volume) to deep bulk inside the silicon (large volume), the boundary condition for the capacitor values must be $C1 < C2 < C3$.

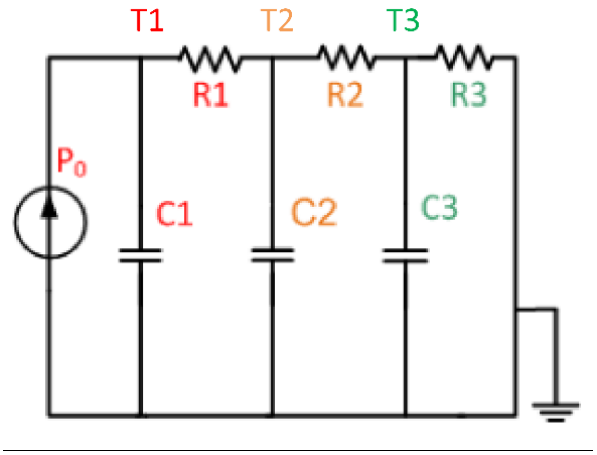


Figure 2-4 Equivalent circuit for the compact junction thermal failure model.

There is always heat dissipation as long as there is temperature differential between two adjacent regions. The values of the resistors define how fast heat can be transferred from inner region to outer silicon. A small resistances indicate fast heat dissipation. It is obvious that a model with more stages can achieve higher accuracy, to balance between accuracy and complexity, we choose a three-stage model here. Given the values of the R's and the C's, the thermal behavior in a p-n junction region can be mapped into temperature variation in the model with discrete stages,

The parameters of the thermal circuit are extracted from experimental measurement under different pulse stresses. Two types of p-n diodes, namely P+/Nwell and N+/Pwell, and one GGNMOS are used for the extraction. The diodes' dimensions are 25 μm × 25 μm square shape and 10 μm ×100 μm rectangle shape for both n-type and p-type diodes. As for the ggNMOS, it is designed as multi-finger device with 200 μm total width. The characterization is carried out by an HPPI transmission line pulse (TLP) tester with 4-pin probe Kelvin configuration. A high

frequency oscilloscope at the back-end of data sampling is used to capture the reflected waveforms. Six pulse widths of 5, 10, 50, 100, 200, and 500 ns have been used to characterize the device.

This time range covers most of the ESD stresses of interest. The TLP tester stresses the device with increasing pulse levels until hard failure occurs. Failure is identified as the current (It2) and voltage (Vt2) where the voltage collapses or the current increases exponentially. Six pulse widths indicate there will be six thermal failure points obtained for each device tested. The product of the Vt2 and It2 yields the power-to-failure (P_f), as shown in Figure 2-5. Six pulse widths yield six P_f waveforms accordingly. Table 2-1 lists the extracted P_f values for model fitting. A square-like waveform approximation is used in this case, for both, forward and reverse bias.

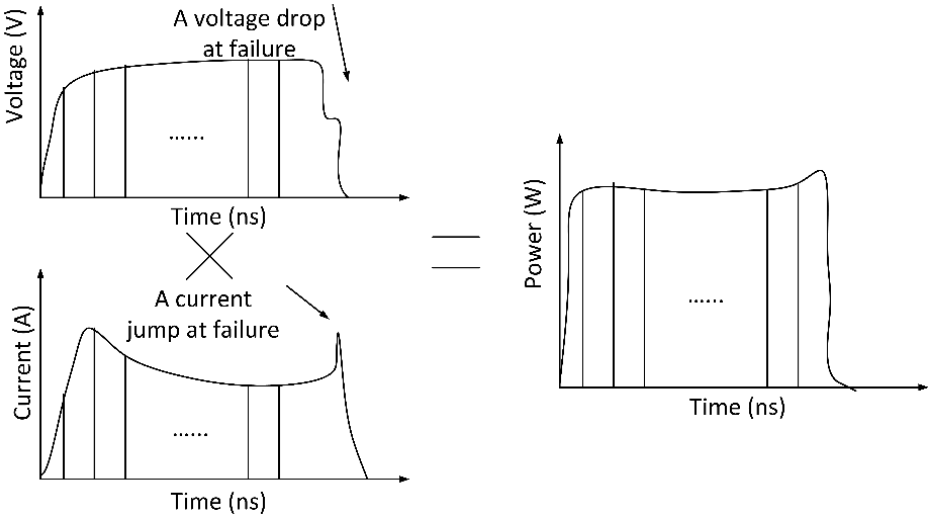


Figure 2-5 Failure power vs. time waveform as the product of Vt2 and It2 waveforms

Table 2-1 MEASURED FAILURE POWER FOR MODEL FITTING

P_f(W)&t(ns)		5	10	50	100	200	500
Reverse	Square	7.50	6.74	5.01	4.35	3.96	3.63
	Rectangle	15.75	15.05	10.71	9.22	8.43	7.48
Forward	Square	24.15	17.53	11.92	9.97	7.78	6.01
	Rectangle	22.32	15.69	11.77	8.56	6.25	4.69
ggNMOS	200μm	83.92	51.96	27.86	21.8	18.9	13.9

The next step in the modeling process is to extract the parameters of components in the equivalent circuit. First, P_f functions as the initial power P_0 in Figure 2-4. With the boundary conditions of P_0 and pulse width, the values of the R's and the C's were extracted by using the fitting tool like ADS to fit the end pulse temperature on C1 to correspond to a fail temperature, T_{fail}. Notice that there is still no conclusive T_{fail} value for thermal failure temperature in silicon, users can define any value as the T_{fail}, but the extracted R and C parameters will change accordingly. Once T_{fail} is reached on C1 in the simulation, the device is permanently damaged marked as the device hard failure.

Assuming that during the stress, device temperature increment is 1,000 from background temperature of 0 and ends at T_{fail}=1,000, the temperature variation in the device can be simulated as three discrete temperatures on thermal capacitors in Figure 2-4.

Figure 2-6 is one example of the fitting setup to extract R and C values using ADS software. An optimization goal is setup that the temperature increment under six different pulse width stresses reaches a consistent T_{fail}=1,000, then the R and C values right for the optimization, namely the parameters for the thermal failure model. To avoid false parameters being extracted, the optimization confinement must be set as C1<C2<C3.

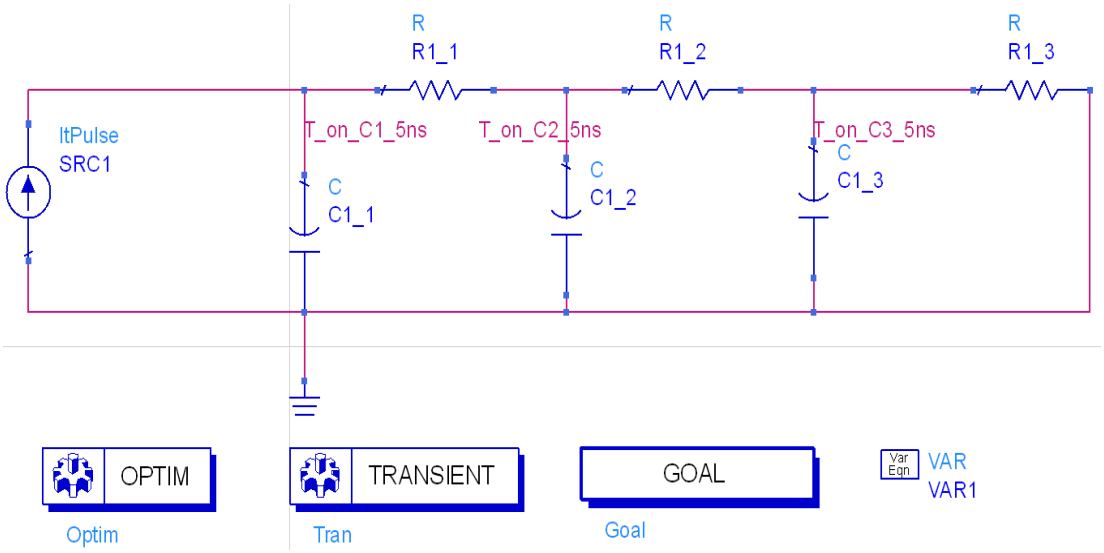


Figure 2-6 ADS software tool setup for parameters extraction

Table 2-1 gives extracted R and C values for the thermal circuit model. Parameters for both forward and reverse directions are extracted. With these parameters, temperature variation during the pulse stress is reproduced as the discrete temperature curves on the three thermal stages. Figure 2-7 shows one such example of a 25 μm × 25 μm square shape diode and its temperature simulation curve.

Table 2-2 EXTRACTED R AND C VALUES

R(Ω)&C(nF)		C1	R1	C2	R2	C3	R3
Reverse	Square	0.0107	131	0.404	108	7.203	71
	Rectangle	0.003	56	0.780	53	7.394	29
Forward	Square	0.072	55	0.498	23	1.828	99
	Rectangle	0.062	64	1.396	176	4.014	13
ggNMOS	200μm	0.32	29.6	3.051	20.4	10.76	57

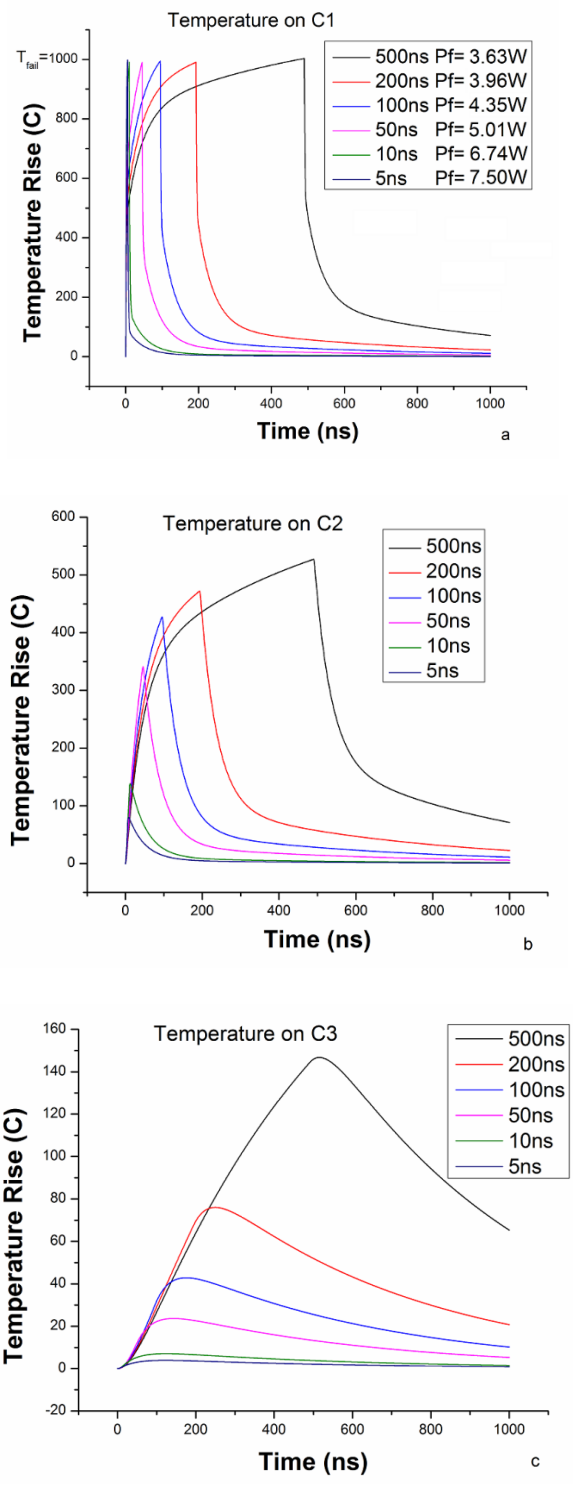


Figure 2-7 Simulated temperature vs, time curves in the three thermal regions of the square-shape diode subject to TLP pulses having different pulse widths and failure power.

Although only the discrete temperature distribution in the silicon region is simulated, it still reflects the correct physical mechanisms within the device. Just as the assumption of the thermal failure, for high power stress, thermal energy is generated fast and dissipation is slow, thus failure temperature is reached in a very short period of time. While during a long pulse with much lower power levels, the temperature rises slowly, leading to a very long period of time before failure temperature is reached.

Figure 2-7 (b) and Figure 2-7(c) also show the simulated temperature variation curves on the C2 and C3 capacitors, which verify the assumption of the thermal failure. They represent the rest of silicon region adjacent to background bulk region. As heat diffuse into deep silicon, the temperature decays gradually from Region-1 to Region-3. For short pulse failure, when C1 reaches the failure temperature, there is little thermal energy dissipate into Region-2 and Region-3, thus temperature increment is very low in these two regions. However, when device fails under long pulse, there is enough time for heat to dissipate into deep silicon, heating Region-2 and Region-3 to higher temperatures. These simulated results correlate well with the numerical TCAD simulations shown in Figure 2-2.

2.4 Model Verification

In the previous mentioned modeling framework, the thermal failure in silicon was evaluated as power-to-failure at a failure time, and when silicon device fails, a collapse in the voltage and an increase in the current transient waveforms can be observed. Based on such an assumption, additional P_f vs. failure time data can be obtained by using different pulse duration

with different power level. These test results can be used for model verification to complement the six fixed pulse widths discussed earlier.

Assuming stressing the same device with higher power level but longer pulse width, we can observe thermal failure will occurs at the mid of a long pulse as voltage waveform drop off or current waveform jump up. In Figure 2-8, the continuous curves are predicting curve of P_f vs. failure time (red curve) and failure energy vs. failure time (blue curve) obtained by the model simulation in previous section. Red dots and blue dots are measured P_f vs. failure time and failure energy vs. failure time data points, respectively. It can be seen that all the measured failure points correlate with the simulated points, the shorter the pulse the higher the failure power. It should be noted that P_f decreases exponentially as pulse width increases, which conforms to the exponential decreasing trend in Figure 2-1.

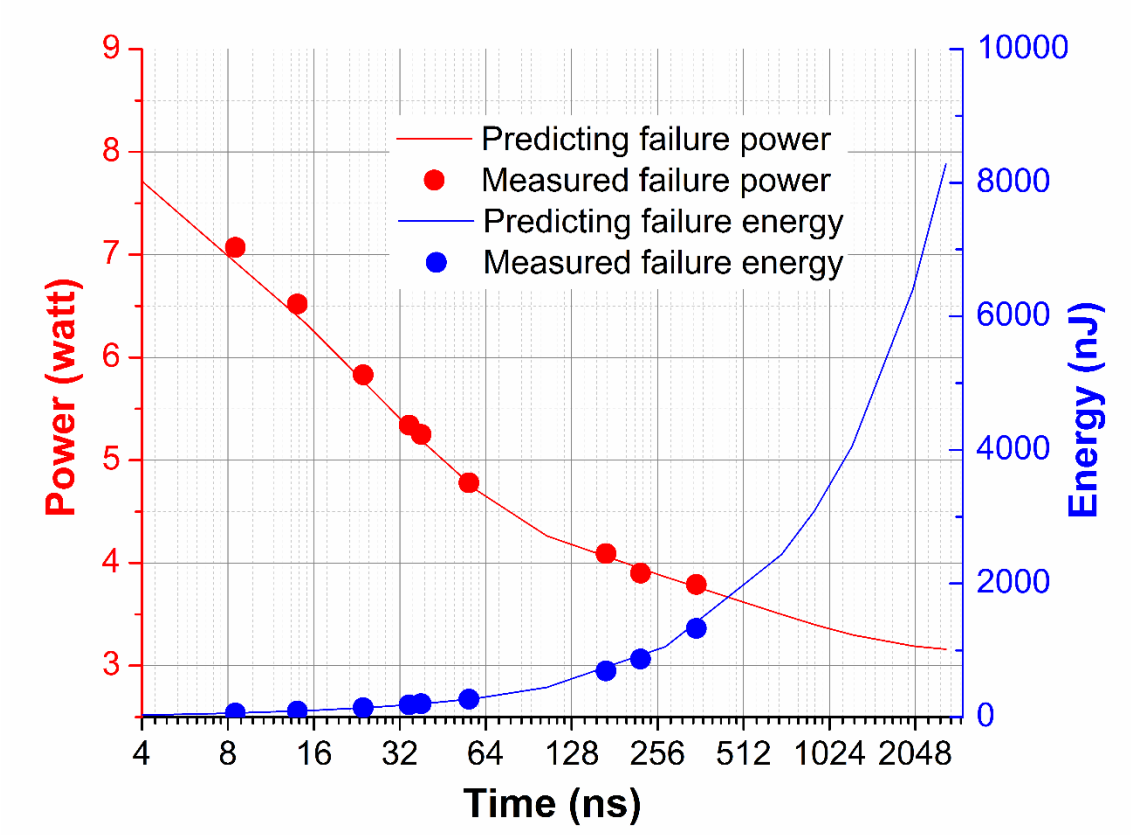


Figure 2-8 Comparison of simulated and measured failure power/energy vs. failure time data for a rectangle shape diode under different pulse widths.

