CHARACTERIZATION, MODELING, AND DESIGN OF ESD PROTECTION CIRCUITS

By

STEPHEN G. BEEBE

March 1998

Technical Report No. xxxxxxx

Prepared under
Semiconductor Research Corporation Contract 94-SJ-116
Semiconductor Research Corporation Contract 94-YC-704

Special support provided by
Advanced Micro Devices, Sunnyvale, California

Abstract

For more than 20 years, susceptibility of integrated circuits to electrostatic discharge (ESD) has warranted the use of dedicated on-chip ESD protection circuits. Although the problem of ESD in integrated circuits (ICs) has received much attention industry-wide since the late 1970s, design of robust ESD circuits remains challenging because ESD failure mechanisms become more acute as critical circuit dimensions continue to shrink. In the past increased sensitivity of smaller devices, coupled with a lack of understanding of ESD phenomena and the consequent trial-and-error approach to ESD circuit design, resulted in design of ESD protection effectively starting from scratch in each new technology. Now, as life cycles of new technologies continue to decrease, better analysis capabilities and a systematic design approach are essential to accomplishing the increasingly difficult task of adequate ESD protection-circuit design.

This thesis reviews the problems of ESD in the IC industry and the standard models used to characterize ESD protection-circuit performance. Previous approaches to ESD circuit design are discussed, including design theory and specific design examples. Transmission-line pulsing (TLP), a relatively new ESD characterization and analysis test method, is presented. This test method offers many advantages over standard characterization techniques, including the ability to extract critical parameters of an ESD protection circuit and to determine the failure level of a circuit over a wide range of ESD stress durations. Dependencies of ESD circuit performance on critical process parameters of a CMOS technology are discussed. Two-dimensional numerical device simulation techniques are presented for modeling ESD in circuits, including electrothermal simulation and a curve-tracing algorithm, detailed in an appendix, used to guide simulations through complex current-voltage (I-V) curves. Results are given for TLP experiments run on parametric ESD structures created in a 0.5μm CMOS technology, including MOSFET snapback I-V

characteristics and failure thresholds. Results of calibrated simulations are also presented and compared to experiments. Details of the simulation calibration procedure are provided.

A design methodology for multiple-fingered CMOS ESD protection transistors is presented. The methodology employs empirical modeling to predict the I-V characteristics and ESD withstand level of a circuit given the circuit's layout parameters. A critical correlation between transmission-line pulse withstand current and human-body model (HBM) withstand voltage is demonstrated. Quantitative prediction is achieved for HBM withstand voltages in a 0.35μm-technology SRAM circuit. Optimization of protection-transistor layout area for a given ESD withstand level is illustrated. The thesis concludes with a discussion of future work and issues pertaining to the impact of ESD on future technologies.

Acknowledgments

This work could not have been accomplished without the assistance of many people. Primary thanks go to Professor Robert Dutton of Stanford University and to Ben Tseng of Advanced Micro Devices for their foresight in recognizing the opportunity to apply recently introduced simulation techniques to a reliability problem which becomes increasingly important to the integrated-circuit industry with every new technology. Forming an industrial partnership with AMD provided an excellent opportunity to conduct research on a leading-edge technology, and during my tenure there as a graduate student I received help from many people I am now happy to call my coworkers. Kurt Taylor took me under his management basically sight unseen, but his increasing interest in ESD generated many enlightening discussions and his effort in laying out test structures made the experimental work in this thesis possible. Kurt must also be credited for steering me towards the design-of-experiments modeling approach. Others deserving thanks include, but are not limited to, Dave Forsythe for discussions regarding simulator calibration, Pete Williams for helping take experimental data, Dave Greenlaw, Ken Hicks, D.H. Ju, Kelvin Lai, and Ian Morgan.

At Stanford, special gratitude goes to Zhiping Yu and Ronald Goossens, both of whom provided not only technical guidance but career advice and friendship. Ronald must also be credited (or blamed) for some of the most grueling runs since my college track days. I would like to thank the members of my reading and orals committees: Drs. Robert Dutton, Bruce Wooley, Kenneth Goodson, and William Dally. Fellow students who provided help include Aon Mujtaba for discussion of mobility modeling, Chiang-Sheng Yao for discussion of impact-ionization modeling, Richard Williams for discussion of any manner of device physics, and Zak Sahul for providing answers to many questions regarding the Unix operating system. Support was also provided by Chris Quinn and Todd Atkins of

EECF/EECNS and by Fely Barrera and Maria Perea. To the other members of the TCADre, I have not forgotten you and will often reminisce on those three character-building softball seasons which taught us that victory is sweeter when tasted less often.

Contents

Abstract **iii**

Acknowledgments **v**

List of Figures **xi**

List of Tables **xvi**

1 Introduction **1**

1.1 ESD in the Integrated Circuit Industry 3

1.2 Characterizing ESD in Integrated Circuits 4

1.3 Protecting Integrated Circuits from ESD 6

1.4 Numerical Simulation 8

1.5 Design Methodology 11

1.6 Outline and Contributions 12

2 ESD Circuit Characterization and Design Issues **15**

2.1 Classical ESD Characterization Models and Industrial Testing 16

2.2 Transmission Line Pulsing 19

2.2.1 MOSFET Snapback I-V Curve 22

2.2.2 Failure Power vs. Time to Failure 29

2.2.3 Leakage Current Evolution 33

2.2.4 Advanced TLP Setup 35

2.3 Overview of Protection Circuit Design 37

2.4 Dependence of Critical MOSFET I-V Parameters on Process and Layout . . . 44

2.5 Design Methodology 49

3 Simulation: Methods and Applications **55**
3.1 Lattice Temperature and Temperature-Dependent Models 57
3.1.1 Mobility and Impact Ionization Models 58
3.1.2 Analysis of Thermal Assumptions 61
3.2 Curve Tracing 64
3.3 Mixed Mode Simulation 67
3.4 Previous ESD Applications 70
3.5 Extraction of MOSFET I-V Parameters 74
3.6 Extraction of MOSFET P_f vs. t_f Curve 77
3.7 Simulation of Dielectric Failure and Latent ESD Damage 86
4 Simulation: Calibration and Results **95**
4.1 Calibration Procedure 96
4.1.1 Structure Definition 96
4.1.2 Calibration of MOSFET Characteristics 99
4.1.3 Calibration of the Snapback I-V Curve 107
4.1.4 Calibration of Thermal Failure 116
4.2 MOSFET Snapback I-V Results 120
4.3 Device Failure Results 125
4.4 Design Example 134
5 Design and Optimization of ESD Protection Transistor Layout **139**
5.1 Methodology 141
5.1.1 Characterization of Test Structures 141
5.1.2 Correlation of TLP to the Human Body Model 143
5.1.3 Development of Second-Order Linear Model 148
5.1.4 Identification of Critical Current Paths 152
5.2 Application 154
5.3 Analysis 156
5.3.1 Model Terms 156
5.3.2 SRAM Model Prediction 158
5.4 Optimization 160
5.5 Summary of Design Methodology 162

6 Conclusion **165**

6.1 Contributions 166

6.1.1 Transmission Line Pulsing 166

6.1.2 Numerical Device Simulation 167

6.1.3 Design Methodology 168

6.2 Future Work 168

6.2.1 Characterization 168

6.2.2 Modeling 170

6.2.3 Design 171

A Tracer User's Manual **173**

A.1 Command Line 174

A.2 Trace File 174

A.3 CONTROL Card 175

A.3.1 Description 175

A.3.2 Syntax 175

A.3.3 Parameters 175

A.3.4 Examples 176

A.4 FIXED Card 177

A.4.1 Description 177

A.4.2 Syntax 177

A.4.3 Parameters 177

A.4.4 Examples 177

A.5 OPTION Card 178

A.5.1 Description 178

A.5.2 Syntax 178

A.5.3 Parameters 178

A.5.4 Examples 181

A.6 SOLVE Card 182

A.6.1 Description 182

A.6.2 Syntax 182

A.6.3 Parameters 182
A.6.4 Examples 183
A.7 Input Deck Specifications 185
A.7.1 Load and Solve Cards 185
A.7.2 Contact Card 185
A.7.3 Method Card 186
A.7.4 Options Card 186
A.8 Data Format in Output Files 186
A.9 Examples 187
A.9.1 BV_{CEO} 187
A.9.2 GaAs MESFET 192

Bibliography **199**

List of Figures

1.1 ESD protection circuits in a CMOS technology 8

2.2 Circuit model for the HBM and MM and SPICE3-generated short-circuit HBM output current waveform 17

2.3 SPICE-generated short-circuit MM output current waveforms 18

2.4 Circuit model for the CDM 20

2.5 TLP schematic and equivalent circuit 21

2.6 Qualitative I-V curve for an NMOS transistor subjected to a positive ESD pulse 23

2.7 Depiction of second snapback in a qualitative transient I-V curve 26

2.8 A screen capture of a Tektronix TDS 684A digitizing oscilloscope for a circuit response to TLP 27

2.9 Screen capture of breakdown and snapback 28

2.10 Screen capture of second breakdown 28

2.11 3D thermal box model. 29

2.12 A qualitative schematic of input power-to-failure vs. time-to-failure predicted by an analytical thermal model 32

2.13 Qualitative plot of device leakage evolution vs. stress-current level of a TLP experiment 34

2.14 Advanced TLP schematic and equivalent circuit 36

2.15 ESD diode protection circuit in a CMOS technology and use of a series resistor in combination with diode protection 39

2.16 CMOS input and output protection 40

2.17 Gate-bouncing techniques. 42

2.18 Combination resistor/transistor ESD input protection circuit. 43

2.19 Layout of a multiple-finger NMOS transistor between input and V_{SS} 44

2.20 Circuit diagram of CMOS input protection using multifinger structures. . . 50

2.21 Qualitative TLP I-V curve for an NMOS multifinger structure subjected to a positive ESD pulse 52

3.22 Qualitative plot of impact-ionization rates for electrons and holes. 60

3.23 Schematic representation of various types of simulator bias specification. 64

3.24 Schematic of a general device with external load and voltage; adapting the load line 66

3.25 Projection and recalibration. 67

3.26 Mixed-mode circuit model for an NMOS transistor subjected to the human-body model 69

3.27 Mixed-mode circuit model for a multiple-finger ESD NMOS structure subjected to the human-body model 69

3.28 ESD protection circuit used for SPICE simulations by Chatterjee et al. [33] 71

3.29 I-V curves for curve-tracing (solid line) and TLP (points) simulations for a 20/0.5μm MOSFET 76

3.30 I-V curves of a single TLP simulation and of points resulting from a group of TLP simulations. 77

3.31 The dependence of the steady-state change in peak temperature, ΔT_{ss}, on b/c 80

3.32 Simulated 1/ΔT vs. time and ΔT vs. time curves for various length/width ratios in a uniformly doped semiconductor region with a constant applied power 81

3.33 Power to failure, normalized by a, κ, and ΔT, is plotted vs. time to failure for the 2D and 3D implementations of the thermal box model 83

3.34 The maximum electric field in the gate oxide of an ESD-protection MOSFET subjected to a square pulse with a 3ns rise time is plotted vs. time 87

3.35 A qualitative plot of time-to-failure vs. stress voltage 89

3.36 Gate current vs. time for 50/0.75μm MOSFETs with (a) 10KΩ gate resistor and (b) grounded gate. 91

3.37 A constant-temperature contour is plotted for every 200K increment in temperature for a simulation structure at the time of peak ESD stress 92

3.38 A contour within which the intrinsic carrier concentration, n_i, is greater than the background doping level is drawn for a simulation structure at the time of peak ESD stress 93

4.39 This example of a MEDICI-generated grid shows the concentration of grid points in the channel, LDD, and junction regions 98

4.40 Qualitative depiction of I-V curves used for MOSFET calibration. 100

4.41 I-V points from the transmission-line pulse sweep of a standard characterization structure 108

4.42 Device current per width vs. device voltage for a dc-sweep simulation of the standard structure. 111

4.43 Simulated I-V sweep for T=297K boundary conditions 114

4.44 Device voltage and current vs. time for a transient simulation of the 100/0.75μm structure 118

4.45 Experimental and simulated snapback resistance, R_{sb}, vs. contact-to-gate spacing for a 50/0.75μm MOSFET test structure 121

4.46 Experimental snapback resistance, R_{sb}, vs. inverse gate width. 122

4.47 Experimental and simulated snapback voltage, V_{sb}, vs. gate length for 20μm-wide test structures 123

4.48 Simulated trigger voltage, V_{t1}, vs. gate resistance, R_{gate}, for the 50/0.75μm-wide test structure 125

4.49 Power to failure, P_f, (a) and current to failure, I_f, (b) vs. device width for 0.75μm test structures 126

4.50 Simulated and experimental power-to-failure, P_f, vs. contact-to-gate spacing for 50/0.75μm test structures 128

4.51 Experimental current-to-failure, I_f, vs. contact-to-gate spacing for 50/0.75μm test structures 129

4.52 Power at second breakdown, P_{t2}, vs. time to breakdown, t_2, for a 25/0.75μm structure 131

4.53 Experimental power-to-failure (a) and current-to-failure (b) vs. time-to-failure, t_f, for 50/0.75μm test structures with varying CGS 133

5.54 Snapback I-V curve for a 50/0.6μm NMOS transistor generated by TLP . . 141

5.55 Layout of a four-fingered ESD structure 143

5.56 Normalized (divided by width) withstand current vs. drain-side CGS 147

5.57 Normalized withstand current vs. number of 50/0.6μm fingers 148

5.58 Example of a complete second-order linear equation modeling the response of a variable with three factors 149

5.59 Schematic of critical ESD protection circuits in a chip with split power supplies and separate clock supply 153

5.60 Catalyst model graph for Lot 1 V_{sb}, R_{sb}, and $I_{TLP,ws}$. 157

5.61 Calculated minimum area of transistor source/drain diffusion needed for 5kV HBM protection 162

5.62 Block diagram of ESD circuit design methodology 163

A.63 The input file, bvceo.pis, for the BV_{CEO} example 188

A.64 The trace file, bvceo.tra, for the BV_{CEO} example 189

A.65 The output file, bvceo.out, for the BV_{CEO} example 190

A.66 Collector current vs. collector voltage for the BV_{CEO} example 191

A.67 The mesh generation and eliminate statements of the file mes.pis for the GaAs MESFET example 193

A.68 The second half of the file mes.pis for the GaAs MESFET example 194

A.69 The input file, mesvg.5.pis, for the GaAs MESFET example 195

A.70 The trace file, mesvg.5.tra, for the GaAs MESFET example 196

A.71 Drain current vs. drain voltage for the GaAs MESFET example 196

A.72 The output file, mesvg.5.out, for the GaAs MESFET example 197

List of Tables

2.1 Dependence of critical I-V parameters on process and layout 46

5.2 Experimental and modeled SRAM HBM withstand voltages 156

Chapter 1
Introduction

Electrostatic discharge (ESD) is one of the most important reliability problems in the integrated circuit (IC) industry. Typically, one-third to one-half of all field failures (customer returns) are due to ESD and other failures known collectively as electrical overstress (EOS) [1-3]. As ESD damage has become more prevalent in newer technologies due to the higher susceptibility of smaller circuit components, there has been a corresponding increase in efforts to understand ESD failures through modeling and failure analysis. This has resulted in a greater industry-wide knowledge of ESD mechanisms and thus a greater ability to design robust ICs which sustain fewer field failures. Despite these efforts, there are still ESD-related problems which are not well understood, especially the phenomenon of “latent damage.”

There are two ways to reduce IC failures due to ESD. One is to ensure proper handling and grounding of personnel and equipment during manufacturing and usage of packaged chips, i.e., to prevent ESD events from occurring. The other approach is to connect protection circuits (almost always on-chip) to the pins of a packaged IC which will divert high currents away from the internal circuitry and clamp high voltages during an ESD stress. A chip manufacturer has limited control over a customer’s handling of its product, so incorporating effective protection circuitry is essential. Since the spectrum of stresses under the label of EOS/ESD is broad and the amplitude of stress is virtually unlimited, it is not possible to guarantee total EOS/ESD immunity. However, through the proper design of protection circuitry the threshold of sustainable stress can be significantly increased, resulting in improved reliability of ICs.

Designing ESD protection circuits becomes more challenging as device dimensions shrink, particularly in MOS technologies [40,57]. As ICs become smaller and faster, susceptibility of the protection circuits to damage increases due to higher current densities

and lower voltage tolerances. Use of lightly doped drains (LDDs) and silicide in newer technologies exacerbates these problems. If the LDD diffusions are shallower than the source/drain diffusions, then for a given current level there is a greater current density in the LDD region, which means there is more localized heating and therefore a greater chance of damage during an ESD stress [4-7]. This effect has been verified with simulations as well as through failure analysis. Similarly, silicided source/drain diffusions lead to current localization by concentrating current flow at the surface of devices as well as reducing the ballasting resistance needed to distribute the current [6-9]. Finally, the thinner gate oxides of newer MOS processes are more susceptible to high-field stress, i.e., dielectric breakdown.

Typically the design of ESD protection is an empirical, trial-and-error procedure in which several variations of a circuit or types of circuits are laid out, processed, packaged, and tested on a simple pass/fail basis. This approach is time consuming and does not facilitate the evolution of protection circuits in future technologies. A better design methodology includes a more complex testing technique and modeling of ESD circuit behavior in order to provide understanding of the functionality of the transistors, diodes, and lumped capacitors and resistors making up the circuit as well as to extract critical parameters of the circuit. In conjunction with a relatively small array of test structures, proper modeling can be used to design an optimum protection circuit as well as predict the performance of similar circuits in next-generation technologies. Recent advances in two-dimensional numerical device simulation have made possible the modeling of ESD events. These simulations predict the device's current-voltage response to an ESD stress and provide analysis capabilities which suggest how and where a protection device will fail.

The focus of this thesis is on characterization, modeling, and design of ESD protection devices in a state-of-the-art silicon CMOS technology using advanced testing techniques and numerical simulation. MOS processes are studied because MOS is prominent in today's advanced technologies. This chapter is meant to create the context in which the project task is undertaken by introducing the phenomena of ESD in the IC industry, classical and novel characterization techniques, various CMOS protection circuits, and the use of numerical device simulation to model ESD phenomena and design ESD protection circuits. An outline of the thesis and a list of its contributions are presented at the end of the chapter.

1.1 ESD in the Integrated Circuit Industry

Electrical overstress is defined as damage to a product caused by exceeding data-sheet maximum ratings [10]. EOS usually leads to gross damage in an integrated circuit resulting from high-energy events such as electrostatic discharge, electromagnetic pulses, lightning, or reversal of power and ground pins. EOS failure mechanisms fall into the two broad categories of thermally induced failures and high electric-field failures [11]. The duration of an EOS event may be anywhere from less than one nanosecond to one millisecond and longer. Long EOS events can lead to damaged areas such as blown metal lines, cavities in the silicon, or discoloration of silicon due to local heating with a characteristic radius of 100μm or greater [10]. This damage leads to either a reduction in IC performance (e.g., increased leakage current on one or more pins) or total circuit failure.

The region of EOS phenomena with stress times of less than one nanosecond up to a few hundred nanoseconds is known as electrostatic discharge. (Although EOS covers a large range of phenomena including ESD, it is common to refer to the time range of 100ns and less as the ESD regime and the time range greater than 1μs as the EOS regime, with a sort of transition region from ESD to EOS between 100ns and 1μs.) ESD is a relatively rapid, high-current event resulting from the high voltage created when electrostatic charges are rapidly transferred between bodies at different potentials. ESD usually leads to relatively subtle, localized damage sites extending less than 10μm.

As stated previously, there are two main dangers of ESD stress. One is the danger of gate oxide dielectric breakdown due to the high voltage seen during an ESD event. In today's MOS technologies, gate oxides are on the order of 100Å thick, which given an SiO_2 dielectric strength of $8X10^6$ V/cm implies that a stress of 8V is enough to cause oxide damage. In a typical CMOS technology, the thin gates of an input buffer are tied directly to the input pin and thus are especially vulnerable to oxide breakdown. Dielectric breakdown is also of concern within the protection circuits since thin-gate MOS devices are commonly used. The other form of damage created by ESD stress is melting of material due to Joule heating, which refers to the resistive heat generated by a current moving through an electric field ($H = \mathbf{J} \cdot \mathbf{E}$, where H is the heat flow or power density). If the high current of an ESD event is sufficiently localized in an area of high electric field, thermal runaway (also called second breakdown) will result [12,13], leading to either

device failure, i.e., shorts and opens, or the more subtle damage of increased leakage. Second breakdown is a positive-feedback process and is a well-known phenomena in power devices. A physical explanation of second breakdown is given in Chapter 2.

Dielectric failure and thermal failure are generally considered to be catastrophic, i.e., the IC is no longer functional after the ESD stress. However, as has been noted above there is another type of ESD damage referred to as latent damage, a phenomenon which is well documented but is not well understood. Latent damage consists of increased leakage current or reduced oxide integrity, without loss of functionality, of a stressed circuit [4, 14, 15]. A latent ESD failure is defined as "a malfunction that occurs in use conditions because of earlier exposure to ESD that did not result in an immediately detectable discrepancy [16]." Latent damage is often bake-recoverable, i.e., reversible. Low-level leakage (an increase in leakage which remains below the failure threshold), also referred to as soft failure, may be due to injection of hot carriers into the gate oxide, which would cause a threshold-voltage shift, or to damage in the silicon resulting from localized melting, or to both. A small damage site could act like a high-resistance filament across a diode junction, thereby increasing the leakage current to a significant but non-catastrophic level. It is certainly possible for second breakdown to occur, and even for melting to occur, without catastrophic failure if there is not enough energy in an ESD pulse to cause widespread damage. Polgreen et al. [8] found this to be true for MOSFETs with widths below a certain critical value. They postulated that a certain amount of total current is needed to cause widespread device damage. In narrow devices, when a hot spot forms all of the available current rushes to the spot, but there is not enough total current to cause catastrophic damage. Extensive damage will not occur until the device is driven deeper into second breakdown by being stressed with a higher current.

1.2 Characterizing ESD in Integrated Circuits

In order to characterize the susceptibility of an IC to ESD damage, the IC must be tested using models which accurately simulate real ESD events. These models should be standardized so that testing is consistent and reliability can be defined quantitatively--attributes which make a figure of merit and design goals possible. Actual ESD stresses occur during wafer fabrication, packaging, testing, or any other time the circuit comes in contact with a person or machine. The majority of stresses occur between two pins of an IC package when the chip is not powered up, a fact reflected in the setup of ESD

characterization tests [58]. Specific tests are designed to model specific events such as human handling, machine handling, or field induction.

The most common industrial tests used to measure ESD robustness are the human-body model (HBM), the machine model (MM), and the charged-device model (CDM) [17,34]. These models will be described in detail in Chapter 2. Briefly, the human-body model, also known as the finger model, consists of charging a capacitor to a high voltage (say, 2000V) and then discharging the capacitor through a series resistor into an I/O or supply pin of a packaged IC with another pin grounded and all other pins floating. The capacitor and resistor values are selected to generate a pulse similar to that generated by an electrostatically charged human touching the pins of an IC, with a rise time of a few nanoseconds and a decay time of about 150ns. After an HBM stress is applied between two pins, the pins are biased at the operating voltage and the consequent leakage current is measured. If the leakage is greater than some predefined level (say, 1μA) then the package has failed the (2000V) HBM test. HBM testing is often the sole means of qualifying ESD reliability because the specifications of the test are standardized industry wide and because several commercial HBM testers are available.

As in the HBM, in the machine model a capacitor is charged up to a high voltage and then discharged through the pins of an IC. Unlike the HBM, however, the MM discharges the capacitor through only a very small, parasitic series resistance, resulting in an oscillatory input pulse comparable to a pulse generated by a charged metal machine part contacting an IC pin. Since the series resistance is very small, parasitic inductances and capacitances of the tester as well as the dynamic impedance of the device under test have a much larger effect on the shape of the pulse, making a standard, repeatable MM test difficult to actualize.

While device heating is the primary failure mechanism in the HBM and MM, dielectric failure is the signature of the charged-device model. Due to the sub-nanosecond rise time of the CDM pulse, protection devices may not be able to turn on and clamp the input voltage to a safe level before high-field oxide damage occurs. The CDM test, which consists of charging a substrate (ground) pin of a package using a voltage source, removing the voltage source, and then discharging the package by shorting a different pin, is meant to simulate the electrostatic charging of a package due to improper grounding and the subsequent discharging when a low-resistance path becomes available. Though much

work needs to be done to understand the mechanisms of the CDM and to develop a standardized test, the CDM is now receiving much more attention in the IC industry as a result of the past focus on prevention and protection of HBM-related ESD.

A relatively new testing technique, transmission-line pulsing (TLP), takes a different approach to characterizing ESD than the classical models described above [8,21-24]. Instead of duplicating a “real-life” event such as electrostatic discharge from a finger or machine, TLP stresses IC pins with square-wave pulses of varying magnitude and length in order to study how a protection circuit responds to stimuli throughout the EOS/ESD spectrum. Short pulse lengths (on the order of 100ns to 1μs) allow extraction of information without causing unintentional thermal damage to the device. The simple square-wave inputs of the TLP method allow easy extraction of the transient current-voltage (I-V) curve of a protection circuit. Additionally, they reveal the pulse power needed to drive a circuit into second breakdown for a given pulse length. Using a spectrum of pulse lengths, a power-to-failure vs. time-to-failure (P_f vs. t_f) curve can be extracted. The I-V and P_f vs. t_f curves are very useful in determining the overall robustness of a protection circuit and in locating the weak point of the circuit. It has been suggested that transmission-line pulsing be used as a qualifier of ESD reliability, but this will probably not happen until correlations are drawn between TLP-generated failures and the classical ESD model-generated failures (TLP-HBM correlation is demonstrated for a range of transistor designs in Chapter 5). The transmission-line pulsing method and its merits will be fully discussed in Chapter 2. Application of TLP to the study of ESD is an important topic of this thesis.

1.3 Protecting Integrated Circuits from ESD

The importance of ESD protection circuits and the increasing difficulty of designing effective circuits for new technologies were discussed at the beginning of this chapter. A protection circuit serves two main purposes: providing a current path during a high-stress event and clamping the voltage at the stressed pin below the gate-oxide breakdown level. Additionally, the protection circuit itself should not become severely damaged during an ESD event. Although the odds of having the same pair of pins stressed more than once is small, it is important that the protection circuit not become leaky and degrade chip performance. Also, in the case of output-protection circuits which double as output drivers, long-term reliability may be reduced if damage is incurred. For example, it has been shown that MOSFETs driven deep into snapback during an ESD stress may suffer

hole trapping in the gate oxide as well as interface-state generation, leading to a shift in the threshold voltage [25]. The hole trapping can increase the susceptibility to time-dependent dielectric breakdown (TDDB) of the gate oxide. (TDDB refers to the observed phenomenon that the higher a stress voltage is, the less time it takes to damage the oxide being stressed.) To avoid being damaged, protection circuits should minimize self-heating by keeping current densities and electric fields in the silicon low and prevent dielectric breakdown of the gate oxides in the protection circuit by minimizing the electric fields across the oxides.

Although this thesis focuses on protection circuits between input/output (I/O) pins and supply pins, ESD phenomena can occur across any pair of pins, e.g., I/O vs. I/O and V_{CC} vs. V_{SS}. Protection circuits are not placed between the I/O pins of a package, and even though a protection diode or transistor is usually placed between the two supply pins, there is no guarantee that an electrostatic discharge will go through this path because the circuits in the chip may provide a lower-resistance path. ESD events between I/Os and between supplies lead to “far-internal” damage, i.e., the discharge paths lead through the actual working circuit, and thus damage can occur in any number of places [20]. Modeling of this behavior and design of protection are difficult because the discharge path is not known *a priori* and more circuit elements are involved.

A few examples of ESD protection circuits are shown in Fig. 1.1. If the circuit of Fig. 1.1a is powered up, diode D1 will turn on and conduct current for any input voltage greater than $V_{CC} + V_d$, where V_d is the forward diode drop. Similarly, diode D2 will clamp any negative voltage below $V_{SS} - V_d$. If the chip is not powered up and an ESD pulse is incident between the input and, say, V_{SS}, the voltage will be clamped at either the reverse breakdown voltage of the diode for a positive pulse or at $-V_d$ for a negative pulse. The PMOS (M1) and NMOS (M2) devices of Fig. 1.1b behave similarly, with the drain-substrate junctions taking the place of the diodes. One major difference is that the drain-substrate junction reverse breakdown triggers the MOS device into a snapback mode in which the drain voltage drops due to the turn-on of the lateral parasitic bipolar transistor formed by the drain, channel, and source regions. Note that the output buffer is self protecting, i.e., the transistors of the output buffer serve as the protection circuit. Finally, Fig. 1.1c is an example of a more complex input protection circuit consisting of two NMOS devices and a well resistor. The merits of this circuit as well as a more complete description of the functionality of all the circuits are presented in Chapter 2.

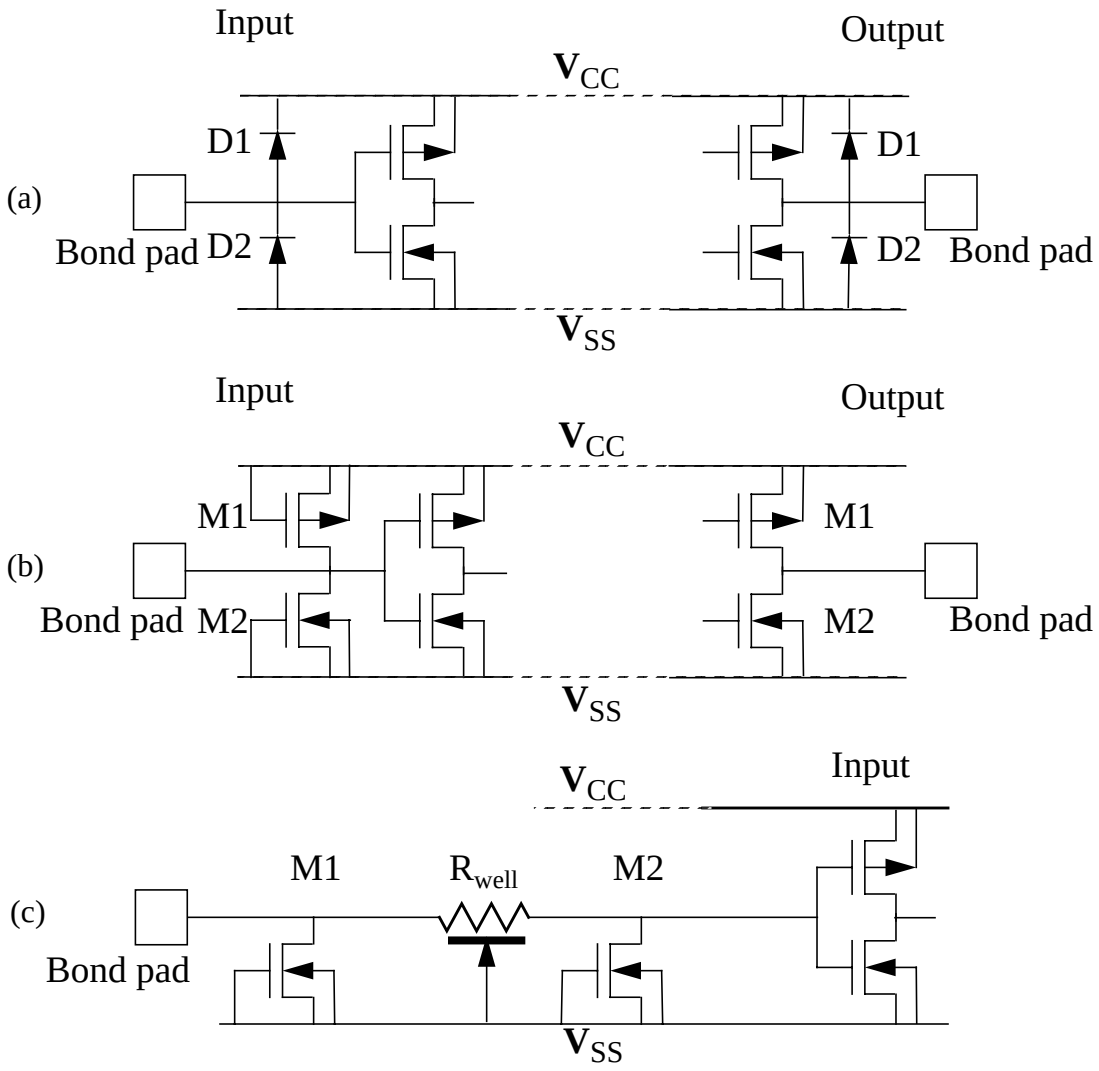

Fig. 1.1 ESD protection circuits in a CMOS technology: (a) diode protection; (b) CMOS transistor protection; (c) combination resistor/transistor protection circuit.

1.4 Numerical Simulation

Numerical device simulation is an excellent tool for studying and designing transistors and diodes in IC technologies. Two-dimensional (2D) numerical device simulators such as PISCES-IIB [26,27] allow a user to create a 2D cross section of a semiconductor device, including definition of silicon and oxide regions, doping profiles, and electrodes, and then

simulate the I-V characteristics of the device. Coefficients for various mobility models, impact-ionization models, and material parameters can be adjusted to calibrate simulations to experimental data, but even uncalibrated simulations offer a qualitative understanding of device performance. Extensive analysis capabilities let the user examine the current density, electric field, impact-ionization generation rate, temperature, and many other properties at any location in the device for any simulated I-V point. If a calibrated simulation accurately models the physics of a device, it can be used to predict the dependence of device performance on process and layout variations. Ideally, simulations take the place of large numbers of process splits and layout structures, thereby reducing the time and cost of technology development.

Simulation of ESD events is attractive because many ESD tests are destructive in nature and thus non-repeatable. In addition to predicting I-V curves, simulations can identify the point of device failure by monitoring the electric field, temperature, and other properties throughout the device during an ESD stress. The simulations can be either transient or steady state (dc). Transient simulations are used to model tests such as the human-body model, charged-device model, and transmission-line pulsing, while steady-state I-V sweeps are useful in predicting junction breakdown voltages and MOSFET snapback voltages.

The ability to model ESD phenomena was greatly expanded by two recent advances in numerical device simulation. A curve-tracing technique [28] (also known as the continuation method) used to automate the steady-state simulation of complex I-V curves was implemented as a C-program wrapper around a device simulator. Automation of complex simulations such as latchup and snapback in MOSFETs saves the time and effort needed to manually change simulation boundary conditions any time there is a sharp turn in the I-V curve. A user manual for the curve tracer is included as an appendix.

The other advance in device simulation is the incorporation of the thermal-diffusion equation [31] into the numerical equations to account for lattice temperature variation due to Joule heating and carrier generation and recombination. With this addition the device simulator solves the discreet thermal-diffusion equation in addition to the Poisson and carrier-continuity equations in either a coupled or decoupled manner. Thermal contacts and boundary conditions can be placed on the edges of the defined device, much as electrical contacts and boundary conditions are, to represent varying degrees of thermal

conduction or insulation. Since device heating occurs as a result of the high currents in ESD devices, and since second breakdown is a thermal process, lattice-temperature modeling is an integral part of simulating ESD devices.

ESD simulations are also facilitated by a mixed-mode capability which allows circuit simulation to be used in conjunction with device simulation. With this feature, device-simulator models of transistors are embedded in a SPICE-like circuit containing lumped elements such as resistors, capacitors, and voltage sources. The circuit determines the terminal voltages for the numerically simulated devices, which in turn provide the currents for the SPICE circuit [29]. Simulations for the human-body model, machine model, and other tests with complex inputs are easy to set up using the mixed-mode feature.

Several investigations have been reported on the use of 2D device simulation of EOS/ESD phenomena, but most of these have been only qualitative (examining trends rather than calibrating an actual process) or have focused on the EOS regime. For example, some studies look at how variations in process parameters (junction depths and profiles, substrate and diffusion doping levels) or layout parameters (gate length, drain contact-to-gate spacing) affect a circuit's ESD performance as measured by peak device temperature, peak $\mathbf{J} \cdot \mathbf{E}$ (power density), or some other failure signature [24,32]. In two of these studies quantitative agreement between simulated and experimental power-to-failure vs. time-to-failure (P_f vs. t_f) curves was attained, but only for one region of the EOS/ESD spectrum and only for one particular device. Other studies focus on topics such as the necessary conditions for second breakdown [13], thermally induced low-level leakage [4], the effect of pulse rise time on the trigger voltage [33], and the relative merits of using peak temperature, peak $\mathbf{J} \cdot \mathbf{E}$, and second-breakdown trigger current as failure criteria in simulated devices. A thorough discussion of past use of 2D simulation to study ESD is given in Chapter 3.

Despite the number of publications on the application of 2D numerical device simulation to ESD, there are significant applications of simulation which have remained unexplored. In general, past studies have dealt mostly with thermal failure mechanisms and not with dielectric failure or latent damage. Additionally, simulation has been used mainly as a research tool and not as a design tool. This thesis investigates the viability of these new applications by using simulation to create ESD models which accurately reflect the electrical and thermal behavior of circuits designed in a state-of-the-art industrial CMOS

process. After calibration of the 2D device models, a circuit's susceptibility to dielectric breakdown can be studied by monitoring the peak electric field in the gate oxide of the MOSFET being protected (or of the MOSFET in the protection circuit) during a simulation. Analysis of hot-carrier injection or non-catastrophic localized heating during a simulated ESD stress may be correlated to low-level leakage (the latter phenomenon has been addressed in [4]). Simulations of TLP experiments can be used to predict the critical parameters of a transient I-V curve (breakdown voltage, snapback voltage, etc.) as well as a power-to-failure vs. time-to-failure curve. Ultimately, 2D device simulation should be used as a design tool to optimize the layout parameters of a protection circuit for ESD robustness for a given CMOS process.

1.5 Design Methodology

Another approach to designing and optimizing protection circuits is to create models for transient I-V and failure parameters using statistical design of experiments. By characterizing a set of protection transistors with variations in layout, models can be developed to describe the TLP I-V parameters, TLP withstand current, and HBM withstand voltage as functions of transistor width, gate length, contact-to-gate spacing, number of poly-gate fingers, and other layout parameters. (The withstand current or voltage is defined as the maximum TLP current or HBM voltage, respectively, a structure can withstand without incurring damage. Thus, the withstand level is always slightly lower than the failure level.) A statistical design-of-experiments program is useful for determining the minimum number of test structures needed and for extracting the model equations. Once models are developed for a given technology, the performance of any ESD circuit designed in that technology can be determined.

In Chapter 5 the design-of-experiments modeling approach is presented as the basis of a complete integrated-circuit ESD design methodology. Second-order linear models are used to relate the I-V and withstand parameters (responses) to transistor layout parameters (factors). Other key parts of the methodology which are addressed include establishing a correlation between TLP withstand current and HBM withstand voltage and identifying an integrated circuit's potential ESD discharge paths. An analysis of measured ESD protection levels for a 0.35μm-technology SRAM circuit verifies that the methodology can achieve quantitative predicition of ESD performance. Chapter 5 also discusses how the second-order linear models may be used for protection-transistor optimization.

1.6 Outline and Contributions

The purpose of this thesis is to demonstrate the power of transmission-line pulsing and 2D numerical device simulation in the characterization, modeling, and design of ESD protection circuits by expanding upon earlier work in these areas and introducing new applications. Emphasis is placed on CMOS technology because it represents the leading edge of the IC industry. Design focuses on layout parameters because the ESD circuit designer is usually given a process with which to work and has no control over the junction depths, junction profiles, doping concentrations, etc. Among the contributions of this thesis are

- a quantitative analysis of the ability of 2D numerical device simulation to model experimental I-V and P_f vs. t_f curves of submicron-technology protection devices in the ESD regime and a demonstration of how simulations can be used to design ESD circuits in a state-of-the-art technology
- an investigation of the use of 2D simulation to study dielectric ESD failures and latent ESD damage
- a demonstration of the unique ESD characterization abilities of the transmission-line pulsing method
- a methodology for layout design and optimization of CMOS ESD protection circuits
- an example of the practical application of Stanford's curve-tracing program
- a calculation, based on an analytical thermal model, of the accuracy of 2D device simulation in predicting thermal failure for a range of ESD pulse times
- confirmation that transmission-line pulse and human-body model withstand levels can be correlated over at least a small transistor design space.

Chapter 2 addresses characterization and design issues of ESD circuits, starting with a detailed discussion of the classical industrial models used to qualify ESD robustness and of the applications of transmission-line pulsing. Next, the functionality of some standard protection circuits is described, including a physical explanation of the transient I-V curve of a MOSFET. The critical parameters of this I-V curve and their dependence on process and layout variables are presented, followed by a discussion of ESD circuit design methodology.

Applications of 2D device simulation to the study of ESD are presented in Chapter 3, starting with a general discussion of simulator features important to ESD modeling and then delineating specific examples. A review of some previous ESD simulation work is also given. Chapter 4 describes the calibration of the simulator ESD models and presents the results of simulations which use these models. Simulation results are compared to TLP experiments, and an example of circuit design using transmission-line pulsing and simulation models is described. In Chapter 5 the concepts of ESD circuit design methodology are re-addressed and developed in detail. Key issues include correlation of TLP and HBM withstand levels, identifying critical discharge paths, and applying design-of-experiments models to transistor optimization. Chapter 6 reviews the contributions of the thesis and discusses future work as well as the limitations of 2D device simulation in studying ESD. The principles of the curve-tracing technique and a user's manual for the curve-tracing program are given in an appendix.

Chapter 2

ESD Circuit Characterization and Design Issues

Although protective circuits were used in MOS technologies before 1970, characterization and design of ESD protection did not receive much attention until the late 1970s [34]. In early MOS processes transient stresses greater than 100V were enough to short out a gate oxide, so simple circuits were designed to shunt such stresses away from the vulnerable gates. The increase in failure thresholds from 100V to about 400V, insignificant by today's standards, was at the time enough to dramatically increase production yields and thus make ESD protection seem like an easily solvable problem. Since ESD was an issue of limited concern, little effort was made to improve ESD reliability. As a result, the increased susceptibility of shrinking technologies led to a dramatic emergence of ESD problems, fostering an industry-wide interest in enhancing ESD control during process and manufacturing, including the design of protection circuits and the development of characterization models which quantitatively test these circuits. This interest was heightened by the beginning of the annual EOS/ESD Symposium in 1979, instituted to increase awareness of electrical overstress and electrostatic discharge failures.

ESD Characterization Methods

Three of the most common industrial models used to test ICs are the human-body model, the machine model, and the charged-device model (others include the field-induced, field-enhanced [35], and capacitive-coupled [36] models). This chapter begins with a discussion of these models, followed by a detailed description of the transmission-line pulsing (TLP) characterization method. In order to understand the many uses of TLP in analyzing protection-circuit MOSFETs, a thorough examination of the MOSFET I-V curve under ESD stress is presented along with the TLP description. The focus of the chapter then shifts to design issues, beginning with an overview of protection circuits used

in CMOS technologies and then a discussion of critical parameters in protection circuits (which can be measured with TLP) and the dependence of these parameters on layout variations. Finally, the concepts of the chapter are brought together to form an ESD protection-circuit design methodology.

2.1 Classical ESD Characterization Models and Industrial Testing

The most popular model used in industry to test ESD robustness is the human-body model (HBM), also known as the finger model. The standardization of this test, first documented in 1980 and most recently updated in 1989 as military standard MIL-STD 883.C/3015.7 [17], is a result of extensive ESD research since the mid 1970s. In this model a 100pF capacitor is charged up to a certain voltage and then discharged through a 1500Ω resistor into an I/O pin of a circuit, with another pin, usually a supply or ground pin, tied to ground (Fig. 2.2a). According to the MIL-STD specification, the resulting waveform must have a rise time less than 10ns and a decay time of 150±20ns into a short-circuit load (Fig. 2.2b). The rise time is dependent upon the parasitic inductance and stray capacitance. For a HBM voltage of 1500V, the peak current would be approximately 1A. This model is meant to represent a discharge from a human finger into a pin of a circuit package. Several commercial testers which meet the military standard are available, e.g., the Hartley Autozap and IMCS 2400C ESD Sensitivity Test System, making HBM testing relatively simple.

In a typical reliability test, all the I/O pins on a package are stressed with respect to all power and ground pins with both polarities of a given HBM voltage using an industrial tester. In addition, I/O pins may be stressed vs. other I/O pins, and supply pins may be stressed vs. ground pins. Current-leakage measurements at a specified reverse voltage (usually the operating voltage) are then performed on the same sets of pins. If a 2kV HBM test is performed on all pins of a package, and the resulting leakage current of all pins is below a certain level, say 1μA, then the IC is said to be resistant to 2kV HBM. Obviously, the HBM failure threshold is dependent upon the chosen failure-current definition. With the use of this model in designing protection circuits, typical HBM failure thresholds have improved from 2kV in the early 1980s to about 6kV in the 1990s [2].

The machine model (MM), also called the Japanese model due to its origin, is similar to the human-body model: a capacitor is charged to a certain voltage and then discharged

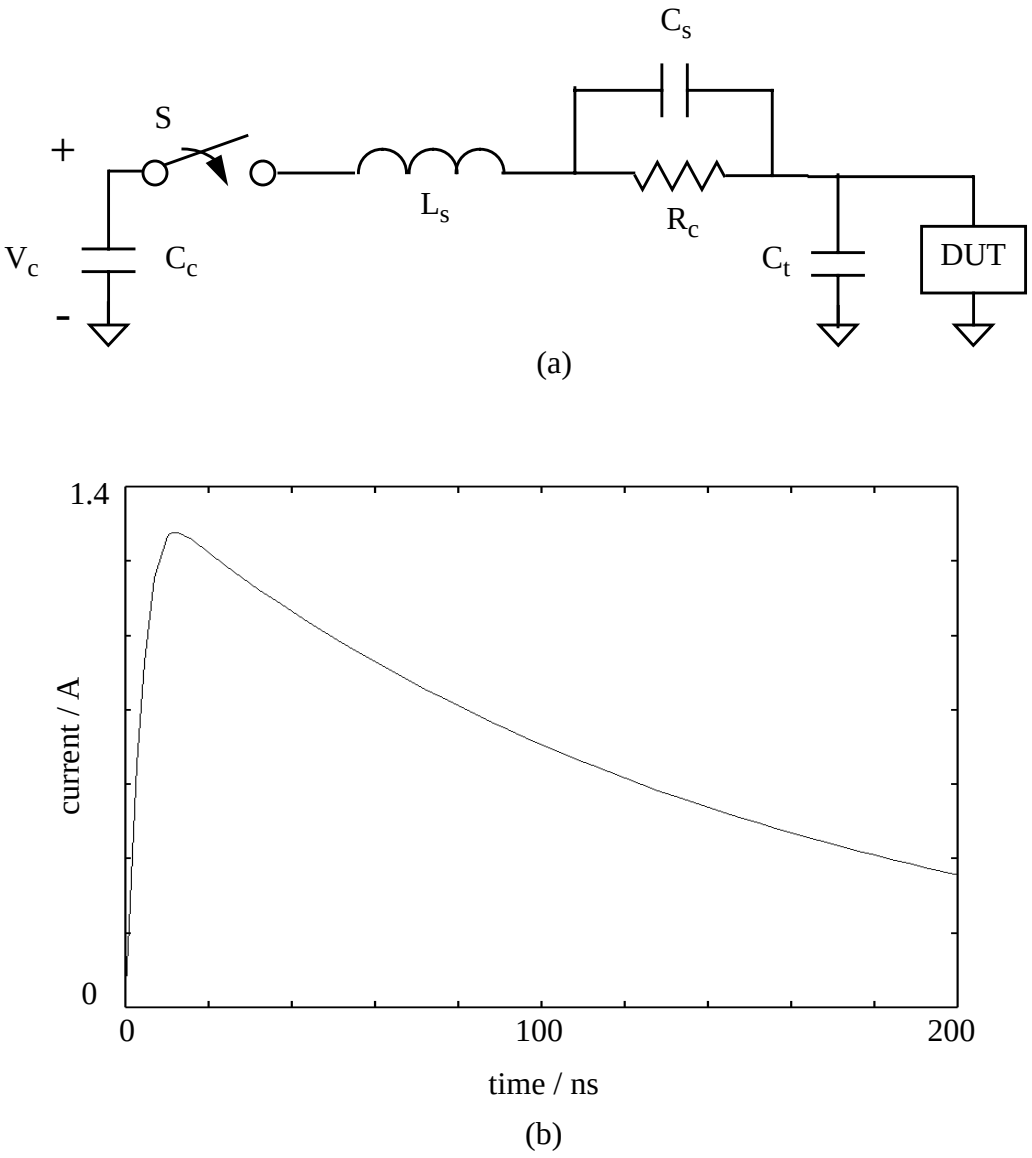

Fig. 2.2 *(a) Circuit model for the HBM and MM. Capacitor C_c is charged to the test voltage, V_c, and then discharged through R_c to the device under test (DUT) by closing switch S. Parasitic circuit elements are represented by series inductance L_s, stray capacitance C_s, and test-board capacitance C_t [18]. (b) SPICE3-generated short-circuit HBM output current waveform for V_c=2000V, C_c=100pF, R_c=1500Ω, L_s=7.5μH, C_s=1pF, and C_t=10pF.*

into a device. In this case, however, the 200pF capacitor is tied directly to the device under test, which means the 1500Ω resistor is replaced by a parasitic resistance of a few ohms and a series inductance of about 1μH. The resulting current waveform is oscillatory in nature (Fig. 2.3), with a rise time on the order of a few nanoseconds. This model simulates the discharge from a tool or machine such as a handler or marker. Unlike the HBM, there is no established standard for the MM. This is most likely because the very low series resistance implies that the dynamic impedance of the device under test and the values of the parasitic capacitance and inductance have a large effect on the MM waveform, making test reproducibility difficult [18]. Fig. 2.3 illustrates the drastic change in rise time and peak current of the waveform when the series inductance is changed from 0.5μH to 2.5μH.

In the integrated-circuit industry, the human-body model test is often the sole means of qualifying EOS/ESD reliability because it is simple to conduct and has been accepted

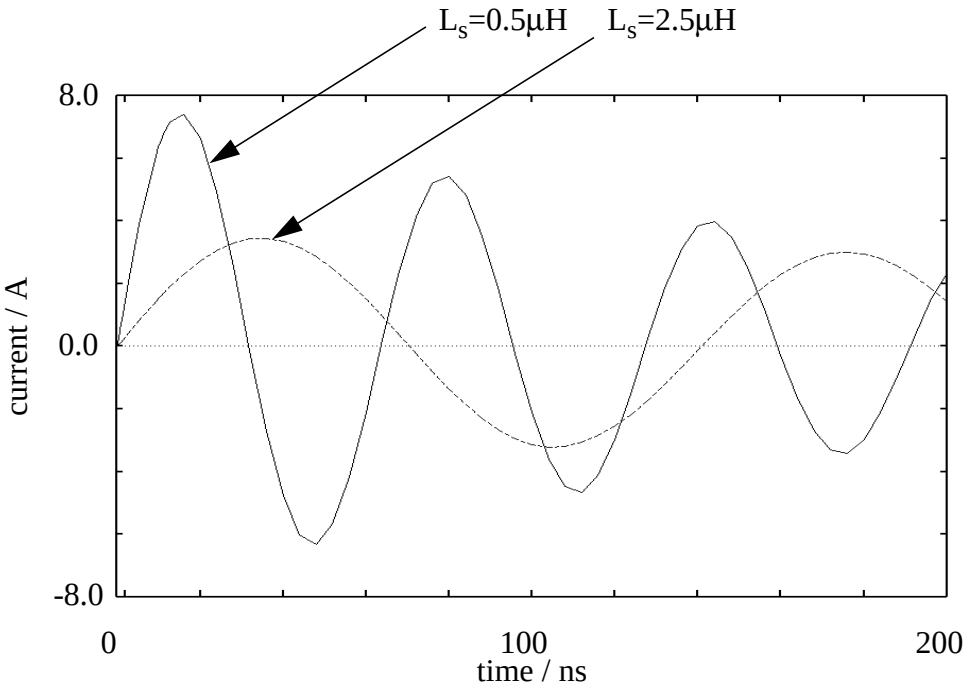

Fig. 2.3 SPICE-generated short-circuit MM output current waveforms for V_c=400V, C_c=200pF, R_c=5Ω, C_s=1pF, and C_t=10pF.

industry wide over a number of years. As long as a packaged product is resistant to HBM tests up to some level of stress, say 4kV, then it is considered to be reliable from an ESD standpoint. However, as the result of an emphasis on preventing ESD damage from human handling during production, e.g., by ensuring proper grounding of personnel and equipment and by using ESD-controlled workstations, the human-body model no longer represents the dominant failure pattern in the industry [10].

Today the main area of concern is shifting to the charged-device model (CDM), which introduces a different failure mode from that of the HBM and MM. In this model, electrostatic charge builds up on a chip due to improper grounding and then discharges when a low-resistance path becomes available. It is meant to simulate ESD phenomena of packaged ICs during manufacturing and assembly. For example, a package connected to the ground pin may be inductively charged up as it is transported along a conveyor belt, then discharged through any pin touched by a metal handler or test socket [18]. The characteristic rise time of a CDM pulse is 1ns or less, with a peak current of several amps. Since the turn-on time of MOS protection circuits is on the order of 1ns, high voltages have a chance to build up across oxides during a CDM event. Thus, damage to thin oxides (of the protection device as well as the internal gates being protected) is the signature failure of CDM events, in contrast with the thermal failure signature of the HBM.

1 A typical CDM test consists of placing a charge on a substrate (ground) pin using a voltage source, then disconnecting the voltage source and connecting a different pin through a low-inductance, low-impedance, 1Ω probe to ground (Fig. 2.4). In another

2 method referred to as the field-induced model (FIM), a charge is induced on the substrate by placing the chip on a conducting surface, then discharged through a pin via a low-impedance probe. Like the machine model, the CDM has no established standard, and there is a need for further understanding of the phenomena underlying the model. The higher ESD sensitivity of shrinking oxides and reduced susceptibility to human handling will provide the incentive for continued development of the CDM.

2.2 Transmission Line Pulsing

It is obvious from the discussion of the classical characterization models that a single type of test or figure of merit is not sufficient to guarantee robustness against all EOS/ESD failures. It is possible for a circuit to pass one type of test, say the human-body model,

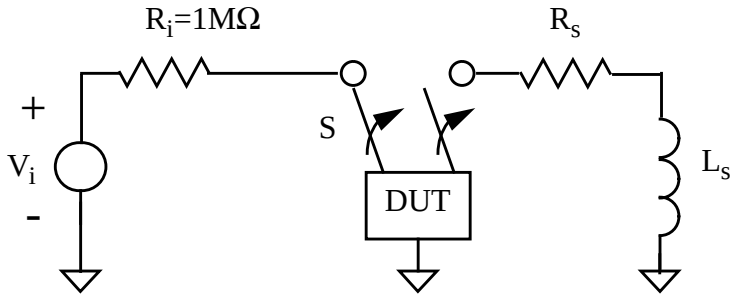

Fig. 2.4 *Circuit model for the CDM. A ground pin of the device under test (DUT) is charged to a voltage V_i, after which the switch S is thrown, removing the voltage source and connecting a different pin of the device to ground.*

while failing another, say the charged-device model [19]. It is even possible for a circuit to survive one level of a test while failing at a lower level of the same test. One well-known case is the failure window of the HBM: a device will pass HBM stresses less than 1kV and greater than 2kV up to 6kV, but will tend to fail at stresses between 1kV and 2kV. Such a case is described in [20].

There are many limitations to using the classical models to characterize ESD robustness of circuits. Foremost is that the models offer only restricted insight as to how the protection circuits work and how and where they fail. The input pulses of the HBM and other models are complex and very brief, so the response of the circuit is also complex and is hard to measure. And although the dependence of increased leakage on the test stress level is tabulated, ESD qualifiers are normally only interested in whether the leakage is above or below a predefined failure level. In short, the classical models are used as a black box with a voltage-level stimulus and a simple “pass or fail” response.

Transmission-line pulsing, a relatively new ESD characterization method, provides a way of opening up this black box. Since this technique was first introduced in 1985 [21] it has become widely used to characterize and design ESD circuits [4,8,22-24]. A schematic of a TLP experiment is shown in Fig. 2.5, in which a coaxial transmission line is charged up to a certain voltage and then discharged into an I/O pin of the device with the ground or

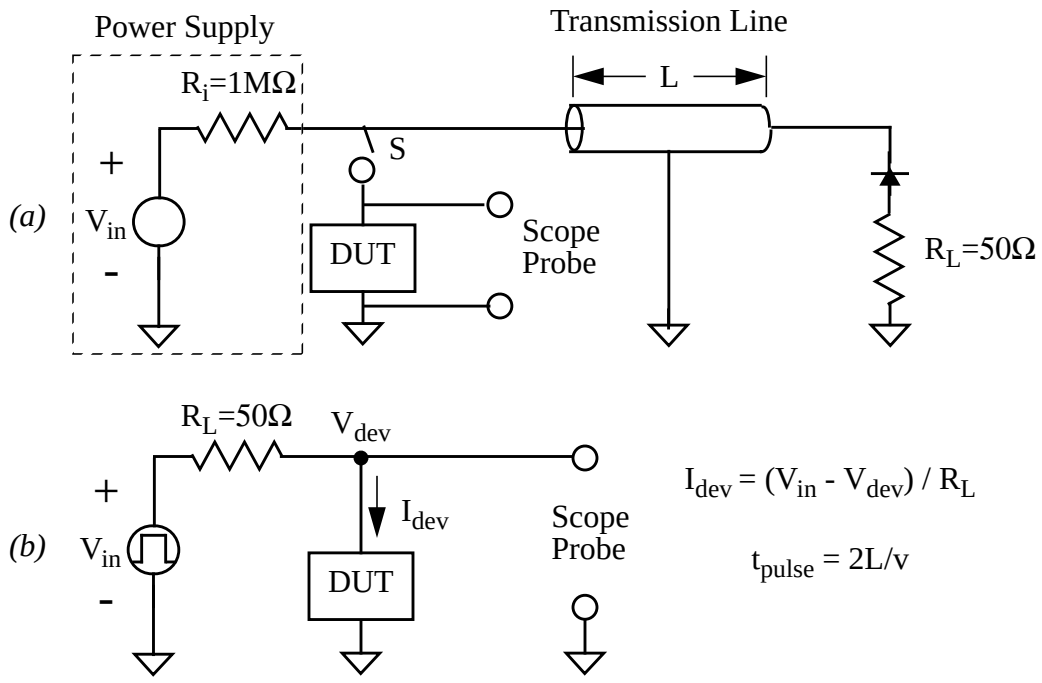

Fig. 2.5 *(a) TLP schematic: a transmission line of length L is charged to voltage V_{in} and then discharged through the device under test (DUT) when switch S closes. An oscilloscope voltage probe across the DUT monitors the circuit response. (b) Equivalent circuit of the TLP setup: the input is a square pulse of height V_{in} and duration 2L/v, where v is the phase velocity of the line.*

DISTRIBUTED CAPACITANCE

supply pin grounded and other pins open. This method is much like the HBM in that a capacitor is charged and then discharged into a circuit. However, for TLP the capacitance is distributed, thus creating a simple square-wave input on the order of 100ns long with a rise time of about 2ns. The height of the pulse is V_{in}, the power-supply voltage, and the width of the pulse is 2L/v, where L is the length of the transmission line and v is the propagation (phase) velocity of the line. If the impedance of the circuit is constant, the transmission line delivers a constant current pulse. An oscilloscope probe measures the voltage across the device; the current may also be probed or may be calculated from the input and device voltages:

$$I_{dev} = (V_{in} - V_{dev}) / R_L . \qquad (2.1)$$

To prevent multiple reflections when the device impedance is less than 50Ω, a diode and resistor are placed on the end of the line opposite the device under test (DUT).

2.2.1 MOSFET Snapback I-V Curve

Transmission-line pulsing is useful for garnering several pieces of information about an ESD protection circuit. The most obvious application is the extraction of transient current-voltage (I-V) curves of protection devices, especially MOSFETs. By pulsing a circuit with a series of increasing input voltages and plotting the resulting device voltage and current points, a characteristic I-V curve is produced. Unlike a conventional curve tracer, which would cause destructive heating with its relatively long stepped stresses, the short pulses of the TLP method allow the extraction of I-V points up to very high current levels without causing thermal damage. Of course, the time between stresses should be enough to allow complete thermal dissipation--one or two seconds is more than enough. The transient I-V curve of a protection device is very informative because it reveals what the device is doing during an ESD stress. Critical parameters of the device such as the turn-on voltage, snapback voltage, and second-breakdown trigger current (all described below) can be read directly from the curve. Although the square-wave input does not precisely model any probable ESD event, parameters of the resulting I-V curve can be correlated with susceptibility to “real” ESD stresses and with tests such as the HBM [23].