The thermal failure model has been implemented using Verilog-A as a failure monitor in a SPICE-type simulator. The monitor uses the current and voltage, the product of which is used to calculate the injected instantaneous power, and it is fed into the thermal equivalent circuit. The monitor gives the temperature rise as its output. A temperature controlled switch is added in parallel with the diode model. The switch is open under normal operation conditions, and once the temperature at C1 reaches the critical value T_{fail}, the switch will become closed. When the

ESD stress causes the diode to fail a low resistance path between the diode anode and cathode is established indicative of a failure.

In that case, we can apply the thermal failure monitor in circuit level ESD simulation. Figure 2-9 demonstrates the simulation results, as well as the measurement data, for the 25 μm × 25 μm diode during a TLP pulse that caused device failure. The ESD stress source is a rectangular pulse with 50 Ω impedance. The pulse width of this specific stress is 50 ns.

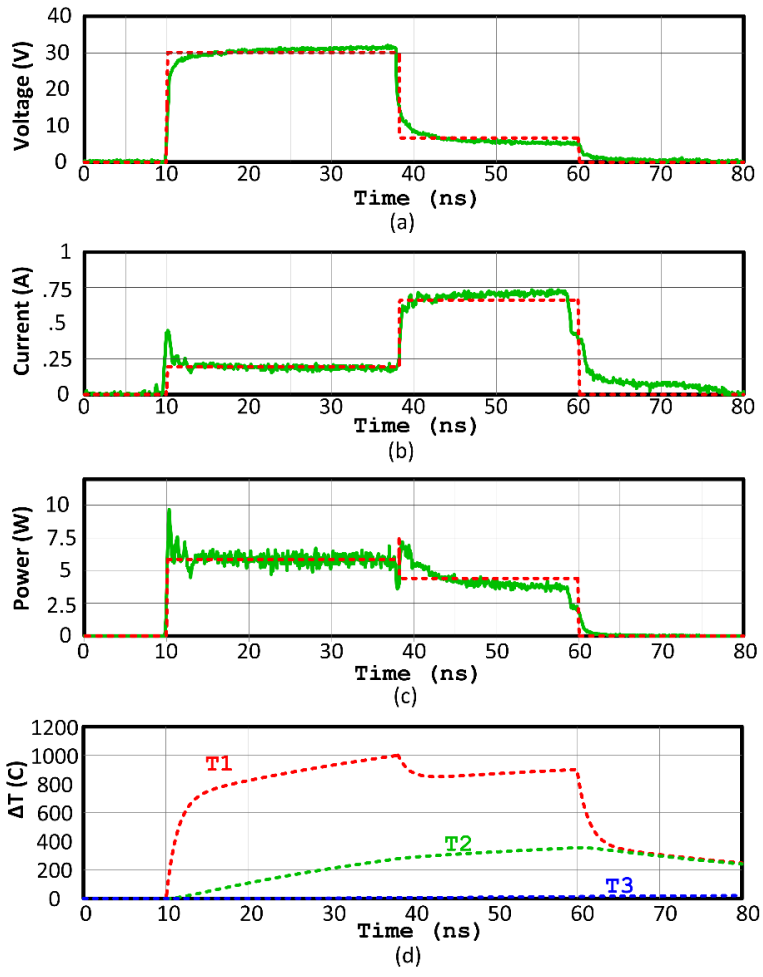


Figure 2-9 Simulation results using a TLP pulse that caused device failure (the solid lines for the measurement and the dash lines for the simulation).

The diode uses the standard SPICE diode model. The measured data showed that when hard failure occurs, the voltage dropped significantly in the middle of the stress pulse and the current became several times higher at the same time point. The power, derived from the product of current and voltage, decreased approximately 25%. The leakage current measured before and after the pulse stress indicated that the device had failed during the stress and the voltage drop and current rise were obviously the result of the failure. The simulation matches the measured voltage, the measured current and the measured power very well (curves in green color). The simulated temperatures are plotted in Figure 2-9(d) from which we can see the temperature at C1 reaching the critical failure point of T_{fail}=1,000 at around 37 ns. Also, we can observe temperature variation in deep silicon regions, represented by C2 and C3, with much lower temperature increment. This thermal failure model has no restriction on stress power profile and can be applied to more complicated transient stress waveforms.

Figure 2-10 shows the SPICE simulation results under HBM stress. The waveform changes immediately before and after device failure for the same diode shown in Figure 2-9. The measurement data were obtained using an HPPI TLP tester. The simulated temperature at C1 was also plotted in Figure 2-10(b). Once failure temperature is reached during simulated stress, the device turns into a short circuit configuration and the power has an abrupt change. Compared to the measured waveform, this simulation accurately predicted the device failure under transient HBM stress. It can be seen that the temperature rises very quickly at the rising edge of the stress pulse and keeps rising even after the power reaches its maximum.

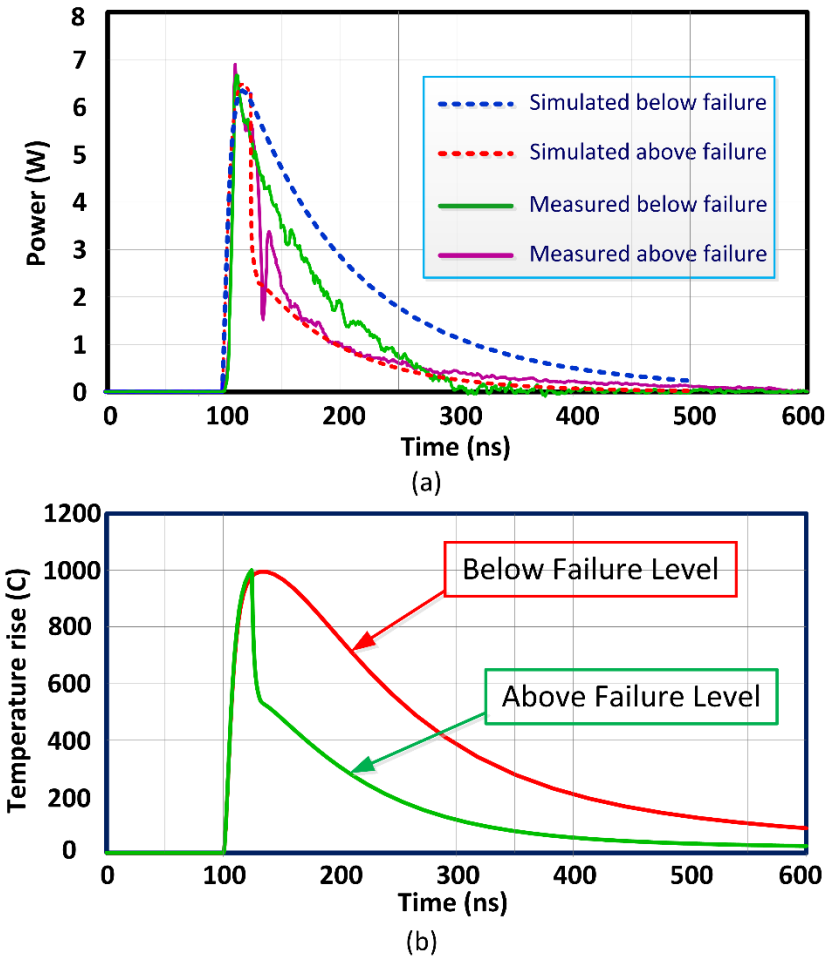


Figure 2-10 HBM power waveforms from measurement and simulation and the corresponding simulated temperature at the stress levels close to device failure.

Functions as a core model only for monitoring temperature rise and thermal failure, this model has no explicit geometric parameters. In other words, it is not a scalable model yet. However, the geometry effects are actually included and embedded in the device models as R and C values. Extracted R and C values will changed with different junction geometries. Combined with existing electrical models for simulation under ESD stress [60], this model can be used as a thermal failure monitor in ESD event simulation for any devices, according to the

different device types or dimensions, the extracted model parameters will change accordingly as proper capacitance and resistance values.

As aforementioned, the thermal failure model can be applied to different device as long as the device's temperature rises under an ESD stress and fails due to thermal failure. Figure 2-11 shows an example of a GGNMOS with silicide block. After the GGNMOS trigger and snapbacks into a low resistance state, ESD current flows through the reversed p-n junction at the drain-body interface and continuously heat the device. With the diode being replaced by a 200 μm GGNMOS model, the voltage and current simulated are convert into stress power. As failure temperature reached, a secondary snapback phenomenon can be observed as device fails and turns into a short circuit state. The NMOS model includes the core MOS component and the parasitic bipolar NPN device [60]. The measured I-V curve was obtained using 100 ns TLP and 10 ns vfTLP measurement, respectively.

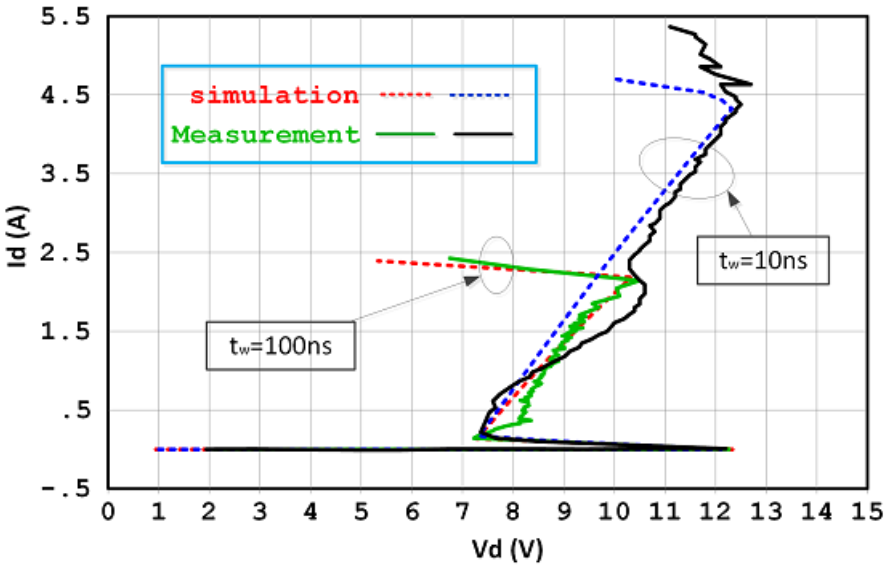


Figure 2-11 Simulation of secondary snapback in a GGNMOS with combination of the thermal failure model and ESD capable MOS device model.

2.5 Conclusion

A compact model designed for predicting the device thermal failure under the ESD stress has been proposed. The model is derived from the physical thermal behavior in device during an ESD stress. Device geometry is accounted for in the effective parameter extraction procedure. It can be implemented SPICE like simulator using Verilog-A, making it compatible with industry-standard circuit-level simulators. Simulations by this model were found consistent with the physics and in good agreement with the measurement. Under rectangular TLP pulse and transient HBM stresses the accuracy of the model simulation is verified. The model can be used as a thermal failure monitor, along with ESD capable compact models of devices, to predict thermal failure in circuit-level simulation for classical ESD events such as HBM, CDM, and HMM.

CHAPTER 3 INVESTIGATION OF FAILURE TEMPERATURES OF SEMICONDUCTOR DEVICES UNDER ELECTROSTATIC DISCHARGE STRESSES

3.1 Introduction

Thermal runaway is one of the major mechanism responsible for semiconductor device's irreversible damage. In-device temperature under test rises in-situ during an electrical stress due to inevitable self-heating. The competition between heat generation and heat dissipation will determine whether the device fails due to reaching a critical failure temperature or entering a thermal equilibrium state. As a common failure criteria, failure voltage and/or current values are use as ESD design parameter, however, extracting the thermal failure temperature gives better physical insight into the degradation of the device. Moreover, the failure criterion of temperature can be potentially valuable for ESD simulation under arbitrary stress waveform. That doesn't mean the two sets of parameters are mutually exclusive, and actually failure current/voltage as well as failure temperature can complement each other when performing ESD simulations [55].

Although with such pros, there is difficulty to apply temperature as a failure criterion, for direct measurement of the failure temperature in a device during an ESD event is very challenging, or even impossible, as the self-heating by an ESD event may be too short to be captured by a thermal detector. Moreover, little thermal energy may escape from the device's surface as most of the heat diffuses into the bulk [56]. This makes a direct observation of temperature very difficult.

On the other hand, different temperature values have been used as boundary conditions for thermal failure analysis, and there is still dispute on which as a failure criterion. It is

commonly accepted that devices fail when the temperature reaches the silicon melting temperature (~1,414 °C). While it has been argued that the device damage threshold is much lower due to the temperature where the intrinsic carrier concentration is equal to the background doping concentration [61], or known as the intrinsic temperature. It was also reported that the onset of thermal breakdown occurs when the thermally generated free carriers become significant in comparison with the avalanche generated free carriers [62]. Besides the junction region through which most current flows, failures may also take place at the silicon-aluminum interface where the eutectic temperature is approximately 550°C. However, none of the failure mechanisms have been decisively verified by experiments. And it is highly possible that failure temperature varies according to device types and stress conditions.

To determine the critical temperature responsible for an ESD-induced thermal failure, an experimental procedure based on the electrical-thermal analogy is proposed in this paper [63]. Such a methodology enables the extraction of the failure temperature for typical device types and, consequently, the determination of the proper failure criterion for the purpose of ESD event simulations. Physical insights into the extracted failure temperature with respect to stress time are also illustrated and discussed.

3.2 Experimental Methodology

Device thermal behavior under ESD stress is difficult to observe directly, and it is highly dependent on stress power and duration. With shorter stress duration, the visibility of thermal emission decreases. TCAD simulations provides one solution to look into the thermal flow in semiconductors. Figure 3-1depicts the thermal profile in a diode stressed by short and long

electrical pulses. By adjusting the stress levels of a 10ns pulse and a 200ns pulse, the same maximum lattice temperature can be achieved at the end of each pulse. Their temperature distribution, however, will be very different. Under a short pulse, the thermal region barely expands, while for the longer pulse the peak temperature is reached with a much bigger corresponding thermal region.

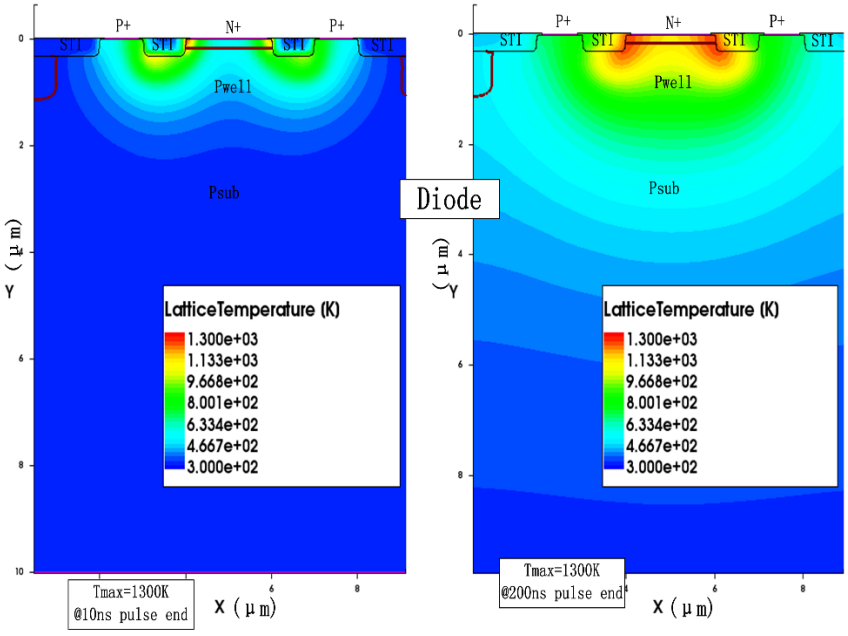


Figure 3-1 Simulated temperature distributions in a diode under a short (10 ns) (left) and a long (200 ns) (right) electrical pulses.

Thermal failure in semiconductor devices can be described by the power-to-failure (Pf) versus time (t) relationship

$$P_f = f\left(T_f - T_0, t^n\right) \tag{3.1}$$

where T_0 is the ambient temperature and also the starting temperature, n is the time index depending on the time range of the ESD event, and T_f is the failure temperature. The equation

can be expressed in different forms depending on the time duration which ranges from the adiabatic region to thermal equilibrium state [48].

Since there is an analogy between heat conduction and electrical conduction [58], the time-dependent thermal diffusion and the expansion of the thermal region in a semiconductor device can be mapped into an equivalent circuit [55, 64], as shown in Figure 3-2. The voltage across the circuit corresponds to the temperature increase in the device, moreover, since there is only one stage in the equivalent circuit, we are assuming a uniform heating of the material.

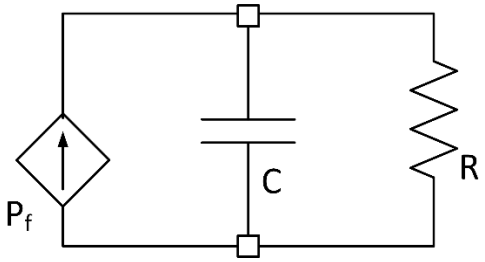


Figure 3-2 One-stage thermal equivalent circuit. The voltage on capacitor C is equivalent to the average temperature in the region under stress.