Since MOSFETs in ESD protection circuits operate in an unconventional manner, it is necessary to discuss the device’s complex I-V curve and the underlying physics to see the advantages of transmission-line pulsing analysis as well as to appreciate the device’s usefulness. Conduction of ESD current does not occur through MOS transistor action but rather via the lateral bipolar transistor in which the drain, channel, and source act as the collector, base, and emitter, respectively. The qualitative I-V characteristic of an NMOS protection device subjected to a positive ESD pulse is shown in Fig. 2.6. In the setup a voltage pulse is incident upon the drain of the device with gate, source, and substrate grounded. As the input pulse rises, the drain voltage rises until the drain-substrate junction breaks down due to impact-ionization (II) and significant current begins to flow from drain to substrate. This breakdown voltage, denoted BV_{DSS} or V_{bd}, is defined as the voltage at which the drain current reaches a critical value, usually 1μA. The substrate current consists of II-generated holes flowing from the junction to the substrate contact. Additionally, some holes will flow to the source. As this current increases, the potential of

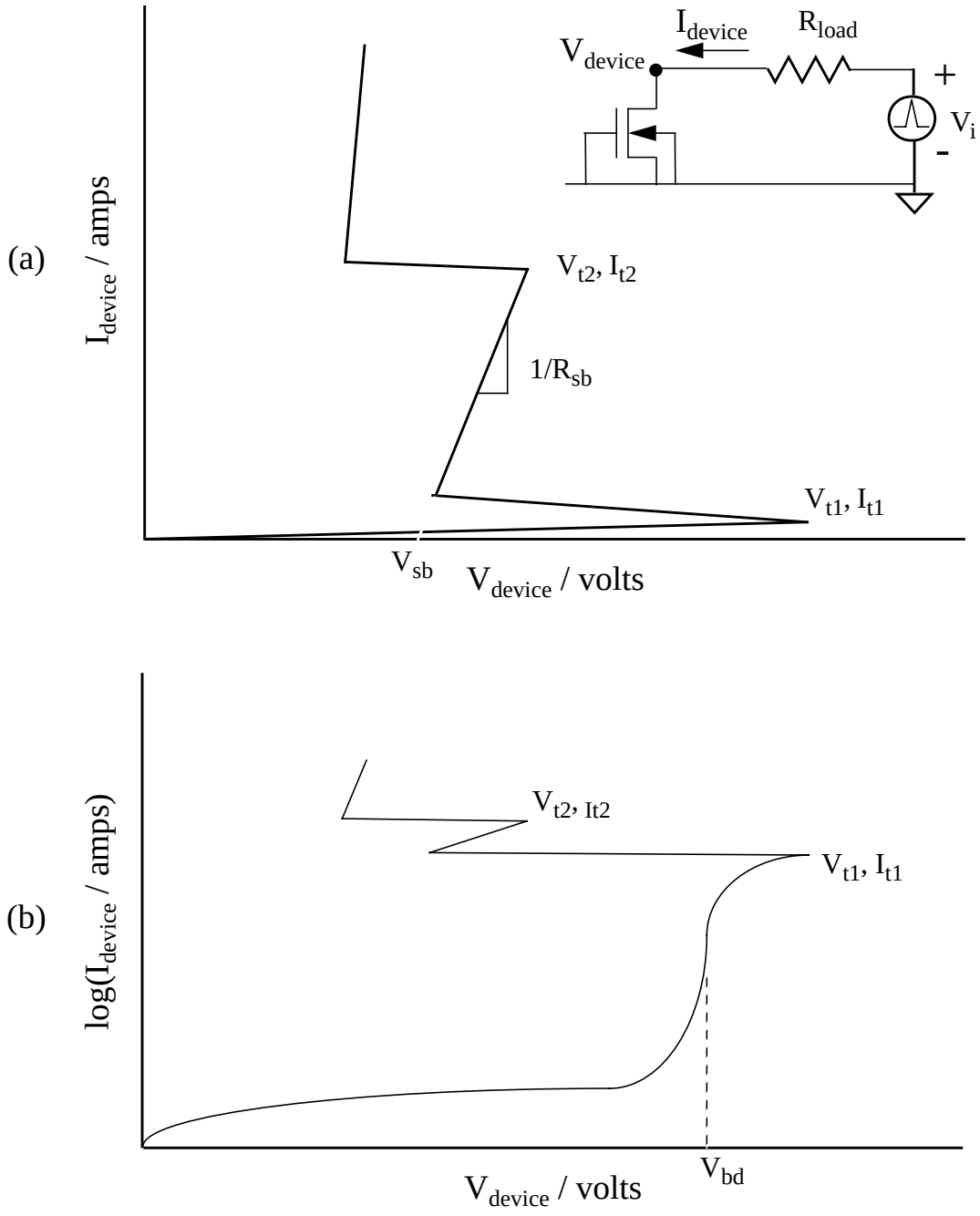

Fig. 2.6 *Qualitative I-V curve for an NMOS transistor subjected to a positive ESD pulse: (a) linear scale shows snapback trigger point (V_{t1}, I_{t1}), snapback voltage (V_{sb}) and resistance (R_{sb}), and second-breakdown trigger point (V_{t2}, I_{t2}), with circuit shown inset; (b) log scale shows the difference between device breakdown voltage (V_{bd}) and snapback trigger point.*

the substrate near the channel builds up due to the voltage drop across the substrate resistance. This resistive drop, combined with possible drops in the drain diffusion and contacts, is observed as a flattening in the I-V curve after the initial steep rise in current. At the trigger point (V_{t1}, I_{t1}) the potential in the channel reaches about 0.6V and the source-substrate junction forward biases, turning on the parasitic bipolar transistor. The suffix t_1 stands for the time it takes to reach the trigger point, which is usually on the order of 1ns but is very dependent upon the pulse height and rise time. Once this transistor turns on the drain current consists mostly of electrons injected from the source, with a small fraction of current still composed of II-generated electrons. Since a high electric field is no longer needed to maintain the current level through impact ionization alone, the drain voltage quickly drops to a level approximately equal to the BV_{CEO} of the lateral bipolar transistor. This "snapback" voltage, V_{sb}, is analogous to the hold voltage of a SCR device. (V_{sb} is actually defined as the x-intercept of the line tangent to the I-V curve near the snapback point.) To first order, the ratio of V_{bd} to V_{sb} is equal to $\beta^{1/n}$, where β is the current gain of the bipolar transistor and n is a constant on the order of 5 [22].

In the snapback mode, the current rises along a line with slope $1/R_{sb}$, where R_{sb} is the dynamic snapback impedance or "on resistance." R_{sb} is equal to the resistance of the source and drain diffusions and contacts and is usually on the order of only a few ohms. The device incurs no damage in the snapback mode unless the current level becomes high enough to trigger thermal runaway (also called second breakdown), a positive-feedback process. At the second-breakdown point (V_{t2}, I_{t2}), which occurs at time t_2, a localized hot spot forms in the region of high Joule heating ($\mathbf{J} \cdot \mathbf{E}$). As the temperature increases at this spot, resistivity increases due to mobility degradation. However, the intrinsic carrier concentration increases with temperature, and when it eventually meets and exceeds the background doping level the silicon resistivity reaches a maximum and then decreases, leading to an even higher current level and thus more heating. In the I-V curve, second breakdown is characterized by a drop in the device voltage, a result of the negative differential resistance. If there is sufficient power in the ESD pulse, enough current will rush into the hot spot to raise the temperature above the silicon melting point, thus damaging the device under stress through diffusion of dopants or formation of polysilicon boundaries upon recrystallization. Beyond the second-breakdown point the current will continue to rise very sharply (indicating very low device resistance) until a short circuit or open circuit is formed by the thermal damage.

In the simplest theory, thermal runaway and device failure follow instantaneously when the intrinsic carrier concentration exceeds the background doping concentration at a certain point in the device [12]. However, this model is too simple because it does not account for spreading resistance and the temperature dependence of mobility and impact-ionization rates. Although the resistivity at the hot spot decreases, the surrounding high-temperature region still has a high resistivity, and the overall device resistance may not decrease until there is a large area in which the intrinsic concentration is larger than the doping. For a very short pulse duration, the temperature at the hot spot may exceed the melting point and create damage without the device entering second breakdown. As mentioned in Chapter 1, even when the current density is high enough to trigger thermal runaway and the device voltage drops, for a narrow-width structure there may not be enough total current to cause major damage, i.e., leakage current greater than 1μA or a short or open circuit. Therefore, second breakdown refers to a drop in device voltage due to the negative differential resistance resulting from device heating and is not synonymous with device failure.

There is one other phenomenon which may occur in LDD MOS protection devices which has received little or no attention. It has been reported that in bipolar technologies making use of an epitaxial layer to form a lightly doped collector region (an n-p-n-n^+ transistor), two non-thermally induced snapbacks may occur during a BV_{CEO} stress [37]. The first snapback is due to the same mechanism described above in which II-generated holes forward bias the base-emitter junction. Beyond the snapback point the current steeply rises, but β goes through a maximum and then falls off rapidly due to the effects of high-level injection (base pushout). Since the gain is decreasing, the level of current must be maintained by increasing the collector voltage (V_{ce}), which increases the II generation by expanding the width of the high-field region further into the epi layer. In this area of operation the I-V curve flattens out due to the additional voltage needed. If the epi layer is thin enough, the peak electric field will move from the lightly doped epi into the heavily doped substrate as V_{ce} continues to increase. Due to the higher doping level the electric field profile becomes higher and narrower. Additionally, high-level injection has made a large part of the epi layer charge neutral, and thus a voltage cannot be sustained across this region. The net result is a drop in V_{ce}, i.e., a second snapback. This phenomenon was predicted with PISCES simulations and was tenably verified by experiments as reported in [37]. In an ESD protection MOSFET, the drain LDD region acts like a lightly doped epi

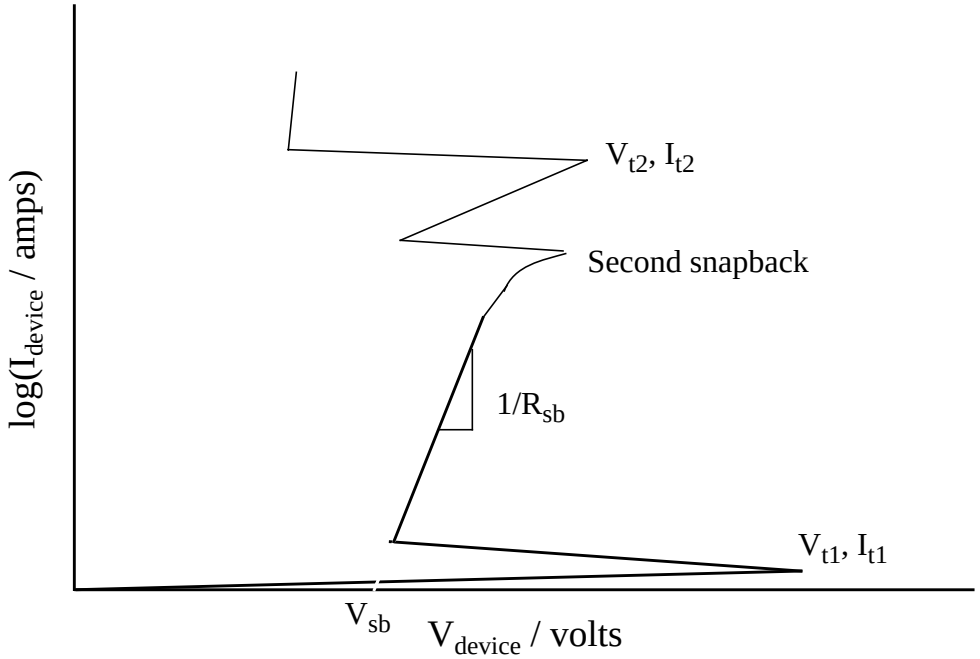

Fig. 2.7 Depiction of second snapback in a qualitative transient I-V curve for an NMOS transistor subjected to a positive ESD pulse.

layer in the lateral parasitic bipolar transistor (the effect of the source LDD region will be ignored here). Since the thin epi layer makes second snapback possible in the bipolar transistor, the LDD region could lead to a second snapback in the I-V curve, as depicted in Fig. 2.7. However, the current level required to trigger second snapback may be higher than the current which triggers thermal runaway and thus second snapback would not be observed.

In stepping through a transient I-V curve with transmission-line pulsing, a curve much like the one in Fig. 2.6a is generated. For initial pulses the device voltage closely follows the input voltage because the device current is very low (Fig. 2.8). Note that the rise time of the device voltage, which is a function of the equipment setup and is independent of the pulse height, is about 3ns. If the resolution of the measurement is high enough, finite current values can be recorded as the device voltage nears the trigger point V_{t1}. When the input voltage exceeds V_{t1}, the device voltage will drop to the snapback level. With a high-resolution oscilloscope the initial rise of the device voltage and subsequent drop to the snapback level can be captured as shown in Fig. 2.9. Beyond this point large steps in input voltage are needed to raise the device voltage because significant device current is now

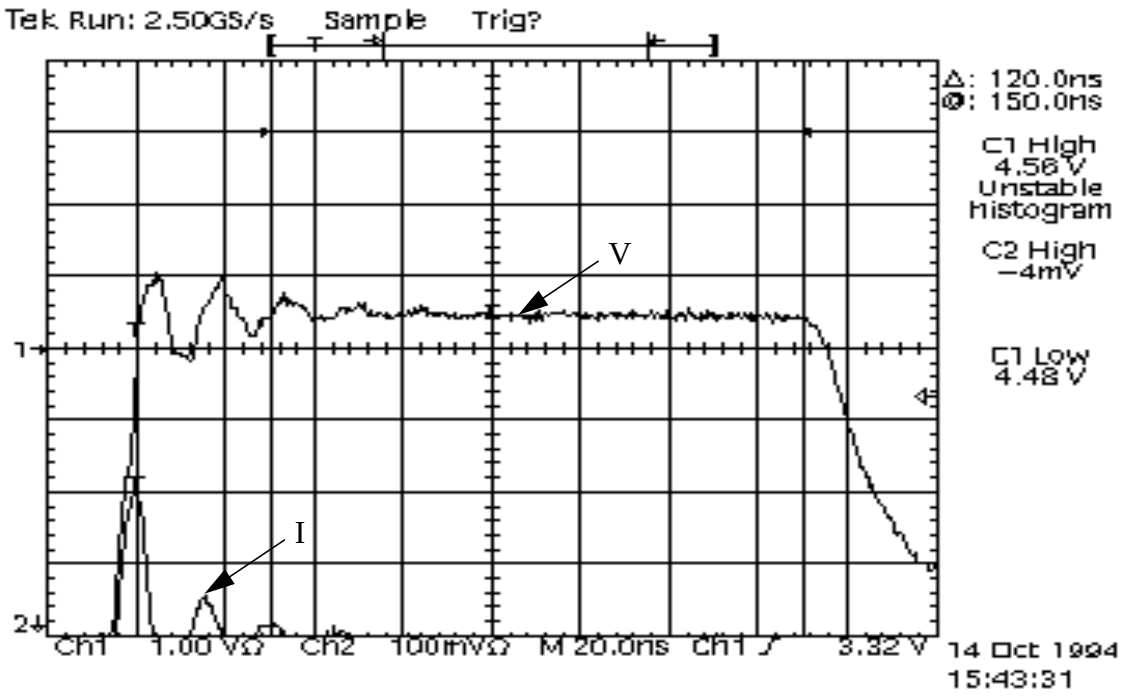

Fig. 2.8 *A screen capture of a Tektronix TDS 684A digitizing oscilloscope displays the device voltage (Ch1) and current (Ch2) response of a 50/0.6μm device to a 4.6V, 150ns input pulse. After some initial ringing, the device voltage settles to a value approximately equal to the input voltage and the device current is very small. The current probe registers 5mV per 1mA of current.*

flowing. If the input voltage is stepped carefully enough, the voltage drop due to second breakdown can also be captured (Fig. 2.10).

It is important to note that beyond snapback, the curve resulting from plotting the current points vs. the voltage points in Fig. 2.9 is different from the overall TLP curve of Fig. 2.6a. Notice that while snapping back the voltage does not drop all the way to V_{sb} and then rise back up to its final level, but rather just drops to the final level. Also, for reasons discussed in Section 2.3, the peak voltage will probably be less than V_{t1} because the voltage rise, as measured in V/ns, is faster for larger pulse heights. In this respect the TLP curve below the second-breakdown point really is a dc curve which doesn't account for device heating. However, it still represents how the device responds to an ESD stress because it reveals the operating points after the initial turn-on transient. Since V_{t1} is dependent on the voltage ramp rate, it is equal to the maximum input voltage during an

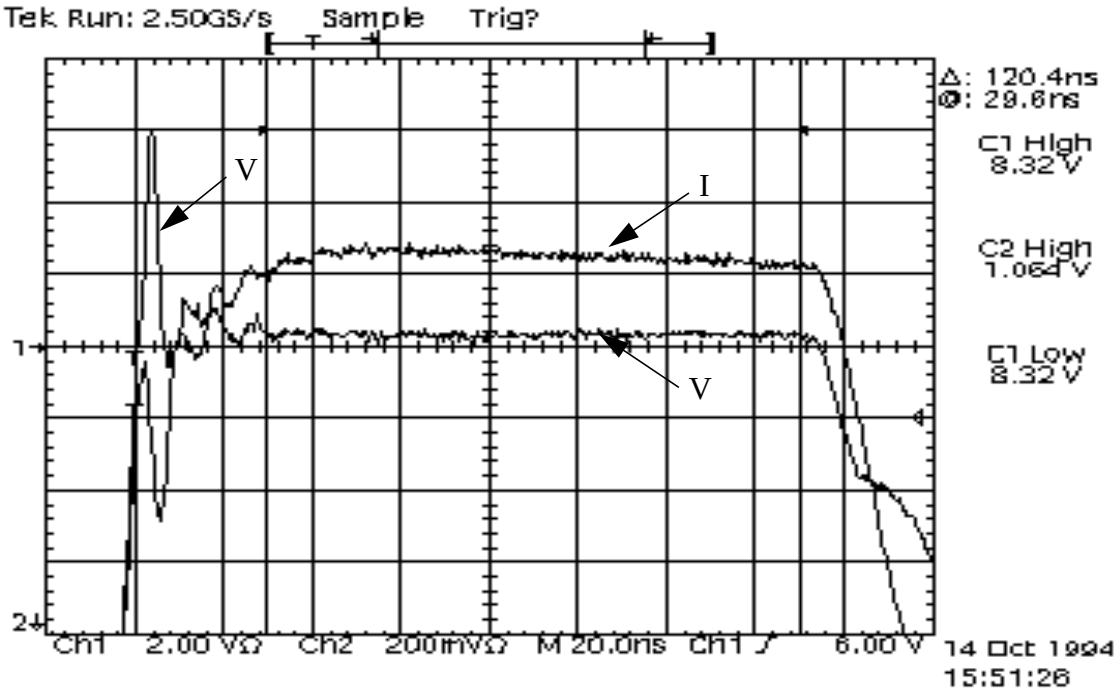

Fig. 2.9 *The screen capture shows the current and voltage response to a 15V, 150ns input pulse. The device voltage breaks down and snaps back in the first few nanoseconds of the pulse.*

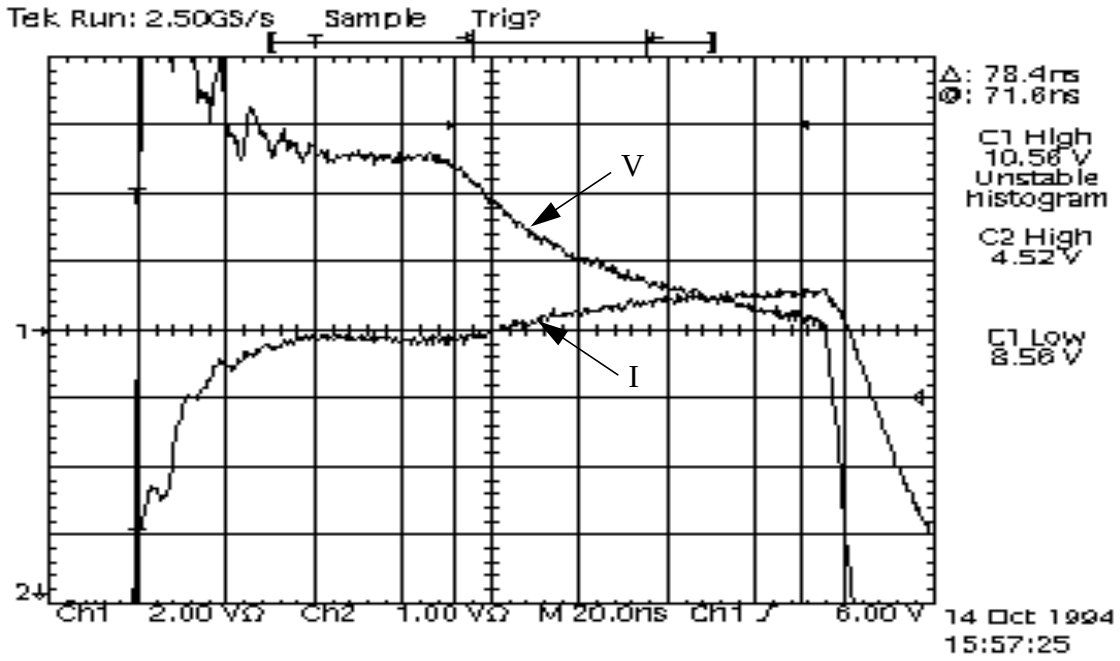

Fig. 2.10 *Second breakdown is observed as a drop in the device voltage and rise in device current about 72ns into the 36V, 150ns input pulse.*

In which case it can go higher

ESD event unless the rise time of the pulse is much longer than that of the TLP pulse. Experimentally the difference in V_{t1} between dc sweeps and TLP pulses with 3ns rise times is only one or two volts, so TLP-measured V_{t1} values are still indicative of the maximum input voltage created by pulses with much longer rise times.

2.2.2 Failure Power vs. Time to Failure

The short-duration pulses used to generate an I-V curve with TLP should be representative of actual ESD events. For example, a 100ns square-wave pulse provides a stress similar to a human-body model pulse, which has a decay time of approximately 150ns. A similar I-V curve can be generated with a well-controlled quasi-steady state current sweep, but the second-breakdown point will occur at a lower current due to the longer time spent at each stress level (there is also a dependence of V_{t1}, I_{t1}, and other parameters on the height and rise time of the input pulse). This is more representative of EOS damage. Intuitively, one expects a device to fail at a lower pulse height if the pulse duration is longer. To quantify this idea, a 3D thermal model has been proposed which defines four distinct regions of power-to-failure vs. time-to failure [23,38,39]. This model assumes a rectangular-box region of device heating in the drain-side junction depletion region of a MOSFET with a spatially uniform, time-invariant power source ($H = \mathbf{J} \cdot \mathbf{E}$ Watts/cm^3); constant-temperature boundary conditions on all sides of the box (an infinite heat sink); and no heating outside the box. As shown in Fig. 2.11, the length of the box, a, is equal to the

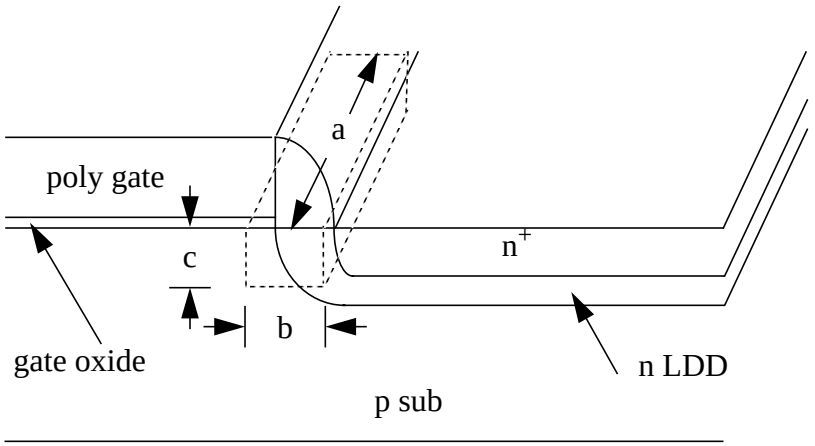

Fig. 2.11 3D thermal box region (dotted lines) of heat dissipation in an NMOS transistor subjected to a positive ESD pulse. The dimension a is equal to the device width, b is related to the gate length, and c is approximately equal to the diffusion depth.

width of the device, the width, b, is related to the gate length, and the depth, c, is approximately equal to the drain diffusion depth. Such a model is reasonable because simulations and experiments show that the junction sidewall is the region of highest electric field and current density and is where most of the potential drop occurs. Although the current density is about the same on the source side, the electric field here is very low.

In the model, failure is defined as the time at which the temperature of the hottest point--the center of the box--reaches a critical value, T_c. This critical temperature could be 1688K, the silicon melting point, or, more accurately, the temperature at which the intrinsic carrier concentration exceeds the doping level (about 1280K for a doping level of $10^{18}cm^{-3}$), i.e., the onset of second breakdown. Initially, the temperature gradient in the box changes in all three dimensions until thermal equilibrium is reached in the shortest dimension, usually c, the junction depth. The time needed to reach equilibrium in the c dimension is $t_c = c^2/4\pi D$, where D is the thermal diffusivity of silicon and is equal to $\kappa/\rho C_p$, where κ is the thermal conductivity, ρ is the density, and C_p is the specific heat capacity (all assumed to be independent of temperature in this model). If at time t_c the peak temperature is less than T_c, the temperature gradient will continue to change in the other two dimensions until thermal equilibrium is reached in the b direction at time $t_b = b^2/4\pi D$. Again, if the peak temperature is less than T_c at time t_b, the temperature gradient in the device width direction will continue to change until time $t_a = a^2/4\pi D$. For times greater than t_a, the temperature profile in the box is constant. This can be seen from the heat flow equation:

$$\rho C_p \frac{\partial T}{\partial t} = H + \nabla(\kappa(T)\nabla T) \ . \tag{2.2}$$

In the steady-state condition the temperature distribution must be constant because the heat source, H, is constant.

By applying the $T = T_0$ (300K) boundary conditions on the sides of the box, the heat equation can be solved to express the power to failure ($P_f = V_{t2} \times I_{t2}$) as a function of the time-to-failure (t_f, synonymous with t_2), the dimensions of the box, and the temperature difference $T_c - T_0$ at the center of the box. As derived in [39], the temperature at the center of the box is

$$T(t) = T_0 + \frac{P}{\rho C_p (abc)} \int_0^t \operatorname{erf}\left(\frac{a}{4\sqrt{D\tau}}\right) \operatorname{erf}\left(\frac{b}{4\sqrt{D\tau}}\right) \operatorname{erf}\left(\frac{c}{4\sqrt{D\tau}}\right) d\tau, \tag{2.3}$$

where P is the input power. By noting that

$$\text{erf}\,(c/4\sqrt{Dt}) \approx \sqrt{t_c/t} \text{ if } t \geq t_c \tag{2.4}$$

$$\text{and erf}\,(c/4\sqrt{Dt}) \approx 1 \text{ if } t \leq t_c \tag{2.5}$$

and setting $P = P_f$ for $T = T_c$, the failure power can be calculated for each of the time ranges described above:

$$P_f = \rho abc C_p (T_c - T_0)/t_f \text{ for } 0 \leq t_f \leq t_c, \tag{2.6}$$

$$P_f = \frac{ab\sqrt{\pi\kappa\rho C_p}\,(T_c - T_0)}{\sqrt{t_f} - \sqrt{t_c}/2} \text{ for } t_c \leq t_f \leq t_b, \tag{2.7}$$

$$P_f = \frac{4\pi\kappa a\,(T_c - T_0)}{\ln(t_f/t_b) + 2 - c/b} \text{ for } t_b \leq t_f \leq t_a, \tag{2.8}$$

$$\text{and } P_f = \frac{2\pi\kappa a\,(T_c - T_0)}{\ln(a/b) + 2 - c/2b - \sqrt{t_a/t_f}} \text{ for } t_f \geq t_a. \tag{2.9}$$

The P_f vs. t_f curve is shown graphically in Fig. 2.12. For times less than t_c, no heat is lost from the box, and a constant energy ($P_f \cdot t_f$) is needed to destroy the device. In the region $t_c \leq t \leq t_b$, failure power is proportional to $1/\sqrt{t}$, then becomes proportional to $1/\ln(t)$ in the region $t_b \leq t \leq t_a$. For times greater than t_a, the failure power approaches a constant value, which means power dissipation is equal to power generation. Using values of 100μm, 1μm, and 0.1μm for a, b, and c, respectively, the values of t_a, t_b, and t_c are approximately 10μs, 1ns, and 10ps, respectively. Thus in the ESD regime we expect to see a $1/\ln(t)$ dependence of P_f. As noted in [23], limitations which affect the accuracy of the model are assumptions that failure follows instantaneously when the temperature reaches T_c and that there is an infinite heat sink outside the rectangular box. If there is little resistance between the depletion region and device contacts, such as in silicided processes, failure should follow quickly after T_c is reached. The main problem with the heat sink assumption is that the SiO_2 layer above the silicon is a thermal insulator and seriously degrades heat dissipation in the vertical direction. This means that the power needed to cause failure is actually lower (by less than a factor of two) than that calculated by the model. Layout parameters which also affect the dissipation of heat are the closeness of the

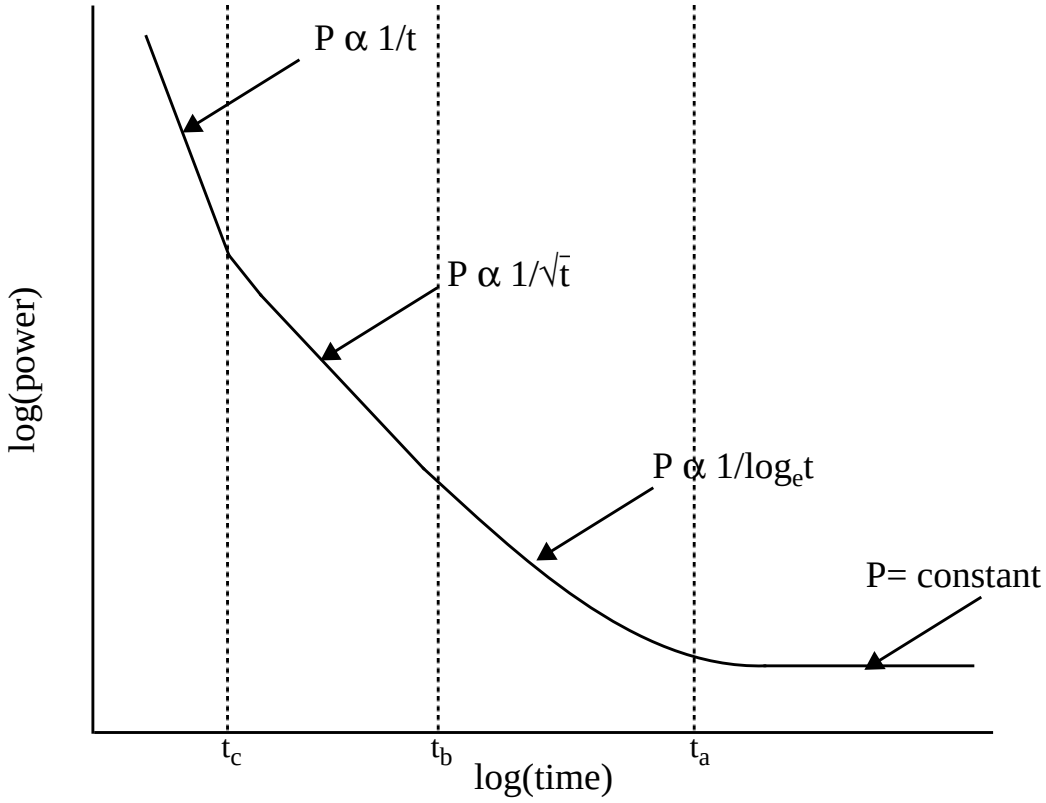

Fig. 2.12 A qualitative schematic of input power-to-failure vs. time-to-failure predicted by an analytical thermal model.

metal contacts, which are good thermal conductors, and the distance between conducting devices (as in a multiple-finger structure). Nevertheless, the model has been shown to agree with experimental results to first order.

By carefully stepping the input voltage and using varying lengths of line, transmission-line pulsing can be used to capture failure points (as in Fig. 2.10) and thus to define a P_f vs. t_f curve of an ESD protection circuit. The failure power level is the product of the device voltage and device current at the point failure occurs (V_{t2} and I_{t2}). Since each test is destructive, several identical devices are needed to extract a curve. If the voltage drop seen on the oscilloscope is second breakdown (thermal failure), there should be a significant increase in the measured leakage current after the stress. As discussed in the previous section, for very short pulse times a significant drop in voltage may not be observed, in which case it is necessary to define failure as an increase in leakage current

above a predefined level. Reasonable time-to-failure measurements can be made down to about 50ns. For times of 1μs and greater, a pulse generator can be used in place of the charged transmission line. After a curve is experimentally determined, the dimensions of the theoretical box can be extracted by fitting the model to the experimental curve.

Since a P_f vs. t_f curve reveals circuit failure thresholds over a wide spectrum of stress times, it suggests how robust a device is throughout the ESD and EOS regimes. It has been suggested that P_f vs. t_f and I_f (failure current, or I_{t2}) vs. t_f curves be used to qualify EOS/ESD reliability in addition to or in place of standard tests such as the HBM because reliability is then defined over a large range of stress events [24]. This attribute is attractive because it may show that a protection-circuit design performs relatively well in one domain of the EOS spectrum but performs poorly in another. Retesting after design modification would reveal what portions of the spectrum are affected by a certain device parameter. Some correlation has been drawn between TLP failure levels and HBM robustness [23], but further qualification must be done before IC manufacturers accept the P_f vs. t_f method as a valid reliability measure. The value of the method ultimately depends on how well the accepted classical models are represented by the constant-current stresses of TLP.

2.2.3 Leakage Current Evolution

The previous section mentioned the measuring of device leakage current after a TLP stress to verify that second breakdown has taken place. If a device exhibited a second snapback, it would probably not create a large increase in leakage and thus could be distinguished from the thermal second breakdown. It is in fact very useful to monitor the leakage evolution after each stress step of a TLP experiment. This can be done by removing the transmission-line connection from the input of the device under test, applying a voltage to the input (typically the supply voltage, V_{CC}), measuring the current with a multimeter in series with the V_{CC} supply, then reconnecting the transmission line. The voltage should be applied as briefly as possible to avoid corrupting the TLP experiment by further stressing the device. In contrast to the single leakage measurement made after a HBM stress, this technique reveals how the increased leakage evolves as a device is stressed through the various levels of the snapback curve. Before snapback, the leakage current is typically in the pA range. A jump in leakage above the 1μA level is usually observed after second breakdown due to diffusion of dopants from source to drain, filament formation across the

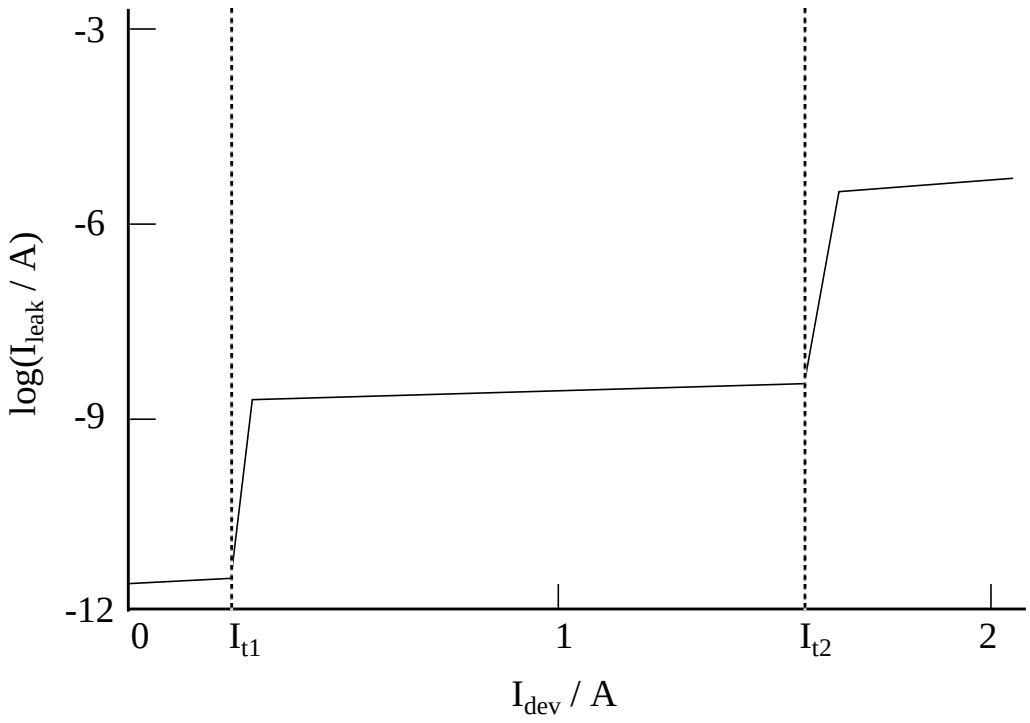

Fig. 2.13 Qualitative plot of device leakage evolution vs. stress-current level of a TLP experiment. Transitions are evident at the snapback and second-breakdown points.

drain-substrate junction, and/or a rupture of the gate oxide. In addition to this transition, sudden increases in leakage from the pA to the nA range have been observed when the device enters snapback [4]. Such a leakage evolution is depicted in Fig. 2.13 by plotting leakage current vs. the device current of the previous TLP stress. Non-catastrophic leakage (also called low-level leakage or soft failure) may be due to a small hot spot forming just before the device snaps back, to a small filament in the gate oxide formed by dielectric stress, or to hot-carrier trapping in the gate oxide which could induce a small channel region by shifting the threshold voltage below zero (for an NMOS device). Although the protection circuit still functions after a low-level stress, the increased leakage may be a signature of a latent failure, i.e., a reduction in lifetime of the circuit due to a "soft" ESD stress. Latent failure is a topic which merits further investigation, and the monitoring of leakage current during stepped TLP stresses is a powerful way to study the phenomenon.

For very short pulse widths in which melting can occur without second breakdown, leakage measurements will be the only way (except for a functionality test) of detecting device failure. If the gate, source, and substrate or well of a protection MOSFET have separate connections, separate leakage measurements can be made between the drain pin and each of these pins. Monitoring the leakage evolution of all pairs of pins would lend even more information about how and where damage is occurring in a device. For example, increased leakage from drain to source or drain to substrate suggest filament formation due to device heating, while increased leakage from drain to gate indicates an oxide failure.

2.2.4 Advanced TLP Setup

To close out the discussion on transmission-line pulsing, we will look at some advanced experimental setup techniques. ESD research data used in this thesis was obtained with a setup created at Advanced Micro Devices (AMD) in Sunnyvale, CA. A schematic of this setup is given in Fig. 2.14. The oscilloscope used to measure the device voltage and current is a 1GHz Tektronix TDS 684A digitizing oscilloscope. A Tektronix P6245 1.5GHz active FET probe is used to monitor the voltage, while a Tektronix CT1 1GHz transformer current probe monitors the current. Notice that a series-parallel resistor combination has been added to the circuit to increase the current resolution of the TLP experiment. Its benefit can be seen by considering what happens as the input voltage is stepped in the original setup of Fig. 2.5. Just before snapback the current, I_{t1}, is approximately zero, so, from Eq. (2.1), $V_{dev} = V_{t1} = V_{in}$. Assuming an infinitesimal increase in V_{in} will cause the device to snap back, just after snapback the device current is

$$I_{dev} = (V_{in} - V_{sb}) / R_L \tag{2.10}$$

$$= (V_{t1} - V_{sb}) / R_L . \tag{2.11}$$

With typical values of 10V for V_{t1} and 4V for V_{sb} and $R_L = 50\Omega$, $I_{dev} = 120mA$ is the minimum current resolution available with this setup, i.e., there is no setting of V_{in} which will yield a device current between I_{t1} and 120mA. For a MOSFET which is only 20μm wide, this current is nearly equal to or greater than the second breakdown level, which means a large portion of the snapback curve cannot be drawn out. This problem is solved with the circuit shown in Fig. 2.14, in which

$$I_{dev} = (V_{in} - 2V_{dev}) / (R_L + 2R_S) . \tag{2.12}$$

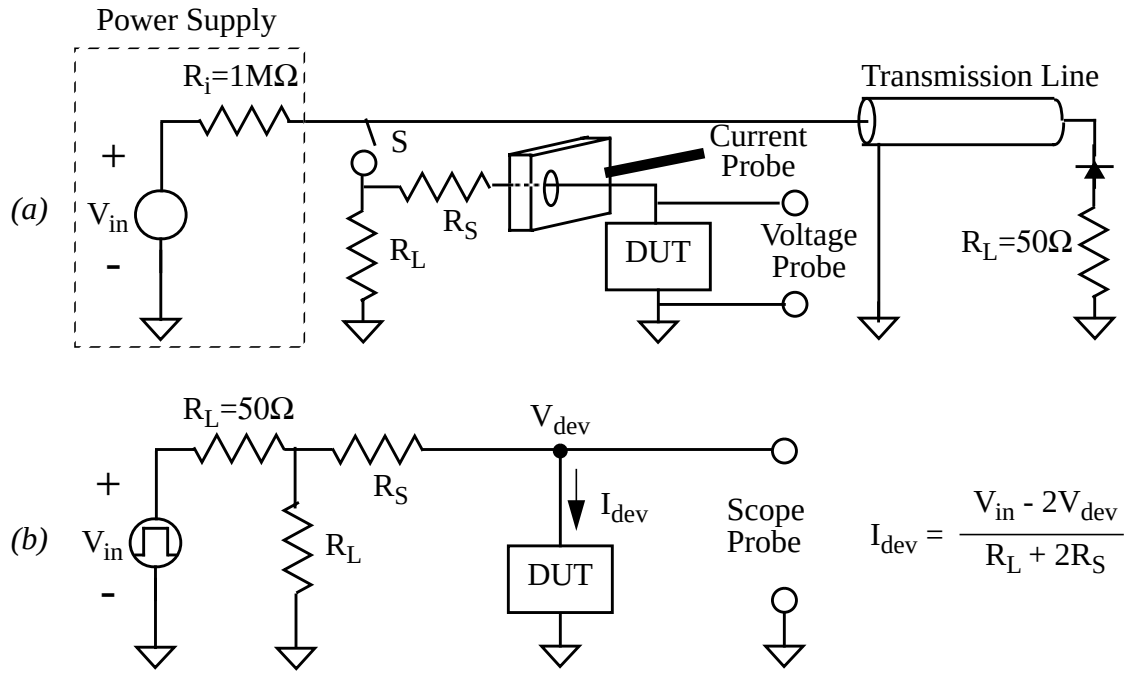

Fig. 2.14 *(a) Advanced TLP schematic: a series-parallel resistor combination has been added to enhance current resolution. A current probe is now used to directly measure the device current on the oscilloscope. (b) Equivalent circuit of the TLP setup.*

Now, just before and just after snapback, V_{in} is approximately $2V_{t1}$, so at snapback

$$I_{dev} = (V_{t1} - V_{sb}) / (R_S + R_L/2) . \tag{2.13}$$

and we see that R_L is replaced by $R_S + R_L/2$. Using a value of R_S = 300Ω, the current resolution is now 18mA. An added benefit of the 50Ω shunt resistor is that it will absorb all the pulse energy and prevent reflections when the impedance of the DUT is high.

In the AMD setup a special high-frequency jig with insulated wires running from BNC connectors to the pins of a low-insertion-force socket was built to minimize noise during measurements of test circuits prepared in dual in-line packages. Additionally, chip resistors are used for the resistor network and all connections are kept as short as possible to minimize parasitic inductances which alter the shape of the measured voltage and

current profiles. The switch S is a normally closed, single-pull double-throw vacuum relay which is opened and closed by applying and disabling a 12V power supply, respectively. During a stepped-stress experiment, a leakage measurement between input and ground is taken after each pulse by switching the input from the transmission line to an ammeter in series with a V_{CC} supply (not shown in Fig. 2.14). This switching is done with a voltage-controlled 10GHz coaxial relay placed between the probes and the DUT input. Currently, an HP3457A multimeter with 1nA resolution is used for leakage measurements, but this will eventually be replaced by an HP4145 parametric analyzer with 1pA resolution. All instruments and power supplies are controlled by a personal computer through either a National Instruments AT-GPIB/TNT IEEE-488 card or a National Instruments PC-DIO-96 digital I/O board. National Instruments' LabVIEW software package is used to automatically run a TLP experiment on a test structure and store all I-V and leakage data from initial device breakdown through device failure. A built-in oscilloscope function which measures the average height of a waveform in a gated region facilitates the automatic extraction of the device voltage and current resulting from each input pulse.

2.3 Overview of Protection Circuit Design

This section is not meant to provide an exhaustive review of all types of on-chip protection but rather to introduce some basic concepts. A thorough discussion of on-chip protection is presented in [34]. Any I/O protection circuit should provide a low-impedance path from input to supply during an ESD event to absorb current but provide a very high impedance during normal operating conditions so as not to affect circuit performance, e.g., through increased leakage current or parasitic capacitance. Additionally, an ESD circuit should clamp input voltages at a safe level, i.e., below the dielectric breakdown voltage of a thin gate transistor. The dielectric threshold electric field is actually time dependent: it must be held across an oxide for a certain length of time before the oxide breaks down, as measured by leakage current [40]. The time to breakdown is lower for a higher stress field. Although the consequences of this time dependence on ESD protection ability are important, for simplicity we will assume that the voltage across a thin gate must not exceed some critical level for any amount of time.

When designing ESD protection circuits, there are some important differences to consider between input protection and output protection. While the high-impedance input pads of a CMOS chip are connected to the thin gates of the input buffer transistors, the low-

impedance outputs are connected to the drains of output-buffer transistors. Design of output protection is thus more restricted than that of input protection because of low output-impedance requirements. For example, a well resistor may be placed between an input pad and the protection MOSFET to reduce the rise time of an ESD pulse, but such a resistor cannot be placed on an output pad because the increased impedance would exceed circuit specifications. Also, since the output-protection transistors often double as the CMOS output buffer, they must meet certain chip-performance specifications. As a result, output protection relies more on the proper layout of one or two transistors than on the use of creative circuit designs.

CIRCUIT I

Fig. 2.15a shows a simple diode protection scheme. The diodes are formed by source/drain diffusions in the p-substrate or n-well. When the circuit is powered up, diode D1 will become forward biased and conduct current for any input voltage greater than $V_{CC} + V_d$, where V_d is the forward diode drop. Similarly, diode D2 clamps any negative voltage below $V_{SS} - V_d$. If the chip is not powered up and an ESD pulse is incident between the input and, say, V_{SS}, the voltage will be clamped at either the breakdown voltage of the diode for a positive pulse (note we are neglecting the voltage drop across the dynamic resistance of the diode) or at $-V_d$ for a negative pulse. The diodes should introduce minimal leakage current and a negligible parasitic capacitance to the circuit since they are normally reverse biased. Series resistors can be used in conjunction with diodes (or other devices) in input protection circuits, as shown in Fig. 2.15b, to create a potential drop from the pad to the diode and thus reduce the voltage at the input gates. Using a diffused resistor distributes the resistance and introduces an additional distributed diode, resulting in a lower gate voltage than that created by a simple polysilicon resistor. Addition of a series resistor aids circuit protection by slowing down transients (e.g., a machine-model waveform would be transformed into a HBM-like waveform), but by the same token it can reduce circuit speed performance by increasing RC time constants.

TRANSIENTS

Although diode circuits are simple to implement and may have provided sufficient ESD protection in the past, there are a few reasons why they are no longer adequate for protecting today's smaller technologies. First, the dynamic resistance of a reverse-biased diode may be too high to keep voltages clamped at a safe level unless the diode area is very large. For example, a 250 μm^2 area of diode with a typical impedance of 5000 Ω-μm^2 has a resistance of 20Ω and will sustain 20V at a stress current of 1A, a voltage well above the dielectric threshold of a thin gate oxide. The potential drop can of course be reduced by

DIODE ESD NOT DONE IN TODAY'S TECHNOLOGIES

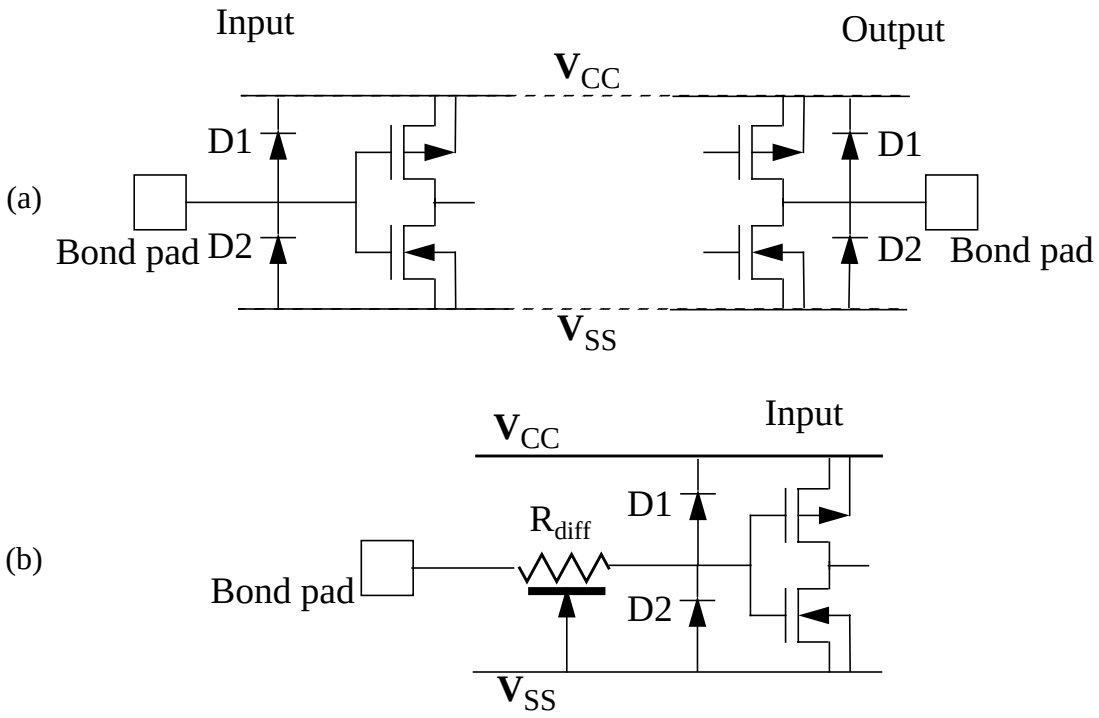

Fig. 2.15 *(a) ESD diode protection circuit in a CMOS technology; (b) use of a series resistor in combination with diode protection. A diffused resistor has a distributed resistance and also forms a distributed diode.*

using larger diode areas, but this takes up valuable chip real estate and may increase the parasitic capacitance to a level no longer negligible compared to the input-gate capacitance, thus degrading high-frequency performance. Reverse diode resistance can also be decreased if a diode with a smaller depletion layer can be processed, but the reduction of the depletion layer again implies a higher capacitance. Another limitation of diodes is that the breakdown voltage itself may be higher than the dielectric threshold of today's thin gate oxides. Finally, a diode often cannot break down quickly enough to protect a circuit from a fast-rising transient pulse such as that created by the charged-device model.

CIRCUIT II

Fig. 2.16 shows a CMOS-transistor protection scheme. These devices, which can be either thin-gate or thick-gate (field) transistors, have the advantage of being built using the standard chip process without additional implant or masking steps. (One exception is that a resist mask which blocks the silicide deposition may be added to increase the drain-to-gate and source-to-gate resistance.) The drain of the NMOS device (M2) is connected to

the I/O with the gate, source, and substrate tied to V_{SS}. A PMOS device (M1) is placed between the I/O pad and V_{CC}, with the drain connected to the I/O and the gate, source, and substrate tied to V_{CC}. Normally, the input protection transistors are turned off because there is no conducting channel. Note that at the output the CMOS buffer doubles as the protection device. If a negative ESD pulse is incident between the I/O and V_{SS}, the drain-substrate diode of the NMOS device becomes forward biased and conducts the high current. If the pulse is positive valued, the NMOS device will conduct current in a parasitic bipolar-transistor mode, with the drain acting as collector, the substrate as base, and the source as emitter. A PMOS device behaves analogously during an I/O vs. V_{CC} stress. If a protection MOSFET has a very short gate length, the device may actually turn on via punchthrough from source to drain rather than through snapback. This is a distinct possibility for devices built with minimal gate length in advanced technologies. A punchthrough device would have a lower V_{t1} than that of a conventional protection MOSFET, but the snapback voltage would be the same because parasitic bipolar action would still dominate at higher current levels. In powered-up CMOS protection circuits, where V_{CC} and V_{SS} both form ac grounds and thus NMOS and PMOS protection circuits may be considered to be in parallel during a transient stress, it is often found that the NMOS absorbs the ESD energy regardless of pulse polarity [18, 21]. This means reverse breakdown of the NMOS device occurs faster than the forward turn-on of the PMOS drain-substrate junction for a positive ESD pulse and that the NMOS drain-substrate junction forward biases before PMOS snapback during a negative input pulse. This makes sense because the gain of the parasitic npn transistor in the NMOS device is much higher

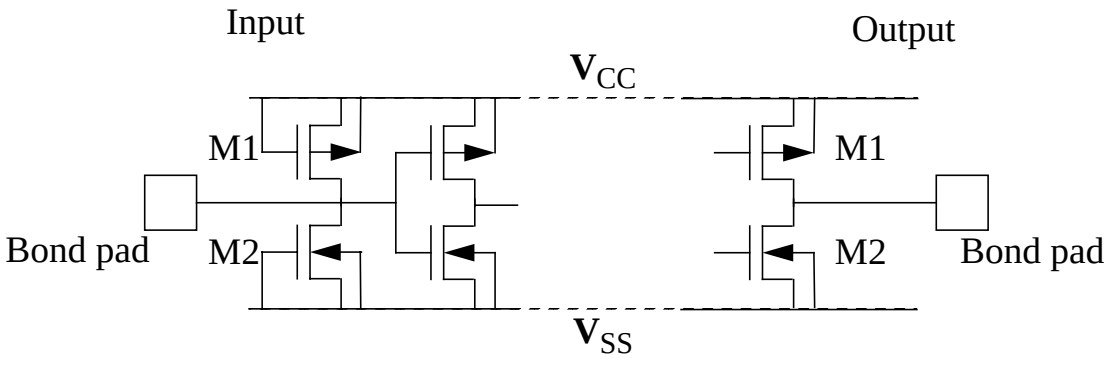

Fig. 2.16 CMOS input and output protection. The input protection transistors protect the thin gates of the input buffer, while the output protection transistors double as the output buffer.

than that of the pnp transistor in the PMOS device due to the lower diffusivity of holes, which means snapback occurs at a lower current for the NMOS device. PMOS transistors are still necessary, however, for protection of unconnected ICs.

CIRCUIT II VARIATION I

A diode created in a CMOS process has the same breakdown voltage as the MOSFET drain-substrate junction (neglecting curvature effects), but the MOSFET can be turned on more quickly and at a lower voltage by using a gate-bouncing technique. As seen in the MOSFET snapback curve, the transistor enters the low-voltage snapback mode when the drain voltage and current generate enough carriers through impact-ionization to forward bias the source-substrate junction. In a dc sweep the voltage V_{t1} is usually two or three volts higher than the breakdown voltage, depending on the gate length of the MOSFET.

Cgd Useful

During a transient pulse event, however, the maximum drain voltage can be held significantly below the dc V_{t1} by coupling of the gate voltage to the input voltage through the drain-gate overlap capacitance. The gate bias is usually created by placing a resistor or tie-off transistor between the NMOS gate and ground (Fig. 2.17a) or between the PMOS gate and supply. An equivalent circuit of this setup is also shown in Fig. 2.17a. Given a ramp input described by $V_{in}(t) = V' \cdot t$, the gate voltage as a function of time is

MOS TRANSISTOR ACTION

$$V_{gate}(t) = V'R_{gate}C_{DG}(1 - \exp(-t/R_{gate}C_{DG})) . \qquad (2.14)$$

Given a gate resistance of 2000Ω, an overlap capacitance of 10fF, and a pulse edge of 100V/ns, the gate voltage should reach a value of $V'R_{gate}C_{DG}$ = 2V during the rise of the pulse, enough bias to create MOS transistor action at the beginning of the pulse. Note that a higher pulse height with the same rise time (higher V′) will yield a higher gate voltage and thus a lower trigger voltage. After the initial bounce the gate bias will decay to zero as the drain voltage reaches a steady value. In protection transistors built using field oxide for the gate, the gate is often tied directly to the drain (input) to bias the gate. This can only be done because the threshold voltage of the field oxide device is higher than normal operating voltages and thus will not turn on during circuit use. Another advantage of the field-oxide device is that the thick oxide is much less susceptible to dielectric breakdown.

CIRCUIT II VARIATION II

Another gate-bounce technique is the relatively new method of dynamic gate coupling [41,43], in which a field-oxide device (FOD) is used to aid turn on of the primary thin-gate (TG) protection device (see Fig. 2.17b). The gate of the FOD is tied to the input pad (as is the thin-gate drain), with the drain of the FOD tied to the TG gate and the source grounded. In this circuit the TG gate is coupled to the input by the drain-gate overlap

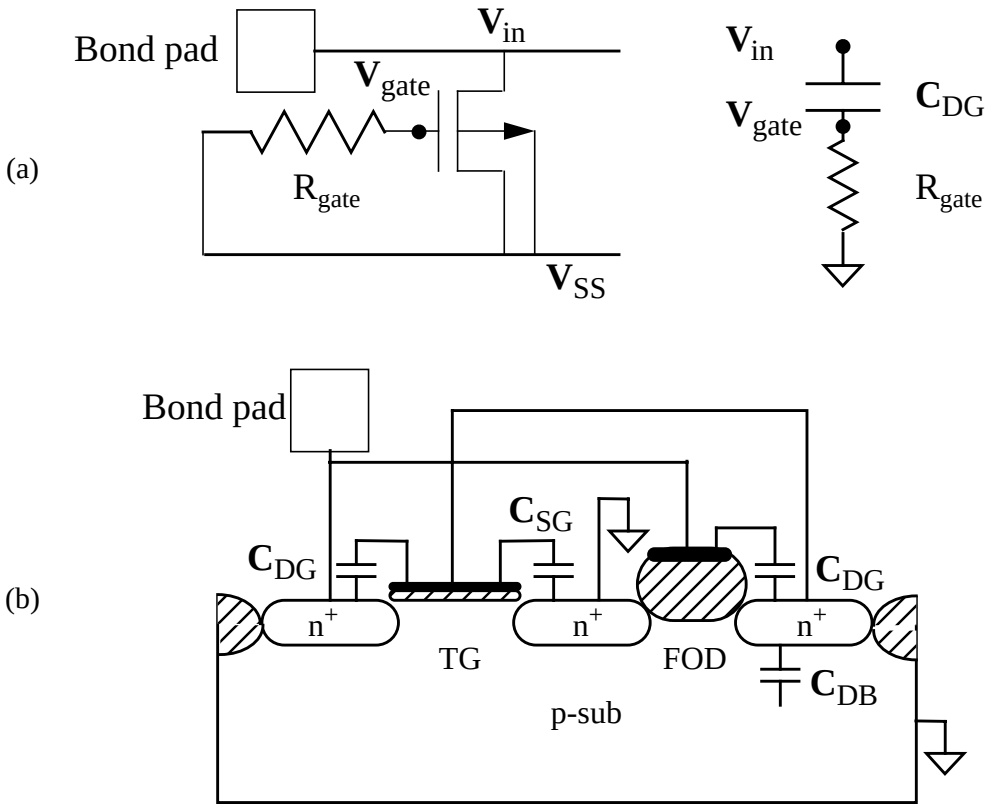

Fig. 2.17 Gate-bouncing techniques: (a) employment of a gate-bounce resistor and equivalent circuit; (b) dynamic gate coupling method.

capacitances (C_{DG}), the TG source-gate overlap capacitance (C_{SG}), and the FOD drain-substrate capacitance (C_{DB}). As in the gate-resistor technique, the coupling of the thin-gate potential to the input voltage helps the thin-gate device quickly enter snapback, but in this technique the amount of coupling is controlled by the ratio of FOD and TG gate widths. The added feature of the dynamic-gate circuit is that as soon as the input voltage reaches the FOD threshold voltage, the FOD turns on, creating a connection from the TG gate to ground. It is beneficial to return the thin-oxide gate to ground after the device enters snapback to avoid localizing the current conduction at the surface, which would lead to premature thermal breakdown.

CIRCUIT III

An example of a more complicated protection circuit, incorporating a wide NMOS transistor (M1), a well (diffused) resistor, and a narrow NMOS transistor (M2), is shown

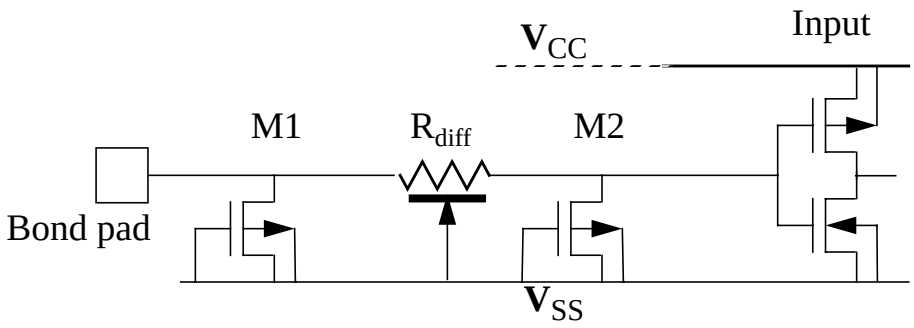

Fig. 2.18 Combination resistor/transistor ESD input protection circuit, featuring a diffused resistor (R_{diff}) between wide (M1) and narrow (M2) NMOS transistors. Resistance from gate to ground is not shown.

in Fig. 2.18. The narrow transistor is designed with a minimal gate length so that its parasitic bipolar transistor will turn on quickly and clamp the input voltage during a short ESD event. During a longer event the wide transistor, which may have a longer gate length and turn-on time, absorbs the majority of ESD current. The well resistor creates a voltage drop which ensures that the drain voltage of the wide transistor will build up to the breakdown value instead of being clamped at V_{sb} of the narrow transistor. This circuit only begins to suggest the creativity that can be used in designing protection circuits, but it exemplifies the implementation of different devices to provide protection across a broad range of the EOS/ESD spectrum.

In closing out this section on ESD circuits, it should be mentioned that a CMOS I/O protection transistor usually consists of several “fingers” of devices in parallel coming off an I/O pad rather than a single, very wide MOSFET (Fig. 2.19). This design method is used because ESD-current robustness increases with device width and multiple fingers furnish a compact way of providing a large effective width on a circuit in which space is at a premium. Also, a single narrow metal finger coming off of the contact pad will have a higher current density than several fingers in parallel and thus will be more susceptible to damage. One important drawback of such “multifingered” devices is that due to random variations between fingers it is almost never the case that all fingers of a protection device will turn on simultaneously during an ESD event. Instead, after one device breaks down and quickly enters the snapback mode, the drain voltage of all the devices is clamped at the snapback voltage since they are all tied to the input. As the current increases, the

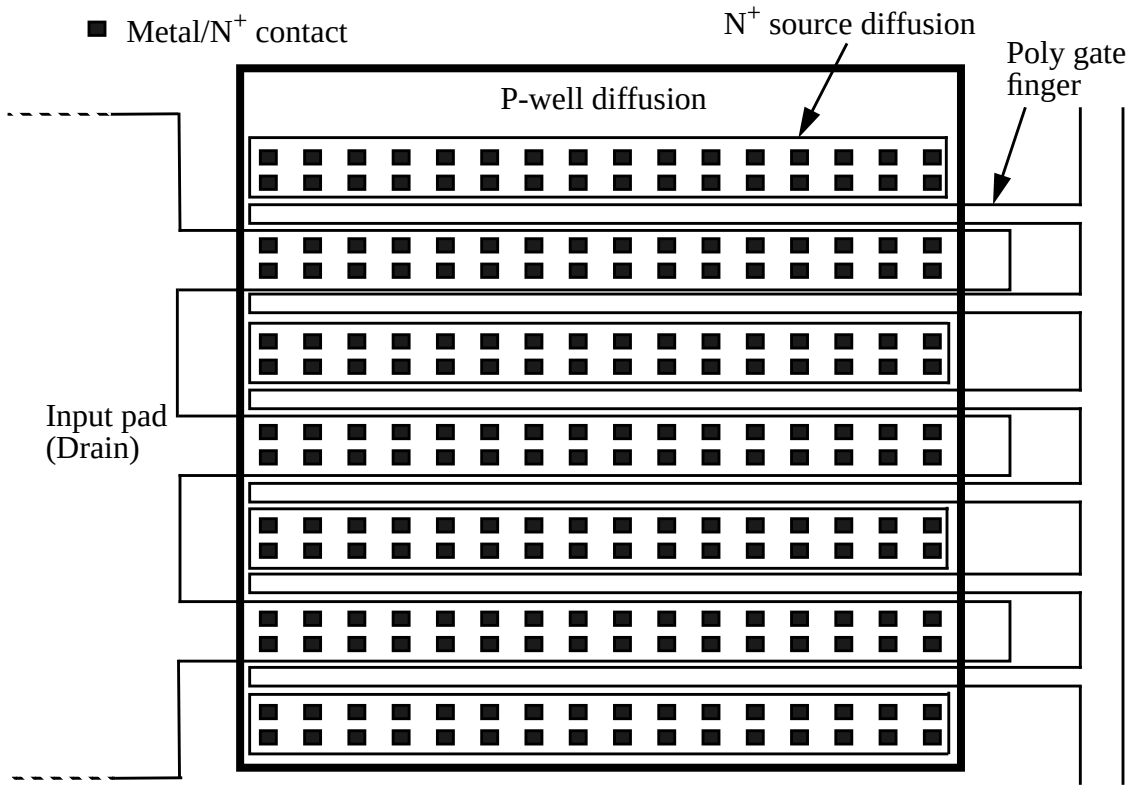

Fig. 2.19 Layout of a multiple-finger NMOS transistor between input and V_{SS}. There are three drain fingers and six poly gates. Another finger (not shown) branches off of the drain pad to the input buffer of the IC.

nonconducting devices will remain off unless the drain voltage of the conducting device increases past the trigger voltage, at which point another device will turn on and the voltage will again snap back (although not as far). If additional fingers do not turn on before the current density in the conducting fingers reaches a catastrophic level (I_{t2}), the robustness of the circuit is effectively reduced.

2.4 Dependence of Critical MOSFET I-V Parameters on Process and Layout

Previous sections of this chapter introduced the transient MOSFET I-V response to pulsed inputs, defined the critical parameters of this curve, and briefly mentioned some of the

effects of these parameters on ESD circuit robustness. Before further discussing these effects and defining a circuit design strategy, we will look at the dependence of V_{bd}, V_{t1}, I_{t1}, V_{sb}, R_{sb}, V_{t2}, and I_{t2} (all defined in Fig. 2.6) on several process and layout parameters. The time to trigger, t_1, and the time to second breakdown, t_2, are also important parameters, but they are really more a function of the incoming pulse profile. As noted before, t_1 decreases as the pulse ramp rate increases, while t_2 decreases as the power in the pulse increases. Given a fixed input pulse, a reduction in V_{t1} and/or I_{t1} implies a reduction in t_1. The effects of process and layout parameters on the critical MOS parameters are discussed below and summarized in Table 2.1. Note that in this discussion the snapback voltage, V_{sb}, will be defined as the minimum voltage after the device is triggered rather than the value extrapolated from the snapback region back to the x-axis as in Fig. 2.6. This is done because the extrapolated V_{sb} depends not only on the minimum voltage in the snapback region but also on the snapback resistance.

- **Gate length** -- Since the gate length, L, is effectively the base width of the parasitic bipolar transistor, it has a strong effect on the I-V curve. As mentioned in Section 2.2.1, the ratio of the breakdown voltage to the snapback voltage is $\beta^{1/n}$, the current gain of the bipolar transistor raised to some power. The breakdown voltage should be determined only by the drain-substrate junction profile and thus be constant vs. gate length, unless the gate length is so short that punchthrough occurs before avalanche breakdown. To first order, $\beta \propto 1/L^2$, so V_{sb} should be proportional to $L^{2/n}$, assuming no potential drops outside of the intrinsic device. For a typical experimental value of n = 5.5, doubling the gate length should increase V_{sb} by 29%. R_{sb} is higher for a longer channel, but this dependence may not be detectable since the series resistance due to the contact-to-gate spacing is usually dominant. V_{t1} and I_{t1}, and thus the turn-on time, also increase with L because the diffusion of holes to the source which triggers snapback becomes less efficient and more impact ionization must be provided by increased current and electric field. Finally, I_{t2} should increase with gate length because there is a larger area over which heat generated in the drain depletion region can dissipate. This is in agreement with the 3D thermal box model.

- **Gate width** -- If a MOS transistor is operating uniformly over its entire width, W, then the current parameters I_{t1} and I_{t2} should scale directly with device width. This means more current is needed to turn on the device, but it also means the device should be more robust since the width of the box in the 3D thermal model is larger. The voltage

parameters should not change, which means R_{sb} should decrease. However, if a gate resistor or other method is used to couple the gate bias to the input, a larger W implies a larger drain-gate overlap capacitance and thus an increase in coupling (see Eq. (2.14)) and a reduction in V_{t1} due to more MOS transistor action. Another point to make is that as discussed in Section 1.1, for very small device widths the failure current appears to be independent of W because some overall current is needed to create severe damage. This is not a contradiction of the $I_{t2} \propto W$ rule because in such cases failure does not follow immediately after second breakdown, so there is a difference between the failure current and I_{t2}. Note that if device operation is not uniform but rather the current and voltage or electric field are concentrated at a corner or edge, second breakdown will occur sooner than predicted, i.e., I_{t2} will not scale linearly with width.

- **Source/Drain (S/D) junction depth and profile** -- Deeper junctions have a larger area over which current is distributed and thus a lower current density for a given current level. In other words, the depth of the box in the 3D thermal model is larger, which

Table 2.1 Dependence of critical I-V parameters on process and layout. An up or down arrow signifies that the I-V parameter increases or decreases, respectively, as the process or layout parameter increases or as otherwise noted. Double arrows indicate a primary dependence, while a single arrow represents a second-order or side effect. ND signifies that there is little or no dependence on the parameter.

Parameter	V_{bd}	V_{t1}	I_{t1}	V_{sb}	R_{sb}	V_{t2}	I_{t2}
Gate length	ND	↑↑	↑	↑↑	↑	↑	↑↑
Gate width	ND	↓[a]	↑↑	ND	↓	ND	↑↑
S/D junction depth (1 / curvature)	↓↓	↓	ND	↓	↓	↑	↑↑
Contact-gate spacing	↑	↑	ND	↑↑	↑↑	↑	ND
Remove silicide	ND	↑	ND	↑↑	↑↑	↑	↑↑
Gate bias/bounce	↓↓	↓↓	↓↓	ND	ND	ND	ND
Block LDD implant	↑↑	↑	ND	↑	↑	↑	↑↑[b]
Substrate resistance	↓	↓	↓↓	ND	ND	ND	↓

a. If gate coupling is used.
b. If LDD junction is shallow compared to S/D junction.

increases I_{t2}. The breakdown voltage, V_{bd}, of a junction increases as the curvature increases, but this effect is much less pronounced in graded junctions than in abrupt junctions [42]. A deeper junction also has a lower resistivity and will significantly decrease R_{sb} if the spacing between the gate and the source/drain contacts is large (see below). This decreased S/D resistance reduces V_{sb} and, to a lesser degree, V_{t1}. I_{t1} is independent of small variations in junction depth because the junction depth does not affect the level of impact-ionization generation needed to induce snapback.

- **Contact-to-gate spacing** -- Increasing the spacing between the gate and the drain contacts increases the series resistance between the input and intrinsic circuit. The main effect of increasing the spacing is an increase in the snapback resistance and snapback voltage. If I_{t2} is not affected, a larger R_{sb} implies a larger V_{t2} (I_{t2} may be affected for longer stress times in which dissipation of heat is important over a wider area extending into the drain diffusion region). There will also be an increase in V_{bd} and V_{t1}, although the current level just before snapback is usually about a milliamp and thus the added potential drop in the drain diffusion may be inconsequential. Increasing the spacing from the gate to the source contacts will have the same effects since it is also just an introduction of more series resistance in the circuit.

- **Silicide** -- Varying the contact-to-gate spacing will only have a significant effect if the S/D diffusions are not silicided. In current MOS technologies a titanium or tungsten silicide is placed over the S/D diffusions to reduce series resistance and thus enhance circuit performance. A typical n^+ diffusion resistance is on the order of 4 Ω/❑, whereas a non-silicided diffusion has a resistivity of about 60 Ω/❑. Silicided diffusions are a disadvantage in ESD circuits because they concentrate current at the surface, which reduces I_{t2} by increasing current density, and because they eliminate the ballast resistance (R_{sb}) between the input and the intrinsic device needed to ensure uniform turn-on in a multifingered structure (see next section). Again, note that a reduction in series resistance implies a reduction in V_{sb} and a slight reduction in V_{t1}. In some technologies a “resist mask” is used to block silicidation of S/D diffusions, and this mask is normally used for ESD protection circuits.

- **Gate bias** -- As discussed in the previous section, biasing the gate by coupling the gate voltage to the input reduces V_{t1} by aiding the onset of snapback through increased drain current; the snapback and second-breakdown regions are unaffected. Maximum reduction in the trigger voltage is attained by biasing the gate just above the threshold

voltage, V_T [41]. The reduction in V_{t1} ranges from a few volts for small gate-length devices to about 50% for larger gate lengths. Beyond V_T, the trigger voltage levels off with increased gate biasing and may actually increase since the reduced electric field in the drain depletion region will reduce impact ionization. If the gate remains biased after a device has entered snapback, I_{t2} can be reduced due to concentration of drain-source current at the surface of the channel, so it is important that the gate be biased only during initial turn-on of the device.

- **LDD** -- It is generally assumed that a lightly doped drain decreases the performance of an ESD protection structure because it has a much lower junction depth than the S/D diffusion, which leads to higher current concentrations in the area of high electric field (i.e., the box depth is smaller in the 3D thermal model) and thus reduces I_{t2}. However, if the LDD depth is not much different than the S/D depth, then there should be little change in I_{t2} unless the accompanying change in the electric-field profile is significant. In a CMOS process the NMOS LDD implant can be blocked simply by covering the NMOS active area with the same oxide used to mask the PMOS active areas during this implant. Of course, the spacer oxide will still be present after the oxide etch, which means the S/D diffusion edges will be separated from the intrinsic channel under the gate contact, i.e., the gate length is effectively increased by twice the spacer width. (Since it is only the drain side of the device which has the high electric field, the source LDD diffusion may be left in the process, meaning the gate length is only increased by one spacer width.) Thus, blocking the LDD implant also effects the same changes as increasing the gate length. These effects may be compensated by reducing the drawn gate length. Although the drain junction may become more abrupt when the LDD is omitted, V_{bd} increases because the net drain doping decreases without the LDD implant, and therefore V_{t1} and V_{sb} also increase. The snapback resistance will also probably be slightly larger due to the increased effective gate length.
- **Substrate resistance** -- Increasing the substrate resistance, either by moving the substrate contact farther away from the drain diffusion or by adding a lumped resistance between the local substrate contact and ground, or floating the substrate accelerates the onset of snapback by creating a higher substrate bias for the same substrate current and by diverting more of the impact-ionization generated holes toward the source to forward bias the source-substrate junction. The reduction in V_{t1} and I_{t1} imply a faster triggering of the device. To first order, the snapback region of operation is not affected by

the substrate resistance. However, I_{t2} will be reduced, especially if the substrate is floating, because the reduced fraction of stress current sunk by the substrate implies a higher concentration of current underneath the gate and thus more device heating.

2.5 Design Methodology

An ESD circuit design methodology should be based on the goal of robust protection from thermal and dielectric failure across a wide range of the EOS/ESD spectrum. In today's environment an IC manufacturer will probably want to guarantee that its packaged devices will perform within specifications after any pins are subjected to some voltage level of the HBM test and possibly of the CDM test because these are the standard ways of measuring ESD robustness. However, it is important to use a broad-range testing method such as TLP to ensure ESD protection not only for specific tests but for any potential stress which can lead to a field failure or customer return. The design methodology presented in this work focuses on multifinger CMOS protection circuits for IC inputs and outputs; this section emphasizes optimization of the individual devices (fingers) before creating and testing the overall circuit. Design and optimization of multifinger circuits is the main topic of Chapter 5. Although ESD circuits are definitely susceptible to failure at contacts, diffused resistors, poly resistors, and other interconnect sites due to excessive heating, this design process is concentrated on MOSFET design and assumes that thermal failure will always occur within a protection device and that dielectric failure is prevented by keeping the voltage at the I/Os of the intrinsic IC below a certain threshold. Only layout parameters will be varied in the optimization process because an ESD designer usually must work within a given process with fixed junction depths, oxide thicknesses, and doping levels. The methodology described below was implemented in an Advanced Micro Devices 0.5μm technology.

The multifinger structure of Fig. 2.19 has three drain fingers coming off of the input pad and four source fingers connected to V_{SS}, yet there are six parallel NMOS transistors because there are six poly gate fingers and each input finger serves as the drain for two devices. A representation of a multifinger input-protection circuit is shown in Fig. 2.20. All NMOS structures are identical, as are all the PMOS structures. Since interaction between devices affects the overall response to an ESD input, it is simpler to analyze a single device at a time while taking into consideration how it will perform once it is placed in the entire circuit. Thus the design process begins with the layout of NMOS and PMOS "single-finger" structures (individual devices) with varying layout dimensions, including

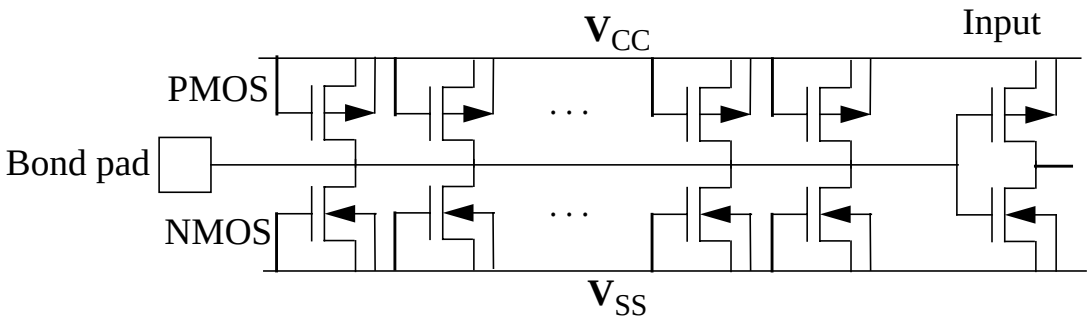

Fig. 2.20 Circuit diagram of CMOS input protection using multifinger structures.

variations in gate length, gate width, and contact-to-gate spacing as well as devices with and without LDD, with and without silicidation, and thin-gate and thick-gate (field) structures. On each test tile there is a common p-well or substrate (V_{SS}) pad for the NMOS devices and a common n-well (V_{CC}) pad for the PMOS devices, but all devices have separate drain, source, and gate contacts to avoid destroying all devices when one device is overstressed. After processing, wafers are diced and the test-tile pads are wire bonded to pins of 24-pin or 28-pin dual in-line packages (DIPs). Gate resistance was not included in the layout of the structures, but a ceramic or chip resistor can be connected externally during testing to investigate gate bouncing. It is debatable whether this lumped-resistor approach accurately represents the use of any type of resistance which can be laid out, and future test tiles may have to include on-chip gate resistors.

In theory, the simplest way to optimize a device is to create an n-dimensional design space, where n is the number of parameters which can be varied, i.e., gate length, gate width, contact-to-gate spacing, etc., and then test all of these devices and note which one performs the best. This procedure would require an impractical number--hundreds or thousands--of devices unless we use a statistical method such as that discussed in Chapter 5. Our approach in this section is to create separate one-dimensional variations of the layout parameters described in the previous section and extract a quantitative dependence of the TLP I-V points as well as HBM and CDM failure thresholds on these parameters. Given these dependences, one or more of the layout parameters can be set to yield optimal device characteristics for robust ESD protection.

The performance of a single protection device is simple to define using the HBM or CDM test because the only characteristic of robustness is the maximum input voltage the device can withstand before the leakage current becomes too high. With TLP analysis, on the other hand, there are several considerations. To prevent dielectric breakdown of the thin input gates or of the thin gate of the protection MOSFET itself, the drain voltage should not exceed the dielectric breakdown threshold, which is about 8V for a 100Å oxide. This means that V_{t1} should not exceed this value during initial turn-on, and V_{sb} and R_{sb} should be low enough that the drain voltage does not move out past the dielectric threshold while in the snapback mode (refer to Fig. 2.6). As mentioned in Section 2.3, there is a time dependence of dielectric failure, so it may be safe for V_{t1} to exceed a steady-state breakdown level as long as the device turns on quickly enough. The MOSFET snapback process occurs on the order of 1ns, so the device should be able to follow any ESD input and clamp it successfully unless the rise time of the pulse is less than 1ns, which may be the case for a CDM stress. V_{t1} should be as low as possible to minimize the chances of dielectric failure and the turn-on time, but it must remain above normal operating voltages so that it does not interfere with the operation of the IC. From the previous section, we expect the gate length and gate-bounce resistance to have the largest effect on V_{t1} and the trigger time, t_1.

To maximize the thermal failure threshold of a single device, the second breakdown current, I_{t2}, or the power to failure, P_f, should be maximized across a range of time to failure, t_f. Since P_f is the product of I_{t2} and V_{t2}, it appears that V_{t2} should also be maximized to raise the P_f vs. t_f curve. However, as just discussed the device voltage should not exceed the dielectric breakdown voltage. Also, if a technique such as increasing the contact-to-gate spacing is used to increase R_{sb} and thus increase V_{t2} for the same I_{t2}, the device has a higher failure power, but the failure current is the same because the extra power is dissipated in the resistance, not in the high $\mathbf{J} \cdot \mathbf{E}$ region, so the device is not really providing any more protection than before. Thus, it has been suggested [24] that an I_f vs. t_f curve is just as valid, if not more valid, for characterizing the thermal robustness of a protection device. Design of a device should focus on maximizing I_{t2} by making the device as wide as possible (within the constraints of the available ESD circuit area), blocking the shallow LDD diffusion, masking silicide deposition, and noting any second-order dependence of I_{t2} on gate length and contact-to-gate spacing.

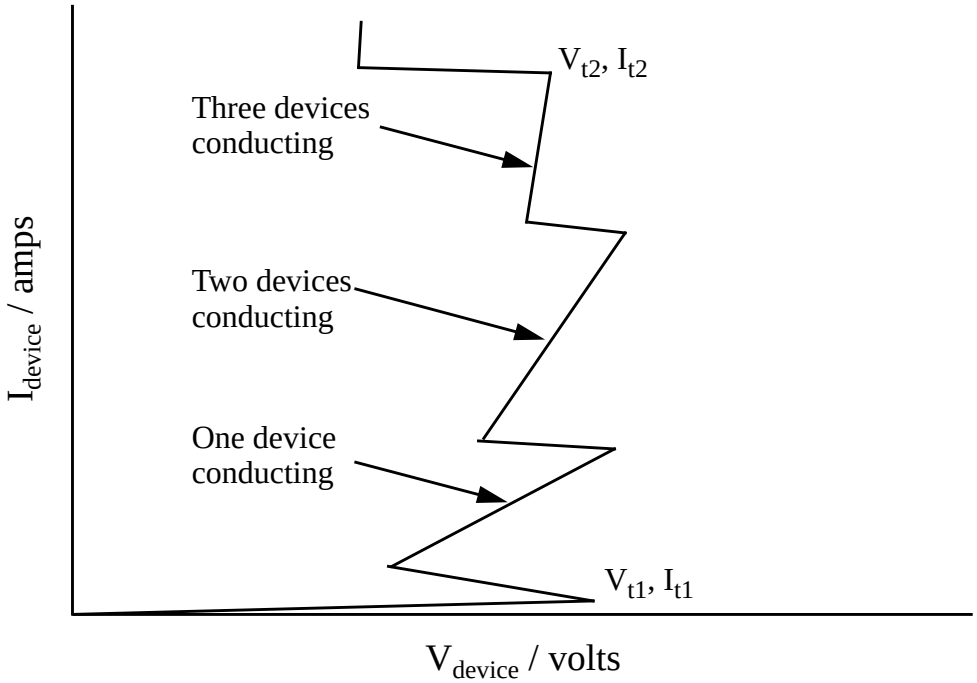

Fig. 2.21 Qualitative TLP I-V curve for an NMOS multifinger structure subjected to a positive ESD pulse (cf. Fig. 2.6). Each snapback indicates another device turning on.

A multifinger protection circuit should have a failure current threshold equal to the sum of the failure currents of the individual transistors, but such circuits have been shown to have a wide distribution of failures as measured by HBM testing [8]. That is, some circuits fail at levels as low as 500V while others are robust out to 2000V. This phenomenon has been traced to nonuniform current flow: in some structures, only one finger turned on and conducted current, leading to thermal failure when the current reached I_{t2} of the single device, while in other identically processed structures two or more fingers conducted, thus raising the failure level. Even though the layouts of all fingers of a circuit are identical, the fingers do not turn on simultaneously due to random variations in processing or the proximity of a finger to the input pad. Once one device turns on, the common drain voltage is clamped at the snapback voltage, so other fingers cannot turn on until the device voltage increases with current beyond V_{t1}. As shown in Fig. 2.21, this process can occur numerous times. After each snapback, the snapback resistance decreases because another device has turned on. If the current level reaches I_{t2} before another device turns on, one of the fingers will

enter second breakdown and incur thermal damage. By designing V_{t2} to be larger than V_{t1}, turn-on of all fingers can be ensured, thus maximizing the thermal-failure threshold.

It is now apparent that optimization of a multifinger MOSFET protection circuit requires more than just optimizing the robustness of the individual devices. The parameters V_{t1}, V_{sb}, R_{sb}, and V_{t2} of the single device must be manipulated so that V_{t2} is greater than V_{t1}. In such multifinger circuits R_{sb} is also called the ballast resistance because it is meant to stabilize the circuit by providing the necessary voltage to turn on all fingers. Adding a poly or diffused resistor between each drain and the common input or increasing the drain contact-to-gate spacing will increase the ballast resistance, but care must be taken not to push V_{t2} beyond the dielectric threshold level. Another way to increase the ballast resistance is to decrease the number or reduce the size of active-metal and interlayer metal-metal contacts, but this technique is dangerous because it increases the current density per contact and thus thermal failure may occur at the contacts. As an alternative to adjusting R_{sb}, V_{t1} can be reduced through gate bounce or, if possible, by floating the substrate. If V_{t1} is reduced to the point where it is less than V_{sb}, i.e., the BV_{ceo} of the parasitic bipolar transistor, then turn-on of all fingers is assured. Again, it is important that V_{t1} not be reduced within the operating level of the IC.

The analysis of one-dimensional layout variations of single-finger structures should suggest which approaches are best for device optimization. Failure analysis, including TLP leakage measurements as well as SEM (scanning electron microscopy) and EMMI (emission microscope for multilayer inspection), should be incorporated in the design process to ascertain where device failures are occurring. As described in the next chapter, numerical device simulations can also be instrumental in designing devices and determining where and how devices will fail. Once the potential single-finger structures have been narrowed down to a few designs, complete multifinger ESD circuits should be laid out and fabricated for testing. The structures should be connected to simple functional circuits which are representative of the actual circuitry being protected in the final IC design to verify that not only is the protection circuit surviving an ESD stress but also is truly protecting the internal circuit from ESD.

Chapter 3

Simulation: Methods and Applications

The general design and analysis capabilities of two-dimensional numerical device simulators were discussed in Chapter 1: semiconductor and oxide regions are defined on a 2D grid, doping profiles and electrodes are specified, coefficients of physical models are set (possibly to calibrate the simulation to an actual process), and electrodes are biased in either a transient or dc mode to simulate the I-V characteristics of the device. Analysis capabilities include 1D, 2D, and contour plotting of the current density, electric field, impact-ionization generation rate, and other position-dependent properties for any solution point[1]. As a result of the introduction of new simulation techniques and the incorporation of lattice temperature modeling into the semiconductor device equations, it is possible to simulate complex, high-current ESD events. The main questions addressed in this chapter are, how can device simulation be used to study ESD phenomena, and what impact can it have on the ESD design process? Qualitatively, simulations of any MOSFET structure can be studied to aid understanding of the physics involved during snapback and second breakdown and suggest how and where a protection circuit will fail. Varying process and layout parameters will reveal the dependences of critical circuit characteristics on the parameters; for design of actual circuits, parametric structures are used to determine these dependences. By calibrating the simulation models to the parametric structures, simulation can be used to verify the experimental trends and optimize the design parameters, thereby replacing costly and time-consuming layout revisions.

1. Current versions of TMA-MEDICI and PISCES-2ET do not provide contour plotting of Joule heating ($\mathbf{J} \cdot \mathbf{E}$), so a C program was written to calculate and plot such contours from printed current and electric field data. The program also can create a contour of $n_i > N$ (intrinsic concentration greater than doping concentration) from lattice temperature and doping data.

One of the most powerful features of device simulation is the ability to examine at any location in the device properties such as temperature, potential, and current density which are not accessible through real measurements. However, the huge quantity of available information is also a drawback because simple results must be extracted from the complex device-simulation models. Although extracting points from a MOSFET snapback I-V curve is straightforward, extraction of a parameter such as the time to failure for a given input power is nontrivial because "failure" is not directly defined in simulation. Instead, it must be determined using some criteria involving the parameters available in the simulation, such as temperature, $\mathbf{J} \cdot \mathbf{E}$ profiles, and sudden drops in device voltage. Interpretation of simulation results is therefore just as important as accurately defining the models. In a way this is the converse of ESD testing, in which a simple leakage measurement determines whether a circuit has failed but the source of the failure cannot be ascertained without extensive testing and failure analysis.

There are of course limitations to the application of 2D device simulation to studying ESD circuits. The accuracy of modeling thermal failure is one of the biggest concerns because there is no way to account for heat dissipation in the third dimension, which becomes important for long stress times. Section 3.6 discusses the implications of 2D modeling on predicting thermal failure. Two-dimensional simulation is also unable to examine edges and corners of devices or to study the susceptibility of semiconductor-metal and metal-metal contacts and interconnects. Mixed-mode simulations can be used to model the separate MOSFET devices of a multiple-finger circuit, but there is no way to model the flow of heat between the closely spaced fingers. For these reasons, the focus of the simulations is on individual devices of an ESD protection circuit, particularly MOSFETs. The following sections present physical models and general simulation techniques which facilitate ESD device simulation and then discuss specific ways in which the models and techniques can be applied. First is a discussion of the facets of simulation which make studying ESD possible: implementation of the thermal diffusion equation, temperature-dependent mobility and impact-ionization models, curve tracing, and mixed-mode simulation. This is followed by a review of published studies on the application of 2D device simulation to ESD. Methods used to model the MOSFET I-V curve, thermal failure, dielectric failure, and latent damage are then discussed.

3.1 Lattice Temperature and Temperature-Dependent Models

The classic heat flow equation (Eq. (2.2)) was presented during the discussion of the 3D thermal box model in Chapter 2. This equation has been coupled with Poisson's equation, the electron and hole current-density equations, and the electron and hole continuity equations to simulate the effects of lattice heating in semiconductor devices (electrothermal simulation) [29,30,44]. The heat-generation term in Eq. (2.2), in W/cm^3, is modeled as

$$H = \mathbf{J_n} \cdot \mathbf{E} + \mathbf{J_p} \cdot \mathbf{E} + H_U, \tag{3.15}$$

where E is the electric field, J_n and J_p are the electron and hole current densities, respectively, and H_U is the recombination contribution and is expressed by

$$H_U = \left(U_{SHR} + U_{Auger} - G^{II} \right) E_g, \tag{3.16}$$

in which U_{SHR} and U_{Auger} are the rates of Shockley-Hall-Read and Auger recombination, respectively, G^{II} is the impact-ionization generation rate, and E_g is the band-gap energy. All four of these parameters are functions of lattice temperature. Since the lattice temperature is no longer spatially constant, the Poisson and current-density equations must be modified. Poisson's equation is now expressed as [45]

$$\vec{\nabla} \cdot \varepsilon \vec{\nabla} (\psi - \theta) = -q (p - n + N_D^+ - N_A^-) - \rho_F, \tag{3.17}$$

where ε is the permittivity, ψ is the electrostatic potential, q is the electron charge, p and n are the hole and current concentrations, respectively, N_D^+ and N_A^- are the ionized impurity concentrations, ρ_F is the fixed-charge density, and θ is the band structure parameter, given by

$$\theta = \chi + \frac{E_g}{2q} + \frac{kT}{2q} \ln\left(\frac{N_C}{N_V} \right), \tag{3.18}$$

where χ is the electron affinity, k is Boltzmann's constant, T is the local lattice temperature, and N_C and N_V are the conduction-band and valence-band density of states, respectively. Additional thermal-diffusion terms are placed in the current-density equations as follows [46]:

$$\mathbf{J_n} = qn\mu_n \mathbf{E} + k\mu_n (T\vec{\nabla} n + n\vec{\nabla} T) \tag{3.19}$$

$$\text{and } \mathbf{J_p} = qp\mu_p \mathbf{E} - k\mu_p (T\vec{\nabla} p + p\vec{\nabla} T), \tag{3.20}$$

where μ_n and μ_p are the electron and hole mobilities, respectively.

To create thermal boundary conditions, thermal electrodes are placed anywhere along the edges of a device in the same manner as electrical contacts and act as infinite heat sinks by enforcing a constant temperature at the contact (Dirichlet boundary conditions). Non-contacted edges obey homogeneous Neumann boundary conditions, i.e., there is no heat flow across non-contacted edges. Lumped linear thermal resistance, in K/W, and capacitance, in J/K, may be placed on a thermal contact to simulate the conduction of heat away from the part of the device defined by the simulation. For example, a lumped resistance may be placed on a thermal contact along the bottom of a structure to simulate the dissipation of heat into the substrate.

3.1.1 Mobility and Impact Ionization Models

Since the lattice temperature is no longer constant throughout a simulated device, the mobility and impact-ionization models must be dependent upon the local temperature. The Lombardi surface mobility model [47] is chosen for low-field and transverse field mobility modeling because it accounts for parallel and perpendicular fields needed to simulate MOSFETs and because it includes lattice-temperature dependence. It is a semi-empirical model with separate terms which account for surface-roughness scattering,

$$\mu_{sr} = \frac{DN}{E_{\perp}^{2}}, \tag{3.21}$$

surface acoustical-phonon scattering,

$$\mu_{ac} = \frac{BN}{E_{\perp}} + \frac{CN \cdot N_{total}^{EN}}{T\sqrt[3]{E_{\perp}}}, \tag{3.22}$$

and bulk mobility,

$$\mu_{b} \neq \text{function}\,(T, E_{\perp}), \tag{3.23}$$

where N_{total} is the local total doping concentration, T is the local temperature, $E_{\perp}$ is the local electric field perpendicular to carrier flow, and BN, CN, DN, and EN are coefficients with different values for electrons and holes. These mobility terms are added in parallel to calculate the overall mobility (Mathiessens's rule) at each point in the simulation space. Other mobility models are available which account for transverse-field and/or temperature effects, but the Lombardi formulation was judged to be the only model which treats both

effects to a reasonable degree. For example, some of the low-field/transverse-field models which do include temperature dependence use only a simple scaling factor to model surface mobility.

In the high-field mobility region, the empirical Caughey-Thomas expression [48] is used to account for velocity saturation. For electrons, the high-field mobility is

$$\mu_n = \mu_{S,n}\left(1 + \left(\frac{\mu_{S,n} E_{||}}{v_n^{sat}}\right)^{\beta_n}\right)^{-1/\beta_n}, \tag{3.24}$$

where $\mu_{S,n}$ is the low-field mobility, $E_{||}$ is the electric field in the direction of current flow, v_n^{sat} is the saturation velocity, and β_n is a fitting parameter. An analogous equation is used for hole mobility. Degradation of mobility at high electric fields is due to high-energy carriers interacting with optical phonons rather than acoustic phonons. Inherent in this situation is that the carriers are no longer in thermal equilibrium with the lattice, i.e., electrons and holes have their own characteristic temperatures. However, since the carrier temperature is related to the local electric field [42], an expression such as Eq. (3.24) allows us to calculate mobility degradation without solving for carrier temperature (such modeling still neglects the non-local effects of extremely high fields on carrier transport). This mobility model is implicitly dependent on the lattice temperature through the temperature-dependent saturation-velocity [29],

$$v_n^{sat} = 2.4\times10^7 / (1 + 0.8\exp(T/600)) . \tag{3.25}$$

Modeling of impact-ionization (II) generation of carriers is essential for the simulation of breakdown and snapback phenomena in ESD protection MOSFETs. The II generation rate can be expressed as

$$G^{II} = \alpha_n \cdot \frac{|\mathbf{J_n}|}{q} + \alpha_p \cdot \frac{|\mathbf{J_p}|}{q}, \tag{3.26}$$

in which α_n and α_p are the electron and hole ionization coefficients, respectively, with units of cm^{-1}. An expression for these coefficients commonly used in numerical simulation is [46]

$$\alpha_{n,p} = \alpha_{n,p}^{\infty} \cdot \exp\left(-\left(\frac{E_{n,p}^{crit}}{E_{||}}\right)^{\beta_{n,p}}\right), \tag{3.27}$$

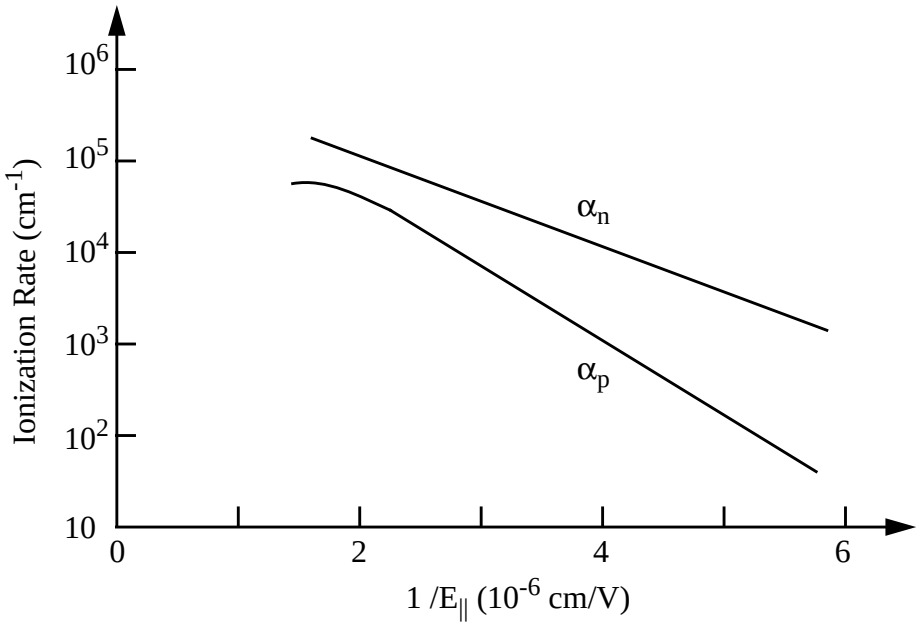

Fig. 3.22 Qualitative plot of impact-ionization rates for electrons (α_n) and holes (α_p) as a function of electric field for silicon at 300K.

where α_i^{∞}, E_i^{crit}, and β_i are fitting parameters (i is n for electrons and p for holes). Note that β_n and β_p, which fall in the range [1,2], are not the same as the coefficients in the Caughey-Thomas mobility expression. Analogous to high-field mobility coefficients, the impact-ionization coefficients are calculated from the local electric field even though impact ionization may best be described as a function of carrier temperature [44]. A qualitative plot of α_n and α_p vs. $1/E_{||}$ for a typical silicon measurement at 300K is shown in Fig. 3.22. Fitting of the II-model parameters is not universal but rather depends strongly on the technology and type of device being tested, and even within the same MOSFET structure much lower ionization-coefficient values will be measured at the surface than in the bulk [49]. By looking at the expression [42]

$$E_{n,\,p}^{crit} = E_g / q\lambda_{n,\,p}, \tag{3.28}$$

where λ_n and λ_p are the optical-phonon mean free paths for electrons and holes, respectively, we see that the lower II generation rate at the surface is most likely a result of the reduced mean free path length due to surface scattering, i.e., the longer a carrier can accelerate in an electric field without a collision, the more probable it is that it will gain enough

energy to create an electron-hole pair via a collision. As temperature increases, the II-generation rate decreases, again as a result of a lower mean free path. This is modeled as

$$\lambda_{n,\,p} = \lambda_{n,\,p}^{300} \cdot \tanh\left(E_p / 2kT\right), \tag{3.29}$$

where λ_i^{300} is the phonon mean free path at 300K and E_p is the optical-phonon energy.

Although our implementation accounts for hot-carrier mobility degradation and impact ionization, it is unable to model other hot-carrier effects such as gate current due to injected carriers. Section 3.7 discusses how hot-carrier induced gate current can be modeled using a “post-processing” simulation tool.

3.1.2 Analysis of Thermal Assumptions

In this implementation the local electron and hole temperatures are set equal to the local lattice temperature, i.e., the carriers are assumed to be in thermal equilibrium with the lattice throughout the device. The consequences of this assumption must be considered for electrical as well as thermal modeling. As discussed in the previous section, high electric fields generate hot carriers, i.e., electrons and holes with a characteristic temperature higher than the lattice temperature, which are responsible for impact-ionization current generation, degradation of mobility, and other phenomena. However, as elaborated in Section 3.1.1, high-field mobility and impact ionization can be modeled as functions of local electric field rather than carrier temperature, and thus the simulations do not have to solve for carrier temperature.

Thermally, electrons and holes must be considered as particles which are separate from the lattice and have their own characteristic heat capacity and thermal conductivity. By setting the carrier temperatures equal to the lattice temperature we are neglecting the carrier heat capacity and thermal conductivity or, at most, we are combining the carrier and lattice contributions. The heat capacity of electrons and holes is $(3/2)nk$, where n is the carrier concentration and k is Boltzmann’s constant. The heat capacity of silicon, which increases with temperature, is 1.63J/K-cm^3 at 300K. Carrier heat capacity is equivalent at a density of $7.9\text{X}10^{22}\text{cm}^{-3}$, about four orders of magnitude higher than the doping level in the high-field region of a MOSFET. Even if the carrier temperature is 100 times greater than the lattice temperature, the heat content in the carriers ($(3/2)nkT_e$, where T_e is the carrier temperature) will only be 1% of the content in the lattice.

Although the heat capacity of the carriers is lower than that of the lattice, it is actually the relative thermal conductivity, i.e., how much heat is transported by the carriers, that is of primary consideration. Heat flux in the lattice at any point is equal to the product of the thermal conductivity of the lattice, κ, and the gradient of the lattice temperature at that point. Similarly, heat flux due to diffusion of carriers is equal to the product of the carrier thermal conductivity, κ_e, and the gradient of the carrier temperature. Carrier thermal conductivity is a function of the carrier temperature [44]:

$$\kappa_e = \frac{3}{2} n k^2 T_e \mu_e / q = \frac{3}{2} n k D_e , \tag{3.30}$$

where μ_e is μ_n for electrons and μ_p for holes and D_e is the carrier diffusion constant. Carriers also contribute to heat conduction via thermoelectric energy current, i.e., heat current due to electrical current. This component of heat conduction is formulated as [44]

$$\mathbf{s}_{n,j} = \frac{3}{2} \cdot \frac{kT_e}{q} \cdot \mathbf{J}_e , \tag{3.31}$$

where J_e is the current density.

To determine the relative contributions of lattice and carriers to thermal conductivity in an ESD application, we analyze a simulation of a 0.5μm-technology NMOS transistor under high-current stress at the time the peak lattice temperature in the transistor has reached the melting point of silicon, 1688K (such simulations are discussed in more detail later in Chapter 3 and in Chapter 4). The peak temperature is in the high-field region of the LDD and the current in this region consists mainly of electrons, which are assumed to have a concentration of $5X10^{18}cm^{-3}$ (the LDD doping concentration). Over the high-field region the average electric field is $4X10^5V/cm$ and the average lattice temperature is 1000K. The saturation velocity, v_n^{sat}, is calculated from Eq. (3.25) as $4.6X10^6cm/s$. Using Eq. (3.24) with a β_n of 2 and a low-field mobility of $140cm^2/V\text{-}s$ (corresponding to a doping level of $5X10^{18}cm^{-3}$ [61]), the average mobility is $11.5cm^2/V\text{-}s$ in the region.

The electron temperature, T_e, can be calculated from the electric field using [42]

$$qEv_n^{sat} = \frac{3}{2} k (T_e - T) / \tau , \tag{3.32}$$

where E is the electric field, T is the lattice temperature, and τ is the energy relaxation time of electrons in silicon and is assumed to be 0.3ps. From Eq. (3.32), the average electron temperature in the high-field region is approximately 5300K which, using Eq. (3.30), yields a κ_e of $5.4X10^{-4}$W/cm-K. By contrast, the silicon lattice has a thermal conductivity of 0.31W/cm-K at 1000K [29]. While the thermal conductivity of the lattice is almost 1000 times greater than that of the electrons, the ratio of lattice to carrier heat diffusion is less than 1000 because the electron temperature gradient is greater than the lattice temperature gradient. The extent of the high-field region is about 0.2μm in the lateral dimension (the direction of current flow, parallel to the silicon surface), and in the center of the region the peak temperature is 1688K for the lattice and, again using Eq. (3.32), about 5950K for the electrons. Assuming the lattice and electron temperatures are 300K at the boundaries of the high-field region, i.e., assuming maximum thermal gradients, the thermal flux in the lateral dimension is $4.3X10^7$W/cm^2 for the lattice and $3.0X10^5$W/cm^2 for the electrons. Therefore, the contribution of heat flux due to carrier diffusion is less than 1% of the total flux.

Heat flux due to electron current must be calculated from the current density in the drain junction. When the lattice temperature reaches 1688K the drain current is about 10mA per μm of device width, of which 60% conducts laterally toward the source and 40% conducts vertically to the substrate. The lateral current conducts uniformly through the high-field region, which has a depth of 0.2μm as determined by the depth of the LDD junction, and thus the current density in the high-field region is $3X10^6$A/cm^2. Using Eq. (3.31) with the average electron temperature of 5300K, the resulting heat flux due to current conduction is $2.0X10^6$W/cm^2, or about 5% of the value of the lattice contribution.

From this analysis we conclude that assuming thermal equilibrium between lattice and carriers leads to an approximately 6% underestimation of thermal dissipation away from the region of heating. One implication of the reduced heat flux is a higher peak lattice temperature in the device at any given time in a simulation, which may be interpreted as a lower failure threshold for the device (simulation of thermal failure is discussed in Section 3.6). However, in light of other uncertainties of simulation discussed in Chapters 3 and 4, a 6% error is reasonably good and thus the assumption of thermal equilibrium between lattice and carriers is valid under most conditions. Electric fields, currents, and carrier concentrations can be monitored during any simulation to quantify the error of the assumption.

3.2 Curve Tracing

Since the thermal-diffusion equation and temperature-dependent mobility and impact-ionization models have been incorporated in 2D device simulation, it is theoretically possible to simulate the MOSFET snapback curve with a dc sweep of the drain voltage. However, this curve (refer to Fig. 2.6b) is complex in the sense that there are very flat regions where the current changes little with voltage, steep regions where the current rises rapidly with voltage, turning points where the slope of the curve changes sign, and multivalued voltage solutions. Simulating this curve with traditional methods is complex because the boundary conditions must be adapted to maintain stability. Experience has shown that a voltage boundary condition (BC) on the electrode being swept is stable if the current does not change "too fast" with the applied bias. On the other hand, a current boundary condition is effective if the I-V curve is very steep, i.e., if the voltage necessary to sustain a certain current is not "too sensitive" to the required current level. These observations are shown graphically in Fig. 3.23, which shows that solving with a voltage BC is equivalent to finding the point on the I-V curve which intersects with the vertical line defined by the voltage, while a current-BC solution is represented by the intersection of the curve with a horizontal line. In general, a solution is stable when the line defined by the boundary condition is perpendicular to the local part of the I-V curve. Thus the load

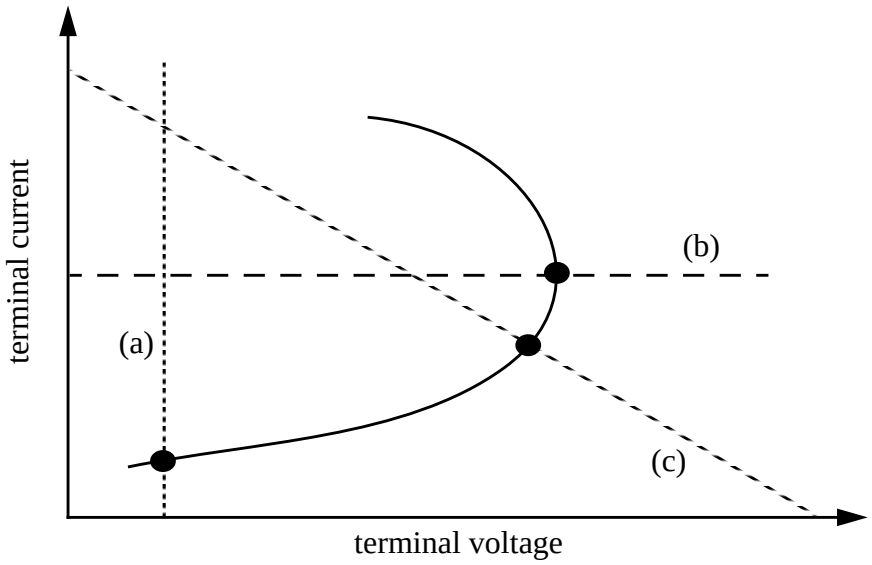

Fig. 3.23 Schematic representation of various types of bias specification: (a) voltage control, (b) current control, (c) load-line control. The simulator will converge to the intersection of the (dashed) constraint line and the I-V curve.

line (c) in Fig. 3.23, which represents a voltage or current source with an internal load resistance, is the ideal boundary condition for the part of the I-V curve with an intermediate slope. This type of boundary condition is available in simulation.

The MOS snapback curve can be simulated by using a voltage BC on the drain during the initial reverse bias, switching to a current BC when the current increases rapidly (on a log scale) after junction breakdown, switching back to a voltage BC at the turning point and stepping the voltage negatively during snapback, then finally switching back to a current BC to trace the curve in the snapback mode. Such a process is time consuming and requires *a priori* knowledge of the curve characteristics since the user of the simulator must know where to change the boundary conditions. A large, fixed load resistor could be placed on the drain with a voltage boundary condition to effectively remove the turning points and multivalued solutions, but this resistance must be greater than the differential resistance at any point in the I-V curve, which again requires knowledge of the curve prior to simulation. The general solution to the curve-tracing problem is to continuously change from pure voltage to pure current control by using a voltage or current source with a load (external) resistor which changes at each solution point to keep the load line perpendicular to the local section of the curve and thus ensure convergence throughout the trace (Fig. 3.24) [28]. This scheme can be automated because its implementation relies only on information readily available from the simulator, *viz.*, the voltage, current, and slope (tangent) of each solution point. In Stanford's 2D device simulator, PISCES-IIB, the tangent information is directly available from the Jacobian matrix and can be printed out for the user when the Newton-projection method is used [44]. If the tangent is not available directly, the local slope of a curve can be approximated by solving at a nearby point for each solution and using the difference method. Note that in this dynamic-load-line method a negative differential resistance implies a negative load resistance, a condition which is totally acceptable from a simulation standpoint.

There are two main steps in curve-tracing simulation. Once a solution point has been found on an I-V curve using an external voltage (a voltage source will be assumed from here on) and a load resistance which yield a perpendicular load line, the solution is *projected* to the next point on the I-V curve via advancement of the external voltage (Fig. 3.25a). Projection along the tangent always provides the best guess for the next solution point. Once this new solution converges, the tangent of this new point is calculated and the point is re-solved using a *recalibrated* load resistance and external voltage which yield a

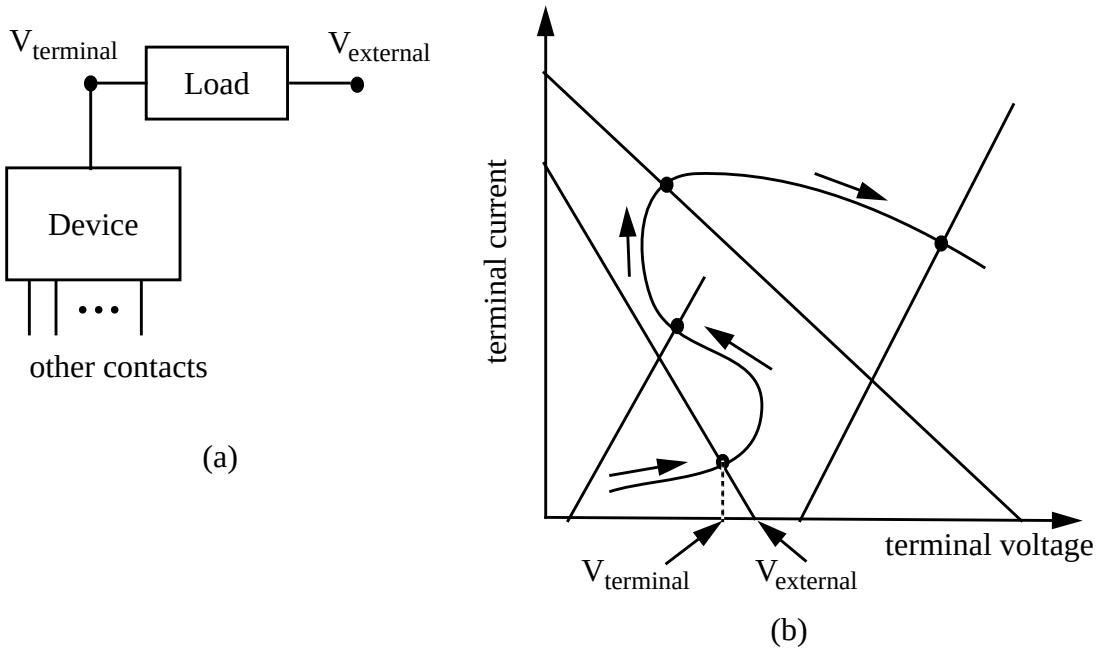

Fig. 3.24 *(a) Schematic of a general device with external load and voltage; (b) adapting the load line (solid lines) along the I-V curve allows for optimal convergence at each operating point.*

load line perpendicular to this new point (Fig. 3.25b). Projection and recalibration are then repeated until the trace is complete. The scheme must keep track of turning points in a curve to ensure that the external voltage is always projected in the right direction. For example, in a trace of a MOSFET snapback, the load resistance and external voltage steps are positive before the trigger point, but when the curve's slope becomes negative at the onset of snapback, the perpendicular load resistance must also become negative and the external voltage must be stepped negatively. Turning points are more fully discussed in [28], as are issues concerning how to keep the curve trace smooth, how projection step sizes are determined, and the necessity of a scaling scheme.

With this method, a simulator can automatically generate any arbitrarily shaped I-V curve given only a user-specified starting point, ending point (maximum voltage or current), and initial step size. The scheme has been implemented as a C program, "Tracer," a virtual instrument which functions as a wrapper around any device simulator which supplies voltage, current, and tangent information, i.e., no modifications need to be made to the simulation code. Tracer communicates with a device simulator by modifying the simulator

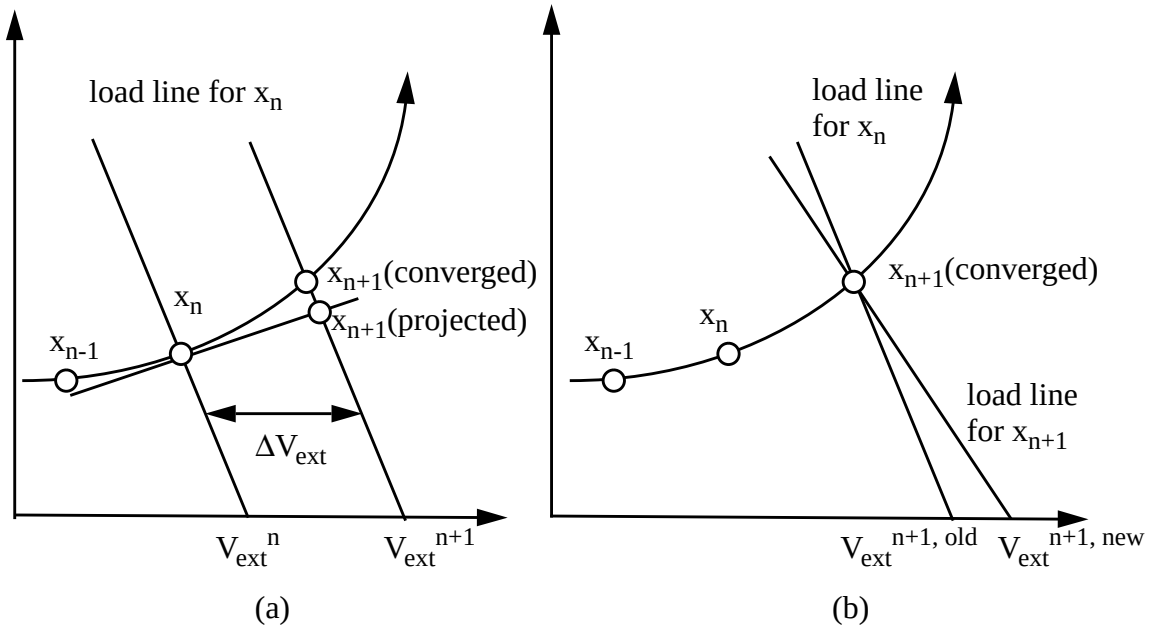

Fig. 3.25 *(a) Projection: the solution is advanced by incrementing the external voltage from V_{ext}^{n} to V_{ext}^{n+1} while the load resistance is held constant; (b) Recalibration: the load resistance is changed so that the load line is perpendicular to the trace at the new solution point. This implies a new external voltage ($V_{ext}^{n+1,new}$).*

input deck (input file) and by parsing information in files generated by the simulator. A user must supply a standard PISCES (or other simulator) input file describing the device to be simulated and a specification file with a PISCES-like syntax delineating the starting point, ending point, and initial step of the node to be swept; fixed boundary conditions used at other device nodes; and which information is to be saved as well as some optional parameters. A complete user's guide for Tracer is presented as an appendix, containing a description of all the parameters in the specification file, requirements of the PISCES input file, and detailed examples which include all input and output files.

3.3 Mixed Mode Simulation

In many numerical device simulators, lumped resistors and capacitors can be placed between the electrodes of the defined 2D device and an external ground. Such elements are useful for simulating the effects of parasitics surrounding the device, e.g., resistance due to inter-layer metal-metal contacts or test probes. Recent advances, however, have

created the capability of embedding one or more numerically simulated devices in a SPICE-like circuit complete with lumped resistors, capacitors, and inductors as well as voltage sources, current sources, and compact models for diodes, MOSFETs, and BJTs. This method is known as mixed-mode simulation. The total circuit model can be solved in either a coupled manner, in which the semiconductor equations (Poisson, continuity, and lattice temperature) describing the devices and the Kirchhoff equations describing the circuit are solved as a coupled set [50], or in a decoupled manner in which an interface is created between SPICE and a device simulator with the device simulator iterating to completion once for each SPICE iteration [51].

Mixed-mode simulations are very useful for transient modeling of ESD tests such as the HBM, MM, and TLP. Using only device simulation, square-wave inputs with variable ramp times can be defined and applied through a series resistor to the drain contact to simulate the simple TLP test shown in Fig. 2.5b. A resistance may also be placed on the gate to study the effects of gate bounce. This type of simulation is all that is needed to generate the I-V points of the MOSFET snapback curve. However, if a more complex setup (Fig. 2.14b) needs to be accurately simulated, mixed-mode simulation is required to define the resistor network. It is also necessary to use mixed-mode for simulations with more complex input waveforms, such as the HBM and MM. In this case, lumped circuit elements are used to create a circuit which yields the proper input current waveform, as in Fig. 2.2a, and parasitic elements can be included. Since the generated waveforms are specified for a short-circuit load, a simple SPICE simulation can be used to verify that the element values yield the proper waveform. This lumped-element circuit can then be defined in the device simulator and a 2D structure can be defined for the DUT. Note that a width must be defined for the 2D device to convert the current units of Amps/μm for the 2D device to Amps for the lumped circuit elements. An example of a human-body model simulation is shown in Fig. 3.26. Notice that if there is no switch model available (as is the case in TMA-MEDICI version 1.1 [29] and in Fig. 3.26), a voltage square-wave source can be placed in series with the 100pF capacitor (C_c) and 1500Ω resistor (R_c). SPICE simulations show that the short-circuit-load waveform generated by this circuit is equivalent to the one generated by the precharged capacitor and switch of Fig. 2.2a provided the square pulse has a very short rise time, i.e., one to two orders of magnitude less than the rise time of the actual waveform. Using such a small rise time ensures that it is the circuit, not the voltage source, which is defining the waveform. Since multiple device structures can be placed in

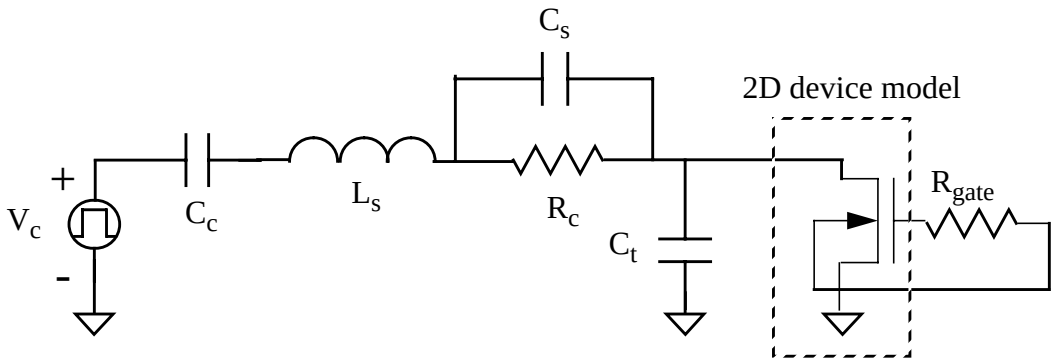

Fig. 3.26 Mixed-mode circuit model for an NMOS transistor subjected to the human-body model. The voltage source and all circuit elements are defined with SPICE models, except for the transistor, which is defined by the 2D device simulator.

a mixed-mode circuit (up to 10 in TMA-MEDICI), two or more MOSFETS could be placed in parallel to simulate a multiple-finger ESD structure (Fig. 3.27). Slight layout variations between the structures can be introduced to model random variations in processing which result in nonuniform turn-on. The circuit can then be modified to ensure turn-on of all fingers, perhaps by incorporating a ballast resistor on each drain.

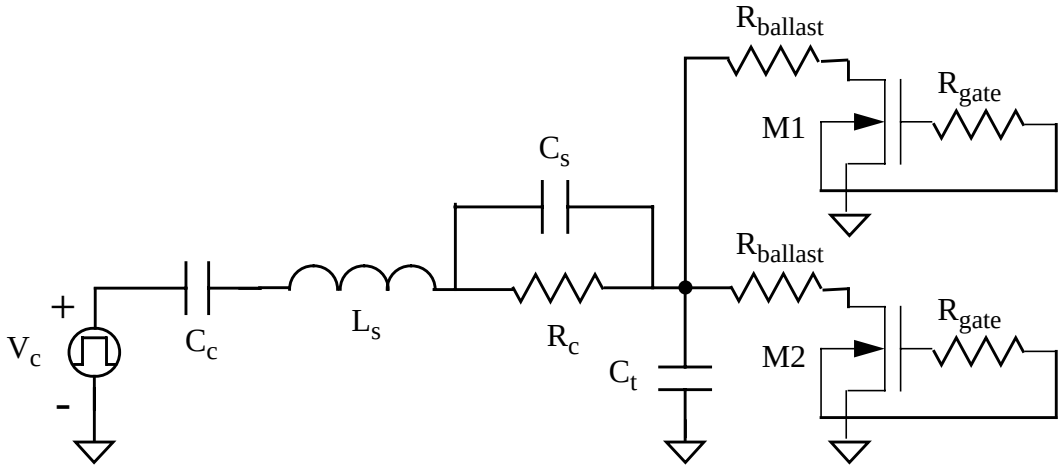

Fig. 3.27 Mixed-mode circuit model for a multiple-finger ESD NMOS structure subjected to the human-body model. Ballast resistors are placed between the output of the HBM circuit and each drain to facilitate uniform turn-on of transistors M1 and M2.

3.4 Previous ESD Applications

Several papers have been published on the subject of ESD in which the use of device simulation is either the main issue or an essential subtopic. Most of these involve the use of electrothermal simulation in order to study thermal-failure mechanisms, while some have looked at the dependence of the trigger point on device layout and the input pulse profile. Very few make use of tools such as curve tracing and mixed-mode simulation. This section reviews the main points of some significant publications on the application of 2D device simulation to the ESD problem in order to show how simulation can be used to study ESD and to highlight areas which have not yet been investigated.