During a transient stress, heat is generated in the device and diffuses from the hotspot to the deep silicon, namely the heat-sink. When a constant power P is applied at time t=0, the temperature change can be expressed analytically as

$$T - T_0 = P \cdot R \cdot \left(1 - e^{\frac{-t}{RC}}\right) \tag{3.2}$$

where P is the power-to-failure, R stands for the thermal diffusion efficiency, C represents the thermal capacity of the thermal region, and t represents the duration of the stress. For the case where a device fails under a stress with a fixed pulse width but at different ambient temperatures

T_0, the temperature T reaches the failure temperature T_f and (2) can be simplified into a linear equation

$$T_f - T_0 = P_f \cdot B \tag{3.3}$$

where $B = R(1-e^{-t/RC})$ is a constant representing for a given pulse width t. In case of the linear form of the equation, it indicates that P_f is lower for a higher T_0 and if the starting temperature T_0 is high enough, stress power could be equal to zero. Therefore, T_f can be extracted from the intercept of the P_f vs. T_0 plot.

Note that there is only one stage in the equivalent circuit, thus we are assuming only one temperature representing the device temperature and the voltage across C in Figure 3-2 is a value representing the average temperature in the heated region. While this is not accurate for actual physical behavior behind thermal failure because the device does not heat uniformly due to the nature of thermal conduction.

A HPPI TLP tester with 4-pin Kelvin configuration was used for device characterization. The HPPI tester, generate a sequence of rectangular current pulses with increasing magnitude. The pulse widths range from 5 ns to 200 ns. Ambient temperatures from -50 °C (223 K) to 250°C (523 K) were used in the design of the experiment.

Figure 3-3 depicts an example of this extraction methodology. It is an N+/Pwell 25 μm × 25 μm diode. First, the reverse bias quasi-static I-V curves of an n-diode are measured at different ambient temperatures, as depicted in Figure 3-3(a). Since the amplitude of the stress pulse increases slowly, the power level of each pulse is just slightly higher than its previous pulse. In that case, once a pulse stresses the device, device always fails at the end of the pulse

duration. P_f is calculated using the last point before secondary breakdown under different pulse widths; namely 5 to 200 ns (see Figure 3-3(b)). Since five pulse widths are applied to the same device, five P_f values are obtained for each ambient temperature. Then plotting all the data dots of P_f in a P_f vs. temperature plot. If connect the data dots of P_f under the same pulse width but different starting ambient, and extend the line to intercept it with temperature axis, we can extrapolate corresponding T_f from the P_f versus temperature curve. The value of T_f also varies due to the expansion of the thermal region with the pulse width.

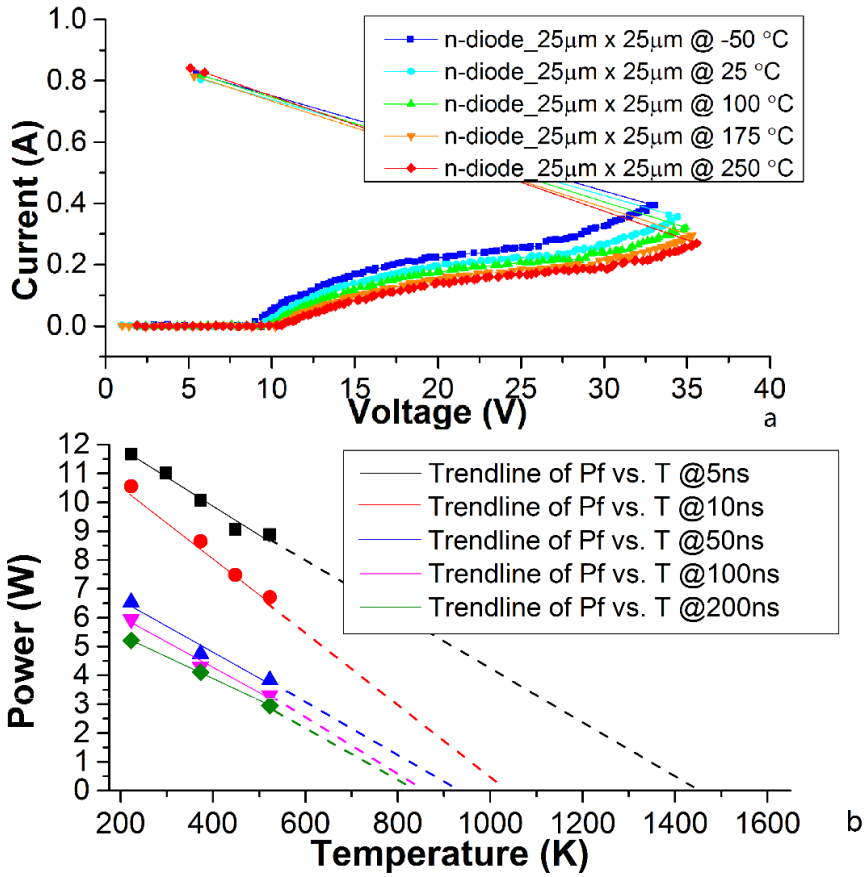


Figure 3-3 (a) I-V curves of a reverse-biased n-diode stressed under 5ns pulses at different ambient temperatures, and (b) failure power vs. temperature lines and extracted failure temperatures (intercept points on the x-axis) for the same diode

3.3 Measurement Results and Analysis

Four different diodes are chosen to testify this method. Applying the same method and extracting T_f vs. stress time (i.e., TLP pulse duration), the results are shown in Figure 3-5. The extracted T_f reaches its peaks at about 1,600 K for the shortest pulse, then with a fast decay temperature lowered to a near constant temperature of 900 K as the pulse duration is increased beyond 50 ns. The decreasing trend of extracted T_f is one proof of the expansion of the thermal region. As shown in Figure 3-4(a), for the thermal failure under a short pulse, failure occurs very quickly, thus there is no time for heat to dissipate into deep silicon. In other words, heat is confined within a small hotspot, with relatively uniform temperature distribution. Whereas for the case of a long pulse (Figure 3-4(b)), the thermal region expands after considerable thermal buildup, and the extracted temperature decreases due to spatial distribution of the temperature beyond the hotspot. In that case, the extracted T_f under the shortest pulse of 5 ns is high and uniform which is closest to the temperature of the hotspot, and therefore is closest to the real failure temperature.

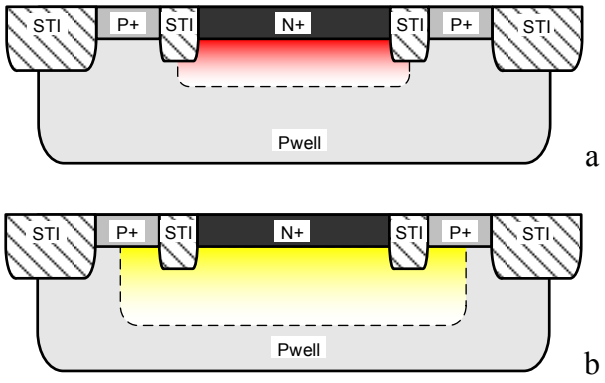


Figure 3-4 Difference of average temperature under short (a) and long (b) pulse stress

The extracted T_f under the 5 ns pulse is much higher than the intrinsic temperature (T_i) of 1,000 K, calculated based on the background doping concentration of 10^{18}/cm3 and conventionally deemed as the temperature responsible for the thermal failure. However, such a high failure temperature corresponds to the finding reported in [62, 65] that, at high current levels, secondary breakdown occurs once heat produces enough minority carriers to support the increasing current without raising the electric field. Although both n-diode and p-diode show similar exponential decreasing trends of the extracted T_f. The trend is somewhat different these two types of diodes because of the different free carrier types. The flat curve seen beyond 50 ns indicates that the expansion of thermal region is slowing down and the temperature distributions in the heated core are basically the same regardless of the pulse widths.

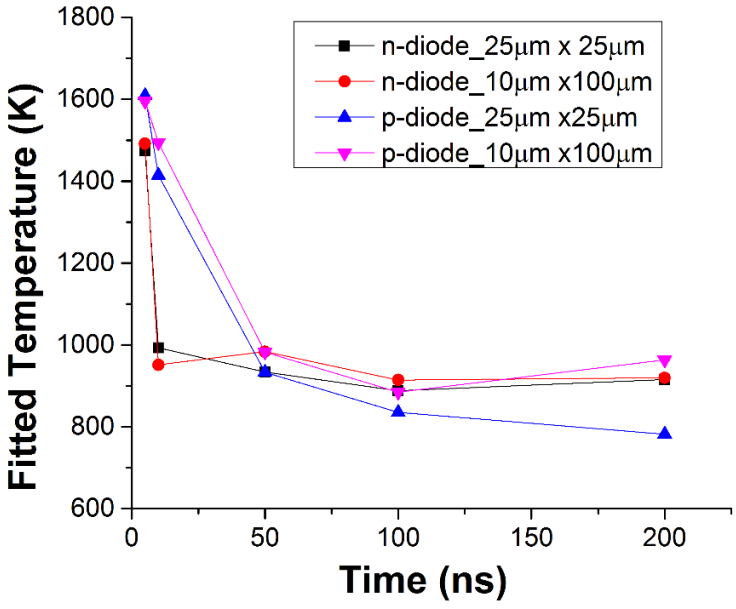


Figure 3-5 Effective failure temperature vs. pulse duration characteristics obtained for four different diodes: single-finger n-diode and p-diode with a square shape (25µm x 25µm) and a rectangular shape (10µm x 100µm).

Still, some may wonder the accuracy of this method, especially for assuming a uniform temperature distribution in the thermal region. To evaluate the accuracy of this estimation method and approach the real failure temperature in the right way, we need to discuss about where the error comes from. The accuracy of this estimation method is affected by two factors. The first is the assumption of uniform temperature distribution within the thermal region, as aforementioned. The effective failure temperature, as extracted based on the model used corresponds to the average temperature in that region, instead of a temperature distribution with gradient. Thus, the effective failure temperature represents all the thermal energy within the thermal region and is always below the highest temperature at the hotspot, i.e., the true failure temperature. In this way, it implies that the method will underestimates the true failure temperature because average temperature is always lower than the highest temperature in it. The second factor is the constant thermal capacity and conductivity used in the model. As we known, thermal capacity and conductivity are not constant during a temperature variation event, but temperature dependent. To estimate the error of our method using a constant C, we first use the re-written equation (1) to calculate the temperature increase ΔT due to power:

$$\Delta T = P \cdot R \cdot \left(1 - e^{-\frac{t}{RC}}\right) \tag{3.4}$$

Using the Taylor series, this equation can be expanded to

$$\Delta T = P \cdot R \cdot [1 - (1 + \frac{-\frac{1}{RC}}{1!} \cdot t + \frac{\left(-\frac{1}{RC}\right)^2}{2!} \cdot t^2 + \frac{\left(-\frac{1}{RC}\right)^3}{3!} \cdot t^3 + \cdots)] \tag{3.5}$$

Considering the first two terms of Taylor series, omitting high order terms, equation (3.5) can be simplified as

$$\Delta T \cdot C = P \cdot t \tag{3.6}$$

This equation only contains C, and thermal resistivity (conductivity) is omit with higher order terms. It indicates that the temperature rise is only related to the thermal capacitor C. Such observation agrees with the Wunsch-Bell theory described in [46] and [48], that when stress during is very short, thermal process is only related to the thermal capacity. In other words, device's thermal behavior is an adiabatic process under the very short pulse.

Now we only need to consider a changing thermal capacity. The Figure 3-6 below shows an empirical curve of C vs. temperature. According to equation (3.7), the integral area between ambient temperature (for example, 223K, 373K and 523K) should be proportional to the measured failure energy with a constant α:

$$\alpha \cdot \int_{ambient\ temperature}^{failure\ temperature} C(T) \cdot dT = P \cdot t \tag{3.7}$$

Using the integral area and measured failure power, one can get the effective failure temperature.

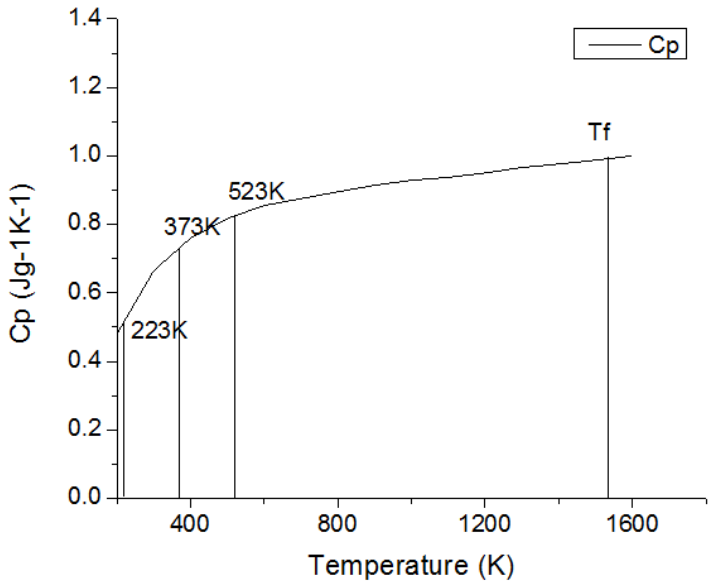


Figure 3-6 Temperature dependence of thermal capacity

For example, in Figure 3-7, measured failure powers for an n-diode with an area of 25 μm × 25 μm under a 5-ns pulse are 11.6688 W @ 223 K, 10.0624 W @373 K and 8.8868 W @ 523 K. By locating the failure temperature and making the integral area proportional to measured failure power, the extracted failure temperature using this method is 1350 K, which is about 100 K below the linear fitting method we proposed. This discrepancy results from the non-constant *C* vs. temperature. However, the increase of *C* saturates quickly as temperature rises, making the increase of integral area at low temperature relatively insignificant compared to the entire integral area.

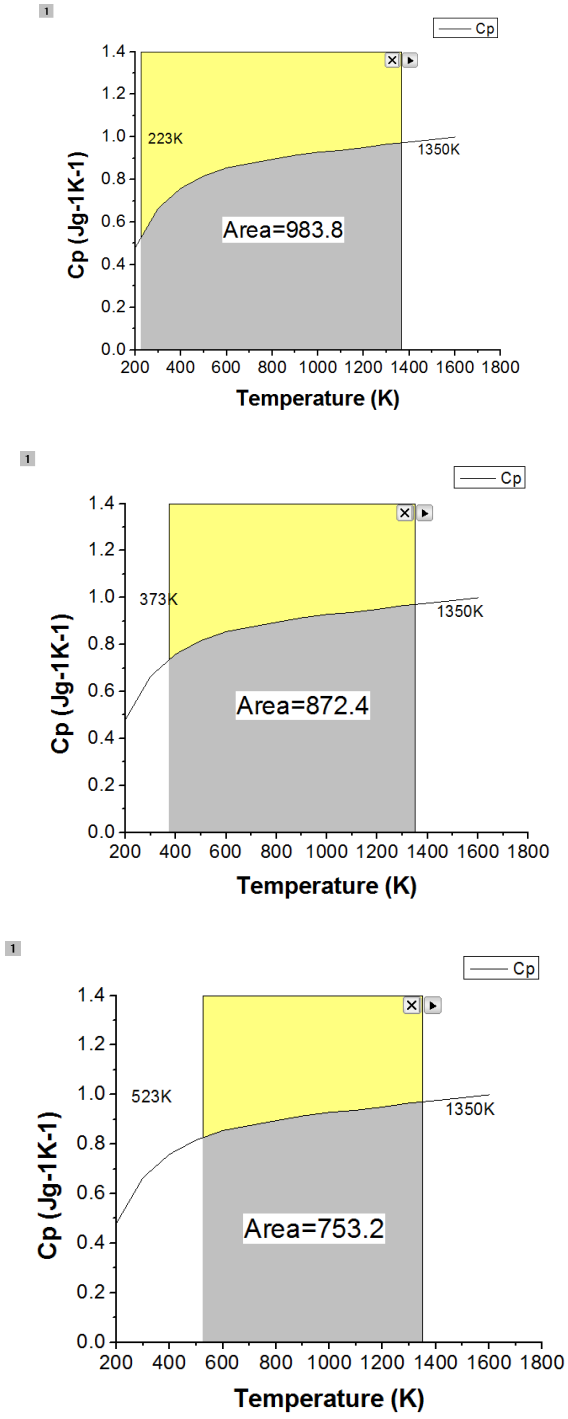


Figure 3-7 Finding failure temperature with a non-constant thermal capacitance

In that case, by assuming a constant thermal capacity *C* will overestimate the failure temperature. Consider the goal of finding approximation of failure temperature, we can assume the estimation under very short pulse, like 5 ns, is reliable, especially combine the overestimation with the underestimation by assuming a uniform thermal region. In other words, the two factors of this method somehow counteract each other, making the effective failure temperature not far from the real failure temperature, and is very close to the real value if only the short pulses data is used. For short pulses, heat dissipation is negligible and devices' thermal behavior is an adiabatic process [47]. In such cases, the temperature rise from a pulse stress depends only on the thermal capacity and is not affected by thermal conductivity. Considering these two opposite effects, it is safe to conclude that the effective failure temperatures obtained under very short pulses are close to the true failure temperatures and are well above the intrinsic temperatures.

Figure 3-9 shows the effective failure temperatures obtained from other types of devices. Two GGNMOS and two NPN bipolar transistors (BJTs). For the GGNMOS, the effective failure temperature starts from a high value of 1,500 K, drops to about 1,200 K and remains flat when the pulse duration is increased beyond 50 ns. This result also demonstrated that $T=T_i$ is not the thermal failure condition for the GGNMOS and BJT under ESD stresses.

Different device types also result in different thermal boundary conditions. The bottom two figures of Figure 3-8 shows the TCAD simulation of a GGNMOS, the temperature distributions under a short (10ns) and long (200ns) pulse are definitely different. Moreover, the thermal boundary conditions for GGNMOS are also quite different from diodes'. As we can see that after reaching a high temperature, heat doesn't spread as deeply as in diodes. Instead, heat spreading stops within a very small region near surface. This phenomenon is explained in [51], which

stems from the specific thermal boundary conditions of the NMOS, the thermal region of which is approximately a rectangular box underneath the gate and the drain sidewall defined by the channel width, channel length, and diffusion depth. Due to the small size of the thermal region, heat diffusion reaches the boundary of thermal region very quickly, resulting in a relatively higher average temperature value. For the diodes and BJTs, on the other hand, the thermal regions are relatively larger and hence the lower average temperatures. As a result, the stabilized temperature for GGNMOS is noticeably higher than those of the diodes and BJTs, as an indirect evidence for the confined thermal boundary for GGNMOS.

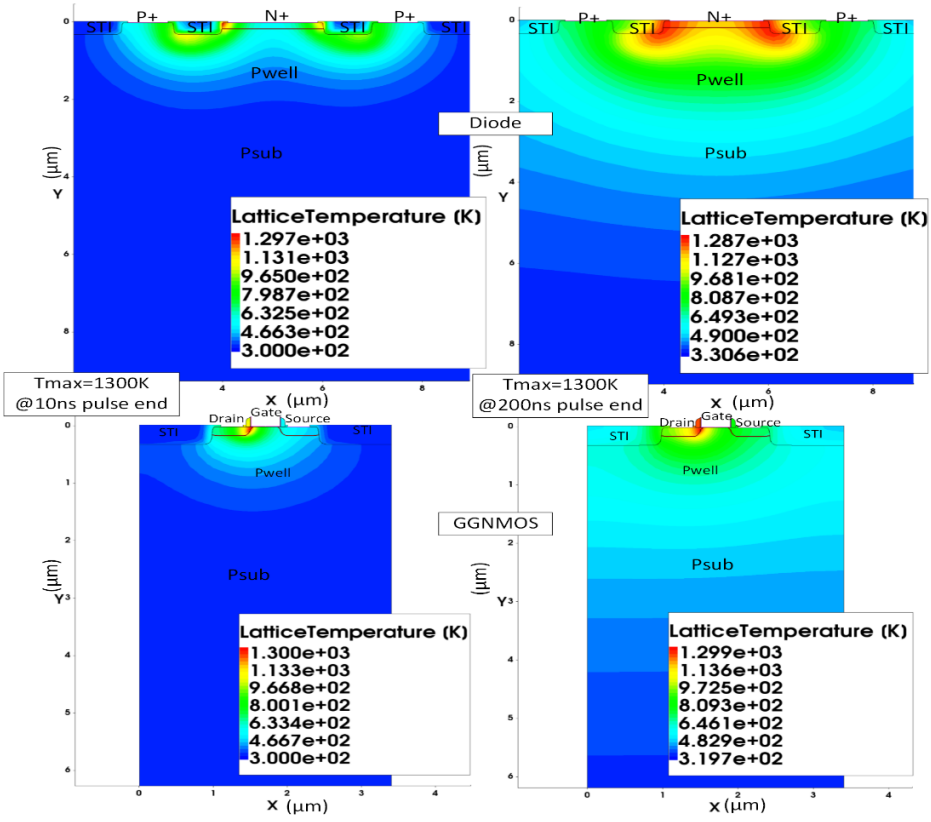


Figure 3-8 Simulated temperature distributions in a diode (top) and in a GGNMOS (bottom) under the short (10ns) (left) and long (200ns) (right) pulses.

Moreover, the non-uniform conduction at corners and edges in multi-stripe devices leads to effective failure power values that do not scale with device dimension. It makes the final temperature curves deviates from each other, which is clearly observed in the case of multi-stripe GGNMOS.

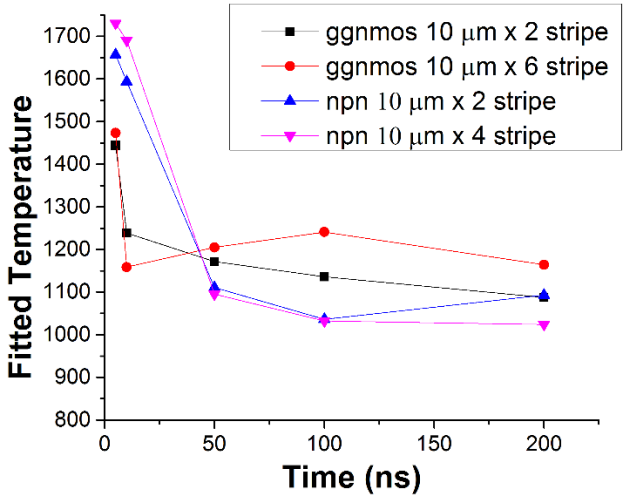


Figure 3-9 Effective failure temperature vs. pulse duration characteristics obtained for four devices: two GGNMOS and two NPN BJTs.

The same heating and failure concept applies to forward biased P/N diodes, but ESD-induced failures in these conditions are more likely to take place at the metal lines due to higher failure power of forward diodes and higher temperature increment in the metal line. When the same approximation method is applied to more complex devices, multiple hotspot may occur, revealing other failure mechanisms accordingly. More research work on this subject will be done in the near future.

3.4 Conclusion

An experimental method to estimate the ESD-induced failure temperature in a semiconductor device was proposed. This development was based on the electrical-heat conduction analogy. This methodology allows for the estimation of failure temperatures of different device types. Extracted failure temperatures as a function of electrical pulse duration were measured on diodes, GGNMOS, and BJTs using the TLP tester. The observed failures were reported and the underlying physical mechanisms were discussed. The results indicate that the failure temperatures of semiconductor devices under ESD stresses are higher than the deemed critical temperature at which the semiconductor becomes intrinsic. This work offers a practical approach for evaluating the ESD-induced failure temperature, which is a better failure criterion for ESD modeling and simulation.

CHAPTER 4 A COMPREHENSIVE DIODE MODEL FOR SIMULATION UNDER ELECTROSTATIC DISCHARGE STRESSES

4.1 Introduction

IC failure caused by ESD and electrical over stress (EOS) is the one of major reliability issues that designer must concern about prior to fabrication. Adding ESD unit is the most common solution to protect IC products from ESD/EOS stresses. However, as transistor shrinking its dimension in advanced technologies, it's getting difficult to design proper ESD unit capable of protecting the inner circuit with moderate layout area. Simulation of ESD event is very valuable to predict possible ESD failure and elevate ESD robustness of circuit design with affordable cost. Although SPICE models are available for validation of IC reliability at circuit level, ESD may push device beyond its normal operation region into a high voltage/current state, even reaching the failure point. ESD simulation based on current models are inaccurate in such extreme conditions. To support accurate simulation of device behaviors during ESD events, compact device models are required.

No single compact model can achieve ESD simulation at full timeframe scale. Instead, it better to combine related compact models together. Figure 4-1 is a typical diagram show how compact ESD models correlate with each other. First, there are ESD electrical models which are used to simulate electrical response of device under ESD stresses, such as overshoot, dynamic resistance of device during stress, reverse breakdown of P-N junction structure. The output of which are usually voltage/current waveforms. These output waveforms then become the input of ESD failure models, these models are monitoring one or several key indices during simulation, like temperature, voltage, current, or electric field strength, to see if a failure criterion is reached.

Certain simulated results are feedback to ESD electrical models to improve the accuracy of simulation, like the temperature, which will influence carrier mobility. With the results of both kinds of models, the ESD capability of the design is evaluated, then improvement can be made.

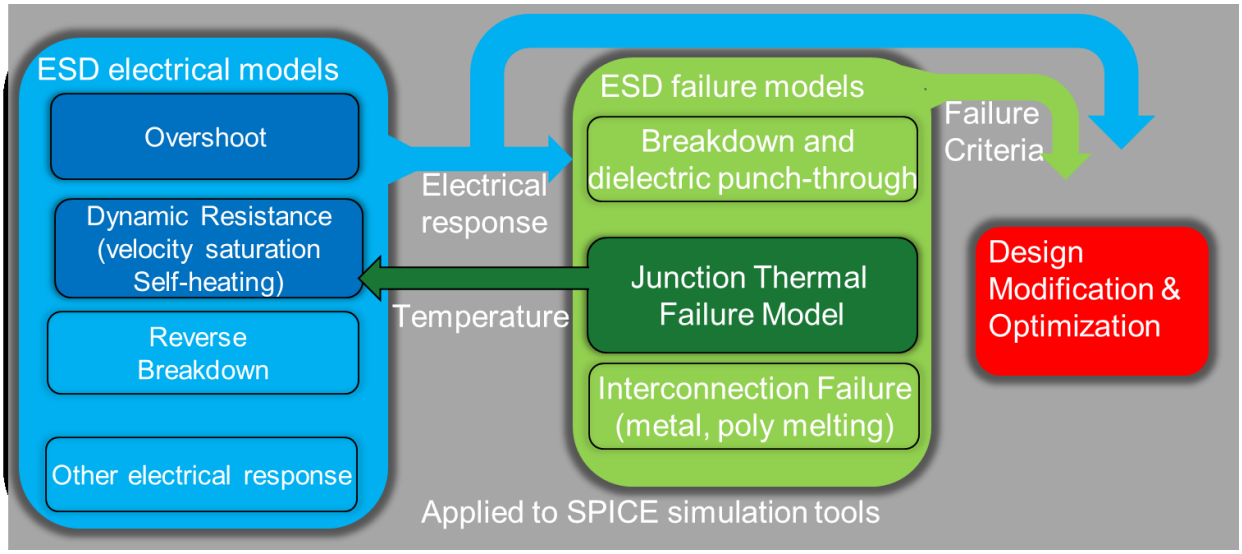


Figure 4-1 A diagram of a comprehensive ESD model consists of several compact models

As one of the major devices used for ESD protection, diode provide basic ESD protection capability and was first to be modeled at circuit level for ESD simulation. Moreover, as the P-N junction widely exists in ESD devices, functions as an ESD trigger structure and sustains most of ESD energy, modeling of P-N junction diode could be substantially useful for modeling other ESD devices with P-N junction structures. Unlike in its normal operation condition, a diode under ESD stress may exhibit uncommon characteristics. At high current region, the ESD performance of diode degrades. An ESD protection scheme is unreliable if the design is based on a conventional diode model which only consists of diode forward turn-on characteristics at low

current region. Due to self-heating and velocity saturation effects, on-resistance of P-N junction diode will increase dramatically when conducting current is pushed to a high value. Corresponding modeling description is required to reflect such inevitable physical phenomenon under high energy ESD stresses. Finally, a hard failure point will be reached and the device is irreversibly damaged. Failure models are required for the purpose of predicting such hard failure caused by ESD. On the other hand, various ESD stresses demands corresponding protection strategies. For transient stresses, especially those with fast rise time, the ESD protection unit must respond quickly enough to prevent high voltage overshoot causing damage to the inner circuit. In that case, modeling of overshoot phenomenon is a necessity.

In this chapter, a comprehensive diode model consists of several compact models is proposed. First model for overshoot or forward recovery phenomenon of diode forward turn-on is discussed. It is a key feature to evaluate diode's ESD capability when handling fast transient ESD stresses. Then, the on-resistance variation phenomenon at high current operation region is attributed to the self-heating and velocity saturation effects. By introducing a failure index related to temperature, a junction thermal failure model is developed to monitor temperature rise during stress and predict possible thermal failure if a critical temperature is reached. These models are abstraction of device physical mechanisms under ESD stresses and simulate different aspects of diode behavior in ESD events. Combined with current diode SPICE models, the complex of models is able to simulate diode I-V characteristics from low voltage turn-on to high current region till it final thermal failure within a wide timeframe.

.

4.2 Modeling of Overshoot at Diode Forward Turn-on

The early version of diode overshoot model stems from diode forward recovery derived or the case of ideal current source, or for DC switch condition [66, 67]. It is questionable when describing a fast transient event, like ESD stress.

A well accepted model for forward recovery of diode turn-on assumes the modulation of diode's quasi-neutral region causes the delay of diode fully turn-on, and such turn-on delay can be modeled as a modulation of series resistor R_S [68]

$$R_S = R_{S0} + \frac{R_{SM}}{1 + {Q_M}/{Q_0}} \tag{4.1}$$

where R_{S0} is the constant part in the total resistance in series representing the finite resistivity of metal interconnection, highly doped P+ and N+ regions. The second term at right hand side of the equation is the modulated part. For a P+/Nwell diode, this modulation region will be in the Nwell, and vice versa for a N+/Pwell diode. R_{SM} is the resistance of the N− region before conductivity modulation begins. Q_0 is the threshold charge for the onset of resistance modulation, and Q_M is the modulated charge for neutralizing the excess electron carriers in neutral region, and it is linked to the diode current I_d by

$$Q_M = \tau \cdot I_d \tag{4.2}$$

The time constant τ is usually fixed based on the duration of ESD stress, like 100 ns for a TLP.

The parameters in this model equation can be extracted from measurement, while if parameters for device fabrication and dimension is available, there is an extension for this overshoot model with

$$Q_0 = \gamma q N_A w A \tag{4.3}$$

and

$$\tau = \frac{w^2}{N_{tt} D} \tag{4.4}$$

where w and A are parameters of diode dimension representing the width and area of the P-N junction. N_A is the doping concentration of the low doped neutral region. q is the elementary electron charge. γ is used for representing the average distance between the highly doped N+ and P+ regions in a lateral diode. N_{tt} is named as the carrier distribution factor, which accounts for the distribution of carriers in the neutral region. D is the diffusion constant. The extension of the model takes into account the saturation of carrier drift velocities and makes use of device fabrication and dimension parameters. When such data is not available for modeling, fitting parameter values extracted from measurement is also an alternative, which provide a flexible option for the overshoot model.

From the equation, it can be assumed that the ideal diode forward turn-on time is τ. For times shorter than τ, the modulated charge Q_M is well below the threshold charge Q_0 in the neutral region, namely $Q_M << Q_0$. As a result, the series resistance R_S equals $R_{S0} + R_{SM}$, which defines the peak value of the overshoot voltage. The higher the series resistance is before modulation, the higher overshoot voltage may reach.

.

4.3 Failure Prediction and Temperature Monitoring

As diode continually conducting ESD current through its P-N junction, in-device temperature will gradually rise due to the inevitable self-heating effect. Come along with the self-heating and temperature rise, many side effects may occur, such as carrier mobility degradation, increased leakage current, etc. More importantly, if ESD stress continues and the in-device temperature keep rising, a failure temperature will be reached at the hotspot of device, which will cause irreversible damage to the device.

Thermal failure by reaching the critical temperature is a common factor for semiconductor hard failure. The other is gate oxide breakdown or dielectric punch-through. Given the reality that P-N junction is one key structure for most ESD device and sustains the majority of ESD current during stress, it is very useful for designing robust ESD protection if the temperature variation at the hotspot of P-N junction can be monitored, and further, thermal failure due to reaching critical temperature can be predicted.

There is a common way to model temperature variation in a single structure by establishing a thermal network. The method in Chapter 2 provide a good balanced version of thermal network that the three-stage thermal circuit achieves both accuracy and low complexity. Thus we can use this three-stage thermal network as failure prediction model, and a temperature monitor to assist simulate thermal behavior in the comprehensive diode model.

4.4 Modeling of On-resistance Variation of Diode at High Current Region

An ideal diode model assumes the current of diode conducting follows equation

$$\mathrm{I}_d = I_S \cdot \left(e^{\frac{V_D}{nV_T}} - 1\right) \tag{4.5}$$

where I_S is the reverse bias saturation current. V_D is the voltage across the diode and V_T as the thermal voltage. n is the ideality factor usually varies from 1 to 2. For V_T as

$$V_T = \frac{kT}{q} \tag{4.6}$$

Where k is the Boltzmann constant, T is the absolute temperature of the P-N junction in Kalvin, and q is magnitude of charge of an electron.