The phenomenon of second breakdown was studied using 2D electrothermal simulations by Mayaram et al. [13]. Temperature and potential profiles in diodes, pn junctions, and MOSFETs subject to transient square-wave pulses (a voltage ramp with a given height and rise time) were monitored to determine the conditions necessary for thermal runaway. The authors determined that the onset of second breakdown, as defined by a drop in the device voltage, has a distinct mechanism in resistive regions and junction regions. In a uniformly doped resistive region the classical definition of the onset of second breakdown, intrinsic concentration (n_i) = doping concentration (N), holds because heating only has an effect on carrier mobility and n_i. In a reverse-biased junction, however, the high-temperature reduction of the impact-ionization rates must also be taken into account, so the classical definition no longer holds. They conclude that “a simple condition for the onset of second breakdown cannot be derived” in a complex device structure with junctions and nonuniform doping, but they did not really examine any conditions other than $n_i = N$. They also remark that 2D simulations underestimate the level of current needed for device failure because the lack of heat flow in the third dimension implies a higher peak in the temperature profile. However, they did not quantify the underestimation or examine their observation to see if it is true regardless of the duration of an ESD pulse.

Chatterjee et al. [33] used TMA-PISCES-IIB, which does not have thermal modeling or mixed-mode capabilities, to simulate ESD protection circuits for a BiCMOS technology in which the vertical npn transistor is the primary protection device, i.e., the pad to be protected is tied to the collector of an npn transistor with grounded emitter which breaks down to absorb ESD current if the input voltage exceeds ~BV_{CEO}. Transient simulations are used with the ESD stress modeled by a voltage ramp of specified rise time, t_r, and peak

value incident at the collector. A resistor, R_b, is placed between the base contact and ground to couple the base voltage to the input voltage via the collector-base junction capacitance. This resistor facilitates turn-on of the transistor by forward biasing the base-emitter junction and thus is similar to the MOS gate-bounce technique depicted in Fig. 2.17a. The purpose of the simulations was to determine the effects of t_r, R_b, and the device geometry on the trigger voltage of the circuit, with a design goal of keeping V_{t1} below a critical value. They found that even when R_b is set to its upper limit (as determined by the required switching time of the circuit), the npn will not turn on if the pulse rise time is greater than about 10ns because the base voltage is not sufficiently biased. To solve this problem extra coupling of the base to the input was provided by placing a MOSFET in parallel with the BJT with the collector tied to the input, source tied to npn base, and gate tied to ground through a large resistance (Fig. 3.28). Similar to the coupling techniques

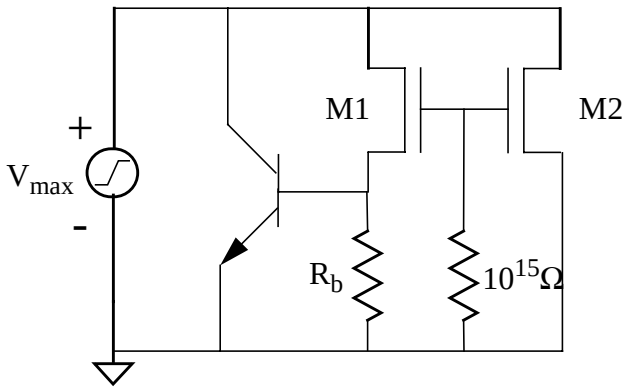

Fig. 3.28 ESD protection circuit used for SPICE simulations by Chatterjee et al. [33]. M1 is an NMOS transistor designed to facilitate the turn-on of the npn transistor during ESD. M2 represents the output NMOS transistor being protected.

described in Section 2.3, the NMOS device will turn on during an ESD pulse to form a channel between the input and npn base to turn on the npn transistor. SPICE simulations were used to verify the design. Since SPICE cannot model the npn breakdown, circuit simulation is only used to determine if the base is sufficiently biased for a given layout and input pulse. (Using contemporary simulators the entire circuit response can be modeled with mixed-mode simulation, using PISCES to model the BJT and either PISCES or SPICE to model the NMOS transistor.) The authors concluded that their modeling

methodology “may be used to achieve a successful first-pass design” and that device simulations are useful for determining qualitative relationships such as the effect of the npn junction capacitances on the trigger voltage.

Use of 2D device simulation in predicting ESD robustness was studied by Amerasekera et al. [32], who investigated the use of simulated peak power density ($\mathbf{J} \cdot \mathbf{E}$), peak temperature, and second-breakdown trigger current, I_{t2}, as relative figures of merit of MOS devices with various source/drain profiles, contact-to-gate spacings, and gate biasing. A Texas Instruments in-house electrothermal simulator was used to generate dc curves which exhibited snapback and, surprisingly, second breakdown (a drop in device voltage due to thermal runaway is usually not observed in 2D dc simulations due to the 3D nature of the phenomenon). Thermal electrodes with a lumped resistance of 10^6 K/W were placed on each of the four electrical contacts. The authors found that reaching a critical temperature is a better figure of merit than reaching a critical $\mathbf{J} \cdot \mathbf{E}$ because the peak electric field is very dependent on the simulation grid, which is different for different structures. Using simulated I_{t2} as a failure criteria was found to agree qualitatively with experiments of varying drain junction profiles and to agree quantitatively with experimental I_{t2} vs. gate bias. On the other hand, simulated I_{t2} did not increase with drain contact-to-gate spacing as it does in experiment, leading the authors to conclude that it is not possible to model the effect of some layout parameters on ESD robustness because the simulation is only two dimensional. It is important to note, however, that they are looking at dc results, i.e., EOS, not ESD. Since ESD events are very brief, the effects of thermal diffusion in the width dimension may not have an impact on the device robustness and no conclusions should be drawn from dc simulations on modeling the ESD regime.

Transient simulations were also run with constant-current pulses used as the ESD input. A good fit of transient simulation points to an experimental P_f vs. t_f curve between 25ns and 200ns of a 0.6μm device was obtained by defining failure as the time at which the peak temperature reaches 1000K. (Experimentally, failure is the point at which a device enters second breakdown.) The analytic thermal model (Section 2.2.2) was also fit to the data using a T_c of 1000K and box dimensions of c = 0.5μm, b = 0.5μm, and a = device width. The good agreement of the P_f vs. t_f results led the authors to conclude that “the concept of a critical temperature for (thermal) breakdown is valid for the devices investigated in this study.”

In a slight departure, or perhaps combination, of the methods used by Amerasekera, Kuper et al. [4] looked at $\mathbf{J} \cdot \mathbf{E}$ contours in the drain region of a MOSFET during a transient simulation for devices with and without an LDD implant. In both drain profiles a hot spot (peak in $\mathbf{J} \cdot \mathbf{E}$) forms deep in the junction, but their simulations predict that a shallow LDD diffusion creates a second hot spot just under the gate which could lead to an "early subsurface second breakdown." This spot may heat up more quickly since it is directly under the insulating gate, but it is more localized and thus will only slightly damage the device. The authors conclude that soft failures in LDD structures, defined as a relatively small increase in leakage (less than 1μA) due to a moderate ESD stress, may be a result of the second hot spot seen in the simulations.

Diaz et al. [24] also used 2D electrothermal device simulations (TMA-MEDICI) to study thermal breakdown, in this case for 0.6μm MOSFETs subjected to square-wave pulses. By running transient simulations with different pulse lengths and monitoring peak device temperature and drain voltage, they constructed simulated P_f vs. t_f and I_{t2} vs. t_f curves between about 50ns and 400μs (a broad range of the EOS spectrum) for devices with various drain and source contact-to-gate spacings and compared the P_f vs. t_f results to experiments. Experimentally, failure was defined as "a change in the device leakage characteristics," while for simulations failure was defined by either a drop in the drain voltage (second breakdown) or the maximum device temperature exceeding the melting point of silicon (1688K), whichever occurred first. Only one thermal contact was placed along the bottom of the simulated device, with a lumped thermal resistance and capacitance to model heat conduction into the majority of the substrate that is not included in the simulation space. Qualitative study of the temperature, potential profiles, and current flow lines in the simulations suggested that device failure was due to second breakdown in the drain depletion region. Peaks in the temperature profiles along the gate oxide-silicon interface at the time of failure were very sharp and narrow for short times but much broader with a large high-temperature region for long stress times. The variation in peak temperature with failure time lead the authors to conclude that "it is not possible to define the onset of device failure, particularly the onset of second breakdown, in terms of a unique temperature value."

Simulated P_f vs. t_f curves were higher for devices with larger contact-to-gate spacing, in qualitative agreement with experiments. However, the simulated failure power was too low for failure times less than about 20μs and too high for times greater than 20μs. The

authors attributed the discrepancy at low times to the two-dimensional nature of the simulation and the discrepancy at high times to the oversimplified lumped thermal elements used to model heat conduction through the bottom of the device, leading them to determine that 2D device simulation is only useful for qualitative studies of thermal failure. They do not consider that the underestimation of the failure power for short pulse times may be a result of using the failure criterion of $T_{peak} > 1688K$, which may not be correct. For instance, it is possible that melting does occur in short-pulse experiments but that the damage is so localized that the measured increase in leakage is not significant. If this is the case, then a simulation should not be considered to have reached failure until a later time, such as when a critical temperature has been exceeded over a "significant" region of the device.

In contrast to Amerasekera's results, Diaz found that the simulated failure current, I_{t2}, does increase when the contact-to-gate spacing is increased. V_{sb} and R_{sb} also increased in transient simulations when the contact spacing was increased. The conflicting results between Amerasekera and Diaz are most likely due to the different types of simulations used, i.e., dc vs. transient, and they underline the importance of considering the time range of interest when qualifying ESD circuits. The fact that one study found that defining a critical temperature for failure is valid while the other study found this to be invalid may also be attributed to the different types of simulations used as well as to the different thermal boundary conditions used. We can conclude from Amerasekera's and Diaz's studies that defining failure in simulation depends not only upon the type of criteria chosen but also on the thermal boundary conditions.

3.5 Extraction of MOSFET I-V Parameters

As discussed in Chapter 2, generating an I-V curve using transmission-line pulsing is an excellent way to study how a device will respond to an ESD stress: the trigger point (V_{t1}, I_{t1}) indicates the maximum voltage allowed at the input of the circuit before the protection device turns on as well as the amount of current needed to turn on the device; the snapback voltage and snapback resistance determine what the input voltage will be when a given amount of current is conducting through the device; and the second breakdown point determines the maximum power the device can absorb before thermal damage is incurred. All of these circuit parameters can be extracted from device simulations to aid the process of device design. Three types of I-V curves can be generated from simulation (or from

experiments, for that matter): the curve of a single transient pulse, the curve produced by a series of TLP simulations with increasing input-pulse heights, and a single dc curve-tracing sweep of the drain. Although the TLP-generated curve yields the most information, comparing and contrasting the other types of curves with the TLP curve offers important insights. If there is no coupling of other electrodes to the device input, i.e., if the gate, source, and substrate are grounded and the pulse rise time is a few nanoseconds or greater, then the TLP points should coincide with the dc curve (see Fig. 3.29a) until heating effects become important. However, when transient effects are important, e.g., by placing a resistor from gate to ground to induce MOS transistor action which aids device turn-on, the trigger point will be reduced in the TLP simulations but remain the same in the dc curve trace (Fig. 3.29b). TLP simulation points for grounded-gate devices are equivalent to dc-sweep points because each point taken from a TLP stress is the quasi-steady state value taken after the settling of turn-on and snapback transients (refer to Fig. 2.9). This point will differ from a steady-state point only at high currents when Joule heating changes the resistivity of the silicon. When possible, a single curve-tracing simulation should be used in place of numerous transient simulations to save significant computation time.

In a single transient TLP simulation, a square wave with a rise time of about 1ns is incident on the drain of the device under test through a lumped series resistor which models the transmission-line impedance. If the pulse travels through a more complex resistor network, such as in Fig. 2.14b, mixed-mode simulation is required. During a single TLP simulation the I-V curve traced out with time does exhibit breakdown and snapback, but as shown in Fig. 3.30 for a 50V input pulse, the drain voltage and current do not follow the path of the TLP curve. This is in accordance with the measured current and voltage of Fig. 2.9 and the discussion in Section 2.2.1. The trigger voltage and current are lower than the (V_{t1}, I_{t1}) point in the TLP curve because of the increased gate bounce: V', the input ramp rate (see Eq. (2.14)), is higher for the 50V pulse than for the pulse used to generate the TLP trigger point, so the gate coupling is higher. After breakdown the voltage does not snap back all the way to V_{sb} but rather simply decays to its final value as determined by the current level. The transient curve of a single TLP simulation is not useful in itself, but the quasi-steady state I-V points of several TLP simulations are needed to create a TLP I-V curve, just as they are in experiment. Individual simulations are needed, however, to examine thermal failure during an ESD pulse.

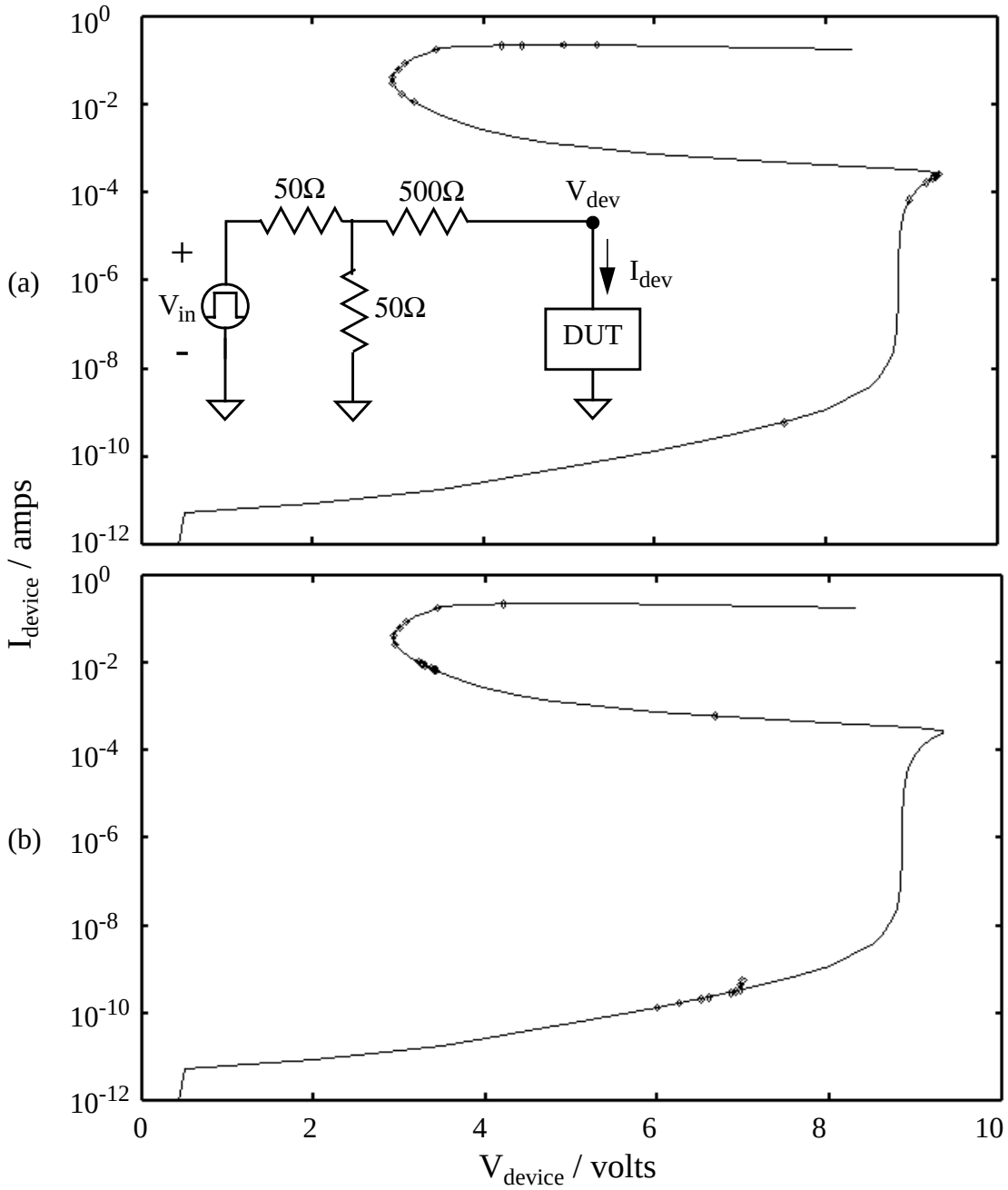

Fig. 3.29 *I-V curves for curve-tracing (solid line) and TLP (points) simulations for a 20/0.5μm MOSFET with (a) a 2000Ω gate resistor and (b) an 8000Ω gate resistor, with the TLP circuit shown inset. Each point represents one non-catastrophic (maximum temperature < 1688K) 100ns TLP simulation with a unique pulse height. The 2000Ω TLP results are virtually identical to the curve-tracing results while the 8000Ω results are markedly different. In the 8000Ω TLP simulations the device current jumps from 1nA to 7mA with only a 0.06V increment in the pulse height.*

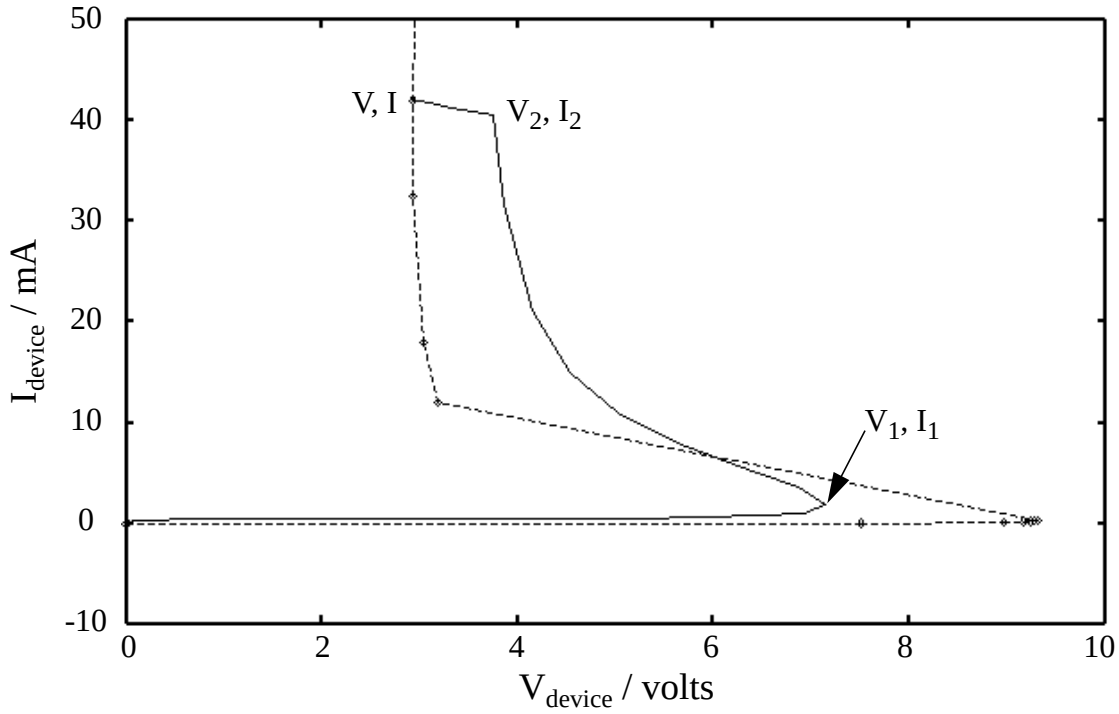

Fig. 3.30 *I-V curves of a single TLP simulation (solid line, V_{in}=50V in the circuit of Fig. 3.27) and of points resulting from a group of TLP simulations (dashed line). Point (V_1, I_1) corresponds to the turn-on of the parasitic bipolar transistor. V_2 and I_2 are the device voltage and current values at the time the input pulse reaches its peak; in this case t_{rise} = 1ns. The quasi-steady state of the pulse is the point (V, I).*

In studying the I-V characteristics of a simulated ESD protection transistor in the next chapter, the general strategy will be to run a dc curve trace to extract the snapback voltage, V_{sb}, and snapback resistance, R_{sb}, and then run transient TLP simulations to extract the trigger point and second breakdown/thermal failure point. Different sets of TLP simulations must be run for each iteration of a gate-bouncing implementation in order to see the effects on the trigger voltage and current. Simulation of the second-breakdown portion of the I-V curve is very important in itself and is the subject of the next section.

3.6 Extraction of MOSFET P_f vs. t_f Curve

Although inclusion of the thermal-diffusion equation in device simulation is useful for studying phenomena such as high-temperature degradation of mobility and impact

ionization, for ESD simulation its most important application is the modeling of gross device heating which leads to thermal runaway. In transient simulations, if the conduction of heat away from a device is accurately modeled by the thermal boundary conditions and if the defined device geometry and doping profiles produce the proper current densities and electric fields, electrothermal simulation should be able to predict at what time thermal failure will occur for a given input pulse and thus to generate a failure power vs. time to failure curve. Thermal runaway is inherently a three-dimensional phenomenon because the hot spot always forms at a point in a device, and after formation current rushes into the spot from all directions. Heat conduction theory predicts that if current is flowing uniformly across the width of a device and the device is surrounded by a spatially invariant heat sink, the hot spot will form in the center of the width dimension because this is the point of peak temperature. (Experimentally, it has been found that thermal runaway may originate at a "weak spot" where the electric field is slightly higher due to the erose drain edge of the gate oxide [52].) In contrast, 2D simulation can only model current rushing in from two dimensions after a hot spot forms. Although it cannot properly model the runaway itself, if current flows relatively uniformly in a device before second breakdown and the simulation cross-section is representative of the real cross-section containing the "weak spot," 2D simulation should be able to predict the *onset* of second breakdown, i.e., the time at which the device voltage drops due to a reduction in overall device resistance. The simulated voltage does fall off with time after the onset of breakdown due to the negative differential resistance, but not as sharply as seen experimentally (e.g., Fig. 2.10) because current cannot rush in from the third dimension.

It is illuminating to apply an analysis like that of the 3D thermal box model in Section 2.2.2 to 2D device simulation. If the assumptions are analogous, i.e., if all power generation occurs uniformly within a rectangle in the drain depletion region and second breakdown follows instantaneously when the peak temperature reaches a critical value, then it appears that the governing equation for peak temperature is just like that of the 3D case (Eq. (2.3)) except there are only two dimensions:

$$T(t) = T_0 + \frac{P'}{\rho C_p (bc)} \int_0^t \operatorname{erf}\left(\frac{b}{4\sqrt{D\tau}}\right) \operatorname{erf}\left(\frac{c}{4\sqrt{D\tau}}\right) d\tau . \tag{3.33}$$

Note that the width dimension, a, is omitted and the power, P, has been replaced by P′, the power per width in W/cm, which is the product of the voltage and the current per width in

A/cm (we may consider P' to be equal to P/a). Presumably, the fact that there is no way to model heat flow in the third dimension is equivalent to setting $a = \infty$, which implies that $\mathrm{erf}(a/(4\sqrt{D\tau})) = 1$, so this term drops out of Eq. (3.33). Solving this equation for times less than the time constant $t_c = c^2/4\pi D$ yields

$$P_f' = \rho C_p bc\,(T_c - T_0)/t_f \text{ for } 0 \le t_f \le t_c, \tag{3.34}$$

which is analogous to Eq. (2.6) for the 3D case. Solving for longer times yields equations equivalent to Eq. (2.7) and Eq. (2.8), except the upper time limit of Eq. (2.8), t_a, is replaced with ∞ since such a time constant has no meaning in the 2D model. As a consequence of this limit, however, note that as the failure time in Eq. (2.8) becomes very large, the power to failure tends to zero, implying that in the 2D case no matter how low the applied power is, if it is applied long enough the peak temperature will eventually reach the critical value T_c. This is clearly nonphysical and is not observed in simulations. Indeed, from a result in Carslaw and Jaeger [31] for the steady-state 2D temperature profile in a rectangle with a constant uniform heat source, constant-temperature boundary conditions at the edges, and no heat flow in the third dimension, it can be shown directly that the 2D steady-state failure power is given by

$$P'_{f,\,ss} = \frac{2\kappa\,(T_c - T_0)\,(c/b)}{1 - \dfrac{32}{\pi^3}\displaystyle\sum_{n=0}^{\infty}\left(\frac{-1^n}{(2n+1)^n \cosh\left[(2n+1)\,\pi c/2b\right]}\right)}. \tag{3.35}$$

Since P_f' does reach a steady state value, the assumption that no heat flow in the third dimension is equivalent to $a = \infty$ is incorrect, and Eq. (3.33) is therefore invalid. Notice in Eq. (3.35) that the failure power is constant for a constant b/c ratio. By numerically solving the equation for varying b/c, it was verified that $P'_{f,\,ss}(b/c)$ is equal to $P'_{f,\,ss}(c/b)$ (as it must to make sense physically), and it was determined that for $b \gg c$ the failure power is described by

$$P'_{f,\,ss} \cong 8\kappa\,(T_c - T_0)\,(b/c). \tag{3.36}$$

This approximation is illustrated in Fig. 3.31, in which $\Delta T_{ss} = (T_c - T_0)$ is plotted vs. b/c for a constant P'_{ss}/κ. The error in the approximation is less than 3% for $(b/c) \ge 3$.

To gain a better understanding of the 2D model of the power to failure as a function of time, transient simulations with varying values of b and c were run on a $b \times c$ μm^2

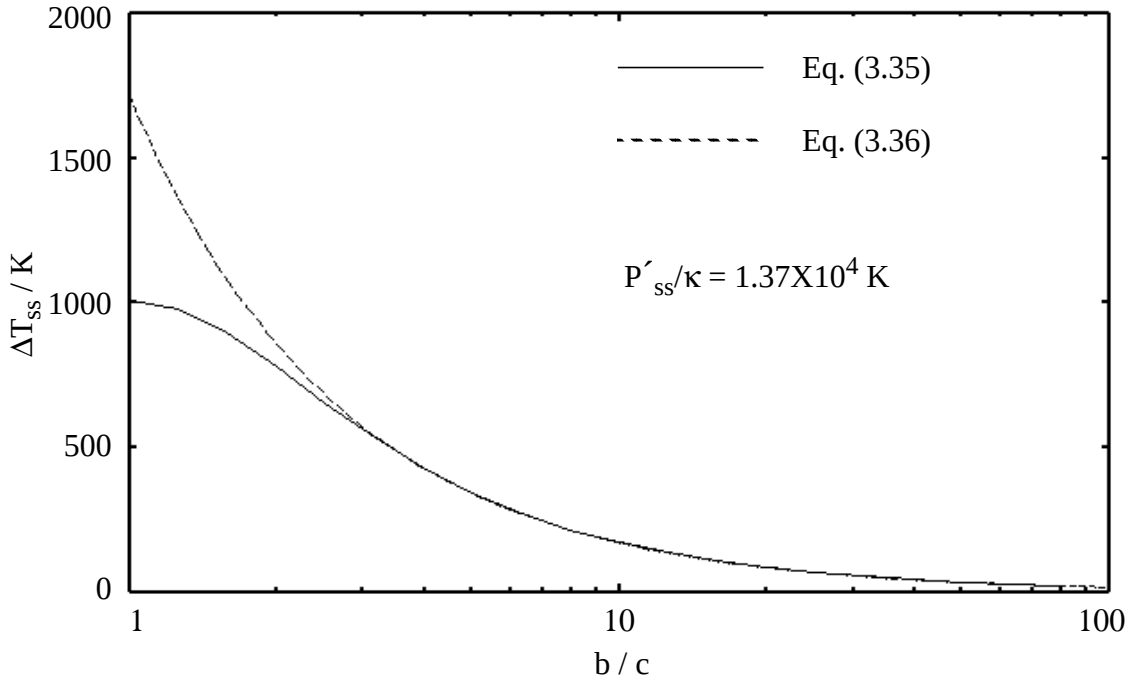

Fig. 3.31 The dependence of the steady-state change in peak temperature, ΔT_{ss}, on b/c *(log scale) as described by Eq. (3.35) is approximated by Eq. (3.36).*

rectangular semiconductor region with uniform doping, thermal boundary conditions of $T_0 = 300K$ applied on the perimeter of the structure, and electrical contacts placed along two opposing sides. In each simulation the applied voltage is ramped up to its steady-state value in 0.01ps to create a uniform, constant power source (V^2/R) in the structure, and the maximum temperature in the structure, T_{max}, is then monitored vs. time from 0.01ps to 1 second. Since the thermal box model assumes heat generation, thermal conductivity, and specific heat are independent of time and temperature, the temperature dependences of κ, C_p, and the band-gap energy are removed in the simulation models and a high doping level of $10^{18}cm^{-3}$ is used to reduce the effect of temperature on carrier concentration, i.e., to keep the resistance constant. Fig. 3.32a shows simulated curves of $1/\Delta T$ vs. time, where $\Delta T = T_{max} - T_0$, for a constant applied power and varying b/c ratios. Note from any of the P_f equations that plotting $1/\Delta T$ vs. time for constant power yields the same curve as plotting power vs. time for constant ΔT. The 2D curves are similar to the 3D P_f vs. t_f curve of Fig. 2.12 except that there are only two clearly defined regions. For times less than t_c (which in Fig. 3.32 is 140ps for $c = 0.25\mu m$ and 9.0ns for $c = 2.0\mu m$),

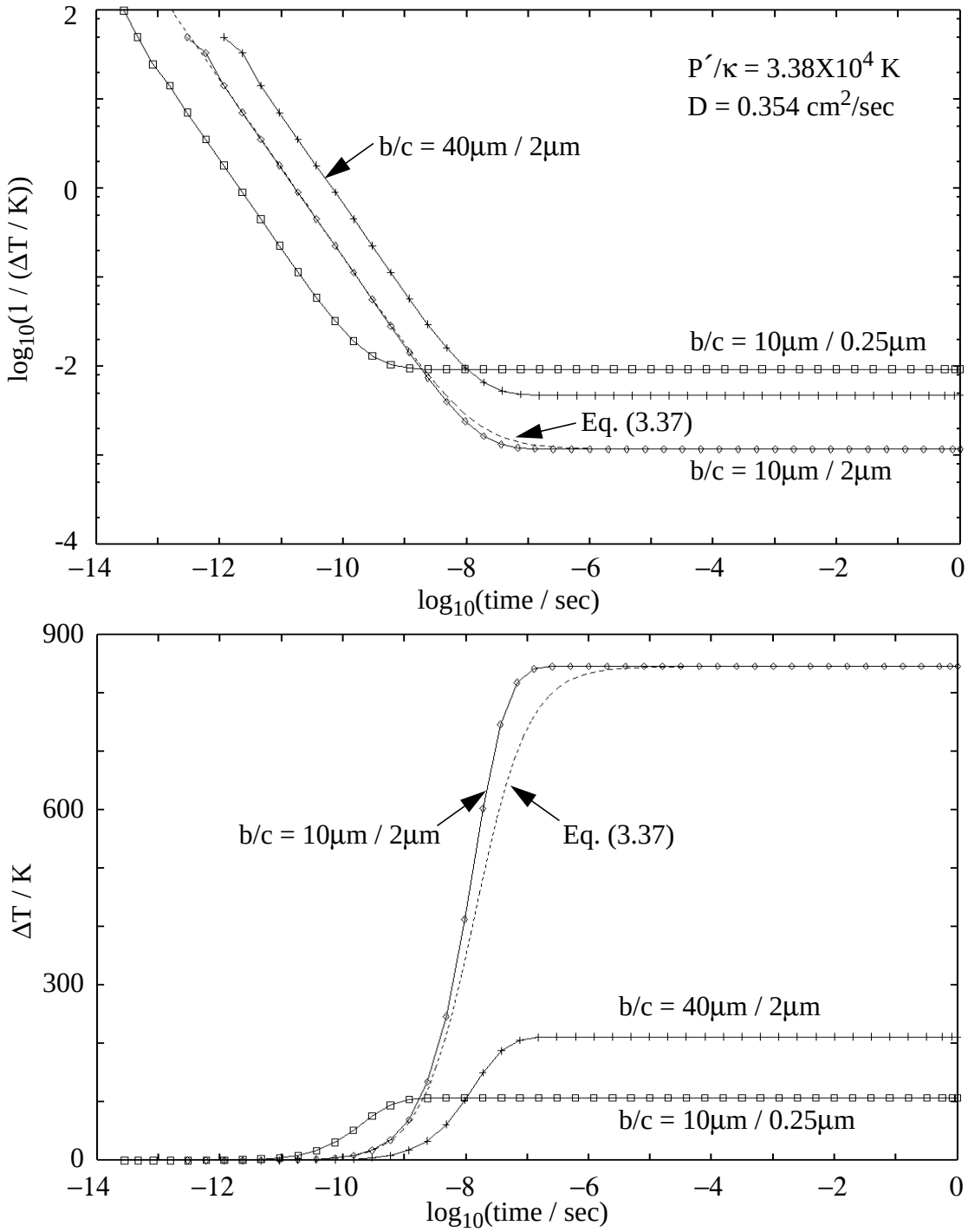

Fig. 3.32 Simulated 1/ΔT vs. time (a) and ΔT vs. time (b) curves for various length/ width ratios in a uniformly doped semiconductor region with a constant applied power. The temperature on the perimeter of the rectangular device is fixed at 300K. Eq. (3.37), the analytic approximation, is plotted for the 10µm X 2µm structure.

$1/\Delta T$ is dependent on time exactly as described by Eq. (3.34), i.e., the dependence is identical to the 3D case. This equivalence is expected because for $t < t_c$ there is no heat transfer outside the box in any direction. For large times, the simulated ΔT reaches a steady-state value in agreement with Eq. (3.36), a result which further establishes confidence in the thermal modeling capability of the 2D simulator. Fig. 3.32b, which plots ΔT vs. time, shows that ΔT is not quite proportional to c/b for the 10μm X 2μm structure, but this is due to a slight reduction in resistance at high temperature rather than to some error in the simulator.

Since four regions of the power-to-failure curve are defined by three time constants in the 3D thermal model, it is logical to expect a third region between t_c and $t_b = b^2/4\pi D$ in the 2D model. Fig. 3.32b does show a linear ($\Delta T \propto \log(t)$) region between $\Delta T = 0$ and steady state, but the this log(t) region is centered about t_c. The time constant t_b, which is 225ns for $b = 10\mu m$ and 3.6μs for $b = 40\mu m$, has no significance in any of the curves. The existence of only one time constant is supported by the finding that regardless of the value of P', κ, b, or c, all simulated $\log(1/\Delta T)$ vs. log(t) curves have the same shape and are merely offset by some $\log(1/\Delta T)$ and some log(t). If there were more than two regions of the 2D P_f vs. t_f characteristic, these curves would have different shapes on a log-log scale.

The overall 2D P_f vs. t_f curve can be approximated by the sum of the equations governing the $1/t_f$ and constant regions:

$$P_f'(t_f) \cong b(T_c - T_0)\left(\frac{\rho C_p c}{t_f} + \frac{8\kappa}{c}\right) = P'_{f,ss}\left(1 + \frac{\pi}{2}\frac{t_c}{t_f}\right). \tag{3.37}$$

Notice that the failure power is proportional to b, which is analogous to the 3D failure power being proportional to a in all time regions (Eq. (2.6) through Eq. (2.9)). In Fig. 3.32, Eq. (3.37) is plotted for the 10μm X 2μm structure. The equation underestimates ΔT (or overestimates P'_f) by up to 16% in the transition region and significantly underestimates ΔT in the steady-state region if b/c is not greater than three, but the equation is still useful for describing the modeled 2D thermal failure behavior. To compare the 2D P_f vs. t_f model to the 3D model, Eq. (3.37) and Eq. (2.3), which was integrated numerically, are plotted in Fig. 3.33 for ΔT = 1000K and $a = 50\mu m$, $b = 0.5\mu m$, and $c = 0.2\mu m$, typical dimensions for the high-field drain junction depletion region in a submicron MOSFET. While 2D simulation does yield the same failure power as the 3D model for times less than t_c,

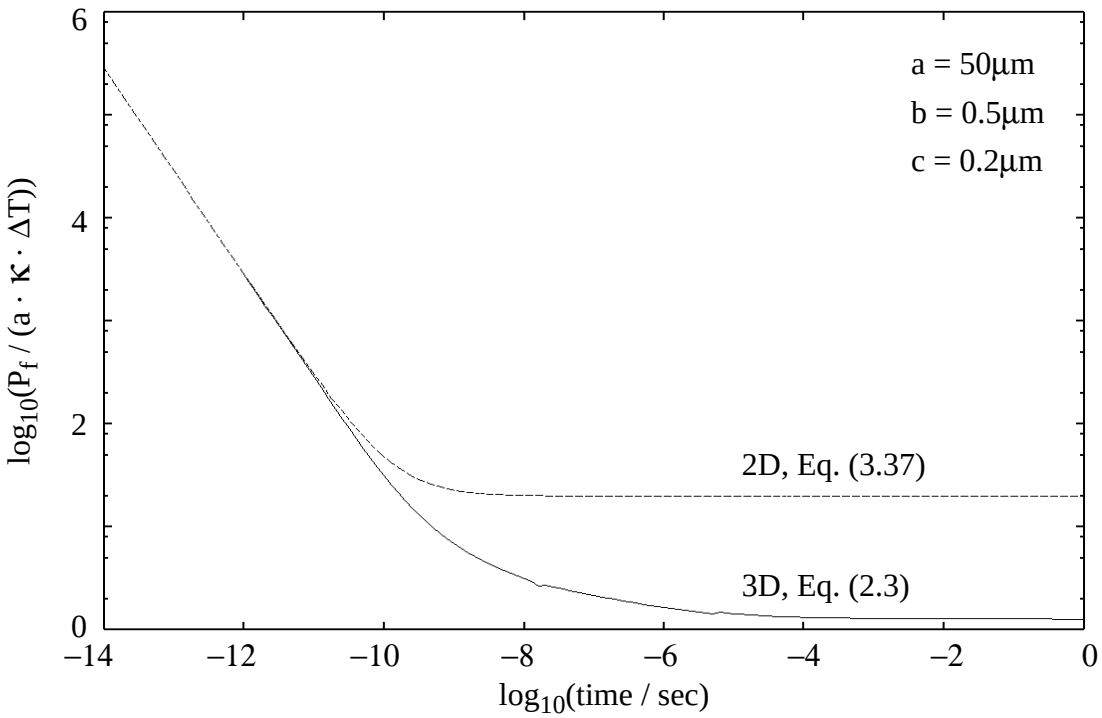

Fig. 3.33 *Power to failure, normalized by a, κ, and ΔT, is plotted vs. time to failure for the 2D and 3D implementations of the thermal box model. The time constants for the given box dimensions are t_a = 5.6μs, t_b = 560ps, and t_c = 90ps.*

this time region (less than 100ps for a leading-edge MOS technology) is of little interest because measurements are not possible and parasitics in any circuit render an ESD pulse of such a short duration impossible. In the region of interest for ESD, say 10ns to 1μs, the power to failure predicted by 2D simulation is too high by about an order of magnitude.

From Eq. (3.36) and Eq. (2.9), the ratio of the 2D to 3D predicted steady-state power to failure is

$$\frac{P'_{f,\,ss}(2D)}{P'_{f,\,ss}(3D)} = \left(\frac{4b}{\pi c}\right)(\ln(a/b) + 2) \quad \text{for } b \gg c, \tag{3.38}$$

which is always greater than unity since $a > b > c$. Eq. (3.38) states that regardless of the value of critical temperature chosen, the power needed to reach this temperature in steady state is greater in the 2D model than in the 3D model, i.e., the 2D model predicts a more

robust device. This directly contradicts the statement made by Mayaram et al. [13] and a similar assumption made by Diaz et al. [24] that 2D ESD simulation overestimates the peak temperature in a device and therefore underestimates its robustness. Eq. (3.38) also may explain why Diaz found that 2D simulations overestimated the power to failure in MOSFETs for times greater than 20μs, although at such long times the high-temperature region has extended well beyond the drain junction depletion region, which means the assumptions of the thermal model no longer precisely hold.

In the next chapter, we will see that in simulations of MOSFET protection devices the capability of 2D simulations to model power to failure for ESD stresses is not nearly as poor as suggested by Fig. 3.33. The ability to overcome the discrepancy between the 2D and 3D thermal models stems from the limitations of the assumptions made in the models when applied to real MOS structures. It was mentioned in Section 2.2.2 that the thermal box model is not completely accurate because the gate oxide at the top of the box acts like an insulator, not a conductor, so heat flow in this direction is greatly restricted and the peak temperature must be higher than predicted by the model. In an actual MOSFET, the reduction in failure power due to the insulating surface of the gate oxide is estimated to be significantly less than a factor of two [32]. By running a few 2D simulations with an insulating thermal boundary condition on one side of the $b \times c$ rectangle, it was determined that due to the insulating surface the peak temperature increases by a factor of two when the sides of the rectangle are equal. For unequal sides, this factor of two is roughly multiplied by the ratio of b/c, where b is the dimension of the side which is insulated. Since the side of the box along the gate is usually longer than the side equal to the drain junction depth, the increase in peak temperature due to the insulating gate may be proportionately greater in 2D MOSFET simulations than in actual structures, thereby reducing the 2D failure power to a level closer to the 3D case.

The other major assumption of the thermal box model which is violated in MOSFET simulations as well as in real devices is that for longer ESD pulse times (greater than a few hundred nanoseconds), the semiconductor region outside the box is no longer fixed at 300K and therefore cannot act as a perfect heat sink. As in the case for the gate oxide, the lack of an ideal heat sink implies that the peak temperature in the box will be greater than predicted by the model, which in turn implies that the power to failure will be lower than predicted. It is not obvious whether the actual boundary conditions surrounding the high-field region increase or decrease the disparity between real structures and 2D simulations.

It is clear, however, that by applying boundary conditions with large thermal resistances around the 2D simulation structure the peak temperature is increased for a given input power, which means the simulated power to failure is reduced. This method will be used in the next chapter to calibrate simulated P_f vs. t_f curves to experimental curves, but it is apparent that caution should be taken against using thermal resistances which are higher than physically justifiable, a definite risk considering the inherent overestimation of the power to failure in the 2D model.

This section has focused on the use of monitoring the peak lattice temperature in predicting thermal failure of ESD protection devices. Presumably, when the peak temperature reaches a critical value, second breakdown occurs and device damage follows instantaneously due to gross melting. If the object of simulation were to correlate simulations with this analytical thermal-model definition of failure, then it would only be necessary to monitor the peak temperature in the simulations. But although device failure, which is really defined by an increase in leakage current above a specified threshold level, correlates well with the occurrence of second breakdown for stress times greater than about 100ns, as mentioned in Chapter 2 for very short pulses device leakage can be increased above the failure level without the device exhibiting second breakdown because the damage site is too localized to reduce the resistance of the entire device. Since there is no unique series of events which leads to thermal failure, various phenomena should be monitored both experimentally and in simulations. During a transmission-line pulse test of a real structure, it is not possible to monitor the transient temperature profile, so thermally induced damage must be inferred by observing second breakdown on an oscilloscope during a pulse and/or confirmed by measuring an increased amount of leakage after the pulse. In contrast, simulations can be used to study not only the voltage drop due to second breakdown but also to study the 2D profiles of the lattice temperature, electric field, heat generation ($\mathbf{J} \cdot \mathbf{E}$), and intrinsic carrier concentration (n_i). Considering the difference between the 2D and 3D thermal models, it may even be beneficial to compare experimental and simulated current to failure rather than power to failure, as suggested by Diaz [24]. Thus, while much effort was devoted to analyzing the thermal model's ability to predict P_f vs. t_f behavior, the larger goal of electrothermal simulation is to be able to predict thermal failure in actual devices using any physical characteristics accessible in a calibrated simulation.

3.7 Simulation of Dielectric Failure and Latent ESD Damage

The previous two sections have addressed simulation of the MOSFET snapback I-V curve, second breakdown, and thermally induced failure. As discussed in Section 1.1, dielectric breakdown and latent damage are also important failure mechanisms in ESD protection circuits. Although the applicability of numerical device simulation to these types of failures is not as apparent as it is for thermal failure, the ability to monitor the electric field in the oxide region and the lattice-temperature profile in the silicon and to calculate hot-carrier injection current affords at least a qualitative examination of dielectric and latent damage. Dielectric breakdown is a threat both in the gate oxides of the input circuit being protected and in the thin-gate protection-circuit transistors which absorb an ESD pulse. Damage of the input gate oxide will most likely occur if the input (gate) voltage is not properly clamped by the protection device during an ESD stress (refer to Fig. 2.16), leading to time-dependent dielectric breakdown (TDDB) [64]. In the protection transistor, oxide damage is more likely due to hot-carrier injection resulting from the high ESD current than from pure high-voltage stress. Oxide damage due to high-voltage stress may occur, but since the protection-transistor oxide area is typically larger than the input-circuit oxide area, and since the input voltage is partially dropped across the n^+ drain diffusion of the protection transistor, the input-circuit oxide is much more likely to fail before the protection-circuit oxide. Nonetheless, it is simplest to study all dielectric failure mechanisms in the same device, so simulations will focus on the protection device while acknowledging that a high-voltage stress on the oxide implies an even higher stress on the input gate being protected. As discussed in Chapters 1 and 2, latent damage, low-level damage which does not cause immediate circuit failure but rather reduces the circuit's operational lifetime, has been attributed to oxide damage as well as to localized silicon melting in MOSFETs. Thus, some of the simulation techniques which apply to dielectric breakdown should also apply to latent failures.

Device simulators model the transport of charge carriers, but there is no way to model the movement or melting of the silicon lattice because the grid defining the structure is fixed and there is no mechanism for modeling the solid-liquid phase change. Instead, it must be assumed that when the modeled temperature exceeds 1688K over some area of a device, melting will occur (TMA-MEDICI allows the lattice temperature to reach 2000K, although the meaningfulness of a temperature greater than the silicon melting point is questionable). For dielectric failure, damage will be inferred from two phenomena:

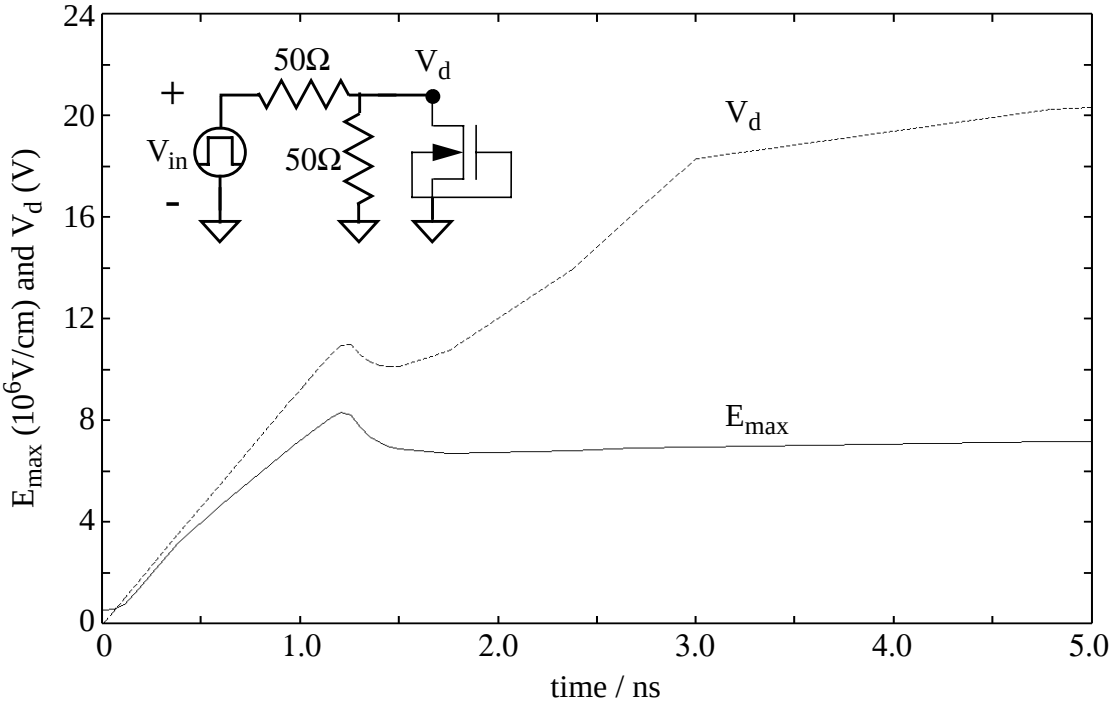

Fig. 3.34 The maximum electric field in the gate oxide (E_{max}, in MV/cm) of an ESD-protection MOSFET subjected to a square pulse with a 3ns rise time is plotted vs. time. As seen from the plot of the input voltage at the drain of the device, V_d, the reduction in E_{max} is due to the device snapping back at 1.2ns.

injection of charge into the oxide and high electric-field stress across the oxide (these are not necessarily mutually exclusive). The simplest analysis of dielectric stress during ESD involves recording the voltage across the gate oxide or the maximum electric field in the oxide of the protection transistor for each solution in a transient or steady-state snapback simulation. The device simulator does not report such voltage and electric-field information directly, but the desired information can be extracted from files containing the 2D potential and electric-field profiles saved from each solution. Fig. 3.34 shows a plot of the simulated maximum electric field vs. time in the 100Å-thick oxide of a protection MOSFET subject to a square-wave pulse with a 3ns rise time. The simulator was instructed to save the solution data for each time point, and the location and value of the maximum electric field in the device were then automatically extracted from each solution using a simple C program. Fig. 3.34 shows that the maximum electric field peaks at a

value of $8.3X10^6$ V/cm just before the drain voltage, V_d, snaps back at 1.2ns. Notice that the electric field appears to be proportional to the drain voltage before snapback, but after the MOSFET turns on the electric field drops because the potential at the drain under the gate drops. As the device begins to conduct current V_d rises but E_{max} remains relatively flat, indicating that there is a significant potential drop along the ballast resistance formed by a large drain contact-to-gate spacing. The peak electric field corresponds to a voltage of 8.3V across the 100Å gate oxide, which is probably not high enough to cause dielectric damage, especially since it is near this peak for less than a nanosecond. On the other hand, if the drain of the protection device were tied to an input-buffer transistor gate with the same oxide thickness, the 20V formed across the gate after the protection transistor turns on would almost assuredly rupture the input oxide.

Calculating the voltage across the protection-transistor oxide by multiplying E_{max} by the oxide thickness is an overestimate of the true value. Extraction of the maximum voltage is more complex than extraction of the maximum electric field because the potential varies along the boundary of the oxide region. Once an algorithm for extracting the maximum voltage is created, the oxide voltage in a transient simulation can be plotted vs. time and then compared to a measured voltage-to-failure vs. time-to-failure TDDB curve of an oxide with the same dimensions (Fig. 3.35). If the simulation accurately models the input-pulse profile and MOSFET dimensions, it should help predict whether the gate oxide will break down during a particular ESD stress.

The other type of dielectric stress considered here is a form of hot-carrier injection (HCI), a reliability issue normally associated with the effects of long-term MOSFET operation on the order of hours or days. Although ESD stress times are very small by comparison, the stress voltage and current far exceed the operational values and thus carrier injection is still a concern. A paper by Doyle et al. [54] reports that different types of oxide damage occur during avalanche breakdown, snapback, and high-current ESD stress. This latent damage may be in the form of interface states and/or oxide traps and is especially critical in output ESD protection transistors which must also function as the output driver of the IC. For ESD stressing, the authors applied a 350V HBM pulse (peak current = 233mA) to silicided NMOS transistors with a W/L ratio of 12.5/1.0μm. Oxide damage was monitored by comparing the measured transconductance (g_m) characteristic, i.e., g_m vs. V_{gate}, before and after each stress. They found that g_m decreases after ESD stress, but the threshold voltage, V_T, does not change. This indicates that there is an increase in the series

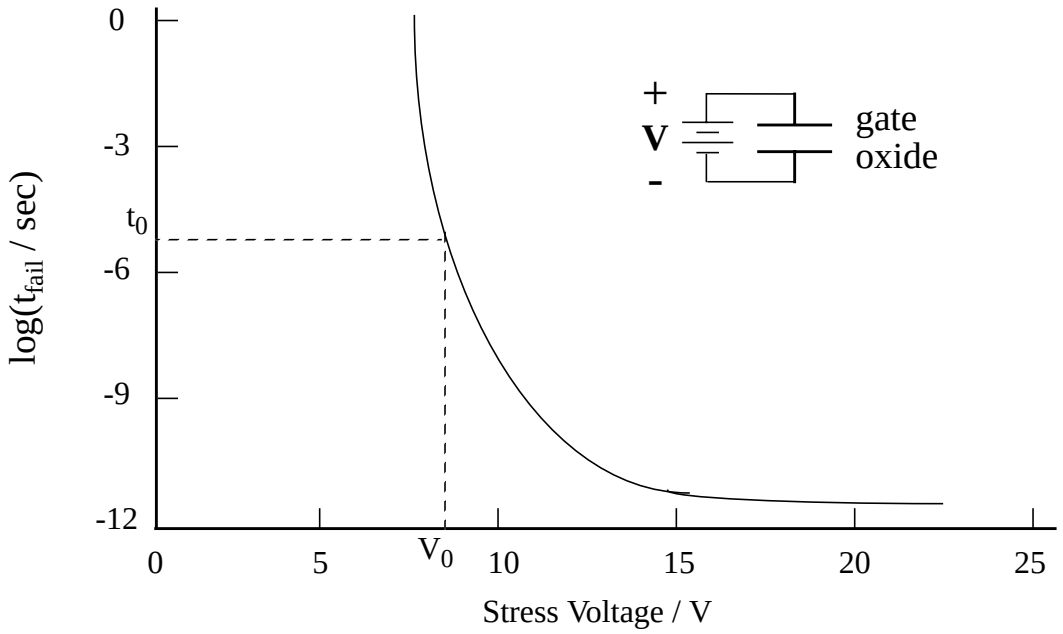

Fig. 3.35 A qualitative plot of time-to-failure vs. stress voltage reveals the time-dependent dielectric breakdown behavior of a gate oxide. If the voltage V_0 is applied across the oxide for a time greater than t_0, the oxide will rupture.

resistance of the device, suggesting that the damage is deep in the drain junction (but still in the oxide, they assert) and that there is no oxide trapping or increase of interface states directly under the gate, which would cause a shift in the threshold voltage. In contrast, HCI stressing, (V_{drain} = 5.9V, V_{gate} = 2.5V for 10,000 seconds) results in an increase in V_T as well as a decrease in g_m, showing that damage occurs directly under the gate. This makes sense because in HCI stressing, the high electric field is not as concentrated at the drain edge of the gate as it is during ESD stress. Other studies have verified that non-catastrophic snapback stress affects drain-current and substrate-current MOSFET characteristics [15] and reduces the dielectric strength of gate oxides as measured by charge-to-breakdown experiments (forcing a current into an oxide until the oxide short circuits) [25]. Two-dimensional simulations have been used to explain how interface states and trapped charges in the gate oxide formed by HCI affect MOSFET characteristics [55,56]. It was found that it if the region of damage is small compared to the gate length, damage due to interface states can be distinguished from damage due to

fixed charges because each has a unique effect on the transconductance and substrate-current characteristics.

Based on the findings of a relation between ESD stress and latent dielectric damage due to charge injection, it should be beneficial to study dielectric damage in ESD simulations. Since models for hot-carrier injection, fixed charge and charge traps at an oxide interface, and fixed charge within an oxide region are implemented in some 2D device simulators [29,30,44], it may be possible to simulate the dielectric damage incurred by a device during an ESD event, although a model of charge trapping within the oxide would also be required. Instead of modeling the change in the amount of trapped oxide charge during a transient ESD simulation, it would be easier and perhaps just as informative to simply look at the calculated hot-carrier gate current for each solution. In TMA-MEDICI, gate current analysis is available as a post-processing tool. That is, gate current is calculated based upon the electric field and current density profiles of a solution, but the resultant value is not fed back into the solver to create a self-consistent solution in which all current sources and sinks sum to zero. Usually this is not a problem because the gate current is several orders of magnitude lower than the source and drain current. The gate-current calculation is based on the lucky-electron model [53], which determines the number of carriers injected into the gate from a product of probabilities that are a function of the local electric field and scattering mean free paths. Since the use of gate-current simulation is only being investigated qualitatively in this section, a detailed discussion of the lucky-electron model is deferred to the TMA-MEDICI manual [29] and default model coefficients will be assumed.

In Fig. 3.36, the gate current is plotted vs. time for two simulated 50/0.75μm MOSFETs subjected to a square-wave pulse with a 3ns rise time, as depicted in the inset of Fig. 3.34. In one structure the gate is grounded, while in the other a 10KΩ bounce resistor, described in Section 2.3, has been placed between the gate electrode and ground to facilitate turn-on of the transistor (normal current through this resistor is not included in the gate-current plot). For both devices, the gate current increases as the electric field and avalanche breakdown build up in the drain-substrate junction and reaches a peak at the time the device enters snapback. In the case of the grounded-gate device, zero potential on the gate favors injection of holes into the oxide. Once this device turns on and the drain voltage drops, the electric field drops and less energy is available for the holes to surmount the oxide barrier, so the gate current falls off. In the case of the device with the gate-bounce

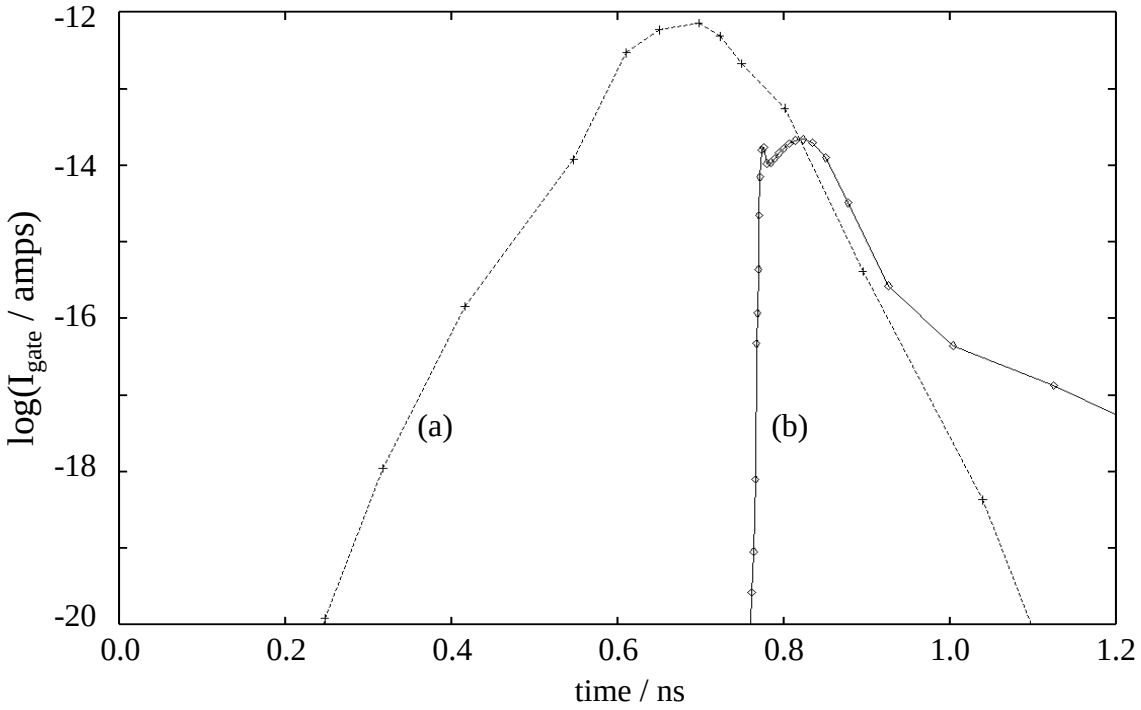

Fig. 3.36 *Gate current vs. time for 50/0.75μm MOSFETs with (a) 10KΩ gate resistor and (b) grounded gate. The drain is subjected to a square-wave pulse with a rise time of 3ns as depicted in the inset of Fig. 3.34. For both structures, the peak in gate current coincides with the time snapback occurs.*

resistor, the coupling of the gate electrode to the input creates a positive bias on the gate, so the injected carriers are electrons. When the device snaps back the gate potential, and thus the favorable electric field, drops and the electron injection quickly falls off. Both simulations show that carrier injection is most prevalent in the short time before a transistor snaps back.

The relationship between the simulated gate current and dielectric damage due to charge injection is not obvious. Presumably, the amount of gate current generated during a simulation correlates with a certain level of gate-oxide degradation, but work needs to be done in this area to determine such a correlation. In the case of ESD stress, experiments could be run in which devices are stressed with HBM or square-wave pulses of various levels and then tested to determine any change in the transconductance or threshold-voltage characteristics or to see if there is a reduction in the gate oxide's charge-to-

breakdown. Simulations of the same ESD stresses and devices could be run on calibrated 2D structures and the resulting levels of gate current could be compared to the measured change in characteristics to determine any correlation between simulated gate current and measured oxide degradation.

In addition to dielectric damage, latent failures may also be caused by local heating, as suggested by Kuper et al. [4]. Experimentally, latent thermal failures may be identified by the measurement of low-level (sub-microamp) leakage after a moderate ESD stress or during the evolution of a transmission-line pulsing experiment. Hypothetically, if a localized hot spot developed at the drain-substrate junction during the stress, the low-level leakage could be attributed to a resistive filament formed by the localized silicon melting. Such a filament would act as a high resistance in parallel with the junction diode and thus the device would become leaky. In a simulation, the latent "failure signature" would be a

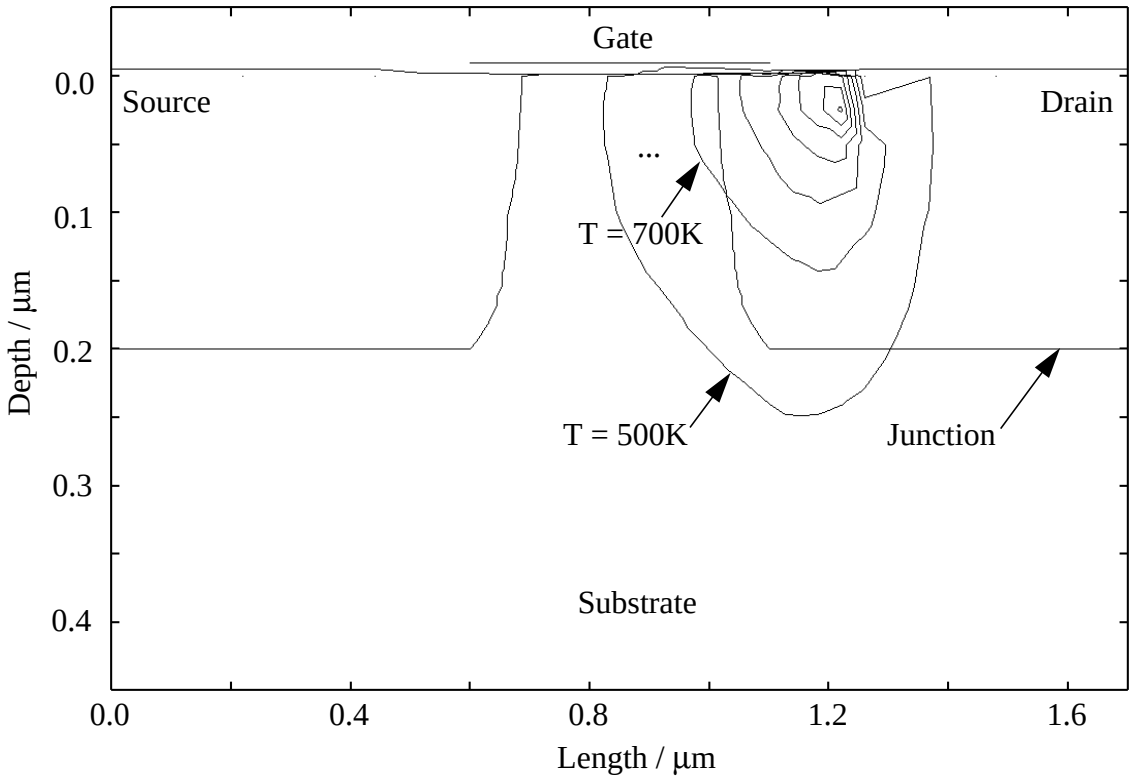

Fig. 3.37 *A constant-temperature contour is plotted for every 200K increment in temperature for a simulation structure at the time of peak ESD stress. Lines are also drawn marking the source and drain junctions of the structure, which is not plotted to scale.*

relatively small area of high temperature with no signs of second breakdown such as a drop in the device voltage or increase in device current. As an example, a transient simulation of a 50/0.5μm MOSFET stressed with a very high (120V), brief (3ns) ESD pulse was run and the solutions were saved for each time point. Using a C program, the temperature profile data was read from the solution file for the time coinciding with maximum device temperature and then used to calculate points along constant-temperature contours, shown in Fig. 3.37. Notice that the smallest contour contains the area in which the temperature is greater than 1700K, demonstrating that melting may occur in a small spot but should not be widespread. This spot is located near the surface at the drain-LDD n^+/n junction. Combining the temperature data with the 2D doping-profile data, a contour was also calculated within which the intrinsic carrier concentration, $n_i(T)$, is greater than the background doping level (Fig. 3.38). Recall that one of the assumed

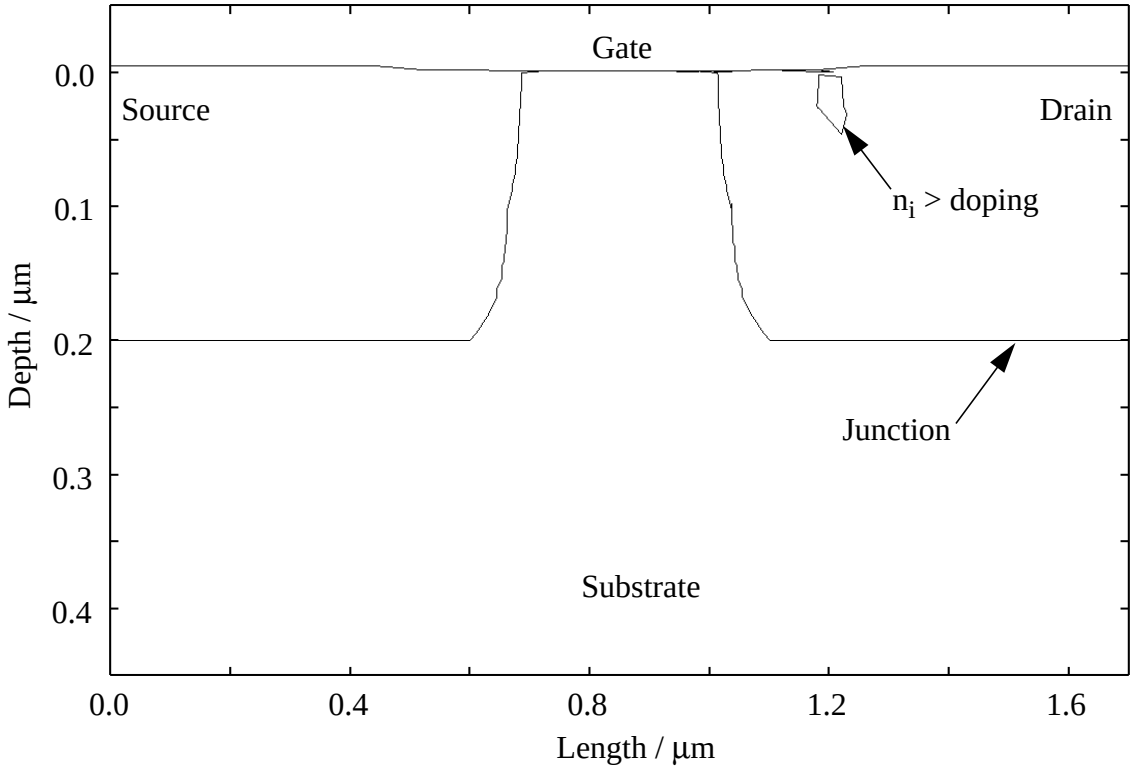

Fig. 3.38 A contour within which the intrinsic carrier concentration, n_i, is greater than the background doping level is drawn for a simulation structure at the time of peak ESD stress. Lines are also drawn marking the source and drain junctions of the structure, which is not plotted to scale.

conditions for second breakdown is the passing of the thermally generated carrier concentration beyond the background doping concentration. In the case of this simulation, Fig. 3.38 shows that this condition is met within a region of the device, but the small size of the region and the fact that the ESD pulse terminates before current can rush into the spot indicate that the device does not enter second breakdown and thus should only exhibit low leakage after the pulse. In practice, simulations of this type could be used to predict the relative susceptibility of different structure layouts to low-level leakage resulting from a particular ESD waveform or TLP pulse height.

Chapter 4

Simulation: Calibration and Results

To apply the concepts of ESD circuit characterization, simulation, and design discussed in Chapters 2 and 3, special MOSFET test structures were laid out in an Advanced Micro Devices 0.5μm, 3.3V CMOS technology and then tested with the transmission-line pulsing setup described in Section 2.2.4. These parametric structures are not designed to protect actual input/output (I/O) circuits but rather to determine the dependence of the ESD circuit parameters on device width, gate length, and contact-to-gate spacing. All structures are single fingered (as opposed to actual protection circuits, which are usually multiple fingered) and make use of a resist mask to block silicidation between the source/drain contacts and the gate. There is one exception: due to space limitations, the structures with varying gate length were not laid out on the special test tiles but rather were taken from a standard, fully salicided (self-aligned silicide) test tile. Software was written and used to garner the TLP data, extract I-V parameters from the data, and perform statistical analysis on the I-V parameters.

Numerical two-dimensional (2D) device simulation of the ESD structures was performed using TMA-MEDICI [29], which was chosen over Stanford's PISCES-2ET [44] because the lattice-temperature code in PISCES was not fully debugged at the time simulations began. The simulation models presented in Chapter 3 were initially calibrated against standard MOSFET characterization curves of two salicided test structures with different gate lengths and then were calibrated against TLP data from the special test structures to model the snapback and thermal effects. Calibration refers to the adjustment of simulation model coefficients which yields simulated device I-V and failure characteristics that match the experimentally determined characteristics of real devices. In the next section,

the calibration philosophy and strategy are discussed in detail. This is followed by sections reporting the experimental and simulation results: the parameters V_{t1}, V_{sb}, R_{sb}, and I_{t2} (refer to Fig. 2.6) are extracted from TLP measurements and compared quantitatively with simulations, as are P_f vs. t_f and I_f vs. t_f failure curves. At the conclusion of the chapter, a design example of an I/O protection circuit based on the parametric results is given in order to demonstrate the applicability of transmission-line pulsing and device simulation to ESD circuit design. Due to limitation of time and resources, only NMOS circuits are studied in this chapter. It is critical to study these devices because it has been observed that the n-channel transistors in a CMOS protection circuit usually absorb the energy of an ESD pulse due to their lower turn-on time [18,21]. A complete circuit design certainly needs to include study of the PMOS transistors, but for purposes of proof of concept it is sufficient to concentrate on NMOS devices in this chapter.

4.1 Calibration Procedure

Calibration of 2D device simulations to the AMD 0.5μm CMOS technology is broken up into three main steps. First, before I-V simulations can begin a 2D structure must be created to model the layout and process characteristics of the technology, including gate length, oxide spacer width, source/drain (S/D) contact-to-gate spacing, gate oxide thickness, and two-dimensional doping profiles. Next, this structure is used for simulations of standard drain, gate, subthreshold, substrate, and breakdown MOSFET characteristics to calibrate the mobility and impact-ionization (II) models. The model coefficients are adjusted until the simulated I-V curves match the experimental curves reasonably well. Finally, TLP-like simulations are calibrated to experimental TLP data. Further adjustment of the II coefficients is performed to match the trigger and snapback voltages while the thermal boundary conditions are set to yield simulated failure levels which parallel those of the actual devices. For proprietary reasons, most of the final model coefficient values and I-V curves will not be explicitly reported for the calibration procedure delineated below.

4.1.1 Structure Definition

The 2D simulation structure was created based on SUPREM-IV [59] process simulations as well as secondary ion mass spectroscopy (SIMS), spreading resistance profile (SRP),

and transmission electron micrograph (TEM) data with the goal of matching the actual structure dimensions and doping profiles. SUPREM-IV simulations were performed by AMD engineers and are based on the technology process flow. Although the 2D gridded structures generated by SUPREM-IV simulations are suitable for use in the device simulations, discrepancies were found between the simulated S/D junction depths and those extracted from SIMS and SRP data, which suggested that the junction depths are about 50nm greater than those of the SUPREM-IV simulations. Also, the simulated spacer width, which is explicitly defined in SUPREM-IV, is about 10nm wider than the spacer oxide seen in TEM photographs. Since there is no way to easily measure the S/D junction abruptness and LDD profile, the junction profiles calculated by SUPREM-IV were assumed to be correct.

Calibrating the junction profiles can be accomplished by adjusting the parameters of the ion-implant and diffusion models in SUPREM-IV and iterating process simulation runs until more accurate results are attained, but this approach has two drawbacks. First, even a partial SUPREM-IV run, starting at the S/D implant, can take on the order of hours of simulation time. Second, the number of grid points needed for accurate process simulation is greater than the number needed for device simulation. Using an unnecessarily large number of grid points for device simulations is a large waste of time, and there is no way to eliminate grid points by "refining" the SUPREM-IV-generated structure file. Therefore, the approach taken in this calibration is to completely define the structure within the MEDICI device simulator. Doping profiles are defined analytically by specifying the peak and characteristic lengths of 2D Gaussian profiles. By using overlapping profiles, the 2D profile of the S/D, LDD, and channel regions can be fit reasonably well, as least within the uncertainty of the SUPREM-IV, SIMS, and SRP data. The gate oxide thickness is explicitly defined, while the spacer width is implicitly defined by the placement of the source-drain/LDD n^+/n junction. When the structure is created, the number of grid points used is controlled by specifying how fine the grid should be in critical areas such as where the doping or electric-potential gradient is steep. Using a template input file to define the layout and profile parameters, MEDICI can create a MOSFET structure in less than five minutes, more than an order of magnitude faster than SUPREM-IV. A MEDICI-generated structure with three doping-profile grid refinements, or regrids, and three electric-potential regrids contains 589 grid points for a 0.5μm-gate-length structure with minimum contact-to-gate spacing, while a 3.0μm structure has 1363 points. In contrast, the 0.4μm structure

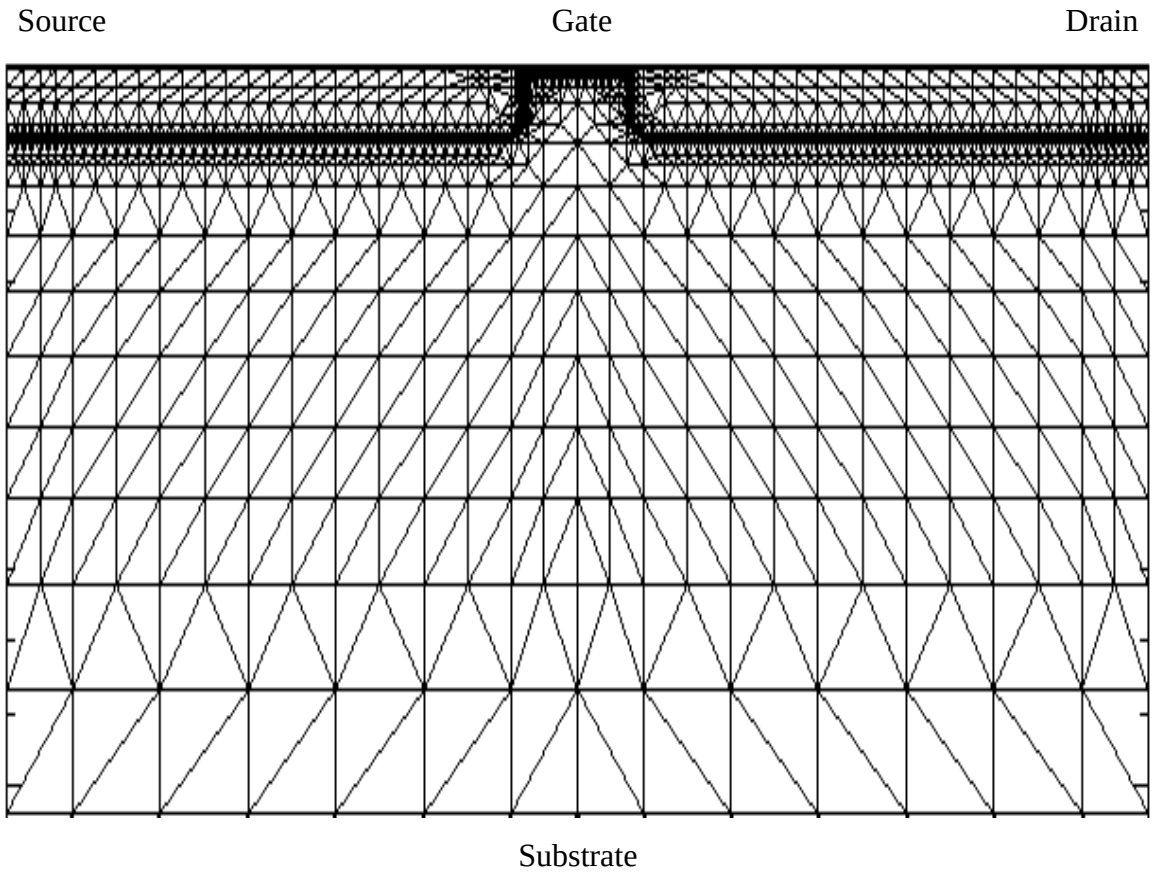

Fig. 4.39 *This example of a MEDICI-generated grid shows the concentration of grid points in the channel, LDD, and junction regions. Several microns of the substrate portion of the simulated structure have been omitted in order to display the grid approximately to scale.*

created by SUPREM-IV has 3827 grid points. Fig. 4.39 shows an example of a MEDICI-generated grid.

For all mobility and impact-ionization model calibration, simulations were run for a 0.5μm-gate structure and a 3.0μm-gate structure to ensure that the models are valid for more than one structure size. Each structure is bounded laterally by the S/D contact edges, with the S/D contact-to gate spacing set equal to that of the test structures. The depth of the device is made large enough that the depletion region does not extend to the bottom edge during any of the simulations. A substrate contact covers this entire bottom edge, neglecting the small substrate resistance between the intrinsic device and the substrate

contact in an actual MOS transistor. The S/D contacts are placed on the top of the structure from the actual contact position all the way up to the spacer edge in order to model the silicide used in the test structures. As mentioned at the beginning of the chapter, the only available test structures with gate-length variations were fully salicided devices. The silicide layer is formed by depositing tungsten or titanium over all active (S/D) areas which reacts with the silicon to form a layer between the S/D contacts and the spacer edge. This layer is a few nanometers deep and has a resistance of a few Ω/❑. Since the silicide's resistivity is low and it is not used in the structures tested with transmission-line pulsing, the layer is simply approximated by an extension of the metal contacts.

4.1.2 Calibration of MOSFET Characteristics

After the correct structure is created, the next phase of calibration is fitting simulated curves to the standard experimental MOSFET characterization curves described in many textbooks [42,61] and depicted in Fig. 4.40. For the AMD structures used in this calibration, data was taken at wafer level using a probe station and HP4145 parametric analyzer. In the simulations, each type of curve is only dependent on certain model parameters. For example, the drain characteristic (Fig. 4.40a) is mainly dependent on parallel-field mobility parameters, while the gate characteristic (Fig. 4.40b) is a function of perpendicular-field mobility parameters. Also, the breakdown voltage (Fig. 4.40e) is determined by the II coefficients of electrons and holes, whereas the substrate current (Fig. 4.40d) really only depends on the electron coefficients since electron current is dominant in this type of stress. Although each curve can be fit individually by calibrating only a few model coefficients, it is important to optimize the mobility and II coefficients over all curves because an ESD event incorporates several physical effects, including junction breakdown, MOSFET action, and bipolar transistor action. Therefore, the philosophy behind the calibration procedure is that separate types of I-V curves can be used to isolate specific model coefficients, but the results of the individual curve fits must yield a set of coefficients which correctly model all device phenomena. Additionally, the calibration should be accomplished while leaving as many model coefficients as possible at their default values, most of which are determined by data published in the literature. Not only does altering a minimum of coefficients save time and effort, it is sensible because although material properties tend to vary across technologies within and between manufacturers, variations in basic physical properties should not be great.

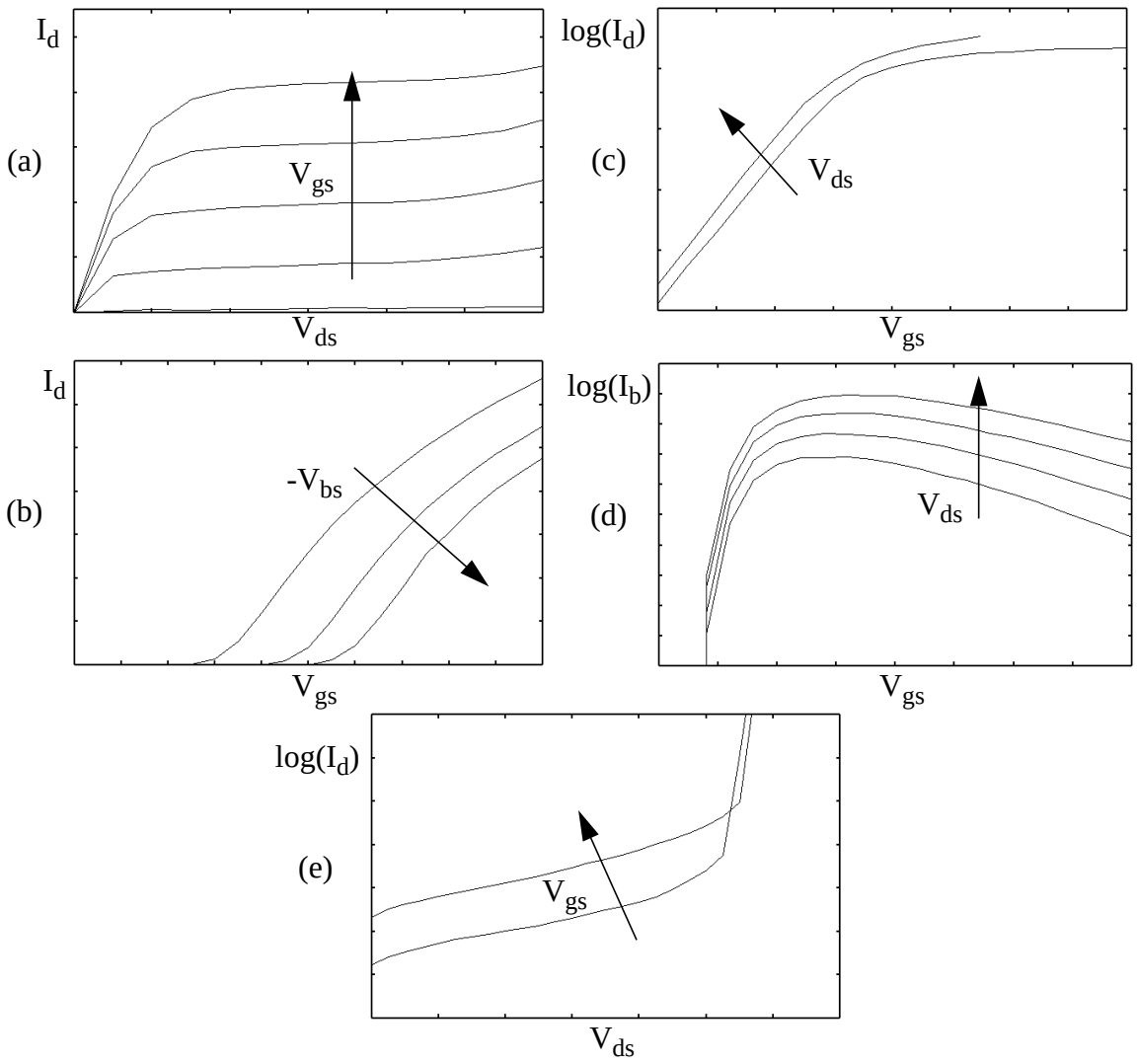

Fig. 4.40 Qualitative depiction of I-V curves used for MOSFET calibration: (a) drain: I_d vs. V_{ds} for stepped V_{gs}; (b) gate: I_d vs. V_{gs} for stepped V_{bs}; (c) sub-threshold: log(I_d) vs. V_{gs} for stepped V_{ds}; (d) substrate: log(I_b) vs. V_{gs} for stepped V_{ds}; (e) breakdown: log(I_d) vs. V_{ds} for stepped V_{gs}. The subscripts d, s, g, and b refer to the drain, source, gate, and substrate, respectively.

Ideally, model coefficients should also be calibrated over a high-temperature range because there is heating during an ESD event which affects the semiconductor properties (refer to Eqs. (3.22), (3.23), (3.25), and (3.29)). For example, at increased temperatures the mobility and II-generation rate degrade due to increased scattering, thereby reducing the saturation drain current of Fig. 4.40a and increasing the breakdown voltage of Fig.

4.40e. Wafer-level I-V data can be taken at temperatures up to around 500K using a hot chuck and then used as a basis for calibrating model coefficients in high-temperature simulations. For calibration of this AMD technology, however, it is assumed that the temperature dependences in the mobility and II models, which are qualitatively correct and have been fit to high-temperature data of other technologies [47,49], are accurate enough using default coefficient values. The benefit of high-temperature calibration is actually limited because for sub-microsecond ESD events the high-temperature region is localized--perhaps covering as little as 10 percent of the simulation space--and thus the temperature dependence of the mobility and II models may not have much effect on the overall I-V curve. Also, these models have only been shown to be valid up to a certain temperature, e.g., 460K for mobility [47], close to the limit of hot-chuck measurements, but critical ESD effects occur at higher temperatures. And even if the mobility and II models are calibrated at high temperatures, other simulation models are suspect. For example, at 900K the band-gap shrinkage model predicts a band gap energy about 40mV higher than the measured value [60]. Instead of calibrating mobility and II coefficients at high temperatures to fit ESD thermal-failure simulations, the approach taken here is to adjust the thermal boundary conditions, i.e., the placement of the thermal contacts and use of lumped thermal resistors and capacitors, to match simulated and experimental data. Since the true thermal boundary conditions are not known exactly, adjusting the thermal contacts and lumped elements to fit simulated thermal failure to ESD data is a reasonable way to determine their values. Discussion of the calibration of thermal effects is not taken up until Section 4.1.4. For all of the MOSFET simulations described in this subsection, the initial lattice temperature is set to 297K and is allowed to increase in regions of heat generation (Eq. (3.15)) as determined by the thermal diffusion equation (Eq. (2.2)). Constant-temperature boundary conditions are placed on the bottom and sides of the simulation structures as a simple way of modeling the large heat sink of the bulk silicon, but these are not really important because the maximum temperature during any of the MOSFET simulations is less than 310K.

Calibration of the Lombardi mobility model began with simulations of the gate characteristic shown in Fig. 4.40b. To reduce simulation time, a one-carrier (electron) solution method was used because hole current is negligible in an NMOS transistor in its normal operating range. This implies that only the electron mobility coefficients are adjusted during calibration. Initial simulations of the 0.5μm and 3.0μm structures using default values

for all model coefficients revealed that the spacing between I_d-V_{gs} curves for different V_{bs} (the subscripts d, s, g, and b stand for drain, source, gate, and substrate, respectively), i.e., the body effect, did not match the experimental data. Since the body-effect parameter [61],

$$\gamma = \frac{\sqrt{2\varepsilon_s q N_a}}{C_{ox}}, \tag{4.39}$$

where ε_s is the permittivity of silicon, q is the electron charge, N_a is the effective channel doping, and C_{ox} is the gate oxide capacitance, is not dependent on mobility but is dependent on the channel doping profile, the doping profile was modified in the 0.5μm and 3.0μm structures until the spacing between simulated I_d-V_{gs} curves matched experiments. This is justified because the change was relatively minor (the peak of the threshold-adjust implant was reduced by a factor of two) and the initial channel profile was not extracted experimentally but rather assumed from the SUPREM-IV simulation and thus was subject to modification. In addition to the channel-doping modification, a fixed-charge density was introduced at the gate oxide-silicon interface to align the simulated and experimental grounded-substrate ($V_{bs} = 0$) curves, i.e., to align the threshold voltage, V_T. The charge-density value used is reasonable in comparison to extracted values from real devices.

In the I_d-V_{gs} simulations V_{ds} is only 0.1V while V_{gs} is swept up to 3.3V (V_{CC}), so the electric field perpendicular to carrier flow, $E_\perp$, is much larger than the parallel field, $E_\parallel$, and only the perpendicular-field mobility parameters in Eq. (3.21) and Eq. (3.22) need to be adjusted to fit the I_d-V_{gs} curves; the bulk term, μ_b (Eq. (3.23)), is left constant. Performing a simple sensitivity analysis by running separate simulations with BN, CN, and DN set to twice the respective default value, and noting the resulting change in the I_d-V_{gs} characteristic, it was found that BN has no discernible effect on the curves while CN and DN each has a significant effect. Therefore, CN and DN were chosen as the coefficients to vary and BN was left at its default value. Also, even though the curves are sensitive to the doping exponent EN in Eq. (3.22), EN was left at its default value because the structures' doping profiles remained fixed after the channel profile adjustment. CN and DN were varied in a full-factorial manner over a simulation design space covering approximately one order of magnitude above and below their default values, and from these simulations a set of values was found which yields an excellent fit for both the 0.5μm and 3.0μm curves. The chosen values are both within a factor of three of their respective default values.

After determining the perpendicular-field mobility coefficients by calibrating the gate characteristic, calibration of the drain characteristic was used to set the remaining mobility coefficients in the bulk mobility term and high-field Caughey-Thomas expression (Eq. (3.24)). Here the advantage of doing the gate calibration before the drain calibration becomes obvious: in the I_d-V_{ds} curves the drain voltage is swept to V_{CC} and the gate voltage is stepped to V_{CC}, so $E_{||}$ and $E_{\perp}$ are both high, but since the $E_{\perp}$ coefficients have already been determined by the I_d-V_{gs} fit, the optimization space is reduced to variation of the $E_{||}$ coefficients. (Actually, a few iterations may need to be performed between gate and drain calibrations because the bulk mobility and saturation velocity do affect the I_d-V_{gs} curves.) As was the case for the gate-characteristic calibration, hole current is not solved for in the drain simulations because its contribution is negligible. In initial I_d-V_{ds} simulations the saturation current, I_{dsat}, as well as the separation between curves at different V_{gs} values (i.e., the transconductance, g_m), were too high for the 0.5μm and 3.0μm structures. To reduce I_{dsat}, the saturation velocity can be effectively lowered by reducing β_n in the Caughey-Thomas expression. The default value for β_n in MEDICI is 2.0, but in this case the default value is too high because it is taken from an old publication [48]. In a more recent publication, Jacoboni et al. report a β_n of 1.11 based on a best fit of several reported curves of drift velocity vs. electric field [62], so the need to reduce β_n was actually expected.

Instead of taking a full-factorial approach to the I_d-V_{ds} calibration, β_n was first individually optimized in an attempt to create a “quick fix” for I_{dsat}. Using one value for β_n, a good fit could be made for the 0.5μm-gate I_{dsat} and g_m, but this resulted in too low an I_{dsat} for the 3.0μm-gate structure. Likewise, a larger value of β_n resulted in a good fit at 3.0μm, but I_{dsat} and g_m are then too high for 0.5μm. Adjusting the bulk mobility does change I_{dsat} and g_m, but it affects the current of both structures proportionately, so μ_b could not be used to remedy the problem. The solution was to adjust β_n to calibrate the 3.0μm-gate structure (the final value of β_n is nearly equal to the value of 1.11 reported by Jacoboni) and then introduce a series source/drain resistance in the structures which effectively reduces I_{dsat} and g_m by dropping part of the drain voltage external to the device. This resistance, added by defining lumped resistors at the source and drain electrodes in the simulations, has a much larger effect on the 0.5μm structure than the 3.0μm structure because the current level is much higher for the shorter gate. Using this method, good fits for both drain curves were attained using a resistance of 12.5Ω on the source and on the drain. The lumped

resistance ostensibly models the contact resistance present in the experiments due to contact vias and/or probe tips. However, 12.5Ω is unreasonably high because the series resistance due to contact vias is typically on the order of 3Ω or less in this AMD technology, and the probe tips used have an area much larger than the effective via area and thus have negligible resistance. Therefore, using such large lumped resistors to complete the drain calibration is not justified. The discrepancy between 0.5μm and 3.0μm structures could probably be resolved by more legitimate means, e.g., further adjustment of all mobility coefficients or of the junction profiles, but such efforts were deferred in the interest of proceeding with the overall calibration, and the source/drain resistance was left at 12.5Ω.

After completion of the gate and drain calibration, simulations of the subthreshold characteristics (Fig. 4.40c) matched the experimental curves very well. The two simulated threshold voltages, defined as the V_{gs} for a certain threshold value of I_{ds} at two values of V_{ds}, were within 5% of the measured values for the 0.5μm structure and within 1% for the 3.0μm structure, a result which is not surprising since V_T was already fit during the I_d-V_{gs} calibration. Furthermore, the subthreshold slopes were also accurate for both gate lengths, with less than 3% difference in mV of V_{gs} per decade of I_{ds}. Since the subthreshold slope is dependent upon the oxide and depletion-layer capacitances [42], the good log(I_{ds})-V_{gs} fit indicates proper modeling of the substrate doping since this determines the depletion-layer capacitance. Due to the good fit of the subthreshold simulations, no adjustments in the models needed to be made, and therefore these curves were not really part of the calibration process.

A good match between experimental and simulated gate and drain characteristics, obtained without changing any of the model coefficients by more than a factor of three (except the source/drain resistance), indicates that the mobility and channel and substrate doping are modeled reasonably well. Accurate modeling of the drain current and 2D doping profile is a prerequisite to simulating impact-ionization-related I-V curves because the II generation rate at any point in the structure is proportional to the local current density (Eq. (3.26)) and to the ionization coefficients, α_n for electron current and α_p for hole current, which in turn are dependent upon the local electric field (Eq. (3.27)). In contrast to the previous simulations, for any simulation involving impact ionization it is necessary to perform a two-carrier analysis because both electrons and holes are involved in the ionization process. In substrate-current testing (Fig. 4.40d) I_{bs} is measured for normal MOSFET operating levels, with the gate voltage being swept from zero to slightly

past V_{CC} and the drain voltage stepped at values around V_{CC}, so prior calibration of the drain current implies that the substrate characteristic should be fit only by adjusting the α_n^∞ and λ_n coefficients (Eq. (3.28)). Similarly, the breakdown voltage, BV_{DSS}, in Fig. 4.40e is dependent upon the drain-substrate junction profile, but calibration of BV_{DSS} should concentrate on adjusting the ionization coefficients because the results of the drain and gate calibrations suggest that the junction model is already accurate. Adjusting the impact-ionization coefficients should not affect the drain, gate, and subthreshold characteristics because relatively high electric fields are not involved. However, introducing the II model to the drain-characteristic simulations does increase the drain current in the 0.5μm-gate structure up to 10% for V_{ds} = 6V (well above V_{CC}) because the electric field is fairly high and the drain sinks most of the electrons generated by impact ionization.

In MEDICI the default II coefficients are based on measurements of impact ionization in bulk silicon [63], but as discussed in Section 3.1 impact-ionization rates in MOSFETs are lower than in bulk silicon because II generation occurs near the surface, where the mean free path is lower, i.e., where the critical electric field of Eq. (3.27) is higher. Therefore, the final fitting values of the electron and hole mean free paths, λ_n and λ_p, are expected to be lower than the MEDICI defaults. In keeping with the philosophy of manipulating as few model coefficients as possible, only λ_n and λ_p were adjusted to calibrate the substrate and breakdown curves while the pre-exponential coefficients, α_n^∞ and α_p^∞, were held constant. This approach works for calibration of the standard MOSFET characteristics, but it has a significant consequence on the snapback simulations that will be discussed in the next subsection.

Calibration of the substrate curves was performed before that of the breakdown curves because the substrate current depends only on the electron II coefficients while BV_{DSS} depends upon the hole coefficients as well as the electron coefficients. In Fig. 4.40d, I_b consists of holes diffusing from the high-field region under the drain side of the gate where they are generated by impact ionization (recall that V_{ds} is around 3.3V during the stress, so the electric field is relatively high in this area). Since the device current consists almost entirely of electrons, only the electron II coefficients affect the level of substrate current. An explanation of the shape of the I_b-V_{gs} characteristic is given in [42]. Basically, the initial increase of I_b with V_{gs} is due to the deepening inversion layer which increases the drain current and proportionately increases I_b. At a critical value of V_{gs}, however, the

effect of increasing drain current is offset by the lowering of the electric field, which is proportional to V_{ds} - V_{gs}. In the initial substrate simulations, I_b was about one order of magnitude too high for the structures of both gate lengths, so simulations were then run with lower values of λ_n until an optimal value was found. For the best-fit case, with λ_n set at a little more than half its default value, the peak log(I_b) for each V_{ds} step is within 2% of the measured value for the 0.5μm-gate structure and within 3% for the 3.0μm-gate structure, and the peak in I_b always occurs at the correct value of V_{gs}. However, for V_{gs} greater than 2.5V the simulated substrate current of both structures rolls off more severely than the measured current, indicating that either the current and electric field profiles in the drain junction region are not correct or that the II model loses accuracy for lower electric fields. It may be possible to correct the latter case by further altering the II coefficients, but it is also possible that there is a limitation in the model. Despite the sharp roll-off, the good fit in the peak I_b region was encouraging enough to allow the calibration to proceed to the breakdown characteristic.

The breakdown of Fig. 4.40e results from avalanche multiplication of carriers caused by reverse biasing the drain-substrate junction. Since the hole current sunk by the substrate is equal to the electron current sourced by the drain, both types of carriers create avalanche pairs and thus λ_n and λ_p both determine the breakdown voltage. Since λ_n was already determined by the I_b-V_{gs} calibration, only λ_p was adjusted to calibrate BV_{DSS}. This is analogous to the gate and drain-characteristic calibrations in which the gate curves were used to fit the $E_\perp$ mobility coefficients and then the drain calibration was used to fit the remaining mobility coefficients. Surprisingly, the default, bulk value of λ_p resulted in a simulated BV_{DSS} less than the measured BV_{DSS}, meaning it had to be increased to fit the curves (structures for both gate lengths have the same breakdown voltage because this voltage does not depend on gate length). This suggests that λ_p had to be adjusted to compensate for a λ_n which is too low or that a majority of the simulated II generation occurs along the drain-substrate junction, where the mean free path is closer to its bulk value, rather than under the gate at the surface. To calibrate the breakdown curve, λ_p only had to be increased about 5% above its default value.

After calibration of the breakdown curves was completed, simulations for all characteristics at both gate lengths were rerun with all of the calibrated coefficients in place. Not surprisingly, adding the impact ionization model to the drain simulations did increase I_{ds} for large V_{ds} in the 0.5μm structure, but it had no effect on the extracted saturation current,

which is measured at $V_{ds} = V_{CC}$. The II model had no effect on the gate characteristic because no high electric fields are present during this type of stress. Finally, as expected changing the hole mean free path did not affect the substrate-current simulations. With all of the MOSFET curves accurately simulated, calibration could move to the next phase.

4.1.3 Calibration of the Snapback I-V Curve

In the final stage of calibration, simulations and experiments focus on ESD phenomena, specifically on transmission-line pulsing. An important assumption of the calibration philosophy is that if the mobility and impact-ionization simulation models accurately describe different simple MOSFET I-V curves, they yield accurate simulations for complex curves such as an ESD-induced snapback curve. For thermal characteristics, however, thermal boundary conditions must be adjusted to calibrate thermal failure of the MOSFET structures. Experimental data was taken using the setup described in Section 2.2.4, with the structures bonded up in dual in-line packages. In each test, the drain of the structure was hit with square pulses with the gate, source, and substrate grounded. A pulse width of 200ns was chosen for the majority of the testing because it is short enough to ensure that stressing is in the ESD regime while still long enough to allow easy extraction of the device current and voltage on the oscilloscope. Fig. 4.41 shows a TLP-generated I-V curve and illustrates the extraction of the parameters V_{t1}, V_{sb}, R_{sb}, V_{t2}, and I_{t2} (defined in Section 2.2.1). The line defining V_{sb} and R_{sb} is the least-squares fit of all I-V points between snapback and second breakdown. Device failure, defined as 1μA of leakage current with the drain biased at V_{CC} with respect to the gate, source, and substrate, usually coincides with the second-breakdown point (V_{t2}, I_{t2}). However, as discussed in Section 2.2.3, second breakdown does not always immediately lead to device failure, and in such cases failure is defined as the point at which microamp leakage is created. Experiments were run on NMOS structures with varying gate length, gate width, and contact-to-gate spacing (CGS), defined as the distance from the edge of the salicided source and drain contacts to the respective edge of the gate. As mentioned at the beginning of the chapter, fully salicided structures had to be used to study gate-length variations, but structures employing a mask to block salicidation between the spacer and S/D contact edges were used for the rest of the experiments. Five to seven tests were run per structure, and the I-V parameter values were extracted for each test. The values used for calibration are the average values of each structure.

A few changes in the simulation structures were made before the final phase of calibration began to more accurately model the non-salicided test structures used for snapback and thermal characterization. Since the lumped source/drain resistance introduced during the calibration of the drain characteristics was unreasonably large, it was removed from the simulation model. This simplifies the simulation-structure specification and is justified because the new, salicide-blocked test structures are at least 2.5 times wider than the previous structures, which implies much more contact area and thus less contact resistance, and because the package leads are ultrasonically bonded to the contact pads, introducing minimal series resistance. Since the new structures make use of a salicide mask, the simulated source and drain contacts are placed at the same distance from the gate as in the actual structures, in contrast to the minimal contact spacing used for the fully salicided structures in the previous subsection. This contact-to-gate spacing varies from 3μm to 8μm on the drain and source sides in the test structures and simulations. The

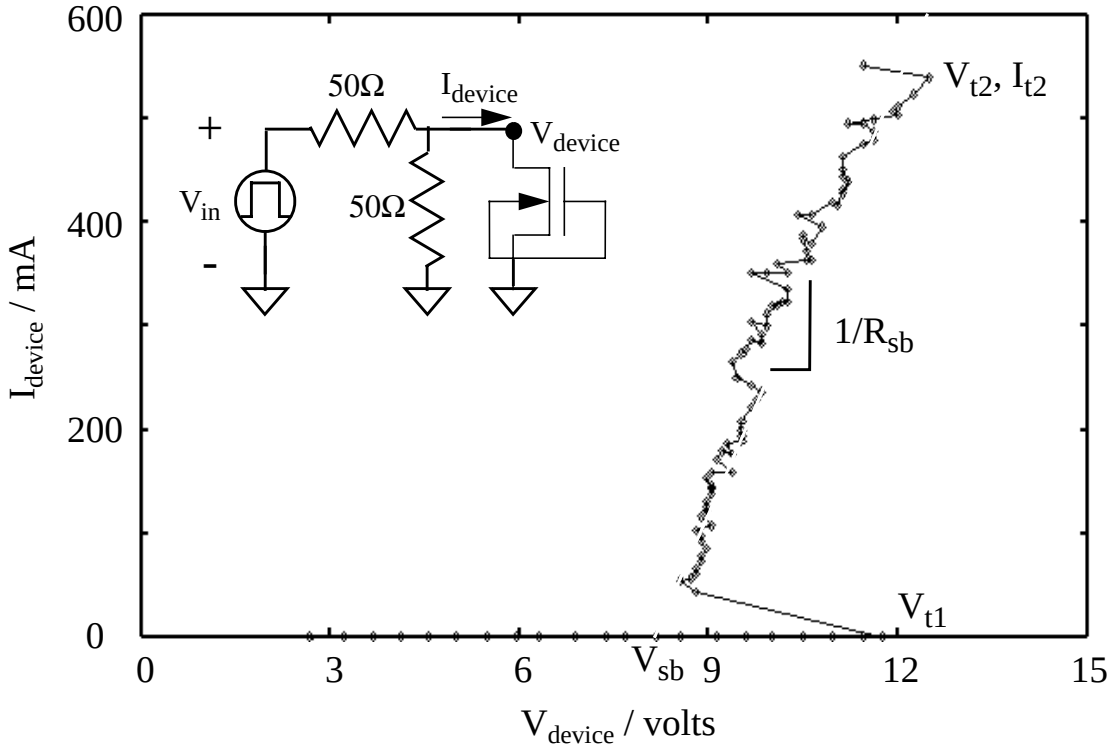

Fig. 4.41 I-V points from the transmission-line pulse sweep of a standard 50/0.75μm test structure (equivalent circuit shown inset). The trigger voltage (V_{t1}), snapback voltage (V_{sb}), snapback resistance (R_{sb}), and second-breakdown point (V_{t2}, I_{t2}) can be extracted from the curve.

simulation gate length was also adjusted to the standard test-structure value of 0.75μm. Since the mobility model coefficients determined in the last calibration phase match characteristics of both 0.5μm and 3.0μm gate-length structures, they should be valid for the intermediate value of 0.75μm.

Initially, the number of doping regrids in the creation of the simulation structures was reduced from three to two in order to decrease the number of grid points and thus reduce simulation time. For the new standard structure, the number of grid points decreased from 3239 to 2073 with the removal of the regrid, resulting in a 30% reduction in the simulation time of the dc snapback I-V sweep. However, a side effect of the coarser grid was an increase in the breakdown voltage (BV_{DSS}) of 0.8V, which meant the simulations no longer properly modeled the AMD technology. This change in breakdown voltage was the result of a change in the electric-field profile along the drain-substrate junction, where the regrid is most critical, which apparently reduced the overall impact-ionization generation rate. (The dependence of the electric-field profile on the simulation grid was also reported by Amerasekera et al. [32].) Due to this drastic change in simulated device characteristics, the third doping regrid was put back into the structure-generation recipe, making it identical to the recipe used in the MOSFET-characteristic calibration. Using this grid-generation method, the breakdown voltage remains approximately constant for varying gate lengths and contact-to-gate spacings. The dependence of the electric field on grid definition is somewhat alarming and should be further examined, but such examination was deferred since the generated structures appeared to work well for the simulations used in this calibration.

In the first part of this calibration phase, dc-sweep snapback simulations were run using the curve-tracing algorithm described in Section 3.2. The goal of the calibration was to match the measured trigger voltage, snapback voltage, and snapback resistance for the silicide-blocked structures with varying contact-to-gate spacings. Matching the dependence of V_{sb} and R_{sb} on gate length was also of interest, but due to the very low series resistance of fully salicided structures (the only test structures available with varying gate lengths), both of these parameters were very small and hard to capture experimentally, so the simulated dependence of V_{sb} and R_{sb} on gate length could not be compared directly with experiment. During the snapback simulations, the lattice-temperature equation (Eq. (2.2)) was not included in the solutions until after the device was well into avalanche breakdown (about 100μA). This procedure saves simulation time

and does not diminish the value of the simulation because the results of interest all occur at current levels above 100μA. The thermal boundary conditions consisted of overlapping the electrical contacts with constant-temperature (297K) thermal contacts with no thermal resistance. Although the simulations examined here are referred to as calibration simulations, if the mobility and impact-ionization models have already been fixed by the MOSFET-characteristic calibration, then comparing the measured and simulated V_{t1}, V_{sb}, and R_{sb} is really a verification procedure rather than a calibration procedure.

An example of the I-V curve of a dc snapback simulation is shown in Fig. 4.42. The horizontal line in the log curve shows where the solutions began incorporating the thermal-diffusion equation. Note that although the lattice temperature does not significantly increase above 300K until after snapback, the breakdown voltage is substantially lower without including the thermal diffusion equation because the temperature-dependent impact-ionization model cannot be used. Two things were immediately noticeable from the initial snapback simulations. First, the snapback resistance appeared to be a reasonable value (compared to experiment) immediately after snapback, but the curve quickly rolled over at higher currents, indicating a much higher resistance than in the experimental structures. Second, even when the snapback voltage was extrapolated from the initial, steep part of the snapback portion of the curve, i.e., using a value of R_{sb} equal to the measured value, the snapback voltage was about 1.8V too high. It was apparent from these simulations that calibration of the mobility and impact-ionization models using the standard MOSFET curves was inadequate for snapback simulations and thus that further manipulation of the model coefficients was needed.

Since the problem regarding the high snapback voltage was the simplest to understand, it was dealt with first. The high V_{sb} value indicates that the impact-ionization generation rate is too low for a given electric field in the snapback region of the I-V curve because the simulated voltage (and electric field) needed to sustain a given current level is too high. As shown by Eq. (3.27) and Fig. 3.22, the impact-ionization rate for electrons is determined by two model coefficients, α_n^{∞} and E_n^{crit} (or λ_n, which by Eq. (3.28) is inversely proportional to E_n^{crit}), assuming β_n is constant. In the calibration of the MOSFET substrate characteristic, α_n^{∞} was held constant and λ_n was varied until the effective II rate resulted in the proper amount of substrate current. A good fit of the substrate characteristic was attained because, as Fig. 3.22 shows, if the spread in peak electric field values throughout the stress conditions of the substrate-current test is relatively narrow, the

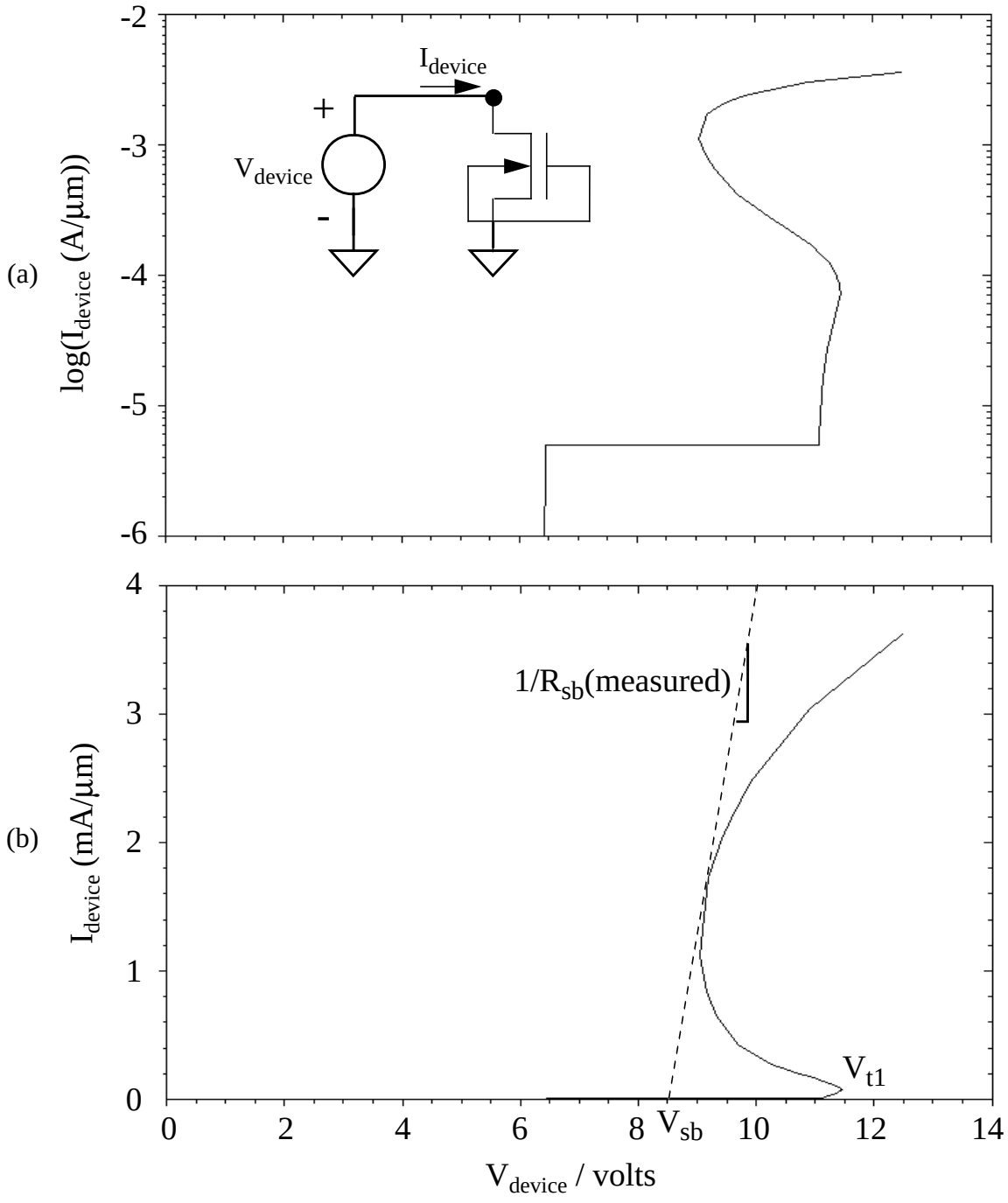

Fig. 4.42 *Device current per width is plotted on a log (a) and linear (b) scale vs. device voltage for a dc-sweep simulation of the standard structure with proper gridding and impact-ionization modeling. The snapback voltage is extracted using a line determined by the measured snapback resistance. To compare the linear curve to Fig. 4.41, multiply the current per width by 50μm.*

ionization rate, α_n, can always be fit by adjusting either α_n^∞ or E_n^{crit}. However, when impact ionization becomes important in a different electric-field regime, both model parameters must be varied to force the α_n vs. $1/E_{||}$ line to go through two (α_n, $E_{||}$) points.

As discussed in the previous subsection, Eq. (3.27) can be used to model substrate current in a MOSFET, but the α_n^∞ and E_n^{crit} coefficients must be altered to reflect the reduced mean free path, λ_n, at the surface of the device where II generation occurs. In an attempt to find II coefficients which would yield better results for the simulated snapback voltage, the substrate-current calibration was redone using a different value of α_n^∞. This value, selected from experimental results reported by Slotboom [49] on II generation at the surface of a MOSFET, is higher than the default value for bulk silicon used in the previous subsection. To compensate for this increase the mean free path had to be reduced, which is consistent with the idea of surface-related impact ionization. Just as before, λ_n was varied until the simulated substrate curves for the 0.5μm and 3.0μm structures matched the experimental curves. A good fit was again attained for both gate lengths. The final value of λ_n was equivalent to an E_n^{crit} 20% higher than the value used in the initial calibration and 46% higher than the value reported by Slotboom for surface II generation. Plotting α_n vs. $1/E_{||}$ for the initial calibration and this calibration yields lines which intersect at $E_{||}$ = $4X10^5$ V/cm, suggesting this is the average level of peak electric field during the substrate-current stress. The new coefficients predict more impact ionization than the old coefficients for electric fields greater than $4X10^5$ V/cm and less impact ionization for lower fields, i.e., the new α_n vs. $1/E_{||}$ curve is steeper. Of course, since the electron II coefficients were readjusted, the hole coefficients also had to be readjusted to refit the breakdown characteristic. Since Slotboom did not report surface coefficients for hole-induced II generation, a value of α_p^∞ was chosen such that the ratio of surface to bulk α_i^∞ was the same for electrons and holes. The hole mean-free path, λ_p, was then adjusted until the simulated breakdown voltage again matched the measured value, resulting in a value equivalent to an E_p^{crit} 50% higher than the initial calibration value.

After the MOSFET characteristic recalibration, snapback simulations were rerun, this time yielding much more accurate values of V_{sb}. The better V_{sb} fit indicates that the peak electric field in the snapback region of the I-V curve is higher than in the MOSFET substrate characteristic because the slope of the α_n vs. $1/E_{||}$ line is steeper for the new coefficients. In Fig. 4.42b, the simulated snapback voltage for the standard structure is extrapolated along the line defined by the measured snapback resistance from the point

where the line is tangent to the simulated I-V curve back to the x-axis. The V_{sb} value extracted from the simulation is still 0.3V greater than the measured value and could be improved with another iteration of substrate-characteristic and snapback-characteristic simulations using a slightly higher value of α_n^{∞}. However, since the simulated V_{sb} is within 4% of the experimental value for the standard structure and the experimental standard deviation is also on the order of 4%, no further V_{sb} calibration was performed. In the simulations, it was found that the minimum voltage during snapback increases by about 1V when the contact-to-gate spacing is increased from 3μm to 6μm, in qualitative agreement with the discussion of Section 2.4 and Table 2.1. Experimentally, however, V_{sb} remains approximately constant (~8.2V) with varying CGS. This disparity is explained by the I-V curve in Fig. 4.42b, which shows that as R_{sb} increases, the difference between the minimum voltage on the curve and the extrapolated V_{sb} also increases. Since R_{sb} increases with contact-to-gate spacing it offsets the increase in the minimum device voltage to keep the extrapolated V_{sb} nearly constant. When the simulated V_{sb} is extrapolated in the various CGS simulations using the respective values of measured R_{sb}, it too remains relatively constant.

Using measured values of R_{sb} to extract V_{sb} from the simulated I-V curves was necessary because the severe roll-off made it difficult to select a snapback resistance value based only on the simulated curve. For the test structures, the dynamic resistance may increase at high current levels due to heating and β roll-off as discussed in Section 2.2.1, but at low currents the snapback region is relatively linear, as evidenced by Fig. 4.41. The simulated rollover is therefore not physical and may be due to a combination of unrealistically high heating, improper modeling of the reduction in mobility and impact-ionization generation with increased temperature, and inaccurate modeling of the electric-field profile in the LDD region. In the simulation of the standard structure, the peak temperature exceeds 400K at a current level around 1.7mA/μm, which is coincident with the beginning of the I-V roll-off (see Fig. 4.42b). As mentioned before, the structures for the dc snapback simulations have 297K fixed-temperature boundary conditions at all the electrical contacts, which means there is no heat transfer through the sides or non-contacted area of the top of the structure. An overestimation of the peak temperature in the device would prematurely reduce the mobility and impact-ionization rates and thus explain the severe increase in simulated device voltage, so simulations were rerun with a fixed temperature of 297K on the entire perimeter of the device to maximize heat dissipation (actual calibration of the

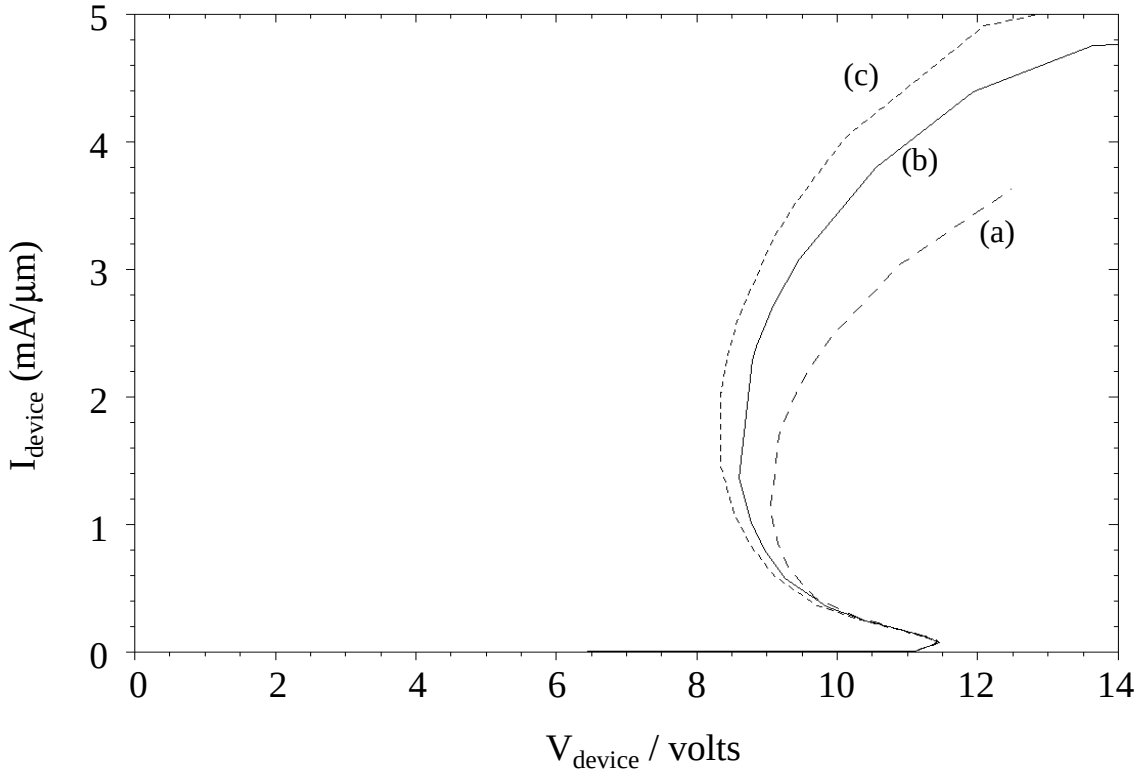

Fig. 4.43 Simulated I-V sweep for T=297K boundary conditions on (a) electrical contacts; (b) perimeter of simulation structure; and (c) perimeter of structure with reduced dependence of impact-ionization rate on temperature.

thermal boundary conditions is discussed in the next subsection). The resulting I-V curve for the standard structure (Fig. 4.43), shows that improper temperature modeling is not responsible for the severe roll-off because although the curvature is lessened around the point of minimum voltage, the roll-off is still present. Notice that the reduction in peak temperature of this simulation, which does not reach 400K until 2.3mA/μm, has definitely affected the mobility and II models because V_{sb} is lower than in the previous simulation.

It is possible that the modeled effect of temperature on the impact-ionization rates is itself incorrect. The dependence of the II rates on temperature is given by Eq. (3.29), which shows that the carrier mean free path decreases as temperature increases. To reduce this effect, the optical-phonon energy, E_p, was increased by 30% and the standard simulation was rerun (λ_n^{300} and λ_p^{300} were reduced to keep the mean-free paths at 297K equal to their values in previous simulations, and the temperature was again fixed at 297K around the

perimeter). As shown in Fig. 4.43, reducing the temperature dependence of the II coefficients has the same effect as reducing the peak temperature in the device, which is not surprising since reducing the temperature has the same effect on the mean free path as increasing E_p. A similar result was obtained for a simulation in which the high-temperature degradation of the bulk mobility was eliminated: the I-V roll-off was reduced or delayed, but it was not eliminated. It can be concluded from these simulations that the mobility and II models could not be modeled so inaccurately as to be solely responsible for the severe roll-off of the I-V snapback curve.

Since the unreasonable roll-over is not explained by any of the theories above, it is most likely due to improper modeling of the electric-field profile in the region of highest II generation, i.e., under the gate in the drain LDD. The layout of the simulation grid partially determines the field profile and thus the II generation, as was already pointed out at the beginning of this subsection when the dependence of the breakdown voltage on the simulation grid was discussed. In simulations run for a MOSFET with no LDD region, the roll-over, although definitely still present, is significantly reduced. One possible reason that an LDD device would be harder to simulate is that the electric-field profile is more complicated in the region of high current density. When the device current is less than about 100μA/μm, the II modeling appears to be correct, but for higher current in the snapback regime the grid problems are disclosed. The problem of grid definition definitely needs more attention, but since modifying the grid layout would require another iteration of calibrating the II coefficients and possibly the mobility coefficients, a solution to the problem was not pursued. As it turns out, the snapback resistance can still be extracted from the simulated I-V curve by measuring the tangent just after snapback, where the peak temperature is not much above 297K. As shown by the curves of Fig. 4.43, the slope is approximately constant for the first 0.5mA/μm above the current corresponding to minimum device voltage. Values for the simulated R_{sb} vs. CGS will be given in the section on snapback I-V results and compared to the experimental values.