These equation is accurate if the operation of diode is confined at low current region, but as the conducting current rise, the measured I-V will deviate from the ideal I-V curve of a diode. In other words, the current of forward conducting diode will saturated at high current region. And such saturation may go even obvious before the second breakdown point that the I-V curve will become almost flat near the end.

The saturation phenomena can be explained by carrier velocity saturation at high current end. The corresponding effect is included in the conduction equation by introducing a mobility reduction factor [69].

$$\mu_r = 1 + \sqrt{\left(\frac{E - E_{C,eff}}{2\,ECRIT}\right)^2 + \frac{DU \cdot E_{C,eff}}{ECRIT}} + \sqrt{\left(\frac{E + E_{C,eff}}{2\,ECRIT}\right)^2 + \frac{DU \cdot E_{C,eff}}{ECRIT}} - \sqrt{\left(\frac{E_{C,eff}}{ECRIT}\right)^2 + \frac{4 \cdot DU \cdot E_{C,eff}}{ECRIT}} \tag{4.7}$$

where $E_{C,eff}$ is given by

$$E_{C,eff} = \sqrt{ECORN^2 + (2 \cdot DU \cdot ECRIT)^2} - (2 \cdot DU \cdot ECRIT) \tag{4.8}$$

and E is the electric field across the resistor body given by

$$E = \frac{V_{21,eff}}{leff} \tag{4.9}$$

$V_{21,eff}$ is effective voltage value across the non-linear resistor.

$$V_{21,eff} = \frac{2 \cdot V_{21} \cdot V_{sat}}{\sqrt{(V_{21} - V_{sat})^2 + 4 \cdot ATS^2} + \sqrt{(V_{21} + V_{sat})^2 + 4 \cdot ATS^2}} \tag{4.10}$$

In the above expressions, *ATS* is the saturation smoothing parameter with a default value of 0. *ECORN* and *ECRIT* are velocity saturation field strength at corner and overall critical value. DU is mobility reduction at *ECORN* with a default value of 0.02.

With the mobility reduction factor μ_r, the conductance of the non-linear resistor model is calculated as

$$g = \frac{1}{R_0} \cdot \frac{\left(1 - f_{depl}\sqrt{DP + V_{21}}\right)}{\mu_r} \tag{4.11}$$

where R_0 is the normalized resistance at room temperature, usually the resistance before an ESD stress is applied. f_{depl} is the depletion factor defined as

$$f_{depl} = \max\left(\frac{\sqrt{DP}}{DP + 1 \times 10^5}, DFINF + \frac{DFL}{len_{depl}} + \frac{DFW}{wid_{depl}} + \frac{DFWL}{len_{depl} \cdot wid_{depl}}\right) \tag{4.12}$$

In the above expressions, *DFL*, *DFW* and *DFWL* are depletion factor coefficient for length, width and area. The default value for these coefficient are 0 for ideal diode assuming depletion region is neglectable. *DFINF* is depletion factor coefficient for wide/long device. Its value depends on device dimension. *DP* is the depletion potential. And finally, the current through this resistor is calculated as

$$I = g \cdot V_{21} \tag{4.13}$$

However, the on-resistance of diode at high current region cannot be explained by carrier velocity saturation alone. Figure 4-2 gives tested I-V curves under different TLP pulse width for a single diode. It can be see that I-V curves of the diode saturated at high current end, and moreover, the on-resistance variations for different TLP pulse width are different. The saturation of long pulse (200 ns, 500 ns) starts at earlier current level than the short pulse (10 ns 50 ns). With longer TLP pulse, the I-V curve bends more to the voltage axis. This cannot be explained by the velocity saturation at high current region theory because the current levels are the same for all pulse widths, but saturation conditions are distinct. In that case, we must re-introduced the self-heating effect, which is also the same effect responsible for device thermal failure, as an additional factor that cause on-resistance variation in the high current region.

It is well accepted that the resistance of semiconductor material will increase with the rise of the temperature. The common equation to describe the temperature dependent behavior of resistance is

$$R = R_0 \cdot (1 + \alpha \cdot \Delta T + \beta \cdot \Delta T^2) \tag{4.14}$$

where R_0 is the normalized resistance before stress is applied, or the resistance increment at background temperature. ΔT is the temperature offset in Kelvin, namely the temperature increment between device temperature and background temperature. α and β are the linear and quadratic temperature coefficient of resistance.

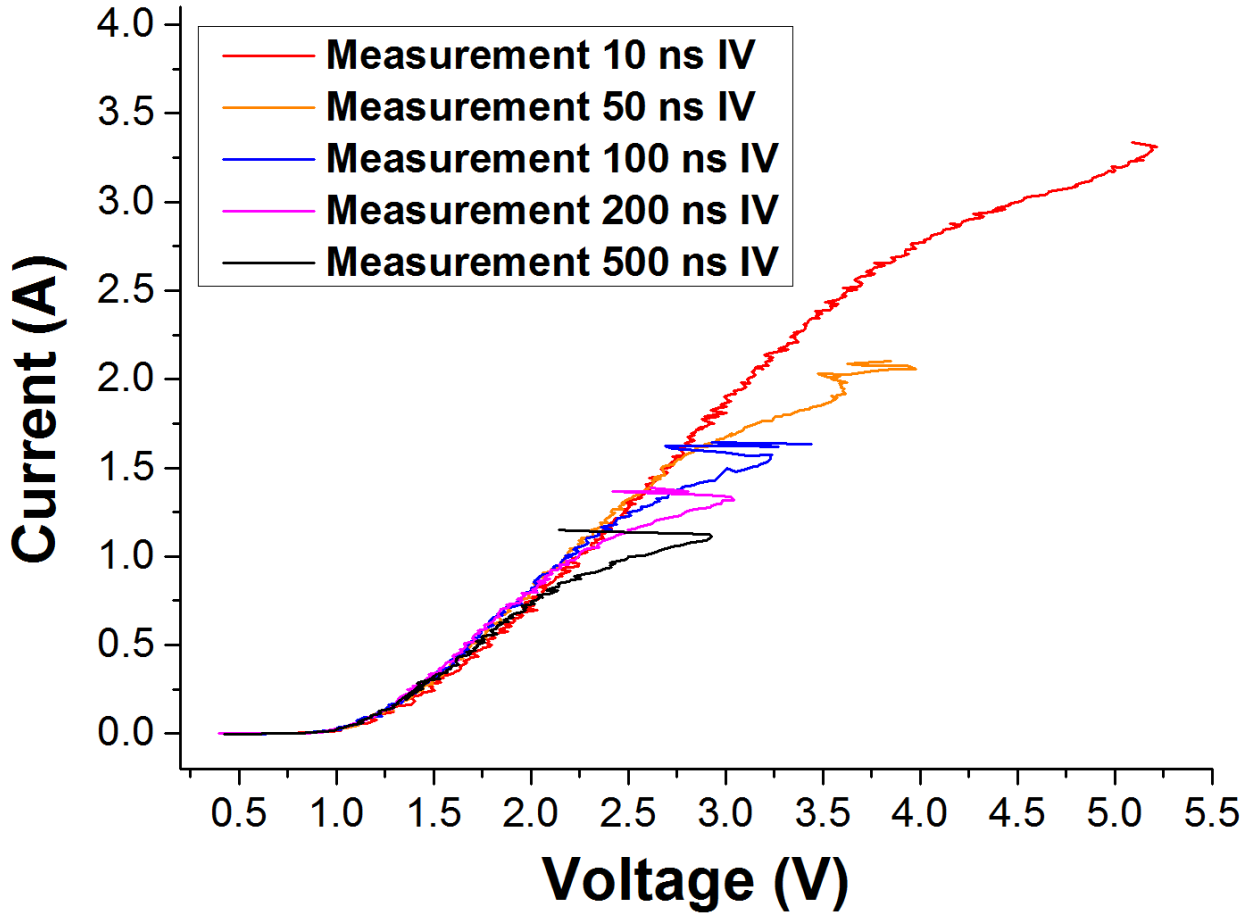


Figure 4-2 Measured TLP I-V curves under different pulse widths, ranging from 10ns to 500ns.

4.5 Parameter Extraction and Simulation Results

The compact models related to different ESD phenomena are incorporated into SPICE based simulation tools. With proper fitting with measurement data, model parameters are extracted and the comprehensive diode model can reproduce device behavior under ESD stresses. The full schematic of this comprehensive diode model is shown below Figure 4-3. First, there is a diode model as a core model to simulate an ideal diode forward turn-on from low current to high current region without any on-resistance variation. In series with which, there is a non-

linear resistor model which adds extra variable resistance to the diode at high current. If failure event need to be simulated, a short-circuit switch can be place in parallel with the diode, which is controlled by the three-stage thermal circuit model, once a critical temperature is reached during simulation, the switch closes, turning the whole sub-circuit into low resistance state, just as the second breakdown point. To enable overshoot simulation in transient simulation mode, the diode model must be modified to add extra overshoot related parameters. Then the transient resistance of a diode at its turn-on stage is no longer constant, but a variable.

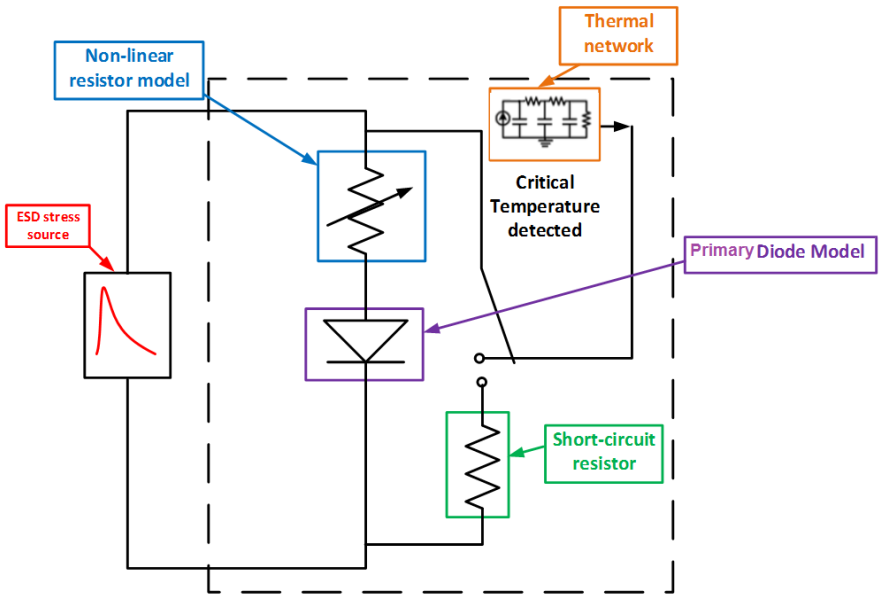


Figure 4-3 Full schematic of comprehensive diode model.

First, the voltage overshoot phenomena can be simulated by fitting related parameters of R_{S0}, R_{SM}, Q_0, Q_M. The extraction of which can come from two ways. If process information is available as described in 4.2, these parameters can be derived by well doping concentration and

device dimension. However, such information is not always available for circuit designer. Thus, by fitting measured overshoot waveform with simulation can also yield correct parameters. From below Figure 4-4 it can be observed the overshoot phenomenon is mostly about the transient voltage appears as a high but short duration peak at the beginning of the transient voltage waveform. Although there is delay of full conducting observed from the transient current waveform, it is believed a diode functions as a current source, which makes the current delay neglectable. More important, due to the thin gate oxide that ESD unit usually protects, voltage overshoot is more concerned about and must be prevented as much as possible.

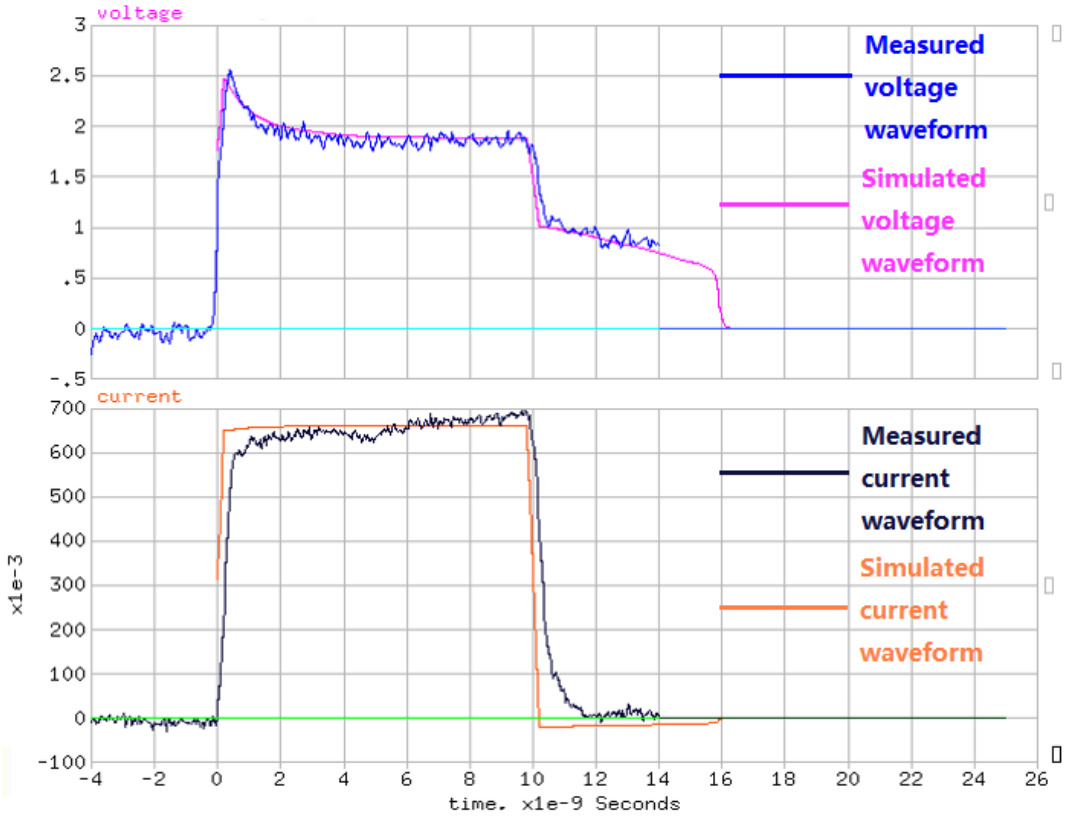
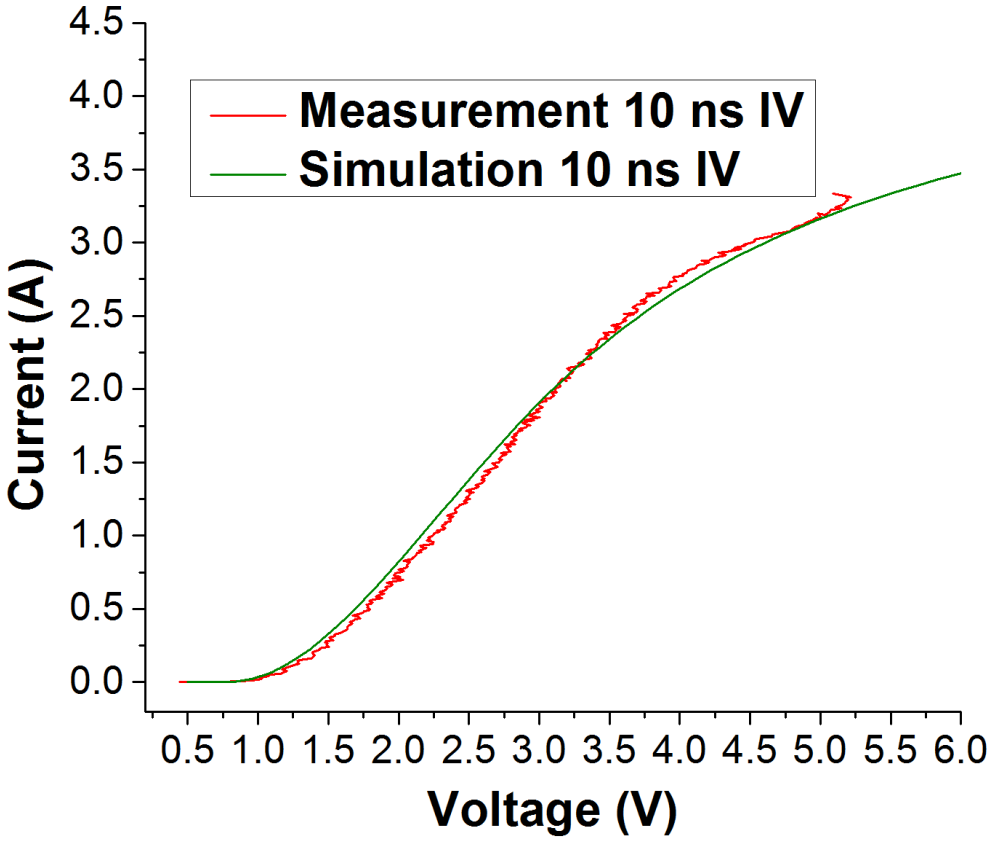
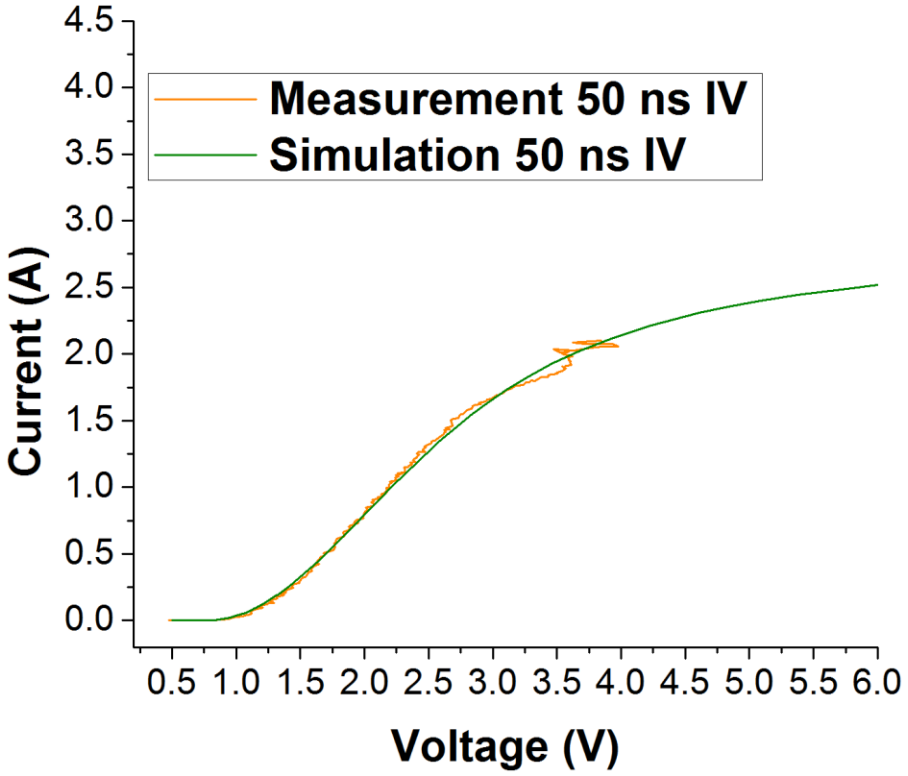
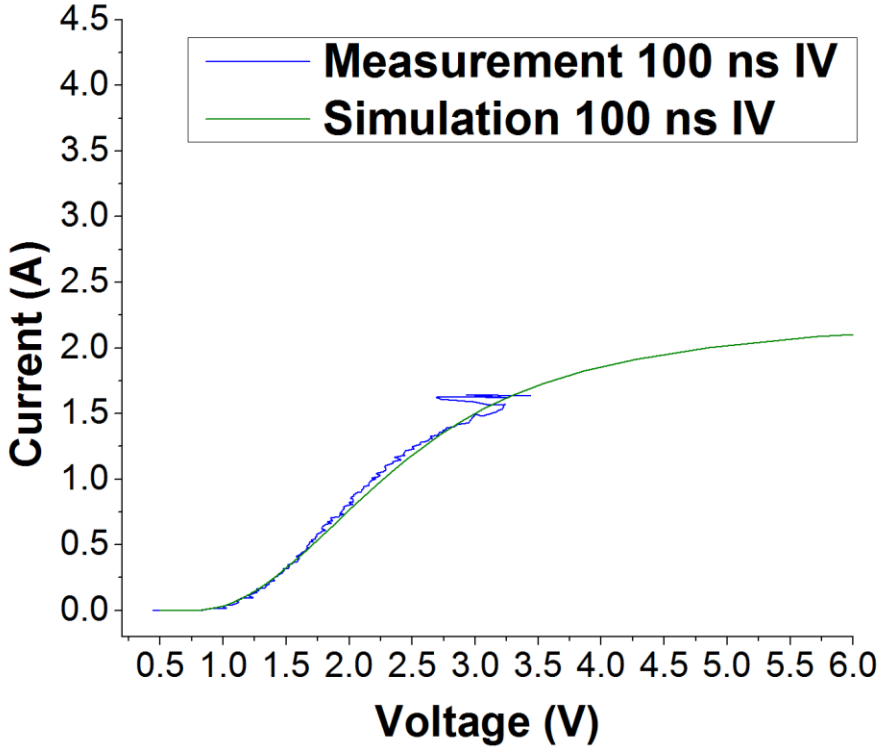