The final parameter to be considered in the dc snapback simulations is the trigger voltage, V_{t1}. In the TLP experiments, a trend could not be seen between variation in the contact-to-gate spacing and V_{t1}. Values ranged from 11.7V to 12.0V (BV_{DSS} is about 11.2V), but the lowest and highest V_{t1} did not correspond to the lowest and highest CGS. The lack of a trend is not surprising. Since the device current before snapback is less than 5mA and the difference in series source/drain resistance between 3μm CGS and 8μm CGS is about 12Ω

for a sheet-resistance of 60 Ω/❑ and width of 50μm, the difference in V_{t1} due to increased CGS should be less than 60mV, a value smaller than the standard deviation of the V_{t1} measurement of any given structure. In the simulations, V_{t1} varies from 11.4V for 3μm CGS to 11.55V for 8μm CGS, a reasonable spread in values. The lower value of V_{t1} in the simulations may indicate that the modeled source/drain resistance or channel resistance is too low. Alternatively, or additionally, the modeled impact-ionization rate may be too high near V_{t1}, requiring less voltage to generate the needed carriers to trigger the MOSFET into snapback. The low V_{t1} would also be explained by an unrealistically high substrate resistance in the simulations which would allow the potential in the channel to build up more quickly and thus facilitate device turn-on, as described in Section 2.4. Since the difference between simulated and measured V_{t1} is only 0.4V, though, the simulations were considered to be calibrated reasonably well.

4.1.4 Calibration of Thermal Failure

The final step in calibrating the NMOS ESD structures is the determination of the thermal boundary conditions which will allow accurate simulation of thermal runaway. To determine these boundary conditions, the placement of thermal electrodes and values of lumped thermal resistances are varied for different transient simulations and the resulting simulated time-to-failure vs. power-to-failure points for a given structure are compared to the measured failure points. As mentioned in the previous subsection, experimental failure points were taken using the TLP setup, which tracks the leakage evolution during a TLP experiment and thus can record the device current and voltage at the point of failure, i.e., when the input pulse produces microamp leakage. In the widest test structure, a 100/0.75μm device, microamp leakage was most often created the first time second breakdown was observed on the oscilloscope. For the narrowest (25μm wide) structure, second breakdown often first occurred without inducing failure, a phenomenon that was explained in Section 1.1. Thus, to avoid confusion in interpreting the experimental results, the failure points used for calibration are taken from the 100μm-wide structure. As with the snapback-curve parameters, the experimental data points used are the average values of a number of tests. Since most of the TLP data was taken using a 200ns pulse, this time frame is the focus of the calibration. The calibration in this subsection covers only the 100/0.75μm device with standard contact-to-gate spacing. In order to calibrate the simulations across a large design space, structures with varying CGS values should also be

simulated. Such simulations were performed, but the results of these simulations are not given until Section 4.3.

Mixed-mode simulations (Section 3.3) were used to model the TLP circuit shown in Fig. 2.14b, using a lumped 50Ω resistor between the square-wave voltage source and the drain of the MOSFET (to simulate the transmission-line impedance) and a 50Ω shunt resistor connected at the drain. Since the 100μm-wide test structures are robust enough that no additional series resistance (R_s) is needed in the TLP circuit, this resistance was left out of the test setup and simulations. The rise time of the simulated square wave was set to 3ns, the average rise time of the pulse in the TLP setup. Just as in the experimental setup, each simulated TLP pulse width used to stress a structure has a unique height which will trigger second breakdown. Thus, multiple simulations with different pulse heights must be run to define a P_f vs. t_f curve. Since the exact relationship between the input pulse height and the time to failure is not known, the simulated square pulses are simply given very large widths and a simulation is discontinued when failure is reached (determination of the failure condition is discussed below).

As a starting point for determining the thermal boundary conditions, thermal electrodes were placed coincident with the source, drain, gate, and substrate contacts just as they were for the dc snapback simulations. This configuration implies that no heat transfer occurs through the sides of the structure or the non-contacted areas on the top of the structure. In the real structures, the substrate electrical contact is on the surface of the source-side of the device, outside the defined simulation space. Therefore, the thermal electrode overlapping the substrate contact along the bottom of the structure is not meant to model the heat sink of the substrate contact itself but rather the heat sink of the entire silicon substrate. As discussed in Section 3.1, by applying a lumped thermal resistance and capacitance to the substrate thermal contact, the contact can be made to approximate the thermal mass of the entire substrate. In simulations of very short ESD pulses, the thermal boundary conditions are not important because the heating is very localized. However, for longer stress times the high-temperature region extends a greater distance and the thermal boundary conditions become more important.

In the initial transient simulations, a lumped thermal resistance of 10,000 K/W (a value loosely based on a calculation by Diaz [24]) was placed on the substrate contact, and in order to simplify the simulations no thermal capacitance was used. For simulations with

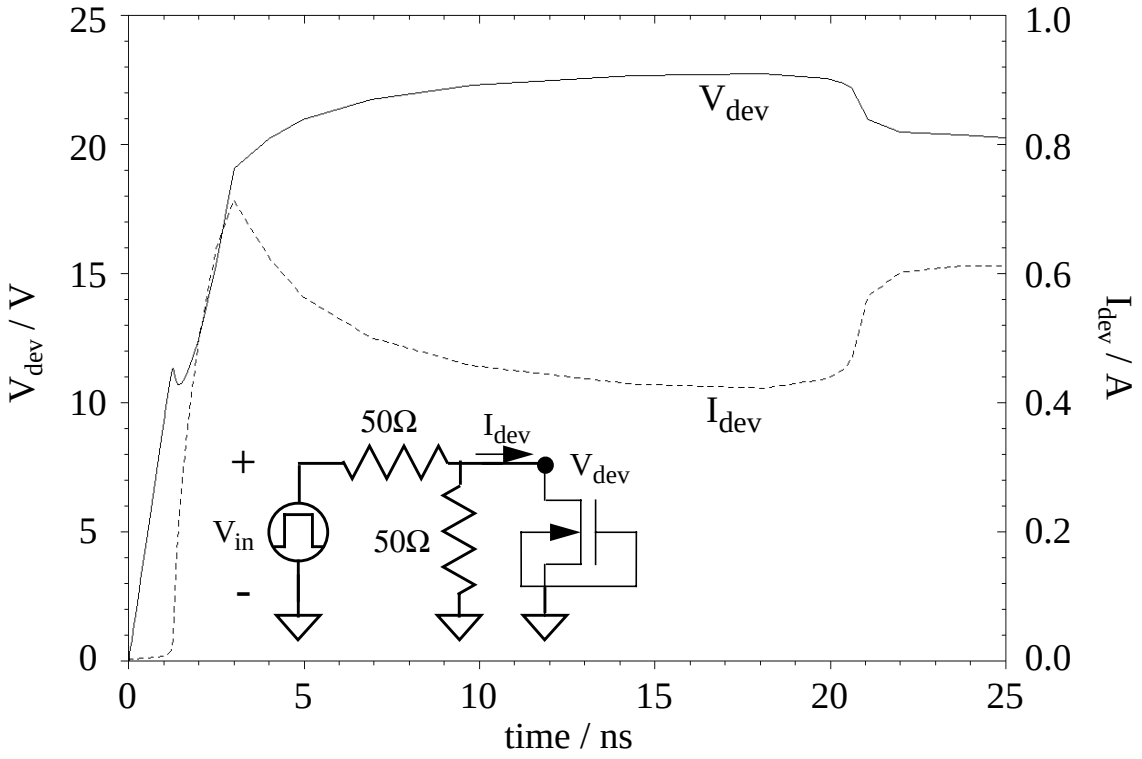

Fig. 4.44 Device voltage and current vs. time for a transient simulation of the 100/0.75μm structure (cf. Fig. 2.10). Second breakdown is observed at 21ns, corresponding to a peak temperature in the device of 1510K. The simulation circuit is shown inset.

relatively high input pulses, a distinct second breakdown was observed, as shown in Fig. 4.44. In the figure, the drop in current and rise in voltage after 3ns are due to the incorrect modeling of the device resistance in the snapback region discussed in the previous subsection. Although the device voltage is too high and the current is too low in the simulation, the power generated in the device is equal to the current-voltage product and thus may still be a reasonable value to use for thermal-failure calibration. In all of the simulations with a second-breakdown time less than 100ns, this time is well defined by a sharp increase in the device current and the peak temperature at this time is around 1500K. The intrinsic carrier concentration at 1500K is about $3X10^{18}$ cm^{-3}, which approximately equals the doping concentration in the LDD region where the temperature is highest. This result is in agreement with the simple theory of thermal failure which states that a critical temperature, in this case 1500K, defines the onset of second breakdown. Since the drop in

voltage and rise in current were more drawn out for failure times greater than 100ns, the time and power to failure (V_{dev} X I_{dev}) were defined as the time and power at which the peak temperature reached 1500K.

For simulations of the 100/0.75μm structure using the thermal boundary conditions described above, the power to failure for a failure time of 200ns was about 4W. In comparison, the average measured failure power using a 200ns TLP pulse was 11.2W, more than twice the simulated value. This underestimate of the power-to-failure indicates that the modeled heat dissipation was too low, i.e., the thermal resistance was too high, forcing the peak temperature to be too high. Thus, for the next iteration of simulations the lumped thermal resistance was removed from the substrate thermal contact to reduce the device heating. As a result, the power-to-failure at 200ns was increased, but only to about 6W, still almost 50% too low. At this point it was recognized that the absence of heat dissipation to the sides of the simulation structure was incorrect. Since no thermal contacts were placed on the sides of the structure, too much heat was being trapped. In the discussion of the 3D thermal box model (Section 2.2.2), it was explained that the linear extent of thermal equilibrium in an area where heating is time-invariant after time t_0 is equal to $\sqrt{4\pi D\,(t - t_0)}$. Assuming a diffusivity of 0.35cm^2/s, a time of 200ns corresponds to a distance of about 9.4μm. This is nearly twice the distance from the heat-generation region under the gate to the sides of the standard structure, and thus the lack of thermal contacts on the sides of the structure drastically increases the peak temperature. In light of this calculation, constant-temperature boundary conditions were added to the sides of the simulation structure with no lumped thermal resistance. The lack of thermal resistance is reasonable because the silicon substrate is an effective heat sink and, as shown by the calculation above, the dissipation of heat for the time scale of interest is not affected by a region much greater than the simulation space.

In simulations using these boundary conditions, the failure power at 200ns again increased, but only to about 8.0W, still 30% lower than the measured value. If the critical temperature for device failure is redefined as 1688K, the melting point of silicon, the 200ns failure power does increase, but only about 10%, still not enough to compensate for the disparity between simulation and experiment. Since the thermal boundary conditions have been set to maximize heat dissipation, it appears that either 2D simulation is not adequate for quantitatively predicting thermal failure or that the inadequate calibration of the snapback I-V curve for currents well above the snapback point renders proper

modeling of thermal failure impossible. In the comparison of the 2D and 3D thermal box models in Section 3.6, the 2D model was found to overestimate the failure power, not underestimate it as in this case. This suggests that the problem lies not in the abilities of 2D simulation but in insufficient calibration of the high-current, high-temperature portion of the I-V curve. More work needs to be done to determine if quantitative power-to-failure vs. time-to-failure simulations can be accomplished using the chosen simulation models. Given the results of the simulations in this subsection, it is clear that the thermal boundary conditions must be set to maximize heat dissipation if the models are to be used with the coefficients determined by the calibration procedures described in this chapter. This is the approach taken in the (qualitative) failure simulations of Section 4.3.

4.2 MOSFET Snapback I-V Results

In this section and the following section, selected results will be presented for snapback I-V curves and device failure, respectively, from transmission-line pulsing tests and TMA-MEDICI 2D simulations. TLP experiments were performed on structures from the AMD 0.5μm-technology described near the beginning of this chapter, and the simulation results are based on the calibrated models detailed in Section 4.1. In the experiments and transient simulations, parametric NMOS transistors were stressed with positive pulses incident at the drain with the source, gate, and substrate grounded (except where noted) as depicted in the inset of Fig. 4.41. In dc simulations, the drain was swept with the source, gate, and substrate grounded, as in Fig. 4.42. The results are presented as a sort of potpourri with the intention of illustrating the uses of TLP discussed in Chapter 2 and the related simulation applications discussed in Chapter 3; comparisons will be made between simulation and experiment where applicable. Many of the individual results will be brought together in Section 4.4 to form the basis of an ESD circuit-design example.

Examples of the I-V curves generated by a TLP experiment and a dc-sweep simulation were already given in Fig. 4.41 and Fig. 4.42, respectively. Section 4.1.3 discussed the relatively weak dependence of the trigger voltage, V_{t1}, and snapback voltage, V_{sb}, on contact-to-gate spacing observed in the TLP tests and simulations. There is a definite dependence of the snapback resistance on CGS, though, and this is shown in Fig. 4.45 for 50/0.75μm devices. Experimental values are the average linear least-squares fit of the I-V points between snapback and second breakdown, while each simulated value is taken as the slope of the dc-sweep I-V curve just after snapback as specified by Section 4.1.3.

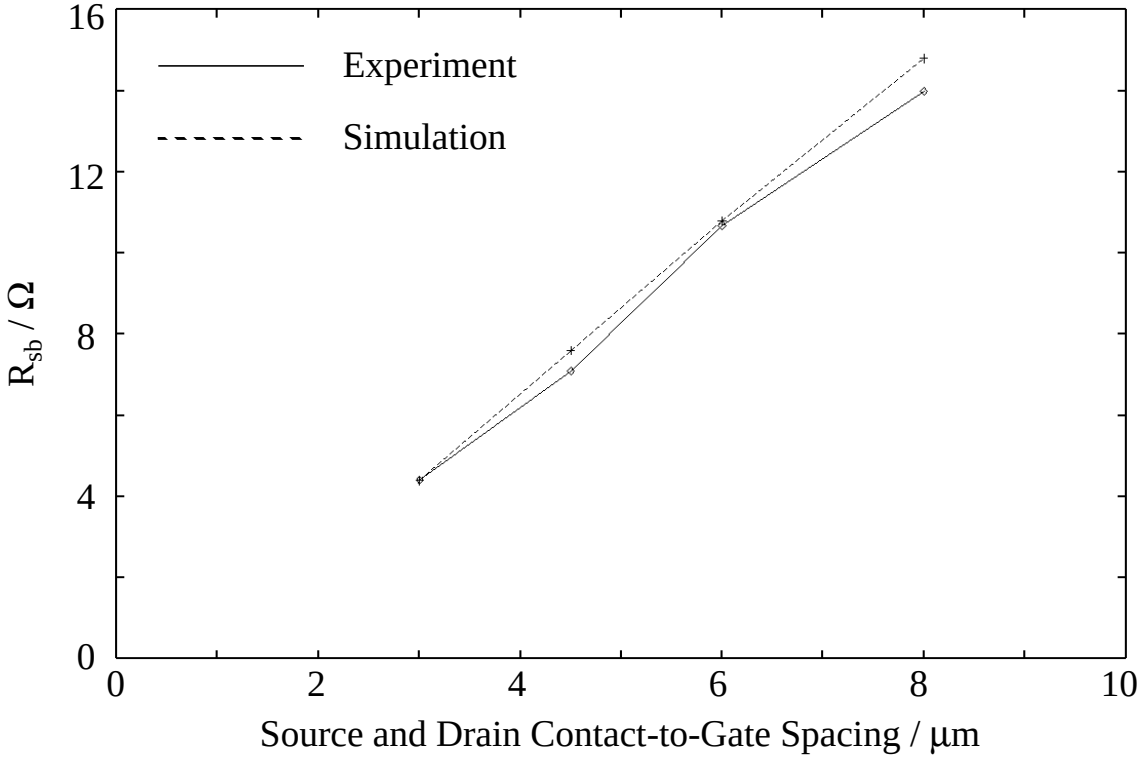

Fig. 4.45 Experimental and simulated snapback resistance, R_{sb}, vs. contact-to-gate spacing for a 50/0.75μm MOSFET test structure. The contact-to-gate spacing refers to the distance from the drain contacts to the gate edge and the source contacts to the gate edge.

There is good agreement between simulation and experiment, and both show that R_{sb} has a linear dependence on CGS for CGS between 3.0 and 8.0μm. This linear dependence might be expected because increasing CGS increases the series resistance from drain to source. However, note that if the line is extrapolated to zero CGS, R_{sb} is negative, indicating that extrapolating linearly to lower CGS values will lead to incorrect results. This could be due to experimental uncertainty and to uncertainty in the simulation extractions, although the agreement between the two curves suggests the values are correct. Heating effects also play a role in determining R_{sb}, as seen in Fig. 4.41, in which the line with slope $1/R_{sb}$, determined by the least-squares fit of the points between snapback and second breakdown, has a smaller slope (greater resistance) than the line formed by the first few I-V points after snapback, a result of the increased resistance at higher currents when device heating becomes significant. If the effect of heating lessens as

CGS decreases, then the slope of the R_{sb} vs. CGS curve should be lower at low CGS, implying that R_{sb} is really positive as CGS approaches zero, as it must be. To determine what parameters do in fact play a role, experiments and simulations need to be run on structures with lower contact-to-gate spacing. However, interpolating values of R_{sb} for CGS between 3μm and 8μm should be a safe practice.

In 2D simulations, any resistance is inversely proportional to device width because the simulations are effectively normalized in the width dimension. However, Fig. 4.46 shows that for real structures the extracted snapback resistance is not proportional to the inverse device width for widths greater than 50μm. Once again, this is a result of device heating and the consequent increase in device resistance at high current levels. For a given current density, heating is more severe in a wider structure because the center of the device is farther away from the structure edges where heat can be dissipated. Therefore, the extracted snapback resistance for wide devices is higher than predicted by the narrow-width line fit.

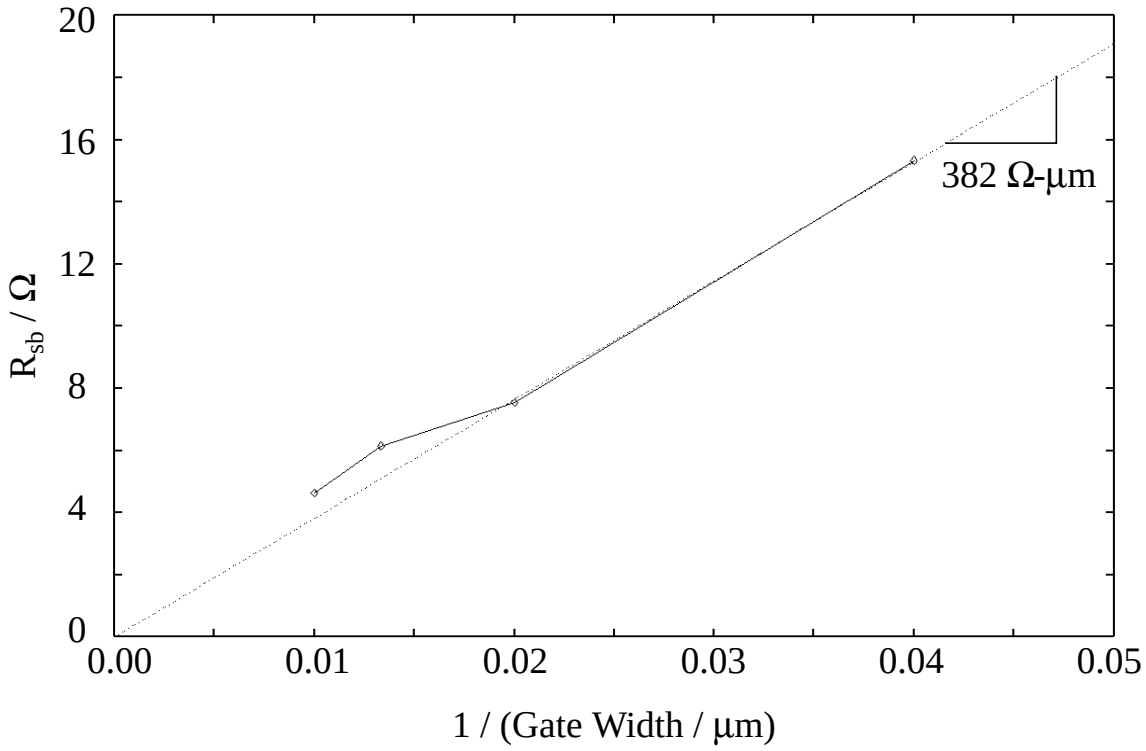

Fig. 4.46 Experimental snapback resistance, R_{sb}, (connected points) vs. inverse gate width, W, for 0.75μm test structures. The dashed line indicates that $R_{sb} \times W = 382$Ω-μm for gate widths less than 50μm.

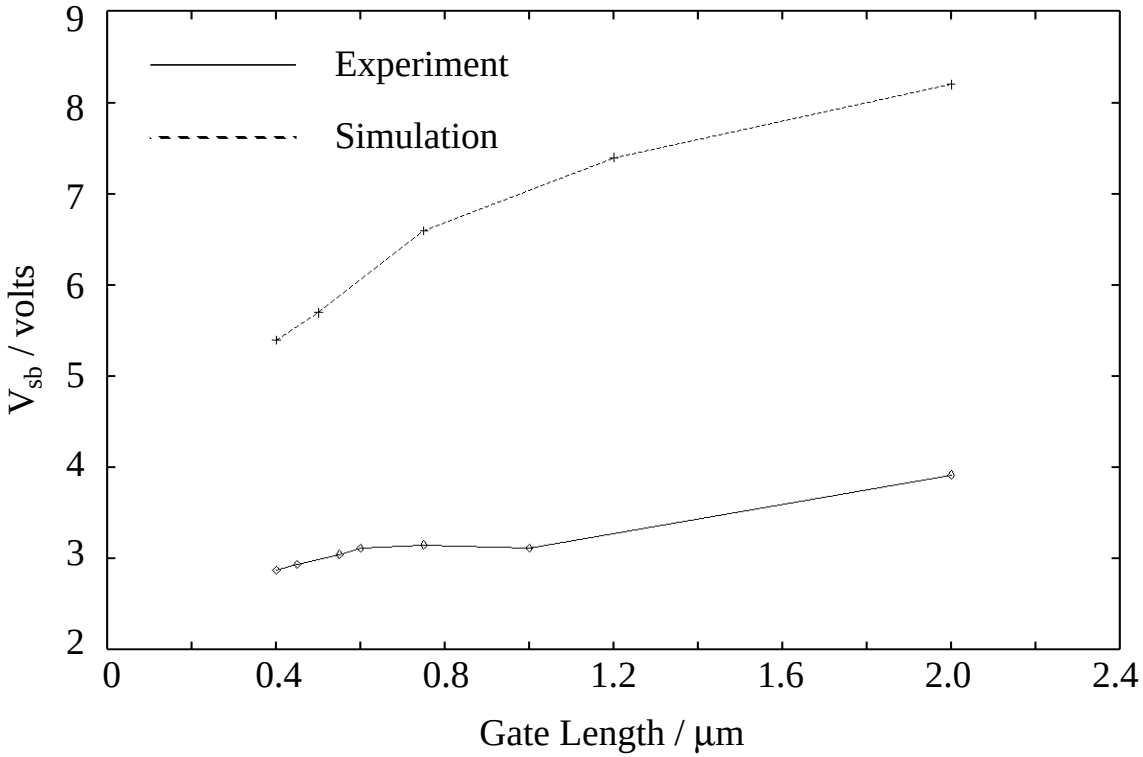

Fig. 4.47 Experimental and simulated snapback voltage, V_{sb}, vs. gate length for 20μm-wide test structures. The experimental results are for fully salicided structures, while the simulation results are for structures with 1.0μm CGS.

Test structures with varying gate length, L, could not be used for calibration in the previous section because the only structures available with varying L were fully salicided structures (due to limited space on the salicide-masked test tiles), for which the snapback portion of the I-V curve is hard to finely capture with TLP due to the very low series resistance and the small size (20μm) of the structures. Extracting a value for R_{sb} is especially hard since it is close to zero, but values for V_{sb} were obtained and are plotted in Fig. 4.47 along with results from simulations. The fact that the extracted snapback voltages are lower than the supply voltage of the technology (5V) indicates that the structures actually snapped immediately into second breakdown.

In the simulation structures, an attempt was made to model the salicide by extending the source and drain contacts right up to the spacer edge, as was done for the first stage of

calibration. However, simulations would not converge for these structures past the snapback region, most likely because the drain contact was so close to the drain depletion region that it was adversely affecting the device physics in this critical region. Therefore, the contact-to-gate spacing was set to 1.0μm on the drain and source sides. Fig. 4.47 does show a reasonable correlation between simulation and experiment, although the simulated V_{sb} is much higher due to the series resistance of the 1.0μm CGS. The gate length will be varied in structures on future test tiles to better determine its effect on V_{sb} and R_{sb} in ESD protection devices.

The last I-V parameter considered in this section is the trigger voltage, V_{t1}, and its dependence on the value of the gate-bounce resistor placed between the gate electrode and the grounded source in an ESD MOSFET structure (see Fig. 2.17a). As described in Section 2.3, placing a resistance between the gate and ground allows a voltage to build up on the gate during the initial stage of an ESD stress which facilitates device turn-on by inducing MOS action. Due to a limited amount of material available for testing, experiments could not be run with several values of gate resistance, R_{gate}, so most of the TLP experiments were run with the gate electrode grounded. A few tests were run on 50μm-wide structures with a lumped resistance of 7kΩ connected between the gate pin and ground (external to the DIP package), but V_{t1} was not significantly lower than in grounded-gate tests, remaining at about 11.8V. Using transient simulations, however, the relationship between V_{t1} and R_{gate} was studied over a wider range of gate resistances. Results of these simulations, plotted in Fig. 4.48, predict that R_{gate} does not significantly affect the trigger voltage until it reaches a value of about 10kΩ, which explains why the 7kΩ resistance used in the experiments had little effect. Using Eq. (2.14) with an input voltage rise of 16V/ns (simulated pulses were 48V with a rise time of 3ns), an overlap capacitance of 17fF (based on a gate oxide thickness of 100Å and an estimated gate-drain overlap of 0.05μm), and a gate resistance of 10kΩ, the calculated gate voltage should reach a maximum of 1.38V. This voltage is well above the threshold voltage of the MOSFET, V_T, and thus MOS transistor action occurs during the initial rise of the ESD pulse. In simulations using a gate resistance of 7kΩ and 10kΩ the simulated peak gate voltages were 1.20V and 1.44V, respectively. Both values are above the MOSFET threshold voltage, but it appears that the peak gate voltage must be significantly above V_T to have an effect on V_{t1}, perhaps because the time to snapback is so brief (about 1.4ns).

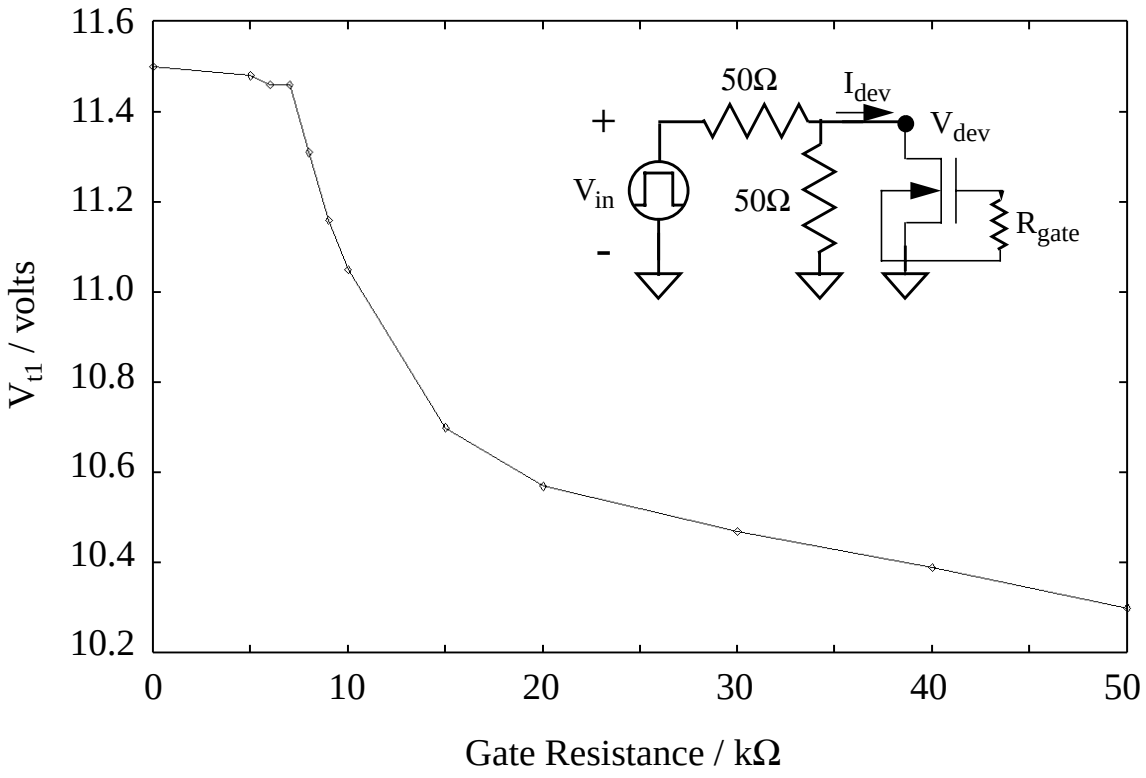

Fig. 4.48 *Simulated trigger voltage, V_{t1}, vs. gate resistance, R_{gate}, for the 50/0.75μm test structure. Simulations predict that R_{gate} does not have a significant effect until it reaches a value of about 10kΩ.*

4.3 Device Failure Results

The transmission-line pulsing simulation and testing procedures used to obtain device failure results were specified in the last section. For studying thermal failure, transient simulations are always used because the time dependence of the power to failure or current to failure cannot be modeled with steady-state I-V sweeps. In any 2D simulation, the modeled failure current and failure power must be directly proportional to the device width because the simulation is normalized in this dimension. The 2D and 3D thermal-box models used to describe thermal failure also predict that the failure power per unit device width is independent of the width. Experimentally, however, the normalized power to failure and current to failure are found to decrease as the device width increases, as shown in Fig. 4.49 for 200ns transmission-line pulses. This discrepancy is explained by the different criteria used to define device failure in the models and experiments and was

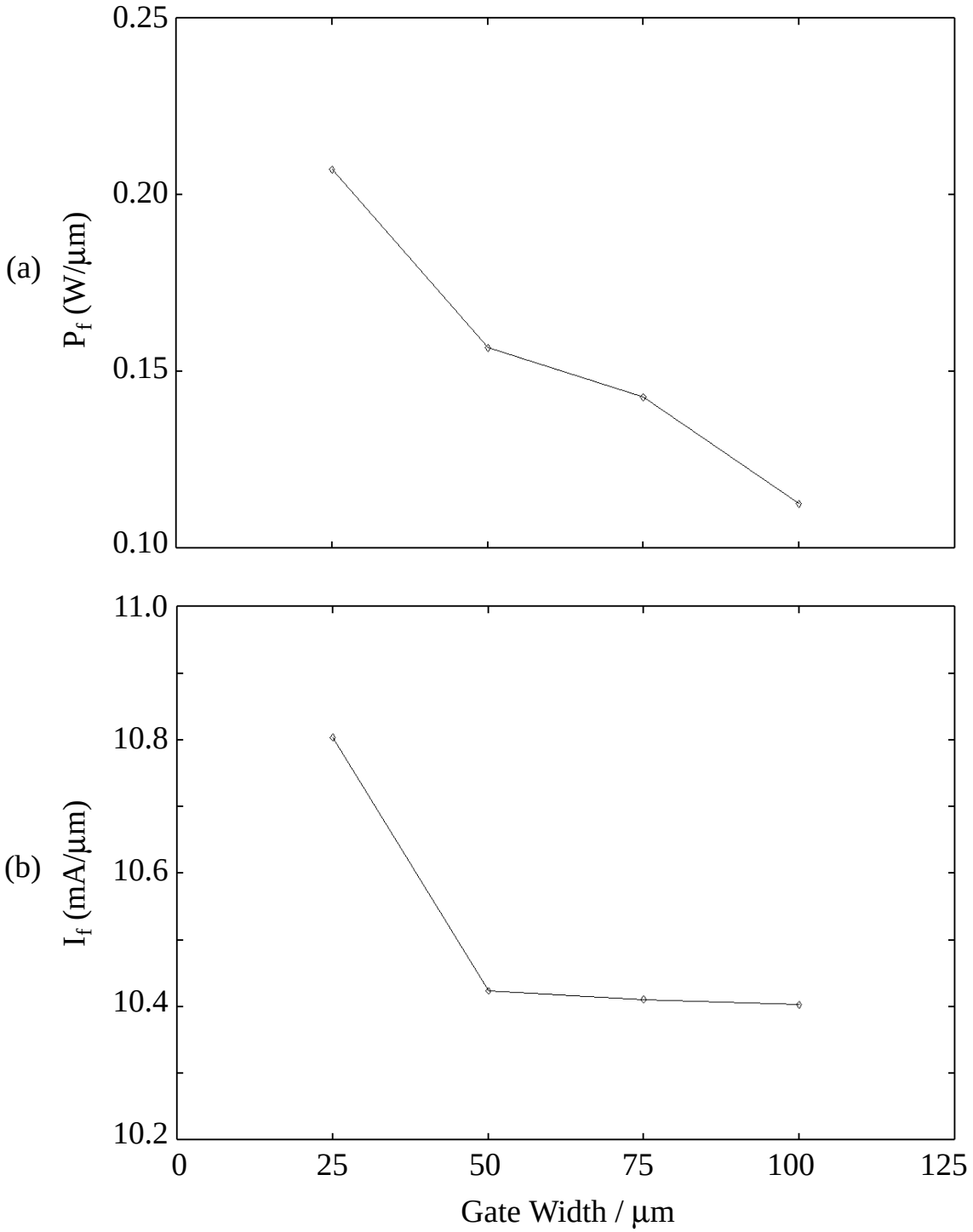

Fig. 4.49 Power to failure, P_f, (a) and current to failure, I_f, (b) vs. device width for 0.75μm test structures subjected to stepped 200ns transmission-line pulses. Each value is divided by the device width to normalize the results.

already discussed in Section 1.1 as well as by Polgreen [8]. In the TLP tests, failure is defined as the point at which device leakages exceeds 1μA, while in the thermal-box model failure is defined as the onset of second breakdown. A certain current density is needed to cause a device to enter second breakdown, but widespread damage does not follow instantaneously in narrow devices because there is not enough total energy in the TLP pulse, and consequently narrow structures must be stressed with higher pulses than predicted before damage is severe enough to create microamp leakage. Of course, the absolute current to failure and power to failure increase with device width, but note that as the width increases beyond 50μm, the failure current per width levels off (Fig. 4.49b) while the normalized failure power continues to decrease (Fig. 4.49a), indicating that the device voltage at failure, V_f, decreases with width. The decrease in failure voltage with width is explained by the fact that the snapback resistance, which is roughly inversely proportional to the width (Fig. 4.46), decreases with width more rapidly than the failure current increases with width. In Section 2.4 and Table 2.1, the width was predicted to have no effect on V_f (V_{t2}), but in Section 2.4 it was assumed that the failure current scales directly with width, which is not the actual case. It would be beneficial to test even wider structures to determine if there is a point at which the normalized power to failure levels off.

In Section 4.1.4, the 100μm-wide structure was used for calibration of thermal failure because microamp leakage was almost always created the first time second breakdown was captured on the oscilloscope and thus there was no ambiguity in defining the failure level. However, as seen in Fig. 4.49b another advantage of using wide structures for calibration is that the measured failure current is proportional to device width for wide devices and therefore more amenable to 2D simulation. In contrast, according to the thermal-box model the intrinsic error between predicted 2D and 3D failure power (or failure current) is independent of device width (Fig. 3.33). Again, the conflicting results are due to the different concepts of failure and underline the importance of consistently defining failure in experiments and simulations.

Experimental and simulated failure power vs. contact-to-gate spacing for 50/0.75μm structures subjected to 200ns TLP stressing are compared in Fig. 4.50. As just stated, the experimental failure level is defined as the power needed to create microamp leakage, but for 200ns pulses this level usually coincides with the power-to-second breakdown. In the simulations failure was defined, as described in Section 4.1.4, either by the time at which

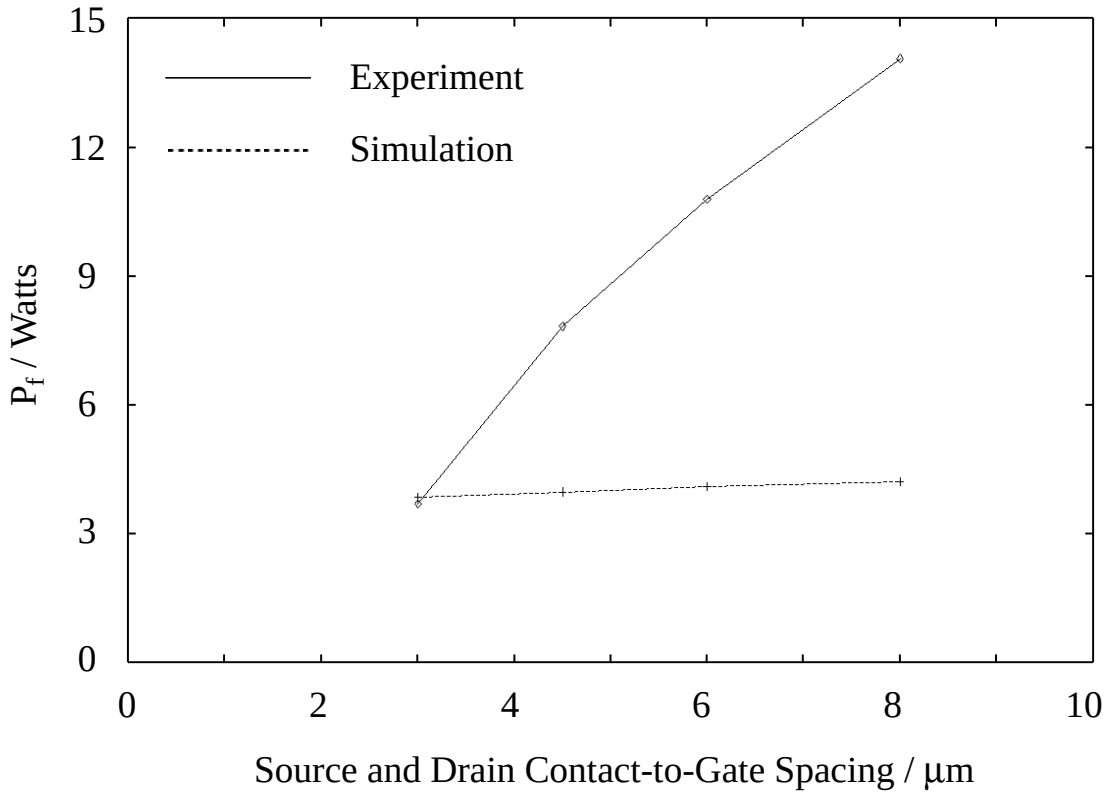

Fig. 4.50 Simulated and experimental power-to-failure, P_f, vs. contact-to-gate spacing for 50/0.75μm test structures subjected to 200ns TLP pulses.

second breakdown was observed or the time at which the peak temperature reached 1500K (the peak temperature is always about 1500K when second breakdown is observed). Since failure ensues immediately upon second breakdown in the experiments, the measured and simulated failure conditions should be consistent. The results reveal the shortcomings of the high-current calibration discussed in Section 4.1.4. As expected, the robustness of the test structures increases with CGS because the added space between the gate and the source/drain contacts provides more area over which to dissipate the energy of a stress pulse. In the simulations the same effect is observed, but it is very abbreviated. The unreasonably large resistance of the intrinsic device at high currents, a result of the improper modeling of the electric field in the LDD region, prevents the current from rising much beyond a certain level, and thus the added resistance of increased CGS only slightly increases the heat (energy) dissipation. Notice that the simulated result for 3.0μm CGS

actually agrees quite well with experiment, in contrast to the standard structure used for calibration, which has a CGS of 4.5μm. This good agreement suggests that structures with lower contact-to-gate spacing may be better suited for use in calibration of the thermal boundary conditions.

While the power to failure appears to continually increase with CGS, Fig. 4.51 shows that the current to failure tends to level off for contact-to-gate spacings greater than about 6μm. This indicates that the added power in structures with larger CGS is being dissipated in the increased active regions of the device (the regions between the gate and the source/ drain contacts). Since the increase in voltage to failure at higher CGS is dropped across the active regions, the results also suggest that the failure point is always in the intrinsic region of the device because the voltage across the drain junction and the current density in the junction--and therefore the power generation in the junction--at the time of failure

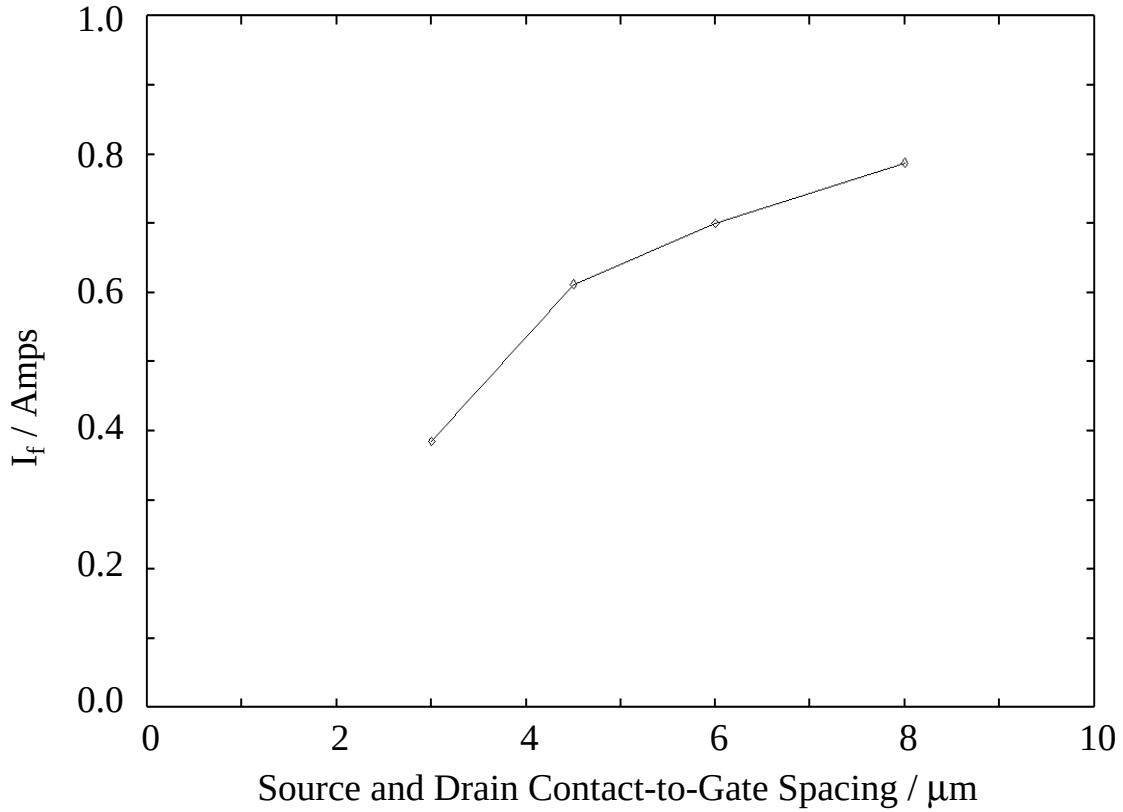

Fig. 4.51 Experimental current-to-failure, I_f, vs. contact-to-gate spacing for 50/0.75μm test structures subjected to 200ns TLP pulses.

are independent of CGS. Simulations also indicate that failure always occurs in the intrinsic device because the point of peak temperature is in the drain LDD regardless of the value of CGS, though the importance of this corroboration is diminished by the inaccuracy of the value of simulated failure power.

The different trends in failure power and failure current with CGS raise the question of which figure of merit is more important, the maximum current a device can sustain without damage, or the maximum power (this question was also raised by Diaz [24]). Since an ESD stress consists of dissipating a certain amount of charge in a certain amount of time, the maximum current a device can withstand for different lengths of time is probably a more important indicator of how well the device will perform under actual ESD stress conditions. Also, even though a protection device with a larger contact-to-gate spacing can sustain a higher input power, the higher voltage at the drain of the device is dangerous because this is the voltage seen by the thin gates of the input circuit being protected. The protection structure with a large CGS may itself survive an ESD pulse while not preventing dielectric damage of the input circuit it was designed to protect.

To determine the effectiveness of a protection structure over a range of stress-event periods, the structure can be tested with transmission-line pulses of several lengths. Fig. 4.52 displays the results of experimental P_{t2} vs. t_2 (power-to-second breakdown vs. time-to-second breakdown) points for 25/0.75μm test structures taken using five different pulse widths between 50ns and 600ns. Each point is the result of capturing the time of second breakdown on the oscilloscope screen and multiplying the current and voltage values just before this time to determine P_{t2}. Although only five pulse widths were used, failure points were captured at several times between 10ns and 600ns, a result of the random TLP stress-step sizes used and the slight dimensional variations from structure to structure. In the oscilloscope display of Fig. 2.10, for instance, the device is stressed with a 150ns pulse, but the captured second breakdown point is at 72ns. Note that P_{t2} is not referred to as the power to failure--if second breakdown occurs right before the end of the pulse, the structure often does not exhibit gross leakage afterwards because only a very short time was spent in the second-breakdown mode and therefore there was not enough energy to create damage.

The P_{t2}-t_2 points of the semi-log scale of Fig. 4.52b suggest that there is a critical time constant equal to about 50ns because for times less than 50ns there is a sharp increase in

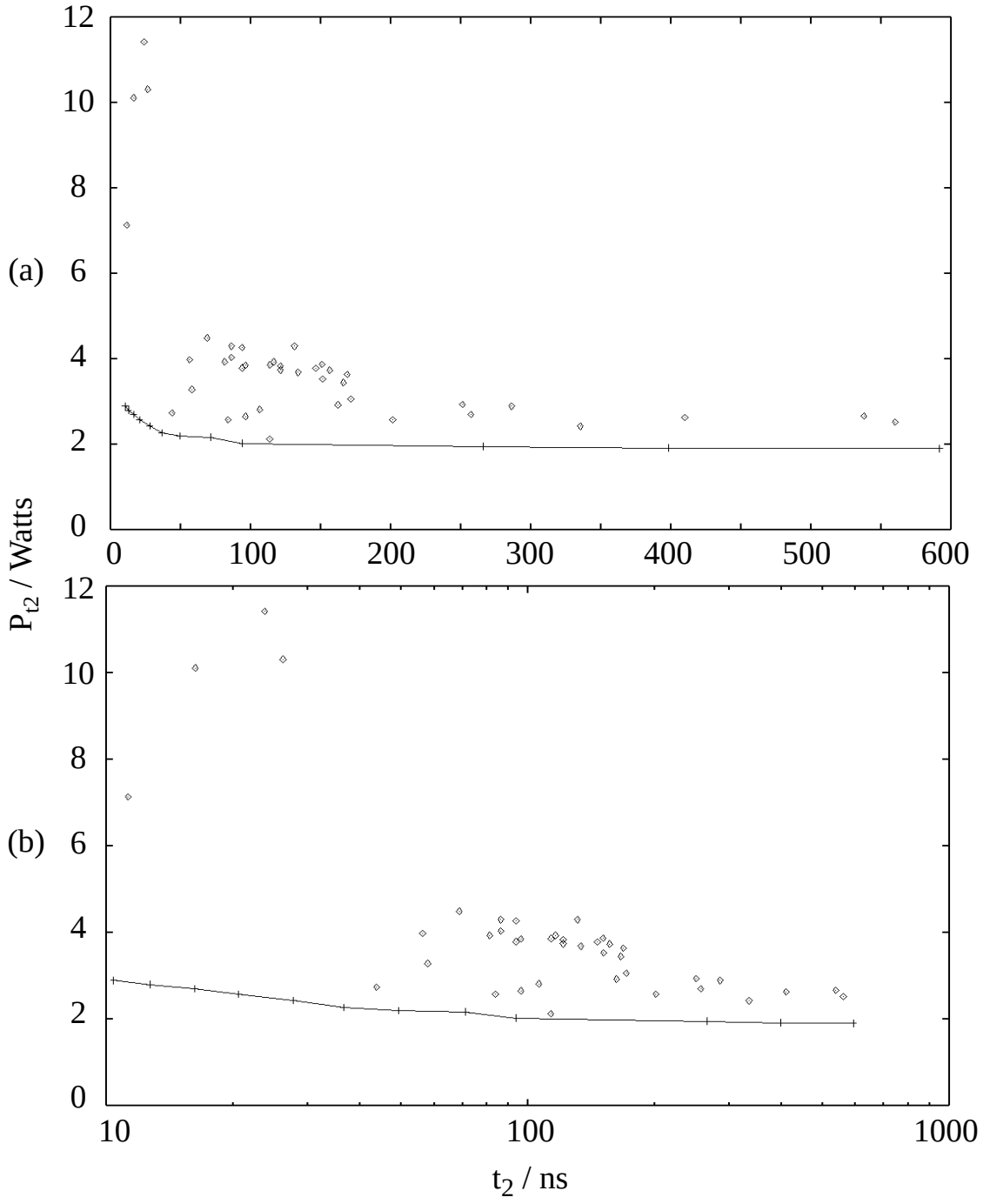

Fig. 4.52 *Power at second breakdown, P_{t2}, vs. time to breakdown, t_2, for a 25/0.75μm structure plotted on linear (a) and semi-log (b) scales. Experimental results (points) are extracted from TLP experiments using various pulse lengths, while simulation points (line) are taken from simulations with varying pulse heights of indefinite width.*

P_{t2}. Assuming a diffusivity, D, of 0.35cm^2/s, the dimension of the 3D thermal-box model corresponding to this time constant is $\sqrt{4\pi D t_2} = 4.7\mu m$. This dimension is too large to be related to the gate length or junction depth, but it is only a factor of five smaller than the device width, so the breakpoint may indicate where the failure power changes from a $1/\log(t)$ dependence to a constant (refer to Fig. 2.12). However, a similar breakpoint time was seen for wider structures, and for times less than about 40ns there is significant uncertainty in the measurements due to circuit noise, so this conclusion is premature. Improvement in the measurement uncertainty can probably be achieved by enhancing the automated algorithm used to capture the second-breakdown points and by further improving the high-frequency characteristics of the test jig. After these tasks are completed we will take more low-end points and try to fit the resulting P_{t2}-t_2 curve to the 3D box model.

Simulated P_{t2}-t_2 points are also plotted in Fig. 4.52 for the 25/0.75μm structure (simulations were actually run on 100μm-wide structures and the resulting powers were reduced by a factor of four). In the various simulations, the pulse length is simply set to a very large value and the pulse height is varied to yield different failure times. Each simulation is discontinued when the maximum temperature reaches 2000K. As in the failure-power results discussed previously, the simulated power to second breakdown is significantly lower than the measured power for all second-breakdown times. However, the simulated points exhibit a break in the P_{t2}-t_2 curve at a time close to that of the experimental results. The significance of this result must once again be questioned because of the unsatisfactory modeling of the high-current regime. Once this modeling issue is resolved, the importance of the simulated breakpoint (if it still exists) can be determined.

To close out this section on ESD device failure analysis using TLP, experimental P_f vs. t_f and I_f vs. t_f failure curves for structures with varying contact-to-gate spacing are plotted in Fig. 4.53a and Fig. 4.53b, respectively. For these plots the time to failure is equal to the TLP pulse width and 1μA leakage is used as the failure criterion. Most of these 50μm-wide structures exhibit a breakpoint between 100ns and 200ns, which again suggests a change in the P_f-t_f relationship theorized by the thermal-box model. For large failure times, the failure points reflect the results of Fig. 4.50 and Fig. 4.51, i.e., the failure power continually increases with CGS but the failure current reaches a sort of saturation point. In contrast, for the smallest pulse width (50ns) increasing the contact-to-gate spacing from 3μm to 8μm does not significantly improve either P_f of I_f (any improvement seen is on the order of three experimental standard deviations of any one structure). This indicates that

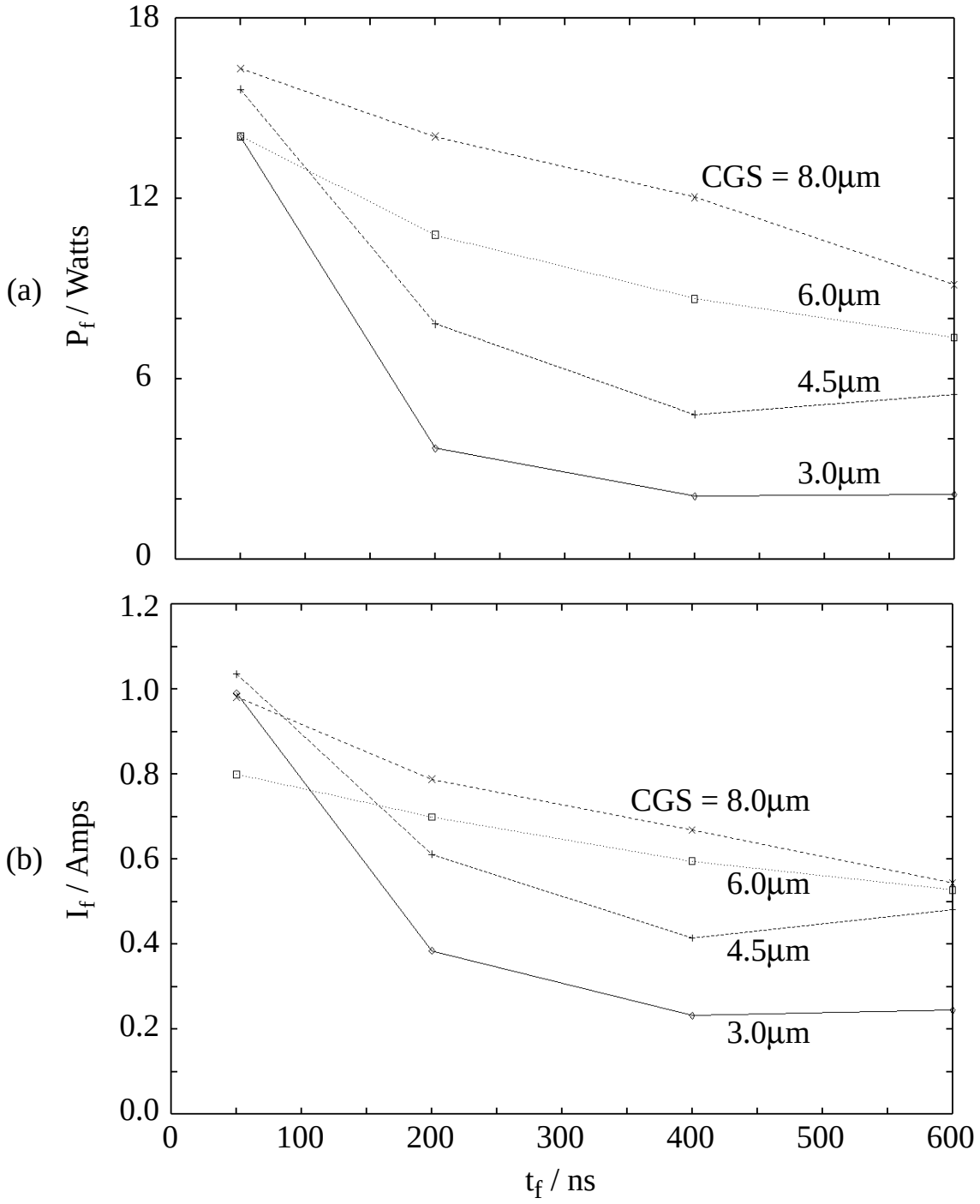

Fig. 4.53 Experimental power-to-failure (a) and current-to-failure (b) vs. time-to-failure, t_f, for 50/0.75μm test structures with varying contact-to-gate spacings (CGS). In these plots, the time to failure is equal to the TLP pulse width and the failure condition is defined as 1μA leakage.

for ESD stress times less than 50ns, the weak point of a structure lies within the intrinsic device. Thus, increasing the contact-to-gate spacing will probably improve EOS performance (stress longer than a few hundred nanoseconds) but will have little impact on the ability of a circuit to survive pulses in the ESD regime.

4.4 Design Example

As a way to unify the results of this chapter with the concepts of Chapter 2 and Chapter 3, the design of a multifingered NMOS input protection device (illustrated in Fig. 2.19) will be outlined based on the measurements and simulations presented in the previous two sections and the design methodology of Section 2.5. The protection structure would be used to protect circuits from stresses between an I/O pin and ground, as depicted in Fig. 2.20. A similar procedure could be followed to design a PMOS protection device between an I/O and supply pins. Design of the NMOS device is guided by certain performance goals:

- The protection device should be able to withstand a 4kV HBM pulse without incurring damage which would result in device leakage above 1μA.
- An effort should be made to make the device robust against EOS (stress time greater than a few hundred nanoseconds) as well as against ESD.
- The input (drain) voltage of the protection structure must not exceed 12V at any time during an ESD event. This will ensure that the gate oxides of the input circuit being protected will not suffer dielectric breakdown.
- Device layout area should be minimized.

To translate the failure thresholds of the structures in Section 4.3 to the HBM specification in the above guidelines, a correlation must be assumed between transmission-line pulse stressing and HBM stressing. Since the HBM capacitor is discharged through a resistance of 1500Ω (neglecting the much smaller device resistance), the 4kV specification translates to a peak current of 2.67A. This current is reached in less than 10ns and then decays exponentially with a time constant of 150ns (see Fig. 2.2). Of the different pulse widths used in the TLP testing, the one closest to the time range of the HBM pulse is 200ns. Thus, the average failure current of structures subjected to 200ns pulses will be directly translated to peak HBM current. This provides a margin of safety because while the current of an HBM pulse decays from its peak value immediately after the peak value is

reached, the current in a TLP pulse remains at its peak value for the entire 200ns and thus applies a greater stress. Since the robustness of the test structures is known in terms of mA of current per μm of device width, once a structure is chosen the total width required is simply the peak HBM current, 2.67A, divided by the mA/μm.

To choose an appropriate structure, a compromise must be reached between the goals of good EOS performance and minimal device area. Fig. 4.53b in the previous section shows that while increasing contact-to-gate spacing does not seem to improve device robustness for stress times on the scale of the human-body model, it definitely improves robustness for longer times, i.e., in the EOS regime. However, increasing CGS increases the total device area, so it cannot be made arbitrarily large. As seen in Fig. 4.51, the gain in failure current with increased CGS seems to level off at about 6μm CGS for 200ns pulses, and Fig. 4.53b shows that this is also true for longer stress times. Thus, a trade-off between EOS performance and device layout area is made by selecting a contact-to-gate spacing of 5μm. Section 4.1.3 reported vales of 11.8V for V_{t1} and 8.2V for V_{sb} for all the test structures. In Fig. 4.45, R_{sb} for a 50μm-wide, 5μm-CGS structure is interpolated as 8.3Ω. Neglecting the nonlinear dependence of R_{sb} on the inverse device width, R_{sb} X W will be assumed to have a constant value of 8.3 X 50 = 415Ω-μm for design purposes. From Fig. 4.53b, the interpolated average failure current of a 50/0.75μm device with 5μm CGS is 641mA, or 12.8mA/μm of device width. Fig. 4.49 indicates that the failure current density for a 50μm structure is approximately constant for fingers wider than 50μm, so the 50μm value will be used regardless of the finger widths chosen. Thus, the total 5μm-CGS device width needed to sustain 2.67A peak HBM current is 208μm.

The value for total required width assumes not only that the failure current density per micron is independent of width but also that when the multiple fingers are placed side by side, each will act exactly as if it were a single-finger structure. This second assumption will not hold for high stress currents because the heat which dissipates from a finger into the substrate in all directions will reduce the heat dissipation in neighboring fingers, thus lowering the effective current per width the device can withstand before failure. This problem is more severe for longer (EOS) stress times than for shorter (ESD) stress times. To quantify the effects of heating in adjacent fingers, multifinger test structures need to be created. For the present case, the fact that there is more energy in a 200ns TLP pulse than in an HBM pulse of the same peak current will be used to justify the calculations. Also, the calculated required width of 208μm will be increased to 250μm. The total device area

will be approximately the same regardless of the number of fingers chosen, so we will choose to build the device with five parallel poly fingers, each 50μm long. Since the total width from the drain contacts to the source contacts of a finger is approximately two times CGS, the total area will be about 50μm X 50μm. With five poly fingers, there will be three fingers coming off of the input pad into the protection device (refer to Fig. 2.19).

Since the measured and simulated grounded-gate trigger voltage of the protection structures is very close to 12V, a gate-bounce resistor should be employed to provide a margin of safety against dielectric failure of the input gates. The simulated results of V_{t1} vs. gate resistance in Fig. 4.48 show that a lumped gate resistance of 50kΩ between the gate electrode and the grounded source will reduce the trigger voltage by 1.2V for a 50μm-wide device subjected to a pulse rise time of 16V/ns. Since the device being designed has five fingers which are each 50μm wide and the drain-gate overlap capacitances add in parallel, a proportionately smaller gate resistance, i.e., 10kΩ, can be used to achieve the same amount of gate bounce. This resistance can most easily be created by placing a well resistor or tie-off transistor with a resistance of 10kΩ between the common gate and the source or substrate pad. The gate bounce should not be made too great because if the gate potential remains significantly high after a finger snaps back, the high current in the finger will be concentrated at the surface and cause severe heating at a much lower current level than if the current is distributed evenly along the vertical junction profile. The reduction in V_{t1} of 1.2V created by the 10kΩ resistor, which makes the value of V_{t1} 10.6V, is probably a reasonable value.

Assuming the fingers turn on one at a time, which is the worst-case scenario but is also the most probable scenario considering the random finger-to-finger variations in layout and the very brief (~1ns) turn-on time, after the first finger turns on the input (drain) device voltage, V_{dev}, will rise with device current, I_{dev}, as (refer to Fig. 4.41)

$$V_{dev} = V_{sb} + R_{sb} \cdot I_{dev}, \tag{4.40}$$

where R_{sb} is the snapback resistance of one finger. For the device to work properly, a second finger must turn on (snap back) before I_{dev} reaches the failure level for one finger, 641mA. In terms of the device parameters,

$$I_{dev} = (V_{t1} - V_{sb}) / R_{sb} < 641\text{mA}. \tag{4.41}$$

Using values of 10.6V, 8.2V, and 8.3Ω for V_{t1}, V_{sb}, and R_{sb}, respectively, I_{dev} will equal 289mA before a second finger snaps back, which is safely below the failure current of a single finger. Equations equivalent to Eq. (4.41) apply when two or more fingers turn on because, to first order, the voltage parameters do not change and the failure current is multiplied by the number of fingers while R_{sb} is divided by the number of fingers.