Figure 4-4 Measured and simulated transient waveforms under a 10ns TLP pulse with voltage overshoot peak.

Based on the transient simulation of voltage and current waveform, a TLP like simulation can be performed by extracting 70%-90% of each waveform and connecting them as a TLP I-V curve. Figure 4-5 gives each simulation result of TLP test under 10ns to 500ns pulses. Compared to measured curves, we can observe the simulation I-V curves are very close to measured I-V curves, namely the simulation well fit the measurement. Such result indicates by introducing velocity saturation and self-heating effects at high current region, the on-resistance variation phenomenon of diode forward conducting can be well simulated.




Measurement 50 ns IV
Simulation 50 ns IV
Current (A)
Voltage (V)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0


Measurement 100 ns IV
Simulation 100 ns IV
Current (A)
Voltage (V)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0

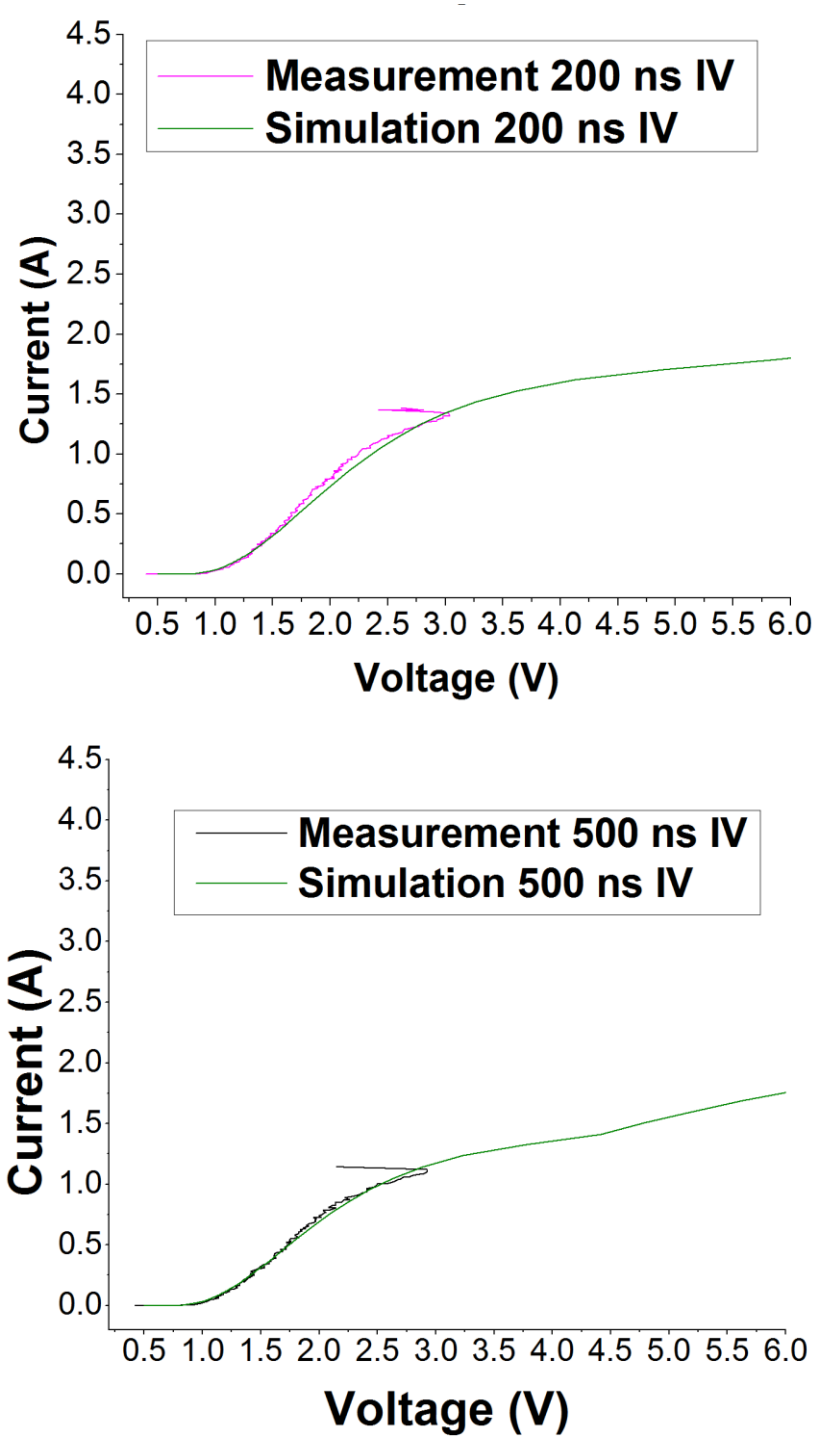


Figure 4-5 TLP I-V simulations fit with measured I-V under different pulse width, ranging from 10ns to 500ns.

4.6 Conclusion

The comprehensive diode model consists of several compact models to simulate device behavior under ESD stress. An ESD pulse will stress the diode from low current turn-on to high current saturation and finally to its thermal breakdown point. In that case, not only conventional device models for normal circuit simulation, but specialized ESD models are required. Moreover, though it is a straight-forward method to use failure current or voltage level as the failure criteria for ESD simulation, to achieve failure prediction under different ESD stress conditions, a thermal circuit model is more suitable to be embedded to monitor the temperature increment during stress and predict the final thermal failure if a critical temperature value is reached. Besides the ESD I-V simulation, due to the transient property of ESD stresses, it is very important to simulate transient voltage and current response as well, especially the overshoot phenomena due to device turn-on delay. Thus, an overshoot model is applied to the comprehensive diode model to reproduce forward recovery of diode and monitor possible early failure due to too high overshoot voltage peak.

CHAPTER 5 INVESTIGATION ON FORWARD TRANSIENT CHARACTERISTICS OF GaN P-N DIODES ON BULK GaN SUBSTRATE

5.1 Introduction

Gallium nitride (GaN) based devices are very promising candidates for various power and high-frequency applications owing to the high breakdown electric field, high electron mobility and high thermal conductivity of GaN [70]. Despite of the difficulty and high cost in fabricating high-quality GaN bulk substrate, vertical GaN power devices possess low density of defects and exhibit better performance than those on sapphire, SiC or Si substrates [71].

As the most fundamental structure, p-n junction diode carries the basic capability and quality of a semiconductor material. To this end, it is continuously optimized to achieve higher operating voltage. In recent years, the reverse breakdown voltage of vertical GaN p-n diode is dramatically increased to over thousands of volts [72]. On the other hand, it is imperative to evaluate its performance in forward bias condition, especially during fast transient event, for possible applications of switching and electrostatic discharge.

In this chapter, vertical GaN p-n diodes are fabricated [73], and the transient characteristics in forward bias condition are revealed [74]. Quasi-static I-V curves obtained by a TLP system exhibit the on-resistance decreases under high voltage stresses. By looking into transient waveforms picked at high current region, an overshoot peak can be observed at the very early portion of pulse stress.

5.2 Experiments and Analysis

The cross-sections of the GaN diodes are shown in Figure 5-1. The cathode is on the back side of a 400 μm thick n-type bulk GaN substrate. Above which, there are 2 μm n+ GaN buffer layer and 10 μm n- GaN drift layer with Si doping concentration of $2\times10^{18}cm^{-3}$ and $1.5\times10^{16}cm^{-3}$, respectively. The p-type regions of the GaN diodes consist of a 500 nm low doped (Mg: $1\times10^{18}cm^{-3}$) p- GaN layer and a 10 nm thick p++ GaN (Mg: $2\times10^{20}cm^{-3}$) layer. Metalized electrodes are formed by wet etch and a Ti/Al/Au stack for both anode and cathode. As shown in right half of Figure 5-1, to reduce reverse leakage current thus improving breakdown voltage, the diode can be covered by a layer of spin-on-glass (SOG), coated with the Ti/Al/Au, namely the field plate. The GaN diode with field plate can achieve breakdown voltage of 1700V. The bottom diameter of the diode is 110 μm, and both vertical GaN diodes with and without field plate share the same n-type substrate.

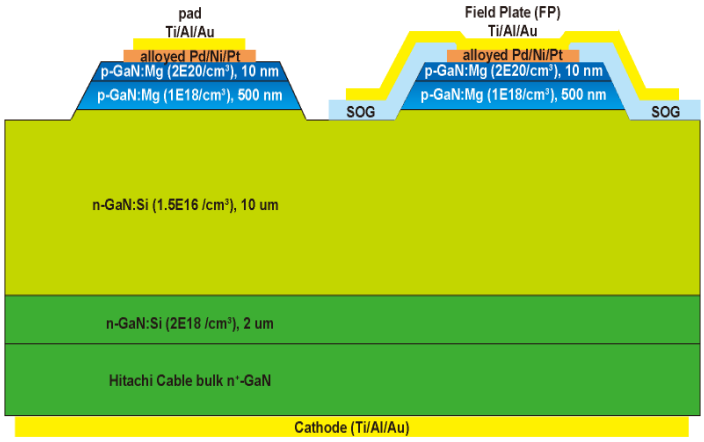


Figure 5-1 Cross-section of the GaN on GaN substrate P-N diodes without field plate (left) and with field plate (right).

Transient characteristics of the vertical GaN p-n diode are investigated as the diode is forward biased by a series of fast pulses. Transient pulses are applied by a TLP system. The TLP system generates a sequence of rectangular current pulses with increasing magnitude. By averaging the voltage/current values at the later portion (typically 70% ~ 90%) of each reflected waveforms, a quasi-static I-V curve is obtained.

In the experiment, two modes of TLP test are employed: the standard TLP generating 100 ns width pulses, and the very fast TLP (VFTLP) which gives much shorter pulses of 10 ns pulse width. Both modes have a steep rise edge of 300 ps to avoid turn-on delay induced by instrument. I-V curves of tested GaN diode are plotted in Figure 5-2. A typical I-V for a forward biased diode is extracted when applying a series of TLP pulses of 100 ns pulse width. The on-resistance is around 2.5 Ω. Before it fully turns on the diode endures a high resistant state (~10 Ω) for a few steps. The I-V curve under short pulses of 10 ns pulse width, however, shows a much longer stage of high resistant. The diode keeps a high resistance of 12 Ω until 33 V and turns into low resistance state. The disparity of on-resistance under different pulse width exists in both vertical GaN diodes with and without field plat.

To figure out the cause for the disparity of I-V under different pulse stresses, Transient waveforms are picked at different stages of I-V curves. Figure 5-3(a) shows transient voltage/current waveforms for 10 ns vfTLP stresses. First, waveforms at the early stage of GaN forward bias are plotted. There is little current flowing through the GaN diode, and both transient voltage and current show a flat waveform. When stressing pulse rises from 20 V to 100V, transient current increased obviously, but on-resistance still stays high. As pulse magnitude increasing, the I-V curve is inflected into a low resistant state of 2.5 Ω. It is much

more obvious on transient waveforms of 200V pulse that current is swelling up while voltage is dropping within 10 ns, which makes a lower on-resistance. It indicates the vertical GaN diode turns on gradually at high stress level. The changing current/voltage gradually decays to a stable value if the pulse is long enough, as shown in Figure 5-3(b). With sufficient stress duration, 100 ns TLP pulses are able to fully turn-on the GaN diode even at low level of 100 V.

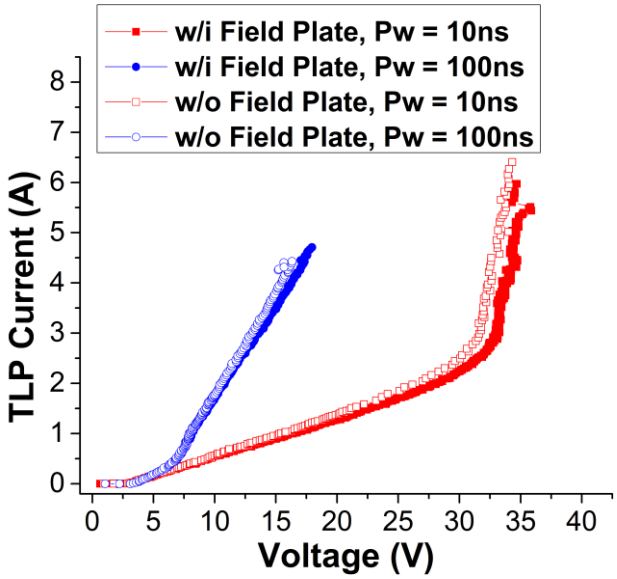


Figure 5-2 Quasi-static I-V curves of vertical GaN diodes, w/i field plate (solid) and w/o field plate (hollow), under 100ns (blue) and 10ns (red) pulse width.