When all fingers are conducting, the device will, according to our design, not undergo thermal failure during an HBM pulse less than 4kV in magnitude. For such a pulse, the peak current is 2.67A. Plugging this value of I_{dev} and an R_{sb} value of $8.3/5 = 1.66\Omega$ into Eq. (4.40), the input voltage at the point of thermal failure is 12.6V, which is greater than the specified dielectric threshold of 12V (it is in fact greater than 12V for HBM voltages above 3.43kV). Although the dielectric-failure design goal was not met, this goal was based on the maximum voltage a 100Å oxide can withstand for any amount of time. For times less than 200ns, a thin gate oxide can withstand a much higher voltage (see Fig. 3.35 for a qualitative understanding), so the protection circuit is most likely still effective in preventing dielectric failure. The final statistics for the proposed NMOS protection-device design are

- five parallel poly fingers, each 50μm wide
- gate length of 0.75μm and symmetric source/drain contact-to-gate spacing of 5.0μm
- a gate-bounce resistance of 10kΩ
- total area on the order of 50μm X 50μm (neglecting area of gate-bounce resistor)
- estimated HBM robustness of 4kV
- input-voltage clamping of 12.6V or less for any period of time.

In this section we assumed certain correlation factors between HBM withstand voltage and TLP withstand current and between single-finger and multifinger withstand levels. Also, the effect of each layout parameter on the I-V and withstand parameters was considered individually, i.e., interactions between the various layout parameters were ignored. The next chater presents a more general design methodology in which multifinger transistors are characterized in order to extract models relating I-V and withstand parameters to layout parameters. The design space covers single-finger and multifinger transistors and the models include interaction terms. Additionally, a more rigorous approach is taken to correlate TLP and HBM withstand levels.

Chapter 5

Design and Optimization of ESD Protection Transistor Layout

To ensure electrostatic discharge (ESD) robustness, a chip designer must follow certain guidelines concerning size and placement of diode and transistor clamps between different power-supply buses as well as between I/Os and supply lines. These guidelines may typically be provided by technology design rules which include minimum transistor width, optimal contact-to-gate spacing (CGS), and examples for placement and hook-up of the various protection circuits. If all of the ESD design rules are followed, the circuit designer presumes that some minimal ESD requirement will be met, typically a human-body model (HBM) withstand voltage of 2000V. However, until actual silicon is packaged and tested, the designer usually does not know what HBM voltage the product will withstand or what quantitative changes must be made in protection-circuit layout parameters to reach a certain level of ESD robustness. The aim of this chapter is to provide circuit designers with a methodology enabling the design of ESD circuitry which meets a product's specific reliability needs. Provided a quantitative model, or layout rules based on this model, a circuit designer can create the optimal design for a given area and have a good idea of how robust the design will be.

As discussed in Chapter 2, numerous papers have analyzed the effectiveness of transmission-line pulsing (TLP) measurements in characterizing the ESD response of CMOS processes and circuits [21,23]. The dependence of MOS snapback I-V characteristics on layout parameters, addressed in Section 2.4, is well known [8]. While layout optimization for ESD circuits has been investigated [65,66], only recently has work been presented on a methodology which uses TLP measurements to quantitatively predict the HBM withstand voltage of any protection transistor for a given technology or to

optimize transistor layout for maximum HBM and/or charged-device model (CDM) robustness, minimum clamping voltage, and minimum area [67]. Such work is of interest because NMOS bipolar snapback will continue to be an effective ESD protection mechanism in future technologies [68].

This chapter explores the use of empirical modeling of ESD protection-transistor performance to optimize transistor layout and quantify the trade-offs in layout parameters. As an example of these trade-offs, suppose that the ESD robustness of a previously designed multiple-finger NMOS clamp must be increased, but there is only limited area for expansion. A designer may choose to either add another poly finger to increase the total transistor width or to increase the contact-to-gate spacing of the existing fingers, thereby presumably increasing the robustness per unit width. It is not obvious which option will yield the greater ESD withstand level, but accurate characterization of a large design space over all critical layout parameters will lead directly to this answer. Chapters 3 and 4 demonstrated how electrothermal simulation is used to study the dependence of ESD robustness on layout parameters, and other work has been published on this application of two-dimensional [24,32] and even three-dimensional [69] simulation. However, in all of these studies the simulations have been of simple circuit elements such as single-finger transistors or diodes rather than of multifinger transistors, mainly because of the greatly increased computation time and resources required for simulating large devices. Therefore, while numerical simulation offers much understanding of the ESD response of individual transistors, empirical modeling of an adequate layout design space may be the best approach to characterizing and optimizing multifingered ESD circuits.

In the next section, an ESD-circuit design methodology is presented by reviewing the TLP characterization of ESD test structures, investigating the correlation between TLP withstand current and HBM withstand voltage, developing second-order linear models of protection-transistor performance, and discussing the importance of identifying critical ESD current paths in an integrated circuit. To verify the methodology, a model is extracted from characterization of a 0.35μm CMOS process and its predicted responses are compared to experimental HBM withstand levels of SRAM protection circuits. These results are analyzed, and optimization of circuit layout is discussed. Conclusions are drawn regarding the effectiveness of the methodology and how it may be enhanced in the future.

5.1 Methodology

Section 2.5 presented general concepts of ESD design methodology, including the procedures for testing single-finger transistors, extracting critical I-V parameters from this testing, and optimizing layout of transistors for use in multifinger protection circuits. A simple, theoretical design example was given in Section 4.4 to demonstrate the application of these ideas. Some of these topics will be readdressed in the following subsections, but they will be expanded upon to form a broader design methodology based on design-of-experiments empirical modeling.

5.1.1 Characterization of Test Structures

Fig. 5.54 shows the transient I-V response, or snapback curve, of a single-finger NMOS ESD protection transistor generated by applying 150ns transmission-line pulses to the drain of the transistor with the source, substrate, and gate grounded (the gate is usually soft-tied to ground through a resistor). This experimental curve is qualitatively similar to the theoretical curve of Fig. 2.6. Critical I-V design parameters extracted from the curve

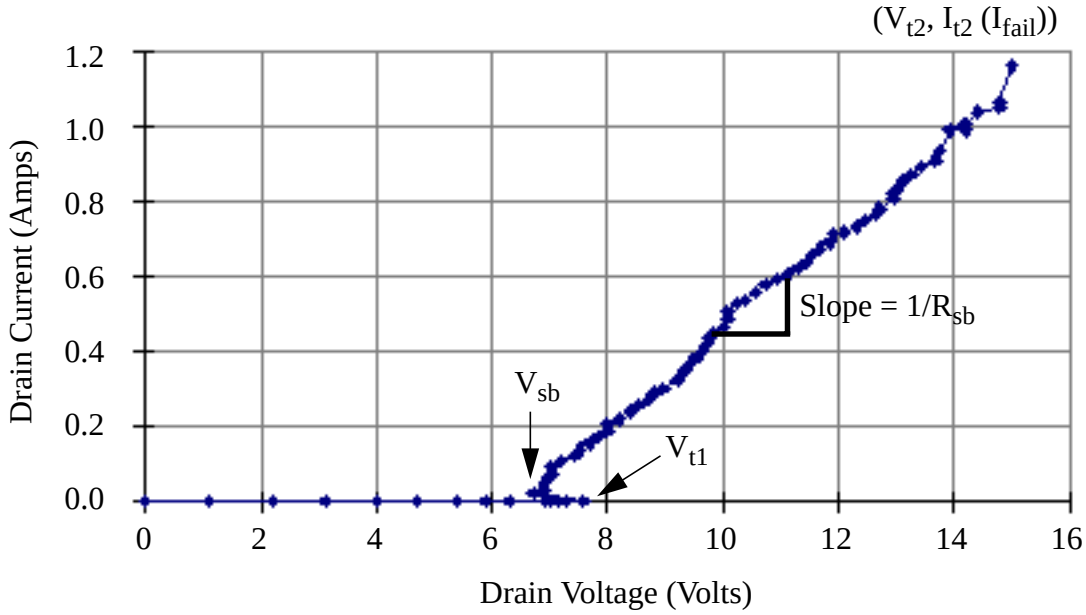

Fig. 5.54 Snapback I-V curve for a 50/0.6μm NMOS transistor generated by TLP. Critical I-V parameters are the trigger voltage (V_{t1}), snapback voltage (V_{sb}), snapback resistance (R_{sb}), and thermal-runaway or second-breakdown point (V_{t2}, I_{t2} (I_{fail})).

are the trigger voltage (V_{t1}), snapback voltage (V_{sb}), snapback resistance (R_{sb}), and second-breakdown (thermal-runaway) point (V_{t2}, I_{t2}). For TLP widths on the order of 100ns, device failure usually follows instantaneously when the second-breakdown point is reached, in which case I_{t2} is equivalent to the failure current, I_{fail}. Failure is defined as 1mA of leakage current when the drain is biased at the technology supply voltage, V_{CC}. Tracking the I-V response of a structure is just as important as determining the failure current because dielectric failure at an input gate oxide will occur if a protection circuit's clamping voltage becomes too high.

Section 2.2 described in detail the equivalent circuit of the TLP setup, the equipment used to monitor the voltage, current, and leakage of the device under test (DUT), and the automated software used to extract the TLP I-V curve of the DUT. For the testing discussed in this chapter, the step size of the transmission-line charging voltage is set to yield current increments of about 30mA per step. In addition to characterizing structures with TLP, test structures are also stressed with HBM pulses using an Oryx Model 700 manual ESD tester. As with TLP, the drain is subjected to pulses with the source, substrate, and gate grounded, but in this case three positive and three negative pulses are applied at each voltage level to parallel the procedure of circuit-qualification HBM testing. The HBM withstand voltage (the maximum HBM voltage a structure can withstand without incurring microamp leakage) is obtained by step stressing the structure in 50-volt increments until the device fails. These 50-volt increments are equivalent to about 33mA increments in peak pulse current since the HBM pulse is discharged through a 1500Ω resistor. Further comparison of the TLP and HBM test methods will be made in the next subsection. To verify that step stressing does not introduce stress-induced hardening, i.e., an artificial increase in withstand voltage due to a burn-in type phenomenon, some structures were also stressed at a single voltage around the failure point determined by the step stressing. Results showed no effect of previous stresses on the failure level of a structure.

To characterize a process, TLP and HBM tests are run on a set of test structures with varying layout parameters, contained on dedicated tiles of a test chip. An example of a multiple-finger test structure is shown in Fig. 5.55 and defines the critical layout parameters: poly finger width (W), gate length (L), drain and source contact-to-gate spacing (DGS and SGS), and number of poly fingers. As discussed in Section 2.4, in fully silicided processes varying CGS has little effect on ESD performance since the silicide

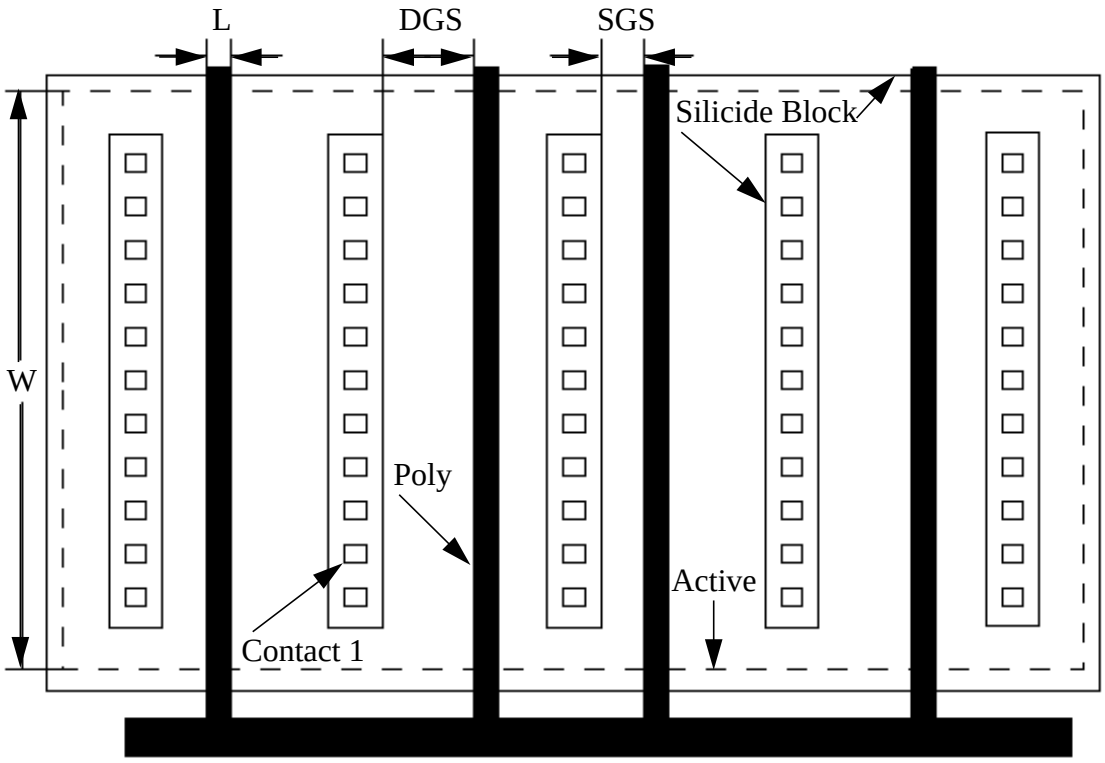

Fig. 5.55 *Layout of a four-fingered ESD structure showing finger width (W), gate length (L), and source (SGS) and drain (DGS) contact-to-gate spacing (actually silicide-to-gate spacing).*

reduces the source/drain resistivity to only a few ohms per square. However, in the CMOS process analyzed here the ESD protection transistors make use of a silicide-blocking technology to maintain a high value of source/drain resistivity which provides design flexibility of the ballast resistance (snapback resistance). Several TLP and HBM tests are run for each structure by testing different die on a wafer or number of wafers. Examples of the dependence of TLP and HBM withstand levels on layout parameters will be given in the next subsection.

5.1.2 Correlation of TLP to the Human Body Model

Transmission-line pulsing provides much insight into device behavior during an ESD event. Actual circuits, however, must pass qualification using the HBM method of testing. In order for TLP to provide useful design-related models, the results of TLP must be correlated to the results of HBM. Although the HBM stress event is characterized by a

certain charging voltage, V_{HBM}, the 1500Ω series resistor of the circuit is usually much larger than the impedance of the device under test, so we can think of both TLP and HBM testers as current sources, with the peak HBM current equal to $V_{HBM}/1500\Omega$. For the Advanced Micro Devices (AMD) 0.35μm technology studied in this chapter, we know from failure analysis that HBM and TLP failures are thermal rather than dielectric in nature. An identical failure mechanism leads us to believe that there may be some TLP pulse width for which the withstand current, $I_{TLP,ws}$, of any structure is equal to the peak current of an HBM pulse at the withstand level of that structure. Note that from this point on the TLP failure current, I_{t2} or I_{fail}, is assumed to be only infinitesimally larger than the withstand current (the maximum TLP current a structure can withstand without incurring damage), so all terms are used interchangeably.

HBM and TLP current waveforms and the equivalent circuits used to generate them were presented in Chapter 2. As one extreme for comparing the HBM withstand voltage, $V_{HBM,ws}$, to $I_{TLP,ws}$, we assume that some total energy is required to create device failure, independent of waveform. This assumes adiabatic thermal boundary conditions, i.e., a hot spot leading to second breakdown which occurs in the device before any generated heat diffuses from the region of heating. In this case, the energy required for failure is

$$E_{fail} = \int_0^\infty I^2(t)\, R_{DUT}\, dt \tag{5.42}$$

where I(t) is the stress current and R_{DUT} is the resistance of the device under test. For a TLP stress, the current is constant for the duration of the pulse, so

$$E_{fail}^{TLP} = I_{TLP}^2 R_{DUT} t_{TLP} \tag{5.43}$$

where t_{TLP} is the width of the pulse.

In the case of the ideal HBM pulse, if we assume that $R_{DUT} << 1500\Omega$, then

$$I_{HBM}(t) = I_{pk} \exp\left(-t / (R_{HBM} C_{HBM})\right) \tag{5.44}$$

where $R_{HBM} = 1500\Omega$ and $C_{HBM} = 100pF$ for an ideal HBM pulse and $I_{pk} = V_{HBM}/R_{HBM}$ is the peak current of an HBM pulse charged to V_{HBM}. Eq. (5.44) neglects the rise of the HBM pulse, which takes less than 10ns, and takes t = 0 to be the time at which the pulse

reaches its peak. This is justified because less than 4% of the pulse energy is contained in the time before the pulse reaches its peak value. Substituting Eq. (5.44) into Eq. (5.42),

$$E_{fail}^{HBM} = I_{pk}^2 R_{DUT} \frac{R_{HBM} C_{HBM}}{2} . \tag{5.45}$$

Equating Eq. (5.45) to Eq. (5.43), we see that for equivalent energies the TLP pulse width must be 75ns for the same peak current ($I_{TLP} = I_{pk} = V_{HBM}/R_{HBM}$).

To determine the validity of the assumed adiabatic boundary conditions, we need to reexamine the three-dimensional thermal-failure model presented in Section 2.2.2. Recall that in this "thermal-box" model for an MOS transistor a uniform Joule heating due to a constant-current stress is assumed to occur in a rectangular parallelepiped whose dimensions are defined by the transistor width, the drain junction depth, and, roughly, the gate length. Failure is assumed to occur when the peak temperature at the center of the box reaches a critical value. The ballast resistances of the non-silicided source and drain regions create additional potential drops and heat sources which affect the boundary conditions. Nonetheless, we still expect the model to serve as a first-order description of device failure.

Using this model the power to failure (P_f) is calculated vs. stress time (t_f), with four regions of the P_f vs. t_f curve bounded by three time constants which are determined by the box dimensions (Fig. 2.12). Each time constant is defined as

$$t_i = i^2 / (4\pi D) \tag{5.46}$$

where D is the thermal diffusivity and i takes on specific values of a, b, or c, which for our technology are assumed to be 50μm for the transistor width (a), 0.5μm for the gate length (b), and 0.2μm for the junction depth (c). Using D = 0.13cm^2/s (based on the calculations from [23]), these result in values of t_a = 15μs, t_b = 1.5ns, and t_c = 0.24ns.

The model allows us to determine that the power to failure, normalized by the transistor width (P_f / a), is inversely proportional to stress time for times less than t_c (Eq. (2.6)). Since the product of the power to failure and the time to failure is constant in this region, a constant energy is needed to induce failure, i.e., this is the adiabatic region. The time

constant of t_c = 0.24ns is much less than the ~100ns stress time of the TLP and HBM testing, so the constant-energy-to-failure assumption is clearly invalid.

The model further predicts that the width-normalized power to failure (P_f / a) is inversely proportional to the square root of the pulse duration for times between t_c and t_b (Eq. (2.7)) and inversely proportional to the log of the pulse duration for $t_b < t < t_a$ (Eq. (2.8)). For stress times greater than t_a, P_f approaches a constant value (Eq. (2.9)). Given our technology dimensions, power to failure for the TLP and HBM stressing is expected to be described by the inverse logarithmic dependence of Eq. (2.8).

This model focuses on power to failure rather than current to failure (I_f), which is the actual parameter of interest. However, these are related by

$$I_f = \sqrt{P_f / R_{DUT}} \ . \tag{5.47}$$

From Eqs. (2.8) and (5.47), the TLP withstand current should be inversely proportional to the square root of the logarithm of the stress time in the time range of interest. While a 150ns transmission-line pulse of height 707mA delivers the same energy as a 75ns pulse of height 1A (a difference in current of 29%), Eqs. (2.8) and (5.47) predict that the current to failure is only 6% lower for the 150ns pulse than for the 75ns pulse. Therefore, while the TLP pulse width is important, the withstand current is not critically dependent on the pulse width over a difference range of 50%.

Although the HBM stress is not a constant-current pulse, we can assume that the thermal-box model describes the first-order dependence between transistor dimensions and peak current in a damage-inducing HBM pulse. By comparing $V_{HBM,ws}$/1500Ω with $I_{TLP,ws}$ for various TLP widths for a set of test structures, a TLP width which best correlates $I_{TLP,ws}$ to $V_{HBM,ws}$ can be determined. Fig. 5.56 plots $V_{HBM,ws}$/1500Ω and $I_{TLP,ws}$ for 75, 100, and 150ns pulse widths vs. DGS (2.2μm SGS) for 50/0.6μm single-finger NMOS structures in the AMD 0.35μm CMOS process. The withstand level increases with DGS since there is more area for dissipation of heat, but there are diminishing returns for DGS above about 6μm. Note that the withstand levels are average values of a number of experiments and are normalized by the total structure width (finger width times the number of fingers), yielding units of mA/μm. Error bars represent the 95% confidence interval of a set of measurements as calculated by the student-t distribution. In Fig. 5.57,

the same withstand currents are plotted vs. the number of 50/0.6µm fingers (4.4µm DGS, 2.2µm SGS) for various multiple-finger NMOS transistors. In this case the normalized withstand level decreases as the number of fingers increases. The flow of heat away from a finger is reduced by heating in adjacent fingers due to the reduced temperature gradient, thus leading to thermal runaway at a lower normalized current level for a multiple-finger circuit.

As seen in Fig. 5.56 and Fig. 5.57, for the standard single-finger structure (50/0.6µm with 4.4µm DGS), shorter TLP pulse widths lead to higher withstand currents, with a range greater than 30%. However, for larger DGS and for the multiple-finger structures, this difference decreases and in many cases the difference is less than the range of the error bars. In both figures the HBM results are seen to follow the same trend as the TLP results, but there is no TLP width for which correlation of $I_{TLP,ws}$ to $V_{HBM,ws}$ is clearly superior. This is somewhat expected since the theoretical difference in withstand currents of 6% is

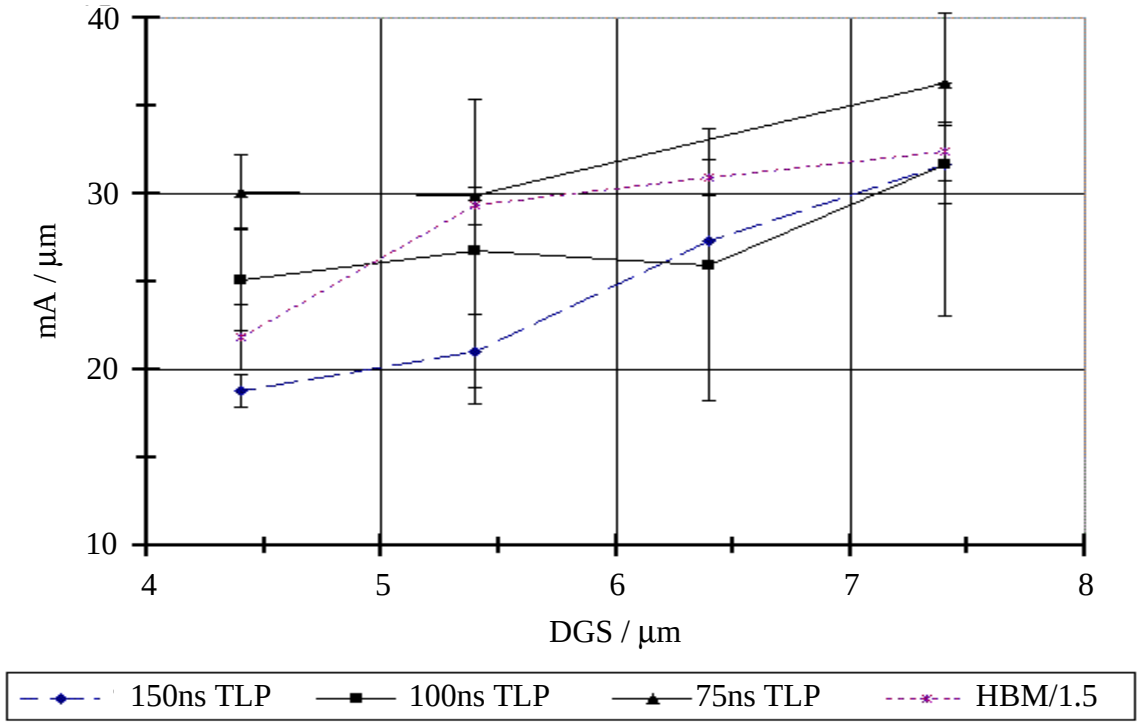

Fig. 5.56 Normalized (divided by width) withstand current vs. drain-side CGS for HBM stressing and 75, 100, and 150ns TLP stressing of 50/0.6µm single-finger transistors. For HBM, the withstand voltage is converted to mA by dividing by 1.5. Error bars represent 95% confidence intervals.

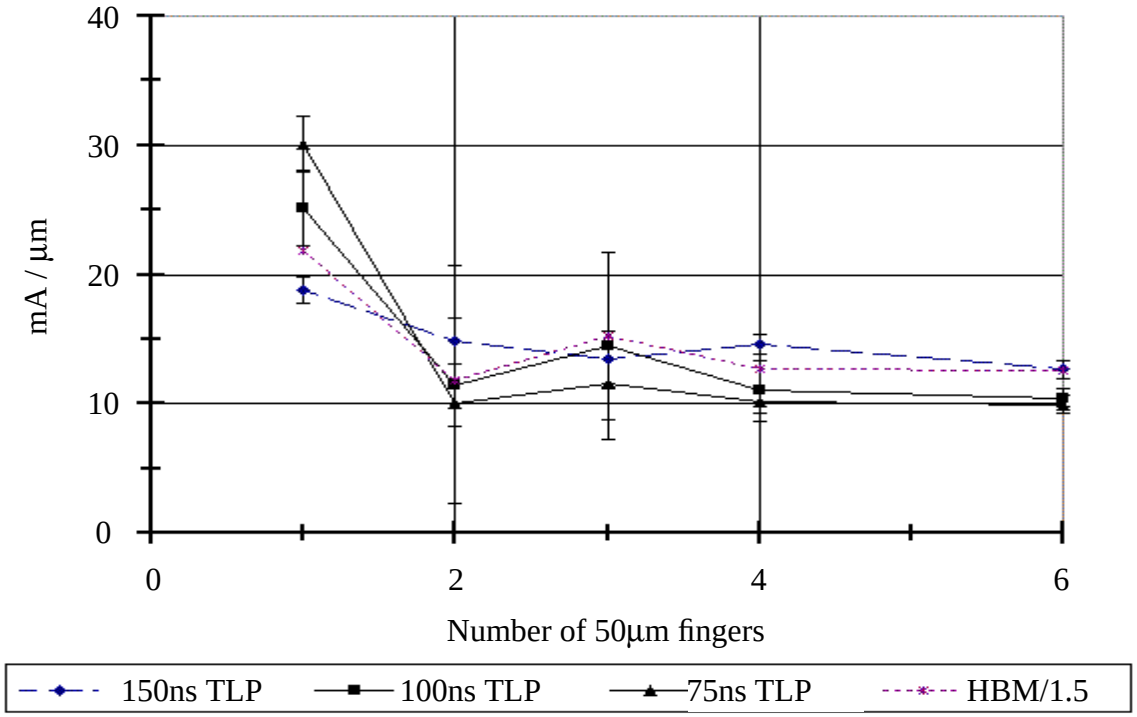

Fig. 5.57 Normalized (divided by width) withstand current vs. number of 50/0.6µm fingers for HBM stressing and 75, 100, and 150ns TLP stressing. Error bars represent 95% confidence intervals.

often less than the experimental range of values for a given pulse width. As a result, 150ns pulses were chosen for characterization of all test structures in the design space since initial turn-on of a structure and inductance in the test setup lead to noise in the first 30ns of a pulse which makes capture of the average voltage and current waveform heights difficult for pulse lengths less than 100ns.

5.1.3 Development of Second-Order Linear Model

A design example based on data from one-dimensional layout variations was already presented in Section 4.4. Ideally, by extracting the TLP I-V and $V_{HBM,ws}$ values from the proper layout-parameter design space, a model can be created which predicts the I-V response and failure level of any protection circuit exhibiting layout parameters within the design space. This concept is implemented with BBN/Catalyst^{TM} design-of-experiments software [70] which, using experimental data, creates linear, second-order models relating

various responses ($V_{HBM,ws}$ and TLP I-V parameters) to a number of factors (layout parameters). Catalyst uses the data to determine optimal constant, linear, quadratic, and two-factor-interaction model coefficients for each response (Fig. 5.58). It provides standard-deviation and residual information to help the user discard ineffective model terms and bad data points. Once a model is developed, a simple graphical interface allows the user to study the effects of varying one or more layout factors (an example is given in Section 5.3) or to create an optimal layout design.

Our model is based on failure occurring within the protection device, which assumes that the protection device turns on quickly enough and clamps to a voltage which is low enough to prevent damage to internal circuitry. Although turn-on time is not characterized, the clamping voltage is easily calculated (see Section 4.4) as

$$V_{device}(I_{device}) = V_{sb} + R_{sb} \cdot I_{device} \tag{5.48}$$

$$R = a_0 + a_1F_1 + a_{11}F_1^2 + a_2F_2 + a_{22}F_2^2 + a_3F_3 + a_{33}F_3^2 + a_{12}F_1F_2 + a_{13}F_1F_3 + a_{23}F_2F_3$$

R = Response

F_i = Factor

a_i, a_j = Model Coefficient

$a_{ij}F_iF_j$ = Model Term

Fig. 5.58 Example of a complete second-order linear equation modeling the response of a variable with three factors.

which is a maximum when $I_{device} = I_{t2}$. If V_{device} of an input pull-down protection device exceeds the dielectric breakdown voltage of a gate oxide before I_{device} reaches I_{t2}, rupture of the gate oxide is expected to occur rather than or in addition to thermal failure of the protection transistor. By including V_{sb} and R_{sb} in the model, the clamping voltage of any circuit is easily monitored.

Another assumption of the model is that all fingers of a multiple-finger circuit participate in current conduction. Since our test structures use only a simple gate-to-source series gate-bounce resistor instead of a more complex gate-bounce scheme [41], in the worst case fingers of a multiple-finger circuit turn on one at a time, with successive fingers triggering into bipolar snapback each time the device voltage reaches V_{t1}. All of the fingers will not turn on before thermal failure unless

$$V_{sb} + I_{t2}'R_{sb}' > V_{t1} \tag{5.49}$$

where the primed values indicate single-finger values. Again, the model is used to predict these values (indirectly, for I_{t2}' and R_{sb}').

The critical part of generating a model is determining the set of factors which have the greatest influence on the targeted responses, which in this case are the trigger voltage, snapback voltage, snapback resistance, $I_{TLP,ws}$, and $V_{HBM,ws}$. Selection of the layout factors should be based on physical reasoning--given the large number of fitting parameters it is easy to create a model which fits all the data yet makes little physical sense. For example, since the snapback resistance is the dynamic series resistance of a structure operating in the snapback mode, it should be inversely proportional to the total structure width and directly proportional to the sum of the source and drain CGS. Thus, the model equation has the form

$$R_{sb} = A_0\left(\frac{1}{Wn}\right) + A_1\left(\frac{DGS}{Wn}\right) + A_2\left(\frac{SGS}{Wn}\right) \tag{5.50}$$

where W is the finger width, n is the number of fingers, and the A_i are the model coefficients (the first term accounts for the resistance of the intrinsic transistor). Note that in the snapback regime significant current still flows from drain to substrate (about 30% according to numerical simulations), but since this parallel resistance is much larger than the resistance of the intrinsic device Eq. (5.50) should be accurate. The layout factors

needed to describe R_{sb} in a linear equation are DGS, SGS, 1/W, and 1/n. However, since only two-factor interactions are represented in the model, a total-width factor, 1/(Wn), must be included as a factor so that it may interact with DGS and SGS in the second and third terms of Eq. (5.50). (An alternative would be to define $W \cdot R_{sb}$ or $W \cdot n \cdot R_{sb}$ as the response.) It is likely that not all layout factors will be needed for all responses. For example, V_{t1} and V_{sb} should have a very weak dependence on DGS and SGS since there is very little potential drop at the low currents from which these responses are extracted. Any of the model terms are easily turned off for any of the responses in the Catalyst program. Model equations for other responses will be discussed in the next section.

Since either $I_{TLP,ws}$ or $V_{HBM,ws}$ data may be used to generate the withstand-voltage model, we should consider which set of data is more valid or which will lead to more accurate modeling. The main issue concerns the differences between the manual HBM tester used to characterize the test structures and the large, automated testers (Verifier) used to qualify circuits in the reliability laboratory. Even though both HBM testers meet rise time, decay time, and ringing specifications for a short-circuit load (MIL STD 883C/ 3015.7), differences in parasitic elements between different HBM testers lead to different withstand voltages for a given device [71]. Specifically, a capacitance in parallel with the DUT due to the test board, C_{TB}, will initially charge to a voltage of V_{t1} (refer to Fig. 5.54) and then partially discharge into the device when the device snaps back. Assuming a constant V_{t1}-V_{sb} difference, smaller structures will be more susceptible to early failure due to this capacitive discharge. Values of C_{TB} extracted from pulse waveforms and SPICE simulations are 32pF for the Oryx manual tester and 20pF for the automated Verifier tester. The large C_{TB} of the manual tester is expected to affect the small test structures and may explain why in Fig. 3.38 the HBM withstand value is lower than the 100ns and 75ns TLP withstand values for the single-finger structure but is more in line with the TLP values for multiple-finger structures.

Although large test structures and the large protection circuits which are the target of the modeling are less susceptible to tester parasitics, artificially low HBM withstand levels of small structures are still a concern since they will skew the model. Therefore, $I_{TLP,ws}$ values will be used to create the models for HBM failure of IC protection circuits. The models will predict $I_{TLP,ws}$ for a circuit, and this value will be multiplied by 1500Ω to arrive at the predicted $V_{HBM,ws}$.

One final modeling issue to consider is that since average values of withstand current or voltage are used to develop the ESD circuit model, the model predicts the average HBM withstand voltage of an actual protection circuit in an IC. However, when an IC is subjected to the reliability qualification process, a limited number of parts are tested at one or more voltages for various pin combinations, and the withstand voltage is taken to be the highest stress voltage for which *all* of the sample parts pass. Furthermore, multiple pins are tested on each part, and even if only one pin fails the part is considered to have failed the test. Therefore, we expect our model's predicted withstand levels to be higher than the qualification withstand voltage because there will likely be a spread in the sample data. It may be possible, through error analysis, to predict the deviation in performance of an IC protection transistor based on the measured deviations of the test-structure design space. In any case, it is necessary to account for the difference between the average withstand voltage predicted by the model and the minimum withstand voltage determined through product qualification.

5.1.4 Identification of Critical Current Paths

Predicting the ESD failure level of an IC presumes knowledge of the discharge current path, so it is important to identify all potential paths between any pair of stressed pins. Fig. 5.59 shows the critical pull-up, pull-down, and supply-clamp circuits in an IC with internal, external, and clock power supplies. For input-only pads, ESD protection is provided by adding a "dummy" CMOS output buffer on the pad to form the pull-up and pull-down circuits, with the gate of each circuit soft-tied to its respective source. For output-only or bi-directional I/O pads, the large output driver doubles as the ESD protection circuitry, with extra "dummy" poly fingers added in parallel if necessary.

In some cases of ESD stress, such as negative voltage on an I/O or V_{CC} pad with respect to V_{SS} or positive voltage on an I/O pad with respect to V_{CC}, the current path is just a forward diode drop across the large drain-substrate junction of a protection circuit. For the opposite stress polarities, however, the current path contains transistors operating in snapback mode and/or diodes in reverse-breakdown mode. Since HBM (or CDM) stressing of both polarities is performed on a given test and forward-biased diodes are found to be very robust in our technology, the focus of the modeling is on bipolar snapback.

The actual path or paths taken during an HBM stress between two pins depends on the trigger and clamping voltages of the various protection circuits, i.e., the parameters which are determined by the model described in the previous subsection. Characterization of PMOS protection transistors in the AMD 0.35μm technology has shown that due to very low gain of the parasitic lateral pnp transistor, V_{sb} is equivalent to the drain-substrate breakdown voltage, i.e., the PMOS transistor does not snap back. Therefore we know that during a negative I/O vs. V_{CC} stress, for example, the discharge path in Fig. 5.59 is through the drain-substrate diode of the pull-down (a) and the parasitic bipolar transistor of the supply clamp (d), not through the drain-well diode or parasitic bipolar of the pull-up (b). Because the sum of the pull-down diode drop (0.7V) and the voltage drop across the supply clamp (~7V) is less than the breakdown or snapback voltage of the pull-up (~10V), damage of the pull-up will not occur. Since the PMOS pull-up structures are not found to break down during any type of ESD stress, only NMOS test structures are examined in this work.

As a final consideration, we must ensure that all I/O-pad and supply-clamp design rules are followed in an IC if the circuit is to have predictive ESD behavior. For example, if guard rings are not used to isolate the pad diffusions from the internal diffusions, substrate current could be diverted to an internal device, thereby circumventing the protection

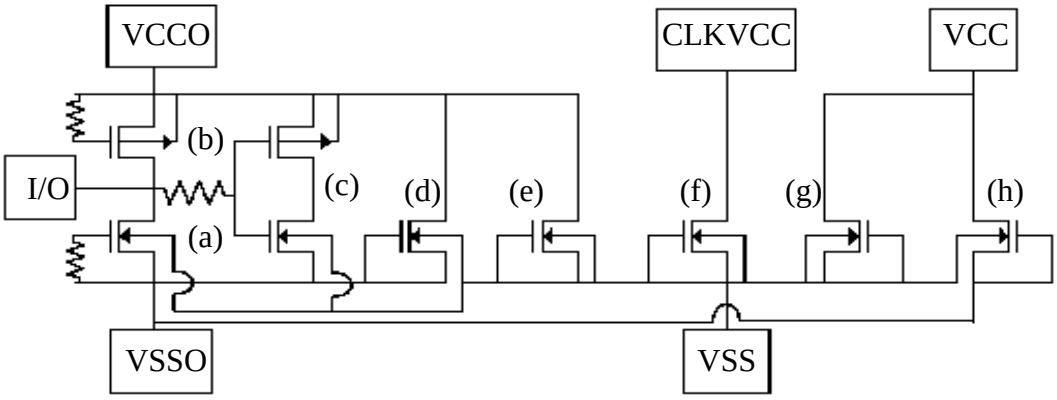

Fig. 5.59 Schematic of critical ESD protection circuits in a chip with split power supplies (VCCO/VSSO and VCC/VSS) and separate clock supply (CLKVCC): (a) n-channel pull-down, (b) p-channel pull-up, (c) CMOS pair representing internal circuitry, (d)-(h) n-channel clamps between various supplies (clamps for VCC-VCCO, VCC-CLKVCC, and VCCO-CLKVCC not shown).

circuit. This would lead to an unpredictable, low-voltage failure to which our modeling cannot be applied.

5.2 Application

NMOS ESD test structures were laid out and characterized using TLP and HBM testing for an AMD 0.35μm CMOS process. The design space covers finger widths between 25 and 150μm, DGS between 4.4 and 7.4μm, and SGS between 2.2 and 4.2μm for single-finger structures and multiple-finger structures with two to six fingers. In order to keep the number of test structures in the design space relatively small, gate length was not used as a factor in this study. The total design space, comprised of 18 structures, is not optimal because layout was not performed with empirical modeling in mind. Catalyst requires 20 structures in order to calculate model coefficients for all linear, quadratic, and interaction terms for four factors. However, since not all possible model terms are needed to describe the responses, our design space is adequate. The responses for which model equations are derived are V_{sb}, R_{sb}, $I_{HBM,ws}$, and $V_{HBM,ws}$. The trigger voltage, V_{t1}, is not modeled because it is mainly dependent on gate length and gate-bounce resistance, parameters which are not varied.

Model terms for each response are chosen based on physical reasoning and observed single-factor dependencies. Examining the snapback voltage first, note that since V_{sb} is the voltage required to sustain parasitic bipolar operation, it should be the sum of the BV_{CEO} of the intrinsic device and the ohmic drops in the source and drain diffusions. The intrinsic device size is a constant in the design space since gate length is not varied, and therefore

$$V_{sb} = a_0 + a_1(DGS) + a_2(SGS). \tag{5.51}$$

The snapback resistance should always be proportional to the total device width, assuming all fingers are conducting. Thus, the R_{sb} response is normalized by the total width and Eq. (5.50) is rewritten as

$$R_{sb} \cdot (Wn) = b_0 + b_1(DGS) + b_2(SGS). \tag{5.52}$$

To determine how to best describe the layout dependence of the withstand current and voltage using a second-order linear model, single-factor trends are examined for DGS, SGS, n, and W. In Fig. 5.56, the normalized $V_{HBM,ws}$ vs. DGS line has a negative curvature, indicating that the $I_{TLP,ws}$ and $V_{HBM,ws}$ model equations should have quadratic as well as linear DGS terms, with the quadratic terms being negative. A quadratic dependence on SGS is also observed, but over the limited range of the design space (2.2 to 4.2μm) a linear term is adequate. As seen in Fig. 5.57, the normalized failure parameters have an inverse dependence on the number of fingers, and consequently these parameters are not well described using linear and quadratic n (number-of-finger) factors. However, if 1/n is chosen as the factor, a good fit is obtained with just a linear term. Since the normalized $I_{TLP,ws}$ and $V_{HBM,ws}$ also have an inverse dependence on width, 1/W is chosen as a factor, but in this case the best fit is obtained by also including a quadratic term. Finally, we assume that SGS does not interact with any of the factors since its value does not vary widely, but the three interaction terms between DGS, 1/n, and 1/W are included. The resulting withstand-current model is

$$\frac{I_{TLP,ws}}{(Wn)} = c_0 + c_1(SGS) + c_2(DGS) + c_3(DGS)^2 + c_4(1/n) + c_5(1/W) + c_6(1/W)^2 + c_7(DGS)(1/n) + c_8(DGS)(1/W) + c_9(1/n)(1/W) \quad (5.53)$$

with an identical equation (with different coefficient values) for $V_{HBM,ws}$. Note that the constant coefficient, c_0, lumps together the constant terms from the separate factor dependencies.

Model coefficients for Eqs. (5.51)-(5.53) were extracted using Catalyst for two development lots with slightly different process recipes. HBM and 150ns TLP characterization of the design space was performed on two wafers per lot and five die sites per wafer, with average response values of each structure used as the Catalyst input. SRAM test circuits from the same wafers were submitted to the AMD Reliability Laboratory for HBM stressing of I/O vs. V_{SS}, I/O vs. V_{CC}, and V_{CC} vs. V_{SS} pin combinations to determine average, i.e., not qualification, HBM withstand voltages.

Results for the two lots are summarized in Table 5.2. For each lot, the layout parameters of each stressed circuit were plugged into the $I_{TLP,ws}$ model equation to determine the $mA/\mu m$ values in Table 5.2. These values were then converted to $V_{HBM,ws}$ values by

Table 5.2 Experimental and modeled SRAM HBM withstand voltages.

Pin Combination	Full I/O vs. V_{SS}	Input vs. V_{SS}	I/O vs. V_{CC}	V_{CC} vs. V_{SS}
Circuit Stressed	1/2 Pull Down	Pull Down	Clamp	Clamp
W X n (μm)	36.2 X 5	36.2 X 10	71 X 5	71 X 5
DGS/SGS (μm)	4.2/2.2	4.2/2.2	4.2/4.2	4.2/4.2
Lot 1				
model mA/μm	19.0	13.9	10.4	10.4
model $V_{HBM,ws}$	5200	7550	5500	5500
exptl. $V_{HBM,ws}$	5200	7500	5400	>10,00
Lot 2				
model mA/μm	19.1	15.1	13.7	13.7
model $V_{HBM,ws}$	5200	8200	7300	7300
exptl. $V_{HBM,ws}$	5400	8000	4600	>10,00

multiplying by the total circuit width and by 1500Ω. The different stress combinations and the model predictions and the SRAM testing, with the exception of I/O vs. V_{CC} testing of the corresponding protection circuits involved will be discussed in the next section, as will the generally slightly higher withstand levels seen in Lot 2 for SRAM HBM testing and for TLP characterization throughout the design space. Good agreement is seen between Lot 2 and V_{CC} vs. V_{SS} testing of both lots. These discrepancies will also be discussed in the next section.

5.3 Analysis

5.3.1 Model Terms

Before further discussion of the SRAM predictive modeling, we will examine the Catalyst model terms in more detail. Fig. 5.60 is the model-graph window generated by Catalyst for Lot 1, which graphically displays the dependence of each response on the four layout factors. Qualitatively similar trends are seen for Lot 2. As a factor changes from its low value to its high value, it affects each response as indicated by the corresponding trend line. In all graphs the error bars reflect typical experimental variations of the responses as determined from the input data. Notice that for V_{sb} and $R_{sb} \cdot (Wn)$ the 1/n and 1/W lines

are flat, a direct result of the independence of these terms on width and number of fingers as dictated by Eqs. (5.51) and (5.52). As expected, V_{sb} and R_{sb} increase linearly with SGS and DGS. However, R_{sb} has a stronger dependence on DGS than on SGS, which may reflect the fact that all stress current flows through the drain but then is split between source and substrate paths. The snapback voltage appears to have a greater dependence on SGS than on DGS, but the large error bars indicate that this difference is within experimental error.

In the withstand current plots, the quadratic model terms for DGS and 1/W result in curved response lines (the negative $I_{TLP,ws}$ vs. DGS curvature agrees with the HBM withstand data in Fig. 5.56), while the interaction terms between DGS, 1/n, and 1/W result in a pair of lines for each of these responses. For each factor the response curve is drawn for the most positive and most negative influence the factor can have on the response as determined by its interaction with other terms. As expected, in all cases $I_{TLP,ws}$ increases as 1/n and 1/W increase. However, for some values of 1/n and 1/W, the model predicts that $I_{TLP,ws}$ will decrease to negative values for large DGS. Although it cannot be directly seen

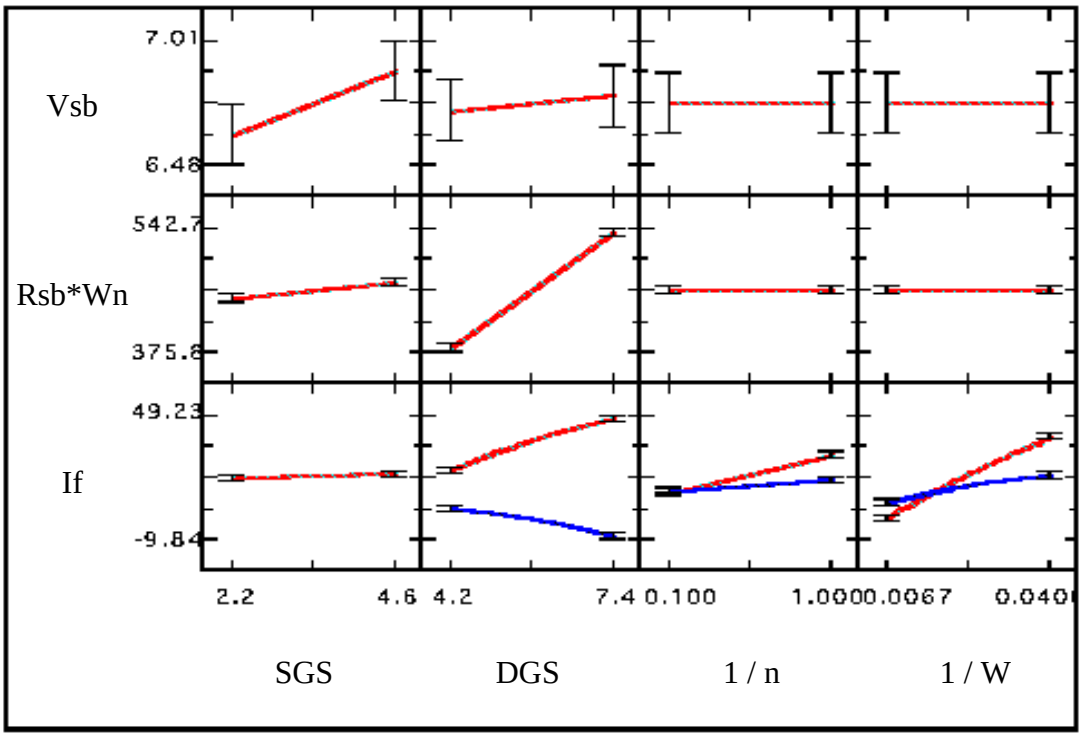

Fig. 5.60 Catalyst model graph for Lot 1 V_{sb}, R_{sb} (multiplied by structure width), and normalized $I_{TLP,ws}$ (If) as a function of SGS, DGS, 1/n, and 1/W.

from Fig. 5.60, the condition for which the model predicts $I_{TLP,ws} < 0$ for large DGS is $1/W < 0.013\mu m^{-1}$ ($W > 76\mu m$). This nonphysical aspect of the model is a result of having to extrapolate beyond the design space, which does not cover the large DGS-large W corner, and could be corrected by expanding the design space to this corner. Fortunately, the largest DGS of any of the SRAM protection circuits is 4.2μm, so the model predictions for the circuits of interest are accurate.

5.3.2 SRAM Model Prediction

As mentioned previously, HBM withstand levels of an IC cannot be predicted unless the stress current paths are known. The SRAM test circuit used for this study has only one V_{CC} and one V_{SS} supply, which simplifies the ESD analysis. For reasons discussed in Section 5.1.4, I/O vs. V_{SS} failures are expected to occur in the NMOS pull-down circuit, while I/O vs. V_{CC} and V_{CC} vs. V_{SS} failures are expected to occur in the V_{CC}-V_{SS} supply-clamp circuit (refer to Fig. 5.59). The observed failure mode for I/O vs. V_{SS} SRAM testing is pin leakage to V_{SS}, while the failure mode for I/O vs. V_{CC} and V_{CC} vs. V_{SS} is increased stand-by current. These failures indicate damage to pull-down and supply-clamp circuits, respectively, confirming the expected failure mechanisms. Emission microscopy was also attempted for failure analysis but no emission sites were seen due to the metal busing over the pull-down and clamp circuits.

Although the pull-down protection circuits of bi-directional (“Full I/O” in Table 5.2) and input-only (“Input”) I/O pins have the same layout parameters, separate HBM stressing of each type of I/O results in higher withstand voltages for the input-only pins. For the input-only pull-down circuits, all 10 gate fingers are tied to a dummy inverter which provides the needed gate bounce to reduce the trigger voltage. For the bi-directional I/Os, however, half of the gate fingers are tied to a dummy pre-driver while the other half are driven by internal circuitry, i.e., they drive the output. Since the two pre-drivers are of different size and thus offer different degrees of gate bounce, we hypothesize that only half of the fingers are turning on due to different trigger voltages, which would explain why the bi-directional I/Os are less robust than the input-only I/Os. For modeling purposes, then, an n value of 10 is used for the input-only stress while a value of 5 is used for the bi-directional I/Os. (Actually, an n value of 1 is used in determining the *normalized* $V_{HBM,ws}$ because in the layout every other finger is tied to the same pre-driver and thus the five fingers are

assumed to be isolated from each other. The final $V_{HBM,ws}$ value is still determined by multiplying by the total width of the five fingers.)

As a result of the different number of fingers used in the model, Table 5.2 shows that the predicted normalized withstand level is different for the full-I/O and input-only circuits even though they have the same finger width and contact-to-gate spacing. Using the proper parameters, the difference between modeled and experimental $V_{HBM,ws}$ values is less than 5% for I/O vs. V_{SS} testing. Note that the model predicts accurate values for the 10-finger device even though this requires extrapolation beyond the design-space limit of six fingers.

A negative-voltage stress on an I/O with respect to V_{CC} will turn on a supply-clamp circuit in the same manner as a positive-voltage stress to V_{CC} with respect to V_{SS} because in the former case the I/O is connected to V_{SS} through the forward-biased drain-substrate diode of the pull-down circuit. However, in Table 5.2 we see that while the withstand voltage of the I/O vs. V_{CC} stress for each lot is within reasonable range of the predicted value, $V_{HBM,ws}$ for the V_{CC} vs. V_{SS} stressing is above the testing limit of 10,000V. Since there are multiple supply clamps laid out at various points along the pad ring of the SRAM circuit, it appears that during V_{CC} vs. V_{SS} stress two or more clamps turn on and act in parallel to dissipate the ESD current. Based on model calculations for the snapback voltage (6.8V for Lot 1, 7.0V for Lot 2) and snapback resistance (1.1Ω, 0.73Ω) of one clamp circuit with all fingers conducting, the second-breakdown voltage (V_{t2}, see Fig. 5.54 and Eq. (5.48)) is 10.8V for Lot 1 and 10.5V for Lot 2. These values are very close to the expected trigger voltage of the clamp circuit, and thus it is reasonable to expect a second clamp to turn on before the first clamp fails.

Turning to the I/O vs. V_{CC} results in Table 5.2, consider that while the experimental $V_{HBM,ws}$ is very close to the model prediction for Lot 1, it is much lower than predicted for Lot 2 and is indeed lower than the Lot 1 experimental value even though the modeling predicts higher performance for Lot 2. This result should make us suspicious of whether the clamp circuit is operating as predicted in Lot 2 SRAMs. Although the snapback voltage for the clamp circuit predicted by the model is about 6.9V for both lots, a lower source/drain diffusion resistance in Lot 2 leads to a lower snapback resistance, with the model predicting 5.5Ω per finger for Lot 1 and 3.7Ω per finger for Lot 2. Thus, one possible explanation for the unexpectedly low experimental value of $V_{HBM,ws}$ in Lot 2 is

that the reduced ballasting effect due to lower R_{sb} prevents all fingers from turning on during the ESD event, resulting in less current-handling capability and reduced withstand voltage. This seems contradictory to the argument just made for the power-supply stressing in which it was determined that the ballasting is good enough in both lots to turn on fingers of multiple clamps. However, the extra diode drop from the I/O pad to V_{SS} in the I/O vs. V_{CC} stress may reduce the rise time of the HBM pulse enough to hinder triggering of the clamp fingers. This is not an issue in the case of V_{CC} vs. V_{SS} stress because there is no diode in the path.

Finally, note that although the modeled mA/μm values for Full I/O vs. V_{SS} stress in Table 5.2 are nearly identical for the two lots, increasing the number of fingers (Input vs. V_{SS}) or finger width (I/O vs. V_{CC} and V_{CC} vs. V_{SS}) more strongly reduces the mA/μm in Lot 1 than in Lot 2 (neglecting the effect of increased SGS for the clamp circuit). This means that the slopes of the $I_{TLP,ws}$ vs. 1/n and $I_{TLP,ws}$ vs. 1/W lines (Fig. 5.60) are steeper for Lot 1 than for Lot 2. Physically, since the source/drain resistance (R_{sb}) is 33% lower in Lot 2 than in Lot 1, less total heat is generated in Lot 2 protection transistors for a given stress current. Thus, the reduced thermal gradient due to increased W or n (discussed in Section 5.1.2) has less of an effect on Lot 2 than on Lot 1, resulting in mA/μm values which are 9% and 32% higher for Lot 2 for the pull-down and clamp circuits, respectively. The mA/μm values are very close for the 1/2-pull-down circuits because heat dissipation is not critical for the five nearly isolated fingers.

5.4 Optimization

Up to this point, the modeling and analysis of ESD circuits has focused on how the protection level of a transistor depends on critical layout parameters. However, in the context of laying out ESD protection for an actual integrated circuit, other factors come into consideration. For example, in a pad-limited circuit layout there is a limited area available for protection circuitry. In the case of an RF circuit, for which speed is critical, the drain-substrate capacitance (C_{DB}) of the I/O buffer needs to be minimized. Fortunately, the factors in our model provide the layout information necessary for calculating the source/drain diffusion area as well as the area and perimeter components of C_{DB}. Thus, the Catalyst modeling can be used to optimize I/O buffer layout for minimum area, minimum capacitance, and maximum ESD withstand level.

Qualitatively, we know from Fig. 5.56 and Fig. 5.60 that as DGS increases, the normalized withstand current increases. Of course, transistor area and C_{DB} also increase, but since the normalized $V_{HBM,ws}$ increases, less total width is required for a certain withstand level. In a similar manner, increasing the number of poly fingers requires lower W values to achieve the same $V_{HBM,ws}$, and if the increase in normalized $V_{HBM,ws}$ for lower W values more than offsets the decrease in normalized $V_{HBM,ws}$ for higher n, less total area will be required for the larger-n transistor.

To study these effects quantitatively, different values of DGS and n were set in the Catalyst model for Lot 1 and W was adjusted to yield a $V_{HBM,ws}$ of 5000V. A lower limit of six was set for the number of fingers since using fewer fingers would require a W much larger than 50µm, which we deem undesirable. An upper limit of 6.2µm was placed on DGS since the data shows that $V_{HBM,ws}$ saturates around this value and thus further increase of DGS would only serve to increase area and capacitance. SGS was held constant at 2.2µm.

Total source/drain diffusion area and C_{DB} were calculated in each case for the minimum W required for 5000V HBM. Calculations for the diffusion area, plotted in Fig. 5.61, show that in the region of interest a reduction in area is always achieved by increasing DGS and/or the number of fingers. Values of W range from 46µm for 4.2µm DGS and six fingers to 7.7µm for 6.2µm DGS and 10 fingers (the model boundaries were expanded to extrapolate $I_{TLP,ws}$ for $W < 25\mu m$). Fig. 5.61 shows diminishing returns for area reduction as the number of fingers is increased, especially for large values of DGS. Although C_{DB} has a perimeter dependence as well as an area dependence, its dependence on layout is very similar to that of the area (including the diminishing returns), with values ranging from 1.4pF for 4.2µm DGS and six fingers to 0.56pF for 6.2µm DGS and 10 fingers. This example illustrates that optimization of layout results in a 60% reduction in area and C_{DB} from the worst-case design.

Other elements can also be considered during optimization. For example, gate delay may be an issue for an RF circuit in which non-silicided, relatively resistive poly gates are used on I/O circuits. In such a case an upper limit on finger width would need to be imposed, and this is easily accomplished in Catalyst by specifying the range of values for the width factor during the model definition phase. Also, each response can be assigned a target value or designated as “larger is better” (e.g., $I_{TLP,ws}$) or “smaller is better” (e.g., V_{sb}).

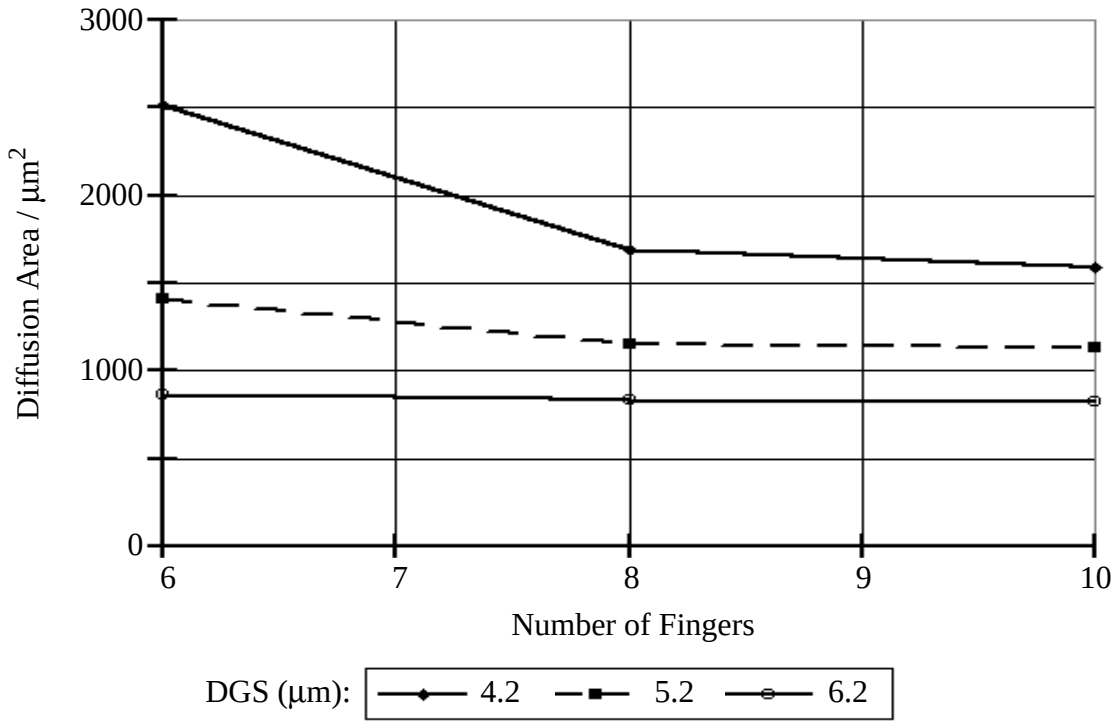

Fig. 5.61 Calculated minimum area of transistor source/drain diffusion needed for 5kV HBM protection for various DGS and number of fingers for Lot 1 with a SGS of 2.2µm.

After calculation of the models Catalyst will run an optimizing routine that attempts to determine a set of factor values which will result in all responses meeting their targets. The program will flag any condition (set of factors) for which a response exceeds specification. This feature could prove useful if a model were added for CDM withstand voltage and a circuit needed to be optimized for CDM as well as HBM performance.