During the turnon of the vertical GaN diode in high injection conditions, the diode appears to be inductive and an overshoot voltage is observed decreasing with time to the normal forward voltage. This phenomenon is called forward recovery. The delay of GaN diode turn-on is due to conductivity modulation and transit time, which can be described as the changing resistance of lowly doped region of a diode [68] as equation (4.1) in chapter 4.

For pulse duration shorter than τ, not enough charge can be established and therefore the diode tends to be high resistance of R_{S0}+R_{SM}. With increasing Id level and longer pulses, the series resistance becomes much lower. On the other hand, due to the thick n- GaN layers with low doping concentration for high voltage application, the vertical GaN diode suffers severe overshoot at the beginning of pulse and requires considerable time to decay to stabilized stage. When extracting the quasi-static I-V by averaging voltage/current value at 70% ~ 90% of pulse duration, long pulses carry sufficient stress energy to fully turn-on the diode, while conductivity modulation may be still underway for short pulses. It well explains different I-V curves obtained by TLP and vfTLP mode in Figure 5-2.

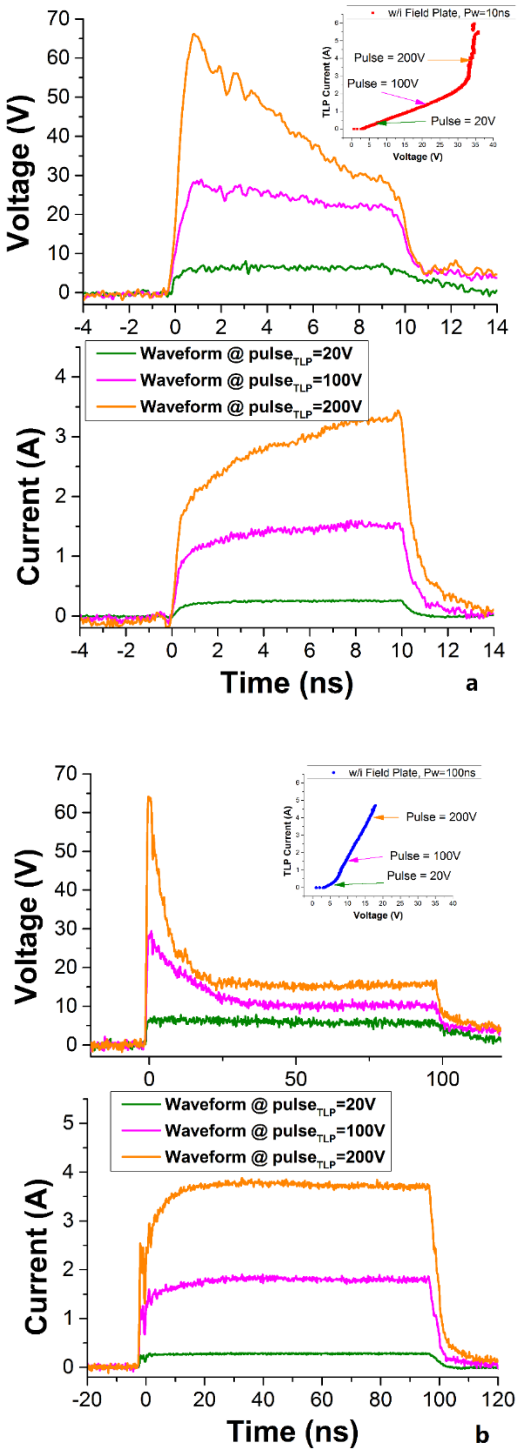


Figure 5-3 Transient voltage (upper figures) and current (lower figures) waveforms for 10ns (a) and 100ns (b) stress pulses.

It isn't hard to explain the difference between 10ns and 100ns I-V curves if we observe the transient waveforms for 10ns pulses and 100ns pulses together. Figure 5-4 overlaps the voltage waveforms for 10ns pulse tests and 100ns pulse tests at pulse voltage = 20V, 100V and 200V. It can be observed the 10ns voltage waveforms overlap with the early portion of 100ns voltage waveform. However, due to the different average windows for 100ns(70ns ~ 90ns) TLP test and 10ns(7ns ~ 9ns) vfTLP test, I-V curves for 100ns and 10ns are different. Since the average window for 10ns pulse is between 7ns and 9ns when the vertical GaN diode is still not fully turned on, the 10ns I-V curve has higher voltage value than the 100ns I-V curve.

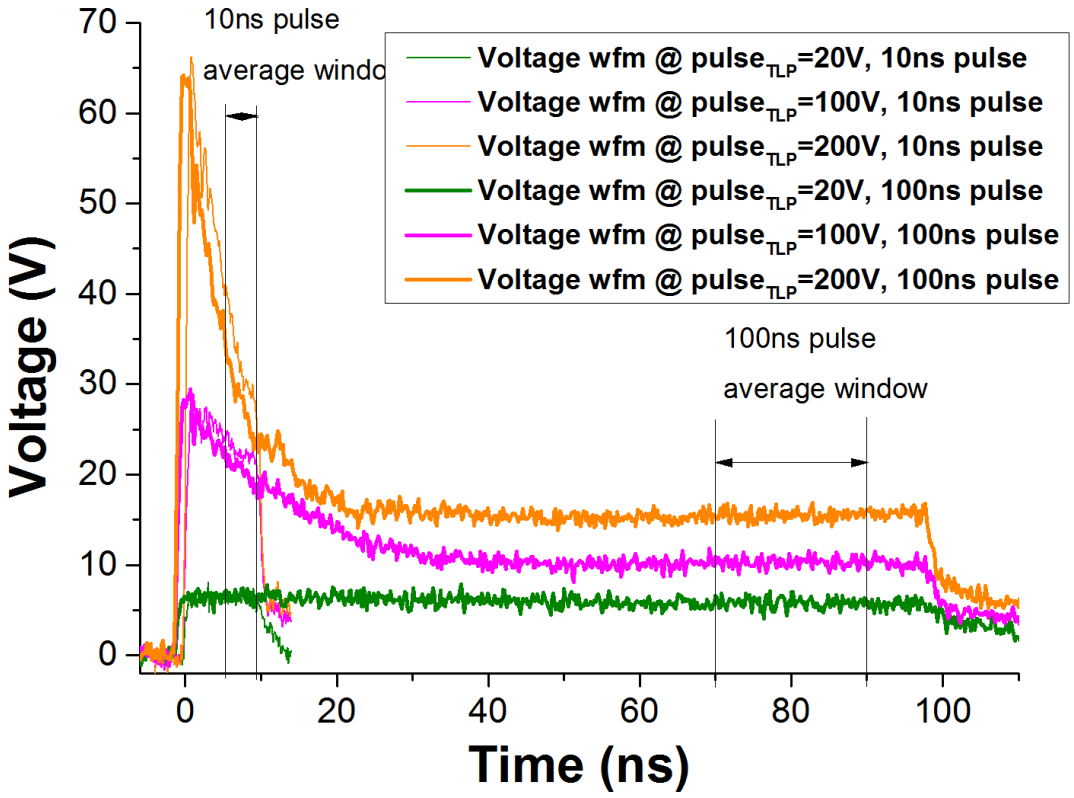


Figure 5-4 Voltage waveform Comparison of 10ns vfTLP test and 100ns TLP test. Waveforms captured at pulse voltage at 20V, 100V and 200V.

5.3 Turn-on Characteristics of Vertical GaN diode under Different Temperature

Further experiment is carried out on how different ambient temperature will influence the turn-on of vertical GaN diode. Since the vertical GaN diode is initially designed for power electronics applications, such as rectification sustain high voltage stress, output drive, which may requires the device to work under high voltage, high current conditions, and inevitably, the temperature of the diode will rise. Figure 5-5 shows the I-V curves for different pulse width (10ns and 100ns) tested under room temperature (25°C) and high temperature (125 °C). It can be seen the I-V curves under high temperature will shift right a little bit.

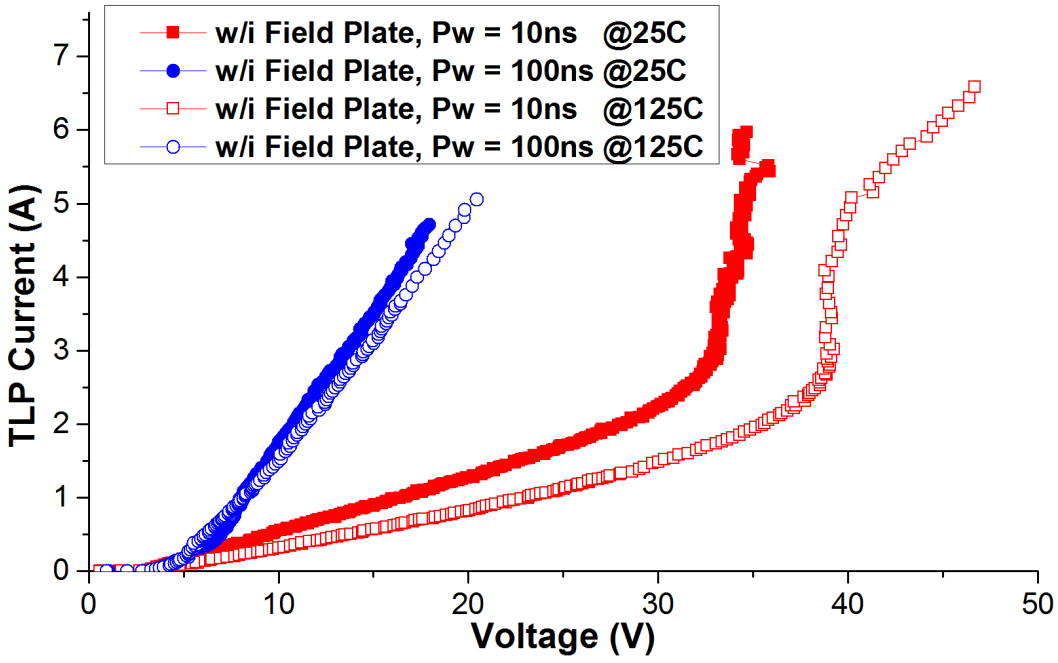


Figure 5-5 TLP I-V curves for a vertical GaN diode with field plate under room temperature (25°C) and high temperature (25 °C), for short pulse (10ns) and long pulse (100ns).

The shifting of I-V curve indicates the on-resistance of vertical GaN has increased after the ambient temperature raised during test. Since the TLP test only capture the average value of transient voltage and current within a portion of the full waveform. It's better to view full transient waveforms under different test temperature. As shown in Figure 5-6, the transient voltage/current waveforms indicates on-resistance under high temperature environment has increased for whole stress duration. For 100ns long pulse, the transient waveforms shows on-resistance increment caused by high temperature leading to lower current and higher voltage values, especially at the early portion of turn-on, which is the 10ns pulse test shows. Such on-resistance increment has the similar causes in the silicon, the carrier mobility reduction and the self-heating effect. Since the 10ns vfTLP averages voltage/current value at the early portion of diode turn-on, the I-V curve shift for high temperature is more obvious than the 100ns TLP I-V curve.

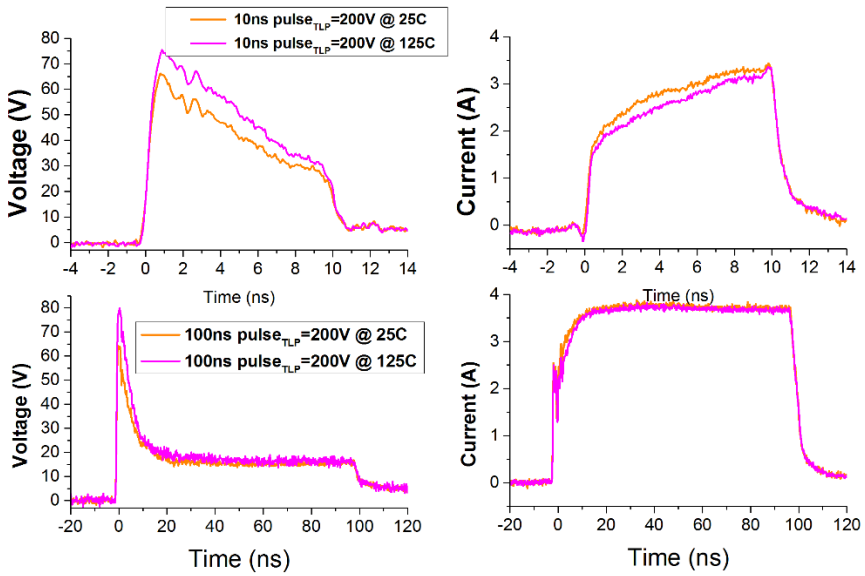


Figure 5-6 Comparison of transient voltage/current waveforms at Vpulse=200V, ambient temperature at 25°C and 125°C, respectively.

5.4 Conclusion

Vertical GaN diode for high voltage applications is fabricated and experiments on its transient characteristics in forward bias conduction are carried out. Quasi-static I-V characteristics are obtained by using fast stress sources of TLP/vfTLP. It shows that vertical GaN diode turns on gradually, and its I-V varies with different pulse width applied. Transient waveforms obtained at different stages of diode forward conducting exhibit an overshoot voltage with significant amplitude. Due to the thick n- GaN layers with low doping concentration, vertical GaN diode is not fully turned on under short pulses, which yields a high-resistant I-V curve.

CHAPTER 6 SUMMARY AND OUTLOOK

ESD failure is a real threat to the semiconductor products, causing major IC failure and malfunction. It is imperative to design proper ESD protection for various IC applications. Due to limited ESD design window and special requirement of specific IC applications, an universal ESD protection for all IC products is impossible. An ideal solution is to design the ESD protection according to the IC chip with ESD evaluation prior to manufacture to optimize its ESD robustness. In that case, circuit level ESD simulation and ESD models for components are needed.

One critical issue about ESD is evaluate device's ESD capability and predict possible failure. For that, specialized failure models are required to be integrated into the simulation tools. Since there are two major damage caused by ESD stresses, one is the dielectric punch-through, basically occurs as gate oxide breakdown in CMOS process, which is related to electric field. The other is thermal failure, like junction thermal failure or interconnection metal melting. This paper focuses on the junction thermal failure, which is very common in semiconductor, and established a simple failure model using temperature increment as the failure criterion. Based on the analogy of electrical heat conductions, the heat conduction in silicon during an ESD stress can be mapped into an electrical circuit. Simulating thermal voltage variation in this thermal circuit will reproduce heat conduction in device, and when a critical temperature is reached, a thermal failure occurs. With this thermal circuit model, it is possible to monitor in-device temperature rise during stress and predict thermal failure, thus the ESD design can be improved prior to manufacture.

To accurately simulate thermal failure in device, the failure criterion, namely the failure temperature must be defined first. Since the type and material of semiconductor devices vary, failure criteria for devices are different. On the other hand, due to the transient characteristic of ESD pulses, it is almost impossible to measure in-device temperature increment within such short time duration. In other words, direct measurement of failure temperature is unavailable as a model parameter extraction method. In that case, a new method to give close estimation of device thermal failure temperature is proposed in this paper.

Based on the assumption that thermal conduction is a adiabatic process during very short timeframe that there is only heat generation and no time for heat to dissipate into deep silicon, a one-pole approximation is established to provide accurate failure temperature estimation under very short pulse. The extraction results show the failure temperature for P-N junction device is a bit higher than the intrinsic temperature, which supports the failure temperature theory that at high current levels, secondary breakdown occurs once heat produces enough minority carriers to support the increasing current without raising the electric field.

As an estimation, the accuracy of this method may be concerned about. In this paper, the discussion of the accuracy of estimation method concludes there are overestimation and underestimation in the approximation which offset each other, and the estimation is more accurate for short pulses.

Further, with the failure model available, it is possible to combine all ESD related compact models as a comprehensive model to simulate device behavior under ESD stress from low current turn-on to high current region, and even its final failure point. In this paper, a comprehensive diode model is proposed. ESD related phenomena, such as overshoot, on-

resistance variation at high current region and thermal failure, are included as compact models in it. Together with conventional diode models, we are able to simulate diode forward conducting under ESD stresses.

The ESD issue is not only a silicon based problem, but it exists widely in all kinds of semiconductor devices. For power electronic device on GaN, the high power capability and high breakdown voltage may largely increase its ESD robustness, but accordingly, the requirement for ESD capability rises. In the last chapter, an investigation on forward transient characteristics of GaN P-N diodes on bulk GaN substrate is proposed. Similar to P-N diodes on silicon, the vertical GaN P-N diodes show overshoot or forward recovery phenomenon under transient ESD pulse, but with larger overshoot peaks and longer turn-on delay. Such phenomenon may attributes to the property of GaN and thicker epitaxy layer for longer time to establish conductivity modulation in diode neutral region.

LIST OF REFERENCES

[1] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectrum,* vol. 34, pp. 52-59, 1997.

[2] A. Z. H. Wang, F. Haigang, Z. Rouying, X. Haolu, C. Guang, W. Qiong*, et al.*, "A review on RF ESD protection design," *IEEE Transactions on Electron Devices,* vol. 52, pp. 1304-1311, 2005.

[3] B. J. Baliga, "Trends in power semiconductor devices," *IEEE Transactions on Electron Devices,* vol. 43, pp. 1717-1731, 1996.

[4] A. Tazzoli, V. Peretti, R. Gaddi, A. Gnudi, E. Zanoni, and G. Meneghesso, "Reliability Issues in RF-MEMS Switches Submitted to Cycling and ESD Test," in *2006 IEEE International Reliability Physics Symposium Proceedings*, 2006, pp. 410-415.

[5] A. Wallash, "ESD challenges in magnetic recording: past, present and future," in *2003 IEEE International Reliability Physics Symposium Proceedings, 2003. 41st Annual.*, 2003, pp. 222-228.

[6] A. Amerasekera and C. Duvvury, *ESD in Silicon Integrated Circuits* vol. 9: John Wiley & Sons, Ltd., 2002.

[7] A. Amerasekera, W. v. d. Abeelen, L. v. Roozendaal, M. Hannemann, and P. Schofield, "ESD failure modes: characteristics mechanisms, and process influences," *IEEE Transactions on Electron Devices,* vol. 39, pp. 430-436, 1992.

[8] J. E. Vinson and J. J. Liou, "Electrostatic discharge in semiconductor devices: an overview," *Proceedings of the IEEE,* vol. 86, pp. 399-420, 1998.

[9] B. L. Gregory and B. D. Shafer, "Latch-Up in CMOS Integrated Circuits," *IEEE Transactions on Nuclear Science,* vol. 20, pp. 293-299, 1973.

[10] K. Ming-Dou and H. Sheng-Fu, "Physical mechanism and device simulation on transient-induced latchup in CMOS ICs under system-level ESD test," *IEEE Transactions on Electron Devices,* vol. 52, pp. 1821-1831, 2005.

[11] V. Chandrasekhar and K. Mayaram, "Analysis of CMOS RF LNAs with ESD protection," in *2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No.02CH37353)*, 2002, pp. IV-799-IV-802 vol.4.

[12] A. Z. Wang, H. G. Feng, R. Y. Zhan, G. Chen, and Q. Wu, "ESD protection design for RF integrated circuits: new challenges," in *Proceedings of the IEEE 2002 Custom Integrated Circuits Conference (Cat. No.02CH37285)*, 2002, pp. 411-418.

[13] A. Tazzoli, V. Peretti, and G. Meneghesso, "Electrostatic Discharge and Cycling Effects on Ohmic and Capacitive RF-MEMS Switches," *IEEE Transactions on Device and Materials Reliability,* vol. 7, pp. 429-437, 2007.

[14] K. Banerjee, H. Li, and C. Xu, "Prospects of carbon nanomaterials in VLSI for interconnections and energy storage," in *2009 31st EOS/ESD Symposium*, 2009, pp. 1-10.

[15] JEDEC, "Electrostatic discharge (ESD) sensitivity testing human body model (HBM)," ed, 2000.

[16] E. S. Association, "Standard test method for electrostatic discharge sensitivity testing – charged device model (CDM) component level," ed, 1999.

[17] K. Muhonen, N. Peachey, and A. Testin, "Human metal model (HMM) testing, challenges to using ESD guns," in *2009 31st EOS/ESD Symposium*, 2009, pp. 1-9.

[18] M. Scholz, D. Linten, S. Thijs, M. Sawada, T. Nakaei, T. Hasebe*, et al.*, "On-wafer human metal model measurements for system-level ESD analysis," in *2009 31st EOS/ESD Symposium*, 2009, pp. 1-9.

[19] IEC, "Methods for simulating of electrostatic effects - Machine model, Component testing. ," in *Part 3.1.*, ed, 2002.

[20] D. o. Defense, "MIL-STD-883E (1989) Electrostatic Discharge Sensitivity Classification," in *Methode 3015.7*, ed, 1989.

[21] ESDA/JEDEC, "ESDA/JEDEC Joint Standard for Electrostatic Discharge Sensitivity Testing – Human Body Model (HBM) – Component Level," ed, 2014.

[22] ESDA/JEDEC, "ESDA/JEDEC Joint Standard for Electrostatic Discharge Sensitivity Testing – Charged Device Model (CDM) – Device Level," ed, 2014.

[23] ESDA, "ESD Association Standard Test Method for Electrostatic Discharge (ESD) Sensitivity Testing – Machine Model (MM) – Component Level," ed, 2012.

[24] IEC, "Electromagnetic compatibility (EMC) - Part 4-2: Testing and measurement techniques - Electrostatic discharge immunity test," ed, 2008.

[25] IEC, "Electromagnetic compatibility (EMC) – Part 4-4: Testing and measurement techniques – Electrical fast transient/burst immunity test," ed, 2012.

[26] IEC, "Electromagnetic compatibility (EMC) - Part 4-5: Testing and measurement techniques - Surge immunity test," ed, 2014.

[27] W. Kai, D. Pommerenke, R. Chundru, T. V. Doren, F. P. Centola, and H. Jiu Sheng, "Characterization of human metal ESD reference discharge event and correlation of generator parameters to failure levels-part II: correlation of generator parameters to failure levels," *IEEE Transactions on Electromagnetic Compatibility,* vol. 46, pp. 505-511, 2004.

[28] ESDA, "ESD Association Standard Practice for Electrostatic Discharge Sensitivity Testing – Human Metal Model (HMM) – Component Level," ed, 2009.

[29] J. E. Barth, K. Verhaege, L. G. Henry, and J. Richner, "TLP calibration, correlation, standards, and new techniques," *IEEE Transactions on Electronics Packaging Manufacturing,* vol. 24, pp. 99-108, 2001.

[30] E. Association, "Standard Practice for Electrostatic Discharge (ESD) Sensitivity Testing, Transmission Line Pulse (TLP) Testing – Component Level " in *ESD SP5.5-TLP.*, ed, 2002.

[31] T. J. Maloney and N. Khurana, "Transmission line pulsing techniques for circuit modeling of ESD phenomena," 1985.

[32] D. Linten, V. Vashchenko, M. Scholz, P. Jansen, D. Lafonteese, S. Thijs*, et al.*, "Extreme voltage and current overshoots in HV snapback devices during HBM ESD stress," in *EOS/ESD 2008 - 2008 30th Electrical Overstress/Electrostatic Discharge Symposium*, 2008, pp. 204-210.

[33] M. X. Huo, Y. Han, S. R. Dong, J. J. Liou, K. B. Ding, X. Y. Du*, et al.*, "Investigation of turn-on speeds of electrostatic discharge protection devices using transmission-line pulsing technique," in *2008 2nd IEEE International Nanoelectronics Conference*, 2008, pp. 601-606.

[34] K. Chatty, D. Alvarez, M. J. Abou-Khalil, C. Russ, J. Li, and R. Gauthier, "Investigation of ESD performance of silicide-blocked stacked NMOSFETs in a 45nm bulk CMOS technology," in *EOS/ESD 2008 - 2008 30th Electrical Overstress/Electrostatic Discharge Symposium*, 2008, pp. 304-312.

[35] G. Boselli, J. Rodriguez, C. Duvvury, and J. Smith, "Analysis of ESD protection components in 65nm CMOS technology: Scaling perspective and impact on ESD design window," in *2005 Electrical Overstress/Electrostatic Discharge Symposium*, 2005, pp. 1-10.

[36] C. Putnam, M. Woo, R. Gauthier, M. Muhammad, K. Chatty, C. Seguin*, et al.*, "An investigation of ESD protection diode options in SOI," in *2004 IEEE International SOI Conference (IEEE Cat. No.04CH37573)*, 2004, pp. 24-26.

[37] K. Chatty, R. Gauthier, C. Putnam, M. Muhammad, M. Woo, J. Li*, et al.*, "Study of factors limiting ESD diode performance in 90nm CMOS technologies and beyond," in *2005 IEEE International Reliability Physics Symposium, 2005. Proceedings. 43rd Annual.*, 2005, pp. 98-105.

[38] W. M. Webster, "On the Variation of Junction-Transistor Current-Amplification Factor with Emitter Current," *Proceedings of the IRE,* vol. 42, pp. 914-920, 1954.

[39] U. Glaser, K. Esmark, M. Streibl, C. Russ, K. Domanski, M. Ciappa*, et al.*, "SCR operation mode of diode strings for ESD protection," in *2005 Electrical Overstress/Electrostatic Discharge Symposium*, 2005, pp. 1-10.

[40] K. Ming-Dou and K. C. Hsu, "Overview of on-chip electrostatic discharge protection design with SCR-based devices in CMOS integrated circuits," *IEEE Transactions on Device and Materials Reliability,* vol. 5, pp. 235-249, 2005.

[41] J. J. Liou, J. A. Salcedo, and Z. Liu, "Robust ESD Protection Solutions in CMOS/BiCMOS Technologies," in *2007 International Workshop on Electron Devices and Semiconductor Technology (EDST)*, 2007, pp. 41-45.

[42] S. H. Voldman, *ESD: physics and devices*: Wiley, 2004.

[43] J. A. Salcedo, H. Zhu, A. W. Righter, and J. J. Hajjar, "Electrostatic discharge protection framework for mixed-signal high voltage CMOS applications," in *2008 9th International Conference on Solid-State and Integrated-Circuit Technology*, 2008, pp. 329-332.

[44] J. A. Salcedo, J. J. Liou, Z. Liu, and J. E. Vinson, "TCAD Methodology for Design of SCR Devices for Electrostatic Discharge (ESD) Applications," *IEEE Transactions on Electron Devices,* vol. 54, pp. 822-832, 2007.

[45] S. Lombardo, J. H. Stathis, B. P. Linder, K. L. Pey, F. Palumbo, and C. H. Tung, "Dielectric breakdown mechanisms in gate oxides," *Journal of Applied Physics,* vol. 98, p. 121301, 2005.

[46] D. C. Wunsch and R. R. Bell, "Determination of Threshold Failure Levels of Semiconductor Diodes and Transistors Due to Pulse Voltages," *IEEE Transactions on Nuclear Science,* vol. 15, pp. 244-259, 1968.

[47] D. M. Tasca, "Pulse Power Failure Modes in Semiconductors," *IEEE Transactions on Nuclear Science,* vol. 17, pp. 364-372, 1970.

[48] V. M. Dwyer, A. J. Franklin, and D. S. Campbell, "Thermal failure in semiconductor devices," *Solid-State Electronics,* vol. 33, pp. 553-560, 1990.

[49] D. F. Ellis, Y. Zhou, J. A. Salcedo, J. J. Hajjar, and J. J. Liou, "Prediction and Modeling of Thin Gate Oxide Breakdown Subject to Arbitrary Transient Stresses," *IEEE Transactions on Electron Devices,* vol. 57, pp. 2296-2305, 2010.

[50] C. Yi-Kan, P. Raha, T. Chin-Chi, E. Rosenbaum, and K. Sung-Mo, "ILLIADS-T: an electrothermal timing simulator for temperature-sensitive reliability diagnosis of CMOS VLSI chips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 17, pp. 668-681, 1998.

[51] A. Amerasekera, L. v. Roozendaal, J. Bruines, and F. Kuper, "Characterization and modeling of second breakdown in NMOST's for the extraction of ESD-related process and design parameters," *IEEE Transactions on Electron Devices,* vol. 38, pp. 2161-2168, 1991.

[52] Y. Cao, U. Glaser, J. Willemen, F. Magrini, M. Mayerhofer, S. Frei*, et al.*, "ESD simulation with Wunsch-Bell based behavior modeling methodology," in *EOS/ESD Symposium Proceedings*, 2011, pp. 1-10.

[53] T. S. Speakman, "A Model for the Failure of Bipolar Silicon Integrated Circuits Subjected to Electrostatic Discharge," in *12th International Reliability Physics Symposium*, 1974, pp. 60-69.

[54] V. M. Dwyer, A. J. Franklin, and D. S. Campbell, "Electrostatic discharge thermal failure in semiconductor devices," *IEEE Transactions on Electron Devices,* vol. 37, pp. 2381-2387, 1990.

[55] Y. Zhou, M. Miao, J. A. Salcedo, J. J. Hajjar, and J. J. Liou, "Compact Thermal Failure Model for Devices Subject to Electrostatic Discharge Stresses," *IEEE Transactions on Electron Devices,* vol. 62, pp. 4128-4134, 2015.

[56] S. Irving, L. Yong, D. Connerny, and L. Timwah, "SOI Die Heat Transfer Analysis from Device to Assembly Package," in *EuroSime 2006 - 7th International Conference on Thermal, Mechanical and Multiphysics Simulation and Experiments in Micro-Electronics and Micro-Systems*, 2006, pp. 1-6.

[57] R. T. Dennison and K. M. Walter, "Local thermal effects in high performance bipolar devices/circuits," in *Proceedings of the Bipolar Circuits and Technology Meeting*, 1989, pp. 164-167.

[58] W. Lang, "Heat transport from a chip," *IEEE Transactions on Electron Devices,* vol. 37, pp. 958-963, 1990.

[59] Y. Çengel, *Heat Transfer: A Practical Approach*: McGraw-Hill, 2003.

[60] Y. Zhou, T. Weyl, J. J. Hajjar, and K. P. Lisiak, "ESD Simulation using Compact Models: from I/O Cell to Full Chip," in *2007 IEEE Conference on Electron Devices and Solid-State Circuits*, 2007, pp. 53-58.

[61] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed.: Wiley, 2006.

[62] A. Amerasekera and J. A. Seitchik, "Electrothermal behavior of deep submicron nMOS transistors under high current snapback (ESD/EOS) conditions," in *Proceedings of 1994 IEEE International Electron Devices Meeting*, 1994, pp. 455-458.

[63] M. Miao, Y. Zhou, J. A. Salcedo, J. J. Hajjar, and J. J. Liou, "A New Method to Estimate Failure Temperatures of Semiconductor Devices Under Electrostatic Discharge Stresses," *IEEE Electron Device Letters,* vol. 37, pp. 1477-1480, 2016.

[64] M. Stockinger and J. W. Miller, "Characterization and modeling of three CMOS diode structures in the CDM to HBM timeframe," in *2006 Electrical Overstress/Electrostatic Discharge Symposium*, 2006, pp. 46-53.

[65] A. Amerasekera, M. C. Chang, J. A. Seitchik, A. Chatterjee, K. Mayaram, and J. H. Chern, "Self-heating effects in basic semiconductor structures," *IEEE Transactions on Electron Devices,* vol. 40, pp. 1836-1844, 1993.

[66] C. L. Ma and P. O. Lauritzen, "A simple power diode model with forward and reverse recovery," *IEEE Transactions on Power Electronics,* vol. 8, pp. 342-346, 1993.

[67] Y. C. Liang and V. J. Gosbell, "Diode forward and reverse recovery model for power electronic SPICE simulations," *IEEE Transactions on Power Electronics,* vol. 5, pp. 346-356, 1990.

[68] Z. Pan, D. Schroeder, S. Holland, and W. H. Krautschneider, "Understanding and Modeling of Diode Voltage Overshoots During Fast Transient ESD Events," *IEEE Transactions on Electron Devices,* vol. 61, pp. 2682-2689, 2014.

[69] C. C. McAndrew, "R3, an Accurate JFET and 3-Terminal Diffused Resistor Model," *Proc. Nanotech WCM,* vol. 2, pp. 86-89, 2004.

[70] Y. Yue, Z. Hu, J. Guo, B. Sensale-Rodriguez, G. Li, R. Wang*, et al.*, "InAlN/AlN/GaN HEMTs With Regrown Ohmic Contacts and f_T of 370 GHz," *IEEE Electron Device Letters,* vol. 33, pp. 988-990, 2012.

[71] I. C. Kizilyalli, A. P. Edwards, O. Aktas, T. Prunty, and D. Bour, "Vertical Power p-n Diodes Based on Bulk GaN," *IEEE Transactions on Electron Devices,* vol. 62, pp. 414-422, 2015.

[72] K. Nomoto, Z. Hu, B. Song, M. Zhu, M. Qi, R. Yan*, et al.*, "GaN-on-GaN p-n power diodes with 3.48 kV and 0.95 mΩ-cm^2: A record high figure-of-merit of 12.8 GW/cm^2," in *2015 IEEE International Electron Devices Meeting (IEDM)*, 2015, pp. 9.7.1-9.7.4.

[73] K. Nomoto, B. Song, Z. Hu, M. Zhu, M. Qi, N. Kaneda*, et al.*, "1.7-kV and 0.55-mΩ•cm^2 GaN p-n Diodes on Bulk GaN Substrates With Avalanche Capability," *IEEE Electron Device Letters,* vol. 37, pp. 161-164, 2016.

[74] M. Miao, J. J. Liou, B. Song, K. Nomoto, H. G. Xing, J. A. Salcedo*, et al.*, "Investigation of forward transient characteristics of vertical GaN-on-GaN p-n diodes," in *2016 IEEE International Nanoelectronics Conference (INEC)*, 2016, pp. 1-2.