5.5 Summary of Design Methodology

The methodology for the design of CMOS ESD protection circuits is effectively summarized in block-diagram form in Fig. 5.62. First, a design space is defined and test structures with varying layout dimensions are laid out for a given technology. Critical I-V parameters and withstand currents are extracted through automated transmission-line pulse characterization. These results are input along with the layout parameters to a software program which generates empirical, second-order linear models relating HBM

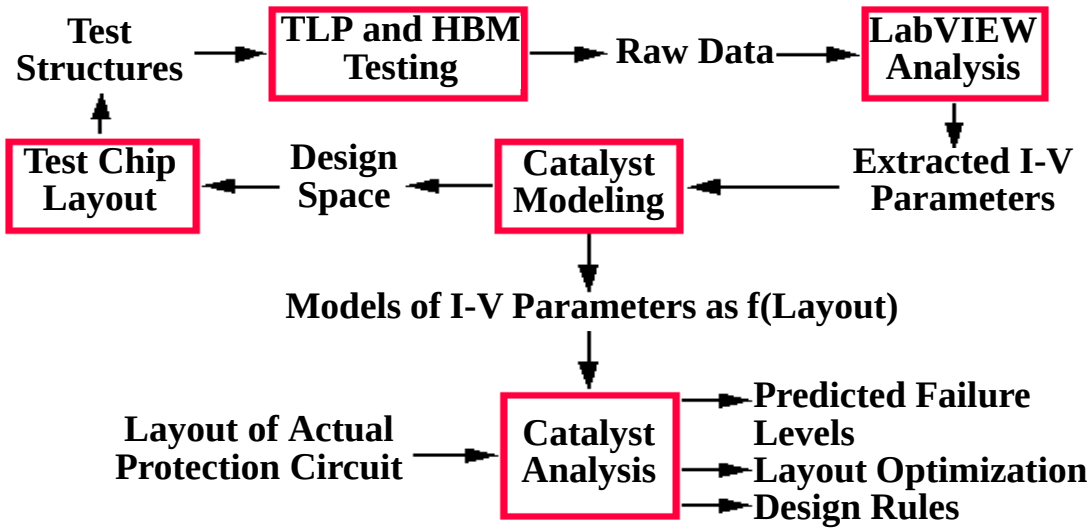

Fig. 5.62 Block diagram of ESD circuit design methodology.

withstand voltage and TLP I-V parameters to circuit layout. As discussed in Section 5.1.2, a key requirement for the implementation of this modeling is good correlation between TLP withstand current and HBM withstand voltage. Experimental and mathematical analysis demonstrated that such a correlation is achievable over at least a limited range of widths and contact-to-gate spacings. Once models have been generated for a technology, they are applied to actual ESD protection circuits to predict HBM performance and optimize circuit design. Note that analysis of the extracted I-V parameters in the Catalyst modeling program may reveal critical regions of the design space, thereby creating a feedback loop in the design-of-experiments process.

Chapter 6
Conclusion

In the integrated-circuit industry, the ceaseless effort to decrease critical transistor dimensions in each new technology guarantees that the prominence of electrostatic-discharge will continue to grow. Devising ways to protect smaller transistors against ESD is just as important as determining how to process and manufacture them because a product with a high susceptibility to damage will not be widely accepted. As a result of its gradually increasing visibility over the last two decades, the problem of ESD is now dealt with by most IC manufacturers on several levels, from designing on-chip protection circuits to properly grounding the furniture and equipment in a fabrication facility to educating all personnel involved with wafer and package handling to minimize the potential for failure. Once an IC is packaged and shipped to a customer, however, the built-in protection circuit is the only means of defense against ESD damage. While circuit designers have successfully created robust ESD protection for past technologies, a lack of understanding of the mechanisms underlying ESD damage limited the amount of transferrable knowledge from one technology to the next.

With continually decreasing technology cycles, which are now less than two years in length, and the probable change in the prominent ESD failure mode from HBM-type damage to CDM-type damage in deep submicron technologies, ESD circuit designers will no longer have time to start designs from scratch or follow a trial-and-error design approach. Characterization and design methodologies, based on an understanding of the failure mechanisms behind ESD and models which accurately describe these mechanisms, must be implemented so that the critical features of a protection circuit can be determined and applied to future technologies. This chapter reviews the contributions of this thesis toward implementing such a methodology and proposes future work to be done in the area of ESD circuit characterization, modeling, and design.

6.1 Contributions

An overview of electrostatic discharge issues in the integrated-circuit industry was constructed to elicit appreciation of the importance of addressing ESD in process development and circuit design. The phenomenon of ESD was defined and its implications to ICs were reviewed. ESD failures fall into three main categories: thermal damage, dielectric damage, and latent failure. Three widely accepted methods used to characterize ESD sensitivity in ICs are the human-body model, machine model, and charged-device model tests. Each of these models represents a potential real-world ESD event, but it was shown that the models offer little insight to the functionality and weaknesses of an ESD protection circuit and thus that a better characterization scheme is desirable. Examples of common ESD protection circuits and the theory behind their design was presented. A review of previous applications of numerical device simulation to the study of ESD illustrated how simulation can be used to design and analyze protection circuits and highlighted previously untried simulation methods. A basic protection-circuit design methodology was outlined and exemplified using results from the transmission-line pulsing characterization method and two-dimensional simulations. This was followed by the description of a more complete design methodology based on empirical models extracted from a fully characterized test-structure design space.

6.1.1 Transmission Line Pulsing

The transmission-line pulsing test method, a relatively new ESD-circuit characterization scheme, was presented. This test method is superior to the classic characterization models because it reveals how a protection circuit functions during an ESD stress and quantifies the failure threshold of a circuit over a wide range of stress times. TLP captures the transient I-V curve of a stressed device by sampling each current level so briefly that damage is not incurred. Using TLP, the evolution of leakage current, which is a measure of the degree of damage, is monitored by measuring the device leakage after each pulse. This feature aids the determination of critical points at which various types of damage are created and is especially important in capturing low-level (sub-microamp) leakage which is a signature of latent failure. The basic setup of a TLP characterization system was detailed along with an overview of some advanced setup techniques.

TLP was shown to be a powerful tool for extracting the critical I-V parameters of ESD test structures fabricated in a leading-edge CMOS technology. A discussion was given on the dependence of these critical I-V parameters on process and layout parameters. Testing focused on structures with varying widths and contact-to-gate spacings, and power to failure and current to failure were measured between 50ns and 600ns. The usefulness of the extracted I-V parameters and failure levels was demonstrated in the application of the ESD design methodology to SRAM circuits.

6.1.2 Numerical Device Simulation

Lattice-temperature modeling in 2D numerical device simulation and the temperature-dependent models required for proper modeling of high-temperature effects associated with ESD were reviewed. New simulation methods were presented, including a general-purpose curve-tracing algorithm, developed and implemented as a C program, which guides a simulator through complex I-V curves. The curve tracer's application to ESD was demonstrated in the control of dc snapback simulations. More general applications of the curve tracer and a user's manual are presented as an appendix. A quantitative analysis was conducted to compare and contrast the 2D and 3D formulations of an analytic thermal model which, to first order, describes the heating of a device during an ESD event. The results of the analysis predict that for stress times in the ESD and EOS regimes, the power to failure modeled in two dimensions will be higher than that of the three-dimensional model or of an actual device. This directly conflicts the conclusions reached in previous studies of electrothermal simulation that 2D simulations underestimate the power to failure. Methods for studying dielectric failure and latent damage with 2D simulation were proposed, including monitoring of hot-carrier injection and hot-spot spreading during an ESD simulation.

A procedure for calibrating simulation models for use in quantitative ESD simulations was delineated, including structure definition and determination of mobility and impact-ionization model coefficients and thermal boundary conditions. I-V and failure characteristics of standard test structures were used as the basis of the calibration. While quantitative modeling of the snapback I-V parameters was achieved, modeling of thermal failure was inadequate due to unresolved issues regarding modeling of the electric field at high current levels in the drain junction region, where the device physics are most critical and most complex. Usefulness of the ESD snapback simulations was nonetheless

demonstrated in the proposed protection-circuit design example. One benefit of the shortcomings of the high-current calibration is the identification of critical obstacles to ESD simulation which can be scrutinized in the future.

6.1.3 Design Methodology

The primary goal of the design methodology is to reduce the design time of ESD protection circuitry by providing quantitative design rules for each process technology. A quantitative model provides IC designers more confidence and flexibility in their ESD protection designs and should reduce the number of design cycles. Aspects of the methodology were presented in detail, including characterization of a test-structure design space; correlation of TLP and HBM failure levels; development of empirical, second-order linear models; and identification of critical ESD current paths.

To verify the methodology, the modeling was successfully applied to explain HBM failures in a 0.35μm CMOS technology. Models were generated from test-structure characterization of two lots with slightly different processing and applied to ESD protection transistors on SRAM circuits from each lot. In general, HBM withstand voltages predicted by the modeling agreed well with experimentally determined levels. In each case for which modeling and experiment differed, analysis of the model-generated circuit I-V parameters suggested that the protection circuit does not function as intended during HBM stress, thereby yielding the different experimental result. Optimization capabilities of the modeling were also examined, demonstrating how optimal design can significantly reduce layout area and input capacitance.

6.2 Future Work

Although new work was presented on specific aspects of ESD such as transmission-line pulsing and 2D electrothermal simulation, all of the topics addressed in this thesis fit one or more of the general categories of characterization, modeling, and design of ESD protection circuits. Thus, future work will be discussed in each of these areas.

6.2.1 Characterization

While the effectiveness of the transmission-line pulsing method was clearly demonstrated, there are unresolved issues regarding this test method which need to be addressed. Since

most ESD qualification procedures in the IC industry are based on human-body model and charged-device model testing, and since the HBM and CDM do represent potential ESD hazards, a complete correlation needs to be drawn between the failure threshold determined by TLP and the thresholds determined by HBM and CDM. TLP is used to examine device failure over a broad time spectrum, and it was demonstrated both theoretically and with a limited number of experiments that a certain pulse width can be identified which yields a failure current consistent with the HBM failure level. If it can be proven that TLP testing predicts the susceptibility of a device to the human-body model over a wide range of circuit designs, TLP should become a more widely accepted test method.

Correlation between TLP and the machine model and charged-device model would be similarly useful. The dependence of the MM and CDM waveforms on circuit parasitics and the very short rise time of the CDM makes such correlation difficult, although some work has already been done on correlation to CDM [72]. On the other hand, transmission-line pulsing is inherently capable of measuring device failure thresholds at stress times associated with EOS. Overall, if agreement can be demonstrated between transmission-line pulsing failures and failures induced by other ESD and EOS testing methods as well as actual field failures, TLP could become part of the qualification process for IC technologies.

In the future, measurement of the turn-on time of a protection circuit will become more important because if a circuit cannot respond to the sub-nanosecond rise time of the charged-device model, the input voltage could easily exceed the dielectric breakdown voltage of the input gate oxide during a CDM stress. In the current TLP setup, the rise time of the pulse at the input of the device under test (DUT) is about 3ns, and noise in the circuit prevents accurate measurement of current and voltage for times less than 40ns. There is room for improvement of the high-frequency characteristics of the TLP setup: connections can be shortened between DUT pins and coaxial or SMA connectors and the inductance can be reduced between the end of the transmission line and the test jig (the rise time of the pulse at the edge of the transmission line is less than 1ns). If the circuit noise can be sufficiently reduced, the effects of certain parameters on the turn-on time, such as gate bounce resistors and substrate resistance, can be fully studied. Improving the quality of sub-50ns measurements will also facilitate extraction of more complete power-to-failure vs. time-to-failure curves which in turn will allow extraction of the thermal-box model parameters.

6.2.2 Modeling

As shown by the results of Chapter 4, simulations may actually provide a more useful method for studying ESD-circuit turn-on time because good agreement between simulated and measured low-current snapback parameters was demonstrated. Of greater concern is the ability to simulate the high-current portion of the MOSFET snapback curve and the onset of thermal failure. It was found that some of the assumptions of the calibration procedure were incorrect. Calibration of mobility and impact ionization using only standard room-temperature MOSFET characteristics is not adequate for simulation of ESD phenomena above the point of snapback. One procedure which was not attempted was the calibration of MOSFET characteristics at higher temperatures. Even if data and simulations are only examined up to 250^{o}C, proper calibration will aid the prevention of the exaggerated increase in snapback resistance observed in present simulations. It may also be worthwhile to measure the temperature-dependent thermal resistance and capacitance of the silicon material to ensure the corresponding simulator models are accurate. Regardless, the most critical issue which must be addressed is the effect of simulation grid on the electric field profile, which was shown to be the main obstruction of proper high-current impact-ionization modeling.

Limitations of 2D device simulation also need to be further quantified. Although the difference between 2D and 3D thermal models was studied, the implications of this study remain unclear due to the incomplete thermal-failure calibration and the deviation of the boundary conditions in a real MOSFET structure from the assumptions of the model. Another concern for future simulations is the validity of the assumption that the electron and hole temperatures are in thermal equilibrium with the lattice. As discussed in Chapter 3, as electric fields increase due to smaller device dimensions and greater stress, hot carrier effects will become more important. During extremely brief, high-field ESD events such as CDM stress, carriers may no longer be in equilibrium with the lattice and full two-carrier-plus-lattice-temperature modeling, such as offered by PISCES-2ET (dual energy transport model), will be needed. Such modeling would require calibration of different mobility and impact ionization models which are dependent on carrier temperature.

Another type of modeling which was not studied in this thesis is compact modeling, i.e., circuit-level or SPICE-level modeling. For ESD simulation, compact modeling is especially useful for determining current paths in circuits subjected to ESD stress.

Significant work has already been done to create compact models for MOSFET snapback and thermal failure [73-75]. Although thermal modeling is best implemented by enhancing the source code of a circuit simulator, parasitic bipolar action, i.e., snapback, can be modeled by adding existing lumped-element bipolar transistor and current generator models in a simulator such as HSPICE. Such modeling is probably adequate for the study of charged-device model stressing: CDM failures are usually dielectric rather than thermal in nature, so failure can be studied by monitoring the voltage across the gate oxides in the simulated circuit.

6.2.3 Design

One obvious way to improve the ESD circuit design methodology presented in Chapter 5 is to increase the range and number of variables in the design space. For the next AMD technology, 0.25μm CMOS, a more complete ESD transistor design space has already been laid out, with gate length included as one of the variables. Gate length is a factor to which CDM robustness may be especially sensitive. One of the shortcomings of the current implementation of the methodology is that the design space is not optimized and not all corners of the space are covered, resulting in nonphysical values of withstand current for the combination of large drain-to-gate spacing and large width. For the 0.25μm technology the design space has been laid out with model extraction in mind by using the Catalyst software's design-of-experiment capability.

Currently, the methodology is undergoing further verification by applying the modeling to protection circuits of other AMD CMOS logic products in the 0.35μm technology. One important product category is RF (high frequency) circuitry, in which I/O capacitance must be kept to a relatively low value in order to meet operating specifications. As demonstrated in Section 5.4, the design methodology allows for optimization under the constraint of a maximum allowable transistor area, i.e., maximum allowable capacitance. Additionally, the I/O gate delay of an RF circuit must not be too large. This translates to a constraint on maximum width of the poly gate fingers, which again can be accounted for during design optimization.

Future plans include expanding the methodology to study special I/O circuits such as those used in ICs with separate internal and external power supplies and in ICs which are "5-volt tolerant." In the former case, the substrate of an I/O pull-down transistor is tied to

the internal VSS supply while the source is tied to the external VSS supply in order to reduce substrate noise. The isolation of the source from the substrate results in different ESD behavior since the discharge current path is altered. In the latter case, a cascoded gate (also called a stacked gate or split gate) pull-down transistor is used at the I/Os because the circuit, although designed to operate using a 3.3V supply, must be able to tolerate a 5V signal on the I/Os in order to meet older circuit-board specifications (a standard pull-down transistor cannot be used in this case because 5V could develop across the transistor gate, which is only designed to withstand a 3.3V signal). Stacking two gates in series affects the ESD response because the snapback voltage and snapback resistance are effectively doubled.

In addition to applying the design methodology to different types of protection circuits, determining the feasibility of modeling CDM withstand voltages is also important because CDM is now the dominant ESD concern in the IC industry. Since CDM stress usually leads to dielectric damage of gate oxides, a different type of test structure may be required. For example, by connecting the input of an inverter circuit to the drain of an NMOS pull-down protection transistor we can determine how effectively the transistor would protect the input gates of an actual integrated circuit during CDM stress. Test structures might also be bonded into different types of packages to model the dependence of CDM robustness on the inductance and resistance of package leads.

An important aspect of the methodology presented in this thesis is that a simple, empirical approach is taken to model ESD protection circuits. However, in the future we would like to integrate two-dimensional electrothermal device simulation and circuit simulation into the process to confirm the trends predicted by the empirical models. In doing so we may find that a more complex model is needed, i.e., something beyond second-order linear equations, in which case a more advanced modeling software package would be required.

Appendix A
Tracer User's Manual

Stephen G. Beebe, Zhiping Yu, Ronald J.G. Goossens, and Robert W. Dutton

In Technology CAD, the use of software to simulate the testing of semiconductor devices is known as virtual instrumentation. A virtual instrument should be able to automatically generate simulation data, e.g., I-V points along a bias sweep, given only the simple specifications a user would input to a real programmable instrument testing a real IC device. Numerical device simulators such as PISCES-2ET provide a means of creating virtual devices and simulating electrical tests on the devices. However, these simulators cannot trace through I-V curves with sharp turns unless the user carefully controls the bias conditions near these turns--a tedious and time-consuming process. This deficiency prompted the creation of Tracer.

Tracer is a C program which automatically guides PISCES and other semiconductor device simulators through complex I-V traces and is ideally suited for device-failure phenomena such as latchup, BV_{CEO}, and electrostatic-discharge protection. Given a PISCES input deck and a specification file with a PISCES-like syntax, a simulation can be run over any current or voltage range without user intervention. Tracer is limited to dc, one-dimensional traces, i.e., only one electrode can be swept per run. It sweeps this electrode by dynamically setting the most stable bias condition at each solution point. Additionally, Tracer has the ability to maintain zero-current bias conditions at one or two electrodes during the trace, even at low device-current levels where such bias conditions are unstable using traditional device simulation. The theory implemented in the Tracer program was introduced in Chapter 3; a complete discussion is given in [28].

A.1 Command Line

Usage: `tracer inputfile tracefile` [`outputfile`]

- `inputfile` is the name of the PISCES input deck which defines the device structure to be simulated and specifies what physical models are to be used. Basically, it contains everything in a normal PISCES deck except the solve card specifications (Section A.7).
- `tracefile` is a file containing instructions on how to conduct the trace as well as specifications for bias conditions on all electrodes (Sections A.2 through A.6).
- `outputfile` is an optional specification of the name of the file where the simulation data is to be written (Section A.8). If `outputfile` is not given, the name of the output file defaults to `inputfile`.out.

A.2 Trace File

The trace specification file, `tracefile`, is similar to a PISCES or SUPREM input deck. Each line begins with a word designating what type of statement, or "card," it is. The four possibilities are **CONTROL**, **FIXED**, **OPTION**, and **SOLVE**. Also, a line may start with a "$" for comments. Such lines are ignored. The cards may appear in any order, and a card may be continued on following lines by placing a "+" at the beginning of each subsequent line. The "+" should be separated from the parameters on the line by at least one space.

Each option in a card should have the following structure: "param = paramvalue". Spaces separating the "=" sign are optional. The parameters for each card are described in the following four sections. As with PISCES syntax, parameter names and values are not case-sensitive and may be abbreviated provided they remain unambiguous. Square brackets, [], enclosing a parameter indicate that it is optional (note that some of these parameters are only optional in the sense that they will default to a certain value if not specified in `tracefile`). A vertical line, |, represents a logical OR--only one of a list of parameters separated by "|" signs can be specified.

All electrodes in the device must have representation in the `tracefile`. Each electrode must appear as one, and only one, of the following: the **CONTROL** electrode, a **FIXED** electrode, or an open contact (**OPENCONT1** or **OPENCONT2**) on the **SOLVE** card.

A.3 CONTROL Card

A.3.1 Description

The **CONTROL** card is used to designate the electrode which will be swept through the trace as well as the boundaries of the trace. This electrode is referred to as the control electrode. To define the start of the simulation range, an initial voltage and an initial voltage step must be specified for the control electrode. The end of the trace is specified by either a maximum electrode voltage, a maximum electrode current, or the total number of simulated points to be found.

A.3.2 Syntax

NUM=<int> **CONTROL=**<char> [**BEGIN**=<real>] [**INITSTEP**=<real>] [**ENDVAL**=<real> | **STEPS**=<int>]

A.3.3 Parameters

- **NUM** is the number of the electrode in the PISCES deck designated as the control electrode, whose voltage or current is swept through the trace. Its integer value must be between 1 and 9, inclusive. Default: none.
- **CONTROL** is either **VMAX**, **IMAX**, or **STEP**. **VMAX** denotes that a maximum voltage on the control electrode, specified by **ENDVAL**, is used as the upper bound on the trace. **IMAX** denotes that **ENDVAL** specifies a maximum control-electrode current for the trace. **STEP** signifies that the trace will proceed for a certain number of simulation points, specified by the **STEPS** parameter. In most cases **VMAX** or **IMAX** will be used because it is not known how many simulation steps it will take to reach a certain voltage or current. Default: none.
- **BEGIN** is the value of the voltage, in volts, at the starting point of the curve trace for the electrode designated by **NUM** (the control electrode). If an initial solution is performed by Tracer, **BEGIN** should be 0.0. If a previous solution is loaded into the input deck at the start of Tracer (see **SOLVE** card below), **BEGIN** should be equal to the voltage of the control electrode in this solution. Default: 0.0V.

- **INITSTEP** is the initial voltage increment, in volts, of the control electrode. Thus, at the second solution point the control electrode will have a voltage of **BEGIN** + **INITSTEP**. A recommended initial step size is 0.1V. The sign of **INITSTEP** determines the direction in which the curve trace will initially proceed. If **INITSTEP** proves to be too large and PISCES cannot converge on the second solution point, Tracer will automatically reduce **INITSTEP** until convergence is attained, then proceed with the trace from this point. Default: 0.1V.
- **ENDVAL** is used when **CONTROL**=**VMAX** or **IMAX**. Tracer stops tracing when the voltage (**CONTROL**=**VMAX**) or current (**CONTROL**=**IMAX**) of the specified electrode equals or exceeds the value specified by **ENDVAL**. Note that it is the absolute values of the voltage or current and of **ENDVAL** which are compared. Default: 10.0V (**CONTROL**=**VMAX**), 10.0A/μm (**CONTROL**=**IMAX**).
- **STEPS** is used when **CONTROL**=**STEP**. It specifies the number of solution points Tracer should find. Default: 10.

A.3.4 Examples

1. Electrode 3 is the control electrode. Tracer will initially proceed in the negative-voltage direction with an initial step of -0.1V. Tracer will proceed until the absolute value of the control current equals or exceeds 3A/μm.

```
control num=3 begin = 0.0 initstep=-0.1 control=IMAX end=-3.0
```

2. Electrode 4 is the control electrode. Tracer will run until 65 solutions are found, starting at v4=0.0V with an initial v4 step of 0.5V.

```
control num=4 begin=0.0 initstep=0.5 control=step steps = 65
```

A.4 FIXED Card

A.4.1 Description

A **FIXED** card is used to designate an electrode whose bias remains fixed throughout the simulation. There should always be at least one **FIXED** electrode and usually there are two or more. The two types of bias conditions available are voltage sources and current sources. The value of the bias is arbitrary, with one exception: a zero-current source (open contact) should be specified through the open-contact option on the **SOLVE** card and not on the **FIXED** card. If non-zero current sources are used for some electrodes in a simulation, in `inputfile` the user must create contact cards with the "current" option for each of these electrodes (see Section A.7).

A.4.2 Syntax

NUM=<int> [**TYPE**=<char>] [**VALUE**=<real>] [**RECORD**=<char>]

A.4.3 Parameters

- **NUM** is the number of an electrode in the PISCES deck. Its integer value must be between 1 and 9, inclusive. Default: none.
- **TYPE** is either **VOLTAGE** or **V** for a voltage source or **CURRENT** or **I** for a current source. Default: **VOLTAGE**.
- **VALUE** is the fixed value of the current or voltage for the electrode specified by **NUM**. **VALUE** has units of either volts or amps/μm, depending on the specification of **TYPE**. Note that the specification of **VALUE** is optional since it is merely for reference and is not used by `Tracer`. Default: 0.0.
- **RECORD** is either **YES** or **NO**. For **RECORD**=**YES**, the simulated current is recorded in the output file for a fixed-voltage electrode, while the simulated voltage is recorded for a fixed-current electrode. Default: **NO**.

A.4.4 Examples

1. In every `Tracer` solution, electrode 1 has a voltage of 0.0V. The current in this node is recorded in `outputfile` at every solution point.

fixed num=1 type = voltage value=0.0 record =yes

A.5 OPTION Card

A.5.1 Description

An **OPTION** card is used to specify convergence criteria and solution-method options for any open electrodes, parameters which affect the smoothness and step-size control of the trace, which PISCES solution files are saved, and whether extra solution data is saved in `outputfile`.

A.5.2 Syntax

Simulations with one or two open contacts:

> [**ABSMAX**=<real>] [**RELMAX**=<real>] [**DAMP**=<real>]
> [**TRYCBC**=<real>]

Smoothness and step-size control:

> [**ANGLE1**=<real>] [**ANGLE2**=<real> [**ANGLE3**=<real>]
> [**ITLIM**=<int>] [**MINCUR**=<real>] [**MINDL**=<real>]

Control of output files:

> [**FREQUENCY**=<int>] [**TURNINGPOINTS**=<char>]
> [**VERBOSE**=<char>]

A.5.3 Parameters

- **ABSMAX** is the maximum current allowed in an open contact and is only relevant when open contacts are used and voltage biases are applied to these contacts. Convergence is satisfied when either the **ABSMAX** or **RELMAX** condition is met. Default: $1.0X10^{-19}$A/μm.
- **RELMAX** is the maximum ratio of open-contact current to control-electrode current and is only relevant when open contacts are used and voltage biases are applied to these contacts. Convergence is satisfied when either the **ABSMAX** or **RELMAX** condition is met. Default: $1.0X10^{-9}$.

- **DAMP** is a number between 0 and 1.0 determining how quickly Tracer will converge on an open-contact solution using voltage biasing. The closer **DAMP** is to 1.0, the more quickly Tracer will converge, but there is also an increased chance of slower convergence due to overshoot. Usually the user should not be concerned with the value of **DAMP**. Default: 0.9.
- **TRYCBC** is used only if there is an open contact. Tracer will only attempt to use zero-current biasing when the current of the control electrode is greater than **TRYCBC**. Otherwise, voltage biasing is used. In most cases the user does not have to worry about this parameter. Default: $1.0X10^{-17}$A/μm.
- **ANGLE1**, **ANGLE2**, and **ANGLE3** are critical angles (in degrees) affecting the smoothness and step size of the trace. They are described in detail in [28]. If the difference in slopes of the last two solution points is less than **ANGLE1**, the step size will be increased for the next projected solution. If the difference is between **ANGLE1** and **ANGLE2**, the step size remains the same. If the difference is greater than **ANGLE2**, the step size is reduced. **ANGLE3** is the maximum difference allowed, unless overridden by the **MINDL** parameter. **ANGLE2** should always be greater than **ANGLE1** and less than **ANGLE3**. Defaults: **ANGLE1** = 5^o, **ANGLE2** = 10^o, **ANGLE3** = 15^o.
- **ITLIM** is the maximum number of Newton loops for a given solution as specified in the method card of the PISCES input deck. The user should make sure that the value of **ITLIM** specified here is the same as that in the input deck. In certain cases, a PISCES solution may be aborted in Tracer because the solution will not converge within the given number of iterations. In some of these cases Tracer will try to redo the solution with a doubled number of iterations. If **ITLIM** is specified on the **OPTION** card, such attempts will be made. If there is no itlim statement or **ITLIM**=0, no attempts will be made. It is recommended that **ITLIM** be set to a low value, around 10 or 15 (or at least high enough to allow convergence of the initial solution). However, for GaAs devices a larger **ITLIM** of 20 or 25 is recommended. Default: 0.
- **MINCUR** is the value of the control current, in A/μm, above which Tracer carefully controls step size and guarantees a smooth trace. Below this current level, the program simply takes voltage steps as large as possible, i.e., as long as numerical convergence can be achieved, without regard for smoothness. If **MINCUR** is set to 0.0, Tracer will not begin smoothness control until it is past the first sharp turn in the I-V curve. This value should be used when the user is only interested in the rough location of a break in

the curve, such as the breakdown voltage of a single-junction device. If smoothness is required, a lower value should be specified. Setting **MINCUR** below $1X10^{-15}$A/μm is not recommended because `Tracer` has problems controlling smoothness at such low currents. Default: 0.0A/μm.

- **MINDL** is the minimum normalized step size allowed in the trace. Usually the user does not need to adjust this parameter. Increasing **MINDL** will reduce the smoothness of the trace by overriding the angle criteria, resulting in more aggressive projection and fewer simulation points. Reducing **MINDL** will enhance the smoothness and increase the number of points in the trace. Default: 0.1.
- **FREQUENCY** specifies how often the binary output (solution) files of the trace are saved. All I-V points are saved in `outputfile`. However, the PISCES solution files corresponding to these points are saved only if they are designated by **FREQUENCY**. If **FREQUENCY**=0, none of the solutions is saved, except perhaps the turning points (see below). If **FREQUENCY**=5, e.g., the solution file of every fifth point will be saved to files named soln.5, soln.10, etc., along with its PISCES input file (input.5, input.10, ...) and output I-V file (iv.5, iv.10, ...). Default: 0.
- **TURNINGPOINTS** is either **YES** or **NO**. If it is **YES**, the binary output (solution) file from PISCES will be saved whenever the slope of the I-V curve changes sign, i.e., there is a turning point. The name of the output file is soln.`num`, where `num` is the number of the current solution. For example, if the 25th point has a different sign than the 24th point, `Tracer` will save a file called soln.25. Default: **NO**.
- **VERBOSE** is either **YES** or **NO**. If it is **YES**, certain information about each solution (which the user may not be interested in) is printed in `outputfile`. The information consists of the external control-electrode voltage, the load resistance on the control electrode, the slope (differential resistance) of the solution, the normalized projected distance of the next simulation I-V point, and the normalized angle difference between the last two simulation points. Default: **NO**.

A.5.4 Examples

1. Step-size control will begin when the control electrode's current exceeds $1X10^{-14}$ A/μm. In the input deck itlim has been set to 12. Only essential information is saved in `outputfile`. The solution file of every tenth point, as well as any turning points, will be saved.

option mincur=1e-14 itlim=12 verbose=no frequency=10 turningpoints=yes

2. In a simulation with one or two open contacts, we want to keep the current through the open electrodes below $1X10^{-16}$A/μm, regardless of the current through the control electrode. Thus **RELMAX** is set to a very low value so that it will not be a factor in determining the current at the open contact(s).

option absmax=1e-16 relmax=1e-25

A.6 SOLVE Card

A.6.1 Description

The solve card is used to specify how the initial solution is obtained, what simulator is used, and whether there are any open contacts (zero-current bias conditions). A `Tracer` run will start either with an initial solution or by loading a solution from a previous PISCES simulation. If such a previous simulation has one or two zero-current electrodes, the user has the option of either specifying the voltages on these electrodes or of simply designating them as open contacts.

A.6.2 Syntax

FIRSTSOLUTION=<char> [**OPENCONT1**=<int>]
[**OPENCONT2**=<int>] [**SIMULATOR**=<char>]
[**VOPEN1**=<real>] [**VOPEN2**=<real>]

A.6.3 Parameters

- **FIRSTSOLUTION** is either **INITIAL**, **LOAD**, or **CURRLOAD**. In all cases a solve statement should be present in the PISCES input deck (`inputfile`). The parameters of this solve card in `inputfile` are not used but rather the card itself is used to mark where a PISCES solve card should be placed by `Tracer` in `inputfile` (see Section A.7).

  If **FIRSTSOLUTION**=**INITIAL**, a solution at thermal equilibrium will be solved by `Tracer` first. This implies that there cannot be any non-zero voltages or currents on a **FIXED** card. If the device has an open contact, i.e., a zero-current source, the user should not specify "current" on the contact line of the PISCES input deck to indicate a zero-current bias condition. Specifying **OPENCONT1** or **OPENCONT2** on the `tracefile` solve card is all that is needed.

  If **FIRSTSOLUTION**=**LOAD**, a load statement should be present directly above the solve card in `inputfile`, and it should designate the infile (see Section A.7). This option is used if the trace is to begin from a previously generated input solution file. The simulation which created this solution file must have used only voltage bias conditions. An open-contact trace can still be generated from such an input solution file

if the voltage bias condition on the open electrode(s) results in near-zero current for that electrode (see **VOPEN1**, **VOPEN2** below). Such an open-contact case would most likely arise if the user wanted to extend a previous `Tracer` run in which voltage bias conditions were used on the zero-current electrodes for the last simulation point.

If the loaded solution is from a simulation using a zero-current bias condition, **FIRSTSOLUTION**=**CURRLOAD** should be used. In this case "current" should be specified on a contact card for each open electrode. As in the **FIRSTSOLUTION**=**LOAD** case, the existing `inputfile` load card is used by `Tracer`, which means the correct "infile" should be specified on a load card directly above the solve card in `inputfile`. Default: none.

- **OPENCONT1** and **OPENCONT2** are the numbers of electrodes (between 1 and 9, inclusive) with a zero-current bias condition. There can be either zero, one, or two open contacts. When a device has an open contact, the user does not have to worry about convergence at low device-current levels. `Tracer` will automatically adapt the bias conditions to guarantee convergence. Default: none.
- **SIMULATOR** is either **PISC2ET** (PISCES-2ET) or **MD3200** or **MD10000** (TMA-MEDICI). It designates the device simulator to be used by `Tracer`. Other additions may be made in the future. Default: **PISC2ET**.
- **VOPEN1** and **VOPEN2** must be used if and only if there is an open contact and **FIRSTSOLUTION**=**LOAD** (voltage bias condition on open contact(s)). The values of **VOPEN1** and **VOPEN2** are the voltages of the open contacts **OPENCONT1** and **OPENCONT2**, respectively, in the loaded solution file designated on the load card of `inputfile`. If there is only one open contact, **VOPEN2** should not be specified. Defaults: 0.0.

A.6.4 Examples

1. The trace starts by solving an initial solution at zero bias and uses PISCES-2ET as the simulator. Electrode 2 is an open contact.

```
solve opencont1=2 firstsolution=init simulator=pisc
```

2. The trace starts with a previous solution using only voltage bias conditions. In this loaded solution the open contacts 2 and 4 have voltages of 0.641V and 0.509V, respectively.

```
solve firstsolution=load simulator=pisc opencont1=2 opencont2=4
+ vopen1=0.641 vopen2=0.509
```

A.7 Input Deck Specifications

As of September 1994, `Tracer` works with PISCES-2ET [44], some in-house versions of Stanford PISCES, and to some extent md3200 or md10000, TMA-MEDICI Version 1.2.2 [29].[1] Use of MEDICI is not yet robust and thus `Tracer` may or may not complete a trace using this simulator; the ability to use MEDICI for simulations with open contacts has not yet been implemented. If `Tracer` is to use simulators which cannot perform ac analysis, the capability for calculating admittances using the difference method must be added (a previous version of `Tracer` had this capability, so it should not be hard to implement).

The input deck used by `Tracer`, `inputfile`, is a standard PISCES file, but `Tracer` has certain requirements. For understanding the basic flow of an input deck, consult the PISCES or TMA-MEDICI manual. The mesh, region, electrode, doping, and model cards must already be present in the input deck. Additionally, the Newton solution method must be specified in the symbolic card. Other requirements are described below.

A.7.1 Load and Solve Cards

In `Tracer`, the user specifies whether to start with an initial solution or to load a previous solution (see Section A.6). In either case, the user must mark a line in `inputfile` where the solve statement should go by starting the line with “solve”. Any parameter specified in this solve statement is irrelevant. If `Tracer` is to start with a previous solution, `inputfile` must contain a standard load statement, above the solve line, containing the name of the input file to be used, i.e., load infil=<solution file name>. In the case of loading a solution with a zero-current bias condition, “current” should be specified on a contact card for the open electrode.

A.7.2 Contact Card

Contact cards are optional in `inputfile` except in the case of electrodes biased with a current source. The case of the zero-current source is noted in Section A.6 above. If there are any electrodes with a finite-current bias condition, a contact card with the “current” option should be placed in `inputfile` for each such electrode, regardless of whether `Tracer` is to begin with an initial solution or a loaded solution.

1. These implementations were developed in connection with Advanced Micro Devices, where TMA software is used, as part of a summer internship.

Even if no contact cards are required in `inputfile`, a line starting with "$contact" must be present so that `Tracer` will know where to add a contact statement. This contact card is necessary because this is where the load resistance of the control electrode is specified by `Tracer`. There is no problem with placing a contact card for the control electrode in the input deck as long as it does not specify a resistance value (which should never happen). Note that at least the first five letters of "contact" must appear for `Tracer` (and PISCES) to recognize it.

A.7.3 Method Card

In order to specify the maximum number of Newton iterations per solution, the itlim statement of the method card must be used in `inputfile`. If no method card is present, PISCES uses a default itlim of 20. However, in order to use the double-itlimit option (see Section A.5.3), a method card must be present in the input deck and itlim must be set to some value.

Another option must be specified in the method card if TMA-MEDICI is used. In this simulator, if a solution is aborted MEDICI will try to solve for an intermediate solution and then retry the original solution. This is not desirable when using `Tracer` since `Tracer` needs to keep track of aborted solutions. Thus, "stack=0" should be specified in the method card of MEDICI so that it does not attempt intermediate solutions. Analogously, the "trap" option should not be specified on the method card in a PISCES-2ET deck.

A.7.4 Options Card

When using PISCES-2ET, "curvetrace" should be specified on the options card so that PISCES will abort nonconverging solutions. Additionally, "nowarning" can be specified to prevent PISCES from printing warning messages which clutter the output, especially the warning issued when the load resistance changes value from one solution to the next. (Note: these options may not be available in early releases of PISCES-2ET.)

A.8 Data Format in Output Files

As each solution is found, it is recorded in `outputfile`. Naming `outputfile` is described in Section A.1. At the start of each line is the number of the solution. The second column of data contains voltage values of the control electrode, while the third

column contains current values of the control electrode. If there is a zero-current electrode, the voltage and current values of **OPENCONT1** will go in the next two columns, followed by the voltage and current of **OPENCONT2** if there is a second open electrode.

Values in the next columns depend on which data are recorded. If requested in the **FIXED** statements of `tracefile`, current values of fixed-voltage electrodes and voltage values of fixed-current electrodes will be recorded for each solution point in `outputfile`. The order from left to right is from low to high electrode number.

After the electrode information is recorded, further columns contain information about each solution if **VERBOSE=YES** in the **SOLVE** card of `tracefile`. These columns are, from left to right, external control-electrode voltage, load resistance on the control electrode, differential resistance, normalized distance of the next projection, and the angle difference between the current and previous solution points (see [28] for a description of these parameters).

The **FREQUENCY** and **TURNINGPOINTS** parameters in the **OPTION** card allow data to be saved for certain specified solutions. In `outputfile`, those points which are saved are marked with an asterisk next to the solution number. The files saved are the input deck, input.*i*; the I-V data file, iv.*i*; and the solution file, soln.*i*; where *i* is the number of the solution in `outputfile`.

A.9 Examples

In each of the `Tracer` examples below, a description of the simulation is given along with the command line used to invoke `Tracer` and figures with the listings of `inputfile` (the PISCES input deck), `tracefile`, and `outputfile`.

A.9.1 BV_{CEO}

The BV_{CEO} experiment is conducted by biasing an npn bipolar transistor's collector positively with respect to the emitter while the base is left open. The PISCES input deck, bvceo.pis, shown in Fig. A.63, defines the mesh, region, electrodes, doping, emitter contact, physical models, and solution method. Even though the contact card is not for the collector, which will be the control electrode, the presence of the card ensures that `Tracer`

```
title NPN Simulation for Toshiba w/ coarse mesh (1/19/92)
options nowarn curvetrace

mesh rect nx=11 ny=12
x.m n=1 l=0 r=1
x.m n=4 l=0.7 r=0.65
x.m n=11 l=2 r=1.2
y.m n=1 l=0 r=1.0
y.m n=3 l=0.2 r=0.7
y.m n=7 l=0.4 r=1.0
y.m n=12 l=2.5 r=1.3

region num=1 ix.l=1 ix.h=11 iy.l=1 iy.h=12 silicon

$electrode 1=emitter 2=base 3=collector
elec num=1 ix.l=1 ix.h=3 iy.l=1 iy.h=1
elec num=2 ix.l=10 ix.h=11 iy.l=1 iy.h=1
elec num=3 ix.l=1 ix.h=11 iy.l=12 iy.h=12

dop ascii n.type infil=npn1.p x.l=0 x.r=2 ra=0.8
dop ascii p.type infil=npn1.b x.l=0 x.r=2 ra=0.8
dop ascii n.type infil=npn1.as x.l=0 x.r=0.6 ra=0.8
dop gauss conc=1e18 p.type x.left=1.9 x.r=2 y.top=0 y.bot=0
+ char=0.3 ra=0.8

contact num=1 surf.rec vsurfn=8e5 vsurfp=8e5

model temp=300 srh auger conmob fldmob bgn impact
symbolic newton carr=2
method itlimit=15

solve
end
```

Fig. A.63 The input file, bvceo.pis, for the BV_{CEO} example.

will be able to find the correct place to insert a contact card for the collector when it needs to. If we did not wish to use the contact card in bvceo.pis, we would still have to insert a line beginning with "$contact" above the model and symbolic cards. Notice that "nowarn" and "curvetrace" are specified on the options card and "newton" is specified on the symbolic card, while nothing is specified on the solve card.

```
fixed num = 1 type=voltage value=0.0 record = no
control num=3 begin=0.0 initstep=0.1 control=vmax end=20
solve opencont1=2 first=init sim=pisc
option verbose=no itlim=15 turnpts=yes freq=5
+ mincur=5e-12 absmax=5e-19
```

Fig. A.64 The trace file, bvceo.tra, for the BV_{CEO} example.

In the trace file bvceo.tra (Fig. A.64), the **FIXED** card sets the voltage on the emitter electrode (num=1, as defined by bvceo.pis) to a constant value of 0.0V and states that the current through this electrode will not be recorded in `outputfile`. Electrode 3, the collector electrode, is designated as the control electrode. The **CONTROL** card states that the first solution will have a collector voltage of 0.0V, while the second solution will have a collector voltage of 0.1V. Tracing will continue until the collector voltage equals or exceeds 20V. If the initial step of 0.1V proves to be too large for convergence, `Tracer` will cut the step size in half, possible more than once, until it converges on a solution, and then will proceed from this solution.

In the **SOLVE** card, we specify that the base electrode (num=2) is to be treated as an open contact during the trace. Also, tracing will begin with a thermal-equilibrium solution and PISCES-2ET will be used for the simulation. Finally, the **OPTION** card specifies that only essential I-V data will be saved in the output file; the PISCES iteration limit is set to 15, agreeing with the PISCES deck in the input file; PISCES solutions will be saved for any turning points as well as for every fifth solution point; smoothness of the I-V curve will not be enforced until the collector current is greater than $5X10^{-12}$A/μm; and while voltage biasing is used on the open base contact, a solution will be accepted only if the current through the base is less than $5X10^{-19}$A/μm (unless the **RELMAX** condition predominates).

To run `Tracer`, the following command is typed at the prompt:

```
machine-prompt% tracer bvceo.pis bvceo.tra bvceo.out
```

While `Tracer` is running, the output of the PISCES runs are sent to the standard output, along with messages announcing when solutions are written to the output file. The output file, named bvceo.out in the command line, is shown in Fig. A.65, and a plot of the

```
#Soln    #Vctrl         Ictrl          Vcurr          Icurr
   1  0.000000e+00   6.640216e-19   0.000000e+00  -1.365566e-18
   2  1.000000e-01   4.536067e-17   1.000000e-01   1.341435e-19
   3  3.000000e-01   1.625653e-14   2.519331e-01   1.110998e-19
   4  7.000000e-01   1.258870e-13   3.047182e-01  -1.057191e-19
  *5  1.500000e+00   5.969134e-13   3.445794e-01  -6.21185e-20
   6  3.100000e+00   9.010139e-13   3.543271e-01   3.507138e-20
   7  6.300000e+00   1.937138e-12   3.725358e-01  -2.823590e-20
   8  1.270000e+01   5.523255e-12   3.960896e-01   4.401873e-20
   9  1.303983e+01   7.789304e-12   4.048689e-01   7.261209e-20
 *10  1.331971e+01   1.233772e-11   4.167027e-01  -2.392157e-20
  11  1.351640e+01   2.144845e-11   4.310039e-01  -3.015821e-20
  12  1.364322e+01   3.968015e-11   4.469627e-01  -2.586306e-20
  13  1.375854e+01   1.126228e-10   4.740828e-01  -3.055604e-20
  14  1.383613e+01   4.044189e-10   5.073686e-01   2.662945e-20
 *15  1.389759e+01   1.571640e-09   5.427516e-01  -6.997169e-20
  16  1.395684e+01   6.240608e-09   5.787376e-01   4.017923e-20
  17  1.401870e+01   2.491678e-08   6.149198e-01   5.800187e-20
  18  1.408638e+01   9.962280e-08   6.512089e-01   2.496370e-19
  19  1.416639e+01   3.984529e-07   6.876080e-01  -7.339886e-20
 *20  1.427994e+01   1.593805e-06   7.241677e-01  -1.364024e-19
  21  1.450307e+01   6.375431e-06   7.610238e-01   1.500092e-21
  22  1.508580e+01   2.550435e-05   7.985911e-01   1.631978e-19
  23  1.653961e+01   1.020878e-04   8.384486e-01  -8.203646e-20
 *24  1.743830e+01   2.563710e-04   8.696470e-01  -1.839253e-19
 *25  1.721139e+01   3.337692e-04   8.797490e-01   2.752857e-21
  26  1.608475e+01   4.874279e-04   8.953096e-01  -6.556564e-20
  27  1.467484e+01   6.389540e-04   9.070167e-01   4.997494e-19
  28  1.349730e+01   7.523351e-04   9.136248e-01  -3.868294e-19
  29  1.275292e+01   8.310109e-04   9.178346e-01   1.061968e-19
 *30  1.208031e+01   9.464580e-04   9.248369e-01   7.411538e-21
  31  1.149536e+01   1.064806e-03   9.311878e-01  -4.402454e-19
  32  1.072725e+01   1.238428e-03   9.391391e-01  -1.234551e-19
  33  1.032238e+01   1.369005e-03   9.444611e-01  -5.772530e-19
  34  1.018224e+01   1.459092e-03   9.480149e-01   3.337310e-19
 *35  1.012632e+01   1.578200e-03   9.526528e-01   5.859350e-19
 *36  1.015785e+01   1.736090e-03   9.586154e-01   1.039733e-19
  37  1.033709e+01   2.050704e-03   9.697170e-01   3.375426e-19
  38  1.082413e+01   2.676637e-03   9.898170e-01  -5.421011e-20
  39  1.216369e+01   3.909572e-03   1.027822e+00  -3.201785e-19
 *40  1.300269e+01   4.379769e-03   1.042119e+00   3.947174e-19
  41  1.372551e+01   4.658421e-03   1.050298e+00  -6.979551e-19
  42  1.482329e+01   4.950367e-03   1.058339e+00   1.677125e-19
  43  1.612039e+01   5.192516e-03   1.064355e+00   1.389134e-19
  44  1.757689e+01   5.415434e-03   1.068990e+00  -4.269046e-19
 *45  1.886187e+01   5.634169e-03   1.071871e+00  -2.778268e-19
  46  2.017690e+01   6.057492e-03   1.073324e+00   6.572976e-19
```

Fig. A.65 The output file, bvceo.out, for the BV_{CEO} example.

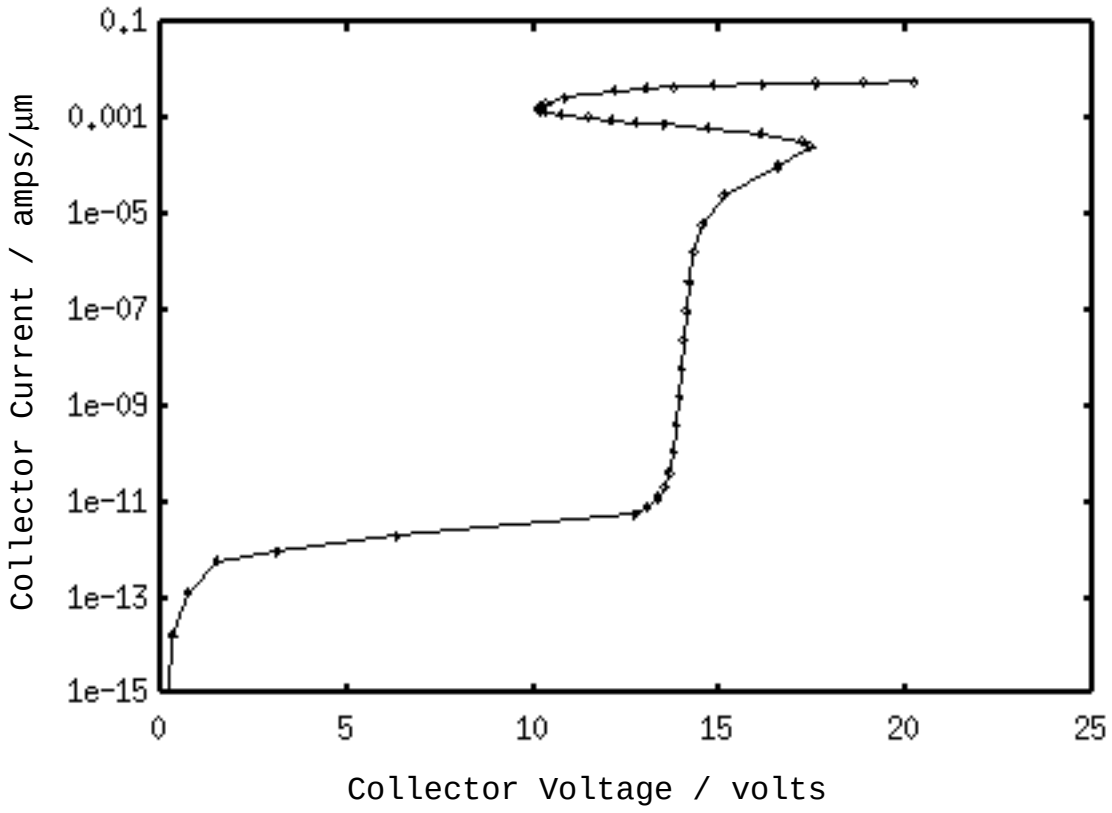

Fig. A.66 Collector current vs. collector voltage for the BV_{CEO} example.

collector current vs. collector voltage is shown in Fig. A.66. In bvceo.out, we see that every fifth solution, along with solutions 24 and 36 (the turning points), has been saved in files named soln.5, soln.10, etc. Additionally, the last solution was saved in the file soln.last, although there is no asterisk marking the last solution in bvceo.out.

At the top of bvceo.out, column headings mark the solution number, control-electrode (collector) voltage, control-electrode current, open-contact (base) voltage, and open-contact current as Soln, Vctrl, Ictrl, Vcurr, and Icurr, respectively. We see that the collector voltages for the first, second, and last solutions are 0.0, 0.1, and 20.18V, respectively. The final solution does not have a collector voltage of exactly 20V, as specified in bvceo.tra, because `Tracer` only guarantees that the curve will be traced out to at least 20V, not exactly 20V.

Other information regarding the trace must be inferred from the PISCES output displayed while Tracer is running (not shown). From this output we can see that voltage biasing was used on the open base contact for the first few solutions, in which the collector current is too small to allow stable use of zero-current biasing. A few PISCES simulations are actually run for each I-V point, with minor adjustments on the base voltage being made until the base current is less than **ABSMAX**. When the collector current is large enough, Tracer places a zero-current bias on the base. We can also see that a variable load resistor is placed on the collector when the collector current exceeds **MINCUR**. After this, the step sizes are regulated to produce a smooth curve.

A.9.2 GaAs MESFET

In this example, the drain of a GaAs MESFET is biased with respect to the grounded source with the gate set at -0.5V and the substrate grounded. Before Tracer can be used to sweep the drain electrode, a solution must be created, using PISCES-2ET, to set up the gate bias. The input deck shown in Fig. A.67 and Fig. A.68 defines the device, finds the thermal-equilibrium solution, and then steps the gate bias to -0.5V while holding the other electrodes at 0V. The mesh and solution files are saved to the files mes.mesh and mesvg.5.ini, respectively.

For Tracer, another PISCES input deck must be created to use as the input file (Fig. A.69). In mesvg.5.pis the mesh file generated by mes.pis, mes.mesh, is read in, preempting the mesh, eliminate, region, electrode, and doping cards. Since Tracer will be starting with a previous solution, the name of the solution file to load must be given in mesvg.5.pis. This load statement appears directly above the solve card with the file name mesvg.5.ini, the solution file generated by mes.pis.

The trace file mesvg.5.tra is shown in Fig. A.70. In the three **FIXED** cards, the voltages of the source and substrate (num=1 and num=4, respectively, as defined by mes.pis) have been fixed at 0V, while the gate voltage (num=2) has been fixed at -0.5V. The current through the gate electrode will be recorded for each solution in the output file. The **CONTROL** card of mesvg.5.tra specifies that the drain (num=3) will be swept from 0.0V to a voltage where the current is greater than or equal to $4.1X10^{-4}$A/μm, with an initial drain voltage step of 0.2V. On the **SOLVE** card, **FIRSTSOLUTION** is specified as **LOAD**, consistent with the input file mesvg.5.pis, and PISCES-2ET is designated as the

```
title mes.pis

mesh nx=53 ny=41 rect diag.fli outf=mes.mesh
x.m n=1 l=0 r=1
x.m n=5 l=1 r=0.85
x.m n=8 l=2 r=1.3
x.m n=11 l=3 r=0.7
x.m n=13 l=3.5 r=1
x.m n=18 l=4 r=0.8
x.m n=24 l=4.5 r=1.15
x.m n=32 l=5 r=0.85
x.m n=40 l=6 r=1.2
x.m n=43 l=7 r=1
x.m n=46 l=8 r=1.35
x.m n=49 l=9 r=0.7
x.m n=53 l=10 r=1.15

y.m n=1 l=-.01 r=1
y.m n=4 l=0.0 r=1
y.m n=7 l=0.01 r=1
y.m n=9 l=0.025 r=1
y.m n=20 l=0.19 r=1
y.m n=26 l=0.36 r=1.15
y.m n=39 l=3.0 r=1.25
y.m n=41 l=6.0 r=1.25

elim y.dir iy.lo=1 iy.hi=3 ix.lo=1 ix.hi=4
elim y.dir iy.lo=1 iy.hi=2 ix.lo=1 ix.hi=4
elim y.dir iy.lo=1 iy.hi=6 ix.lo=19 ix.hi=31
elim y.dir iy.lo=1 iy.hi=5 ix.lo=19 ix.hi=31
elim y.dir iy.lo=1 iy.hi=4 ix.lo=19 ix.hi=31
elim y.dir iy.lo=1 iy.hi=3 ix.lo=19 ix.hi=31
elim y.dir iy.lo=1 iy.hi=3 ix.lo=50 ix.hi=53
elim y.dir iy.lo=1 iy.hi=2 ix.lo=50 ix.hi=53
elim y.dir iy.lo=23 iy.hi=41 ix.lo=2 ix.hi=52
elim y.dir iy.lo=29 iy.hi=41 ix.lo=2 ix.hi=52
elim y.dir iy.lo=33 iy.hi=41 ix.lo=2 ix.hi=52
elim y.dir iy.lo=40 iy.hi=41 ix.lo=2 ix.hi=52

elim x.dir iy.lo=2 iy.hi=40 ix.lo=2 ix.hi=52
elim x.dir iy.lo=2 iy.hi=40 ix.lo=2 ix.hi=52
elim y.dir iy.lo=2 iy.hi=40 ix.lo=2 ix.hi=52
elim x.dir iy.lo=2 iy.hi=40 ix.lo=2 ix.hi=52
elim y.dir iy.lo=2 iy.hi=40 ix.lo=2 ix.hi=52
```

Fig. A.67 *The mesh generation and eliminate statements of the file mes.pis for the GaAs MESFET example.*

```
$*** regions
region num=1 ix.lo=1 ix.hi=53 iy.lo=4 iy.hi=41 gaas
region num=2 ix.lo=1 ix.hi=53 iy.lo=1 iy.hi=4 oxide
region num=2 ix.lo=16 ix.hi=34 iy.lo=1 iy.hi=7 oxide

$*** electrodes: 1=source 2=gate 3=drain 4=substrate
elec num=1 ix.lo=1 ix.hi=5 iy.lo=1 iy.hi=4
elec num=2 ix.lo=18 ix.hi=32 iy.lo=1 iy.hi=7
elec num=3 ix.lo=49 ix.hi=53 iy.lo=1 iy.hi=4
elec num=4 ix.lo=1 ix.hi=53 iy.lo=41 iy.hi=41

$*** doping
dop ascii x.l=0.0 x.r=10.0 inf=mei.dop
dop gaus x.l=-1 x.r=3.0 dos=5.0e13 cha=0.0607 peak=-0.0709
+ n.t erfc.lat lat.cha=0.0866
dop gaus x.l=7.0 x.r=11 dos=5.0e13 cha=0.0607 peak=-0.0709
+ n.t erfc.lat lat.cha=0.0866

$*** material
material num=1 eg300=1.42 affinity=4.07 vsat=10.0e6
+ permi=13.1 nc300=4.35e17 nv300=8.35e18
interface qf=-1e12 x.min=0.0 x.max=10 y.min=-.01 y.max=6.0

$*** contact
contact num=2 alu workf=4.84 surf

model conmob fldmob srh
symb newton carrier=0
method itlim=30 trap

solve ini

symb newton carrier=2
solve v2=-0.25
solve v2=-0.5 proj outfil=mesvg.5.ini

end
```

Fig. A.68 The second half of the file mes.pis for the GaAs MESFET example.

simulator to use. Since **VERBOSE** is **NO** on the **OPTION** card, only the essential I-V data will be recorded in the output file. The iteration limit is 30, consistent with mesvg.5.pis, and every ninth solution, as well as those corresponding to turning points, will have its solution file saved.

To run `Tracer`, the following command is typed at the prompt:

```
machine-prompt% tracer mesvg.5.pis mesvg.5.tra mesvg.5.out
```

Fig. A.72 shows the output file, mesvg.5.out, in which the solution number, drain voltage, drain current, and gate current have been recorded as Soln, Vctrl, Ictrl, and I2, respectively. The solution files of points 9, 18, 27, and 29 (a turning point), as well as of the last point (not marked in the output file) were saved as soln.9, soln.18, soln.27, soln.29, and soln.last, respectively. A plot of the drain current vs. drain voltage is shown in Fig. A.71.

```
title mesvg.5.pis

option nowarn curvetrace

mesh inf=mes.mesh

material num=1 eg300=1.42 affinity=4.07 vsat=10.0e6
+ permi=13.1 nc300=4.35e17 nv300=8.35e18
interface qf=-1e12 x.min=0.0 x.max=10 y.min=-.01 y.max=6.0

contact num=2 alu workf=4.84 surf

model conmob fldmob srh hypert impact
symb newton carrier=2
method itlim=30

load infil=mesvg.5.ini
solve

end
```

Fig. A.69 The input file, mesvg.5.pis, for the GaAs MESFET example.

```
fixed num = 1 type=voltage value=0.0 record = no
fixed num = 2 type=voltage value=-0.5 record = yes
fixed num = 4 type=voltage value=0.0 record = no
control num=3 begin=0.0 initstep=0.2 control=imax end=4.1e-4
solve first=load sim=pisc
option verbose=no itlim=30 turnpts=yes freq=9
```

Fig. A.70 The trace file, mesvg.5.tra, for the GaAs MESFET example.

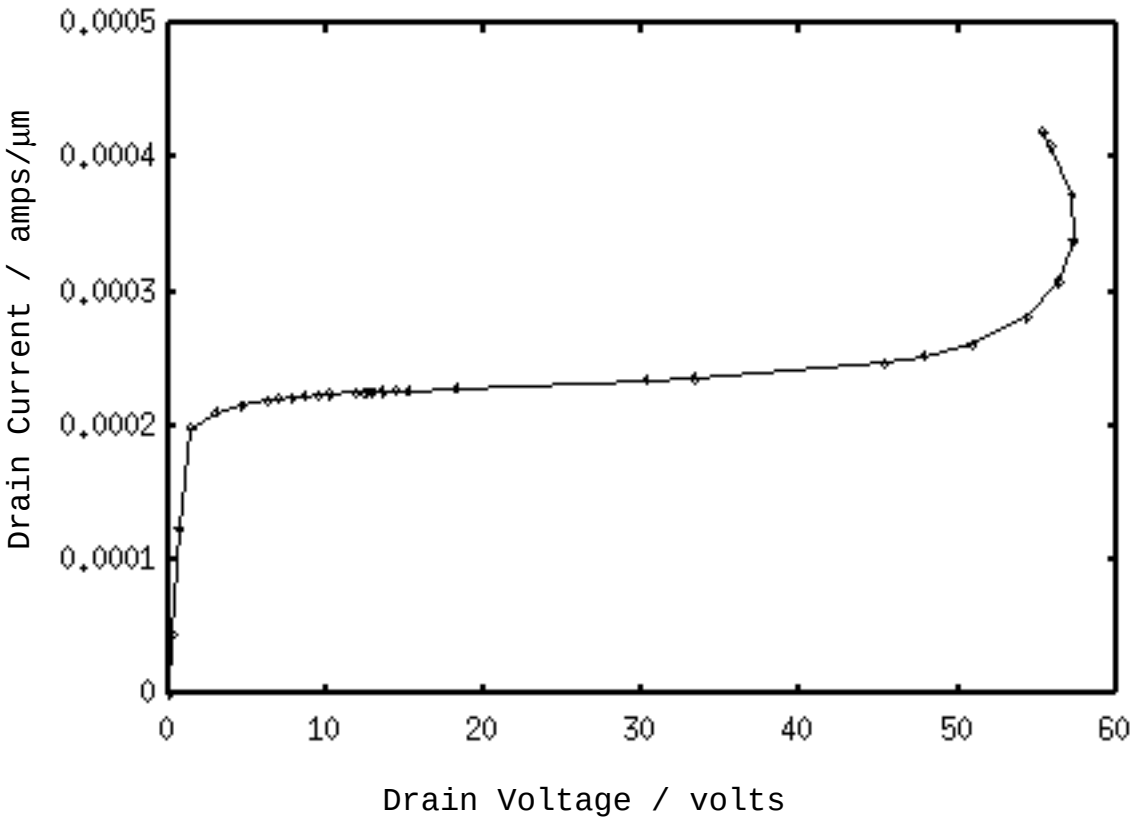

Fig. A.71 Drain current vs. drain voltage for the GaAs MESFET example.

```
#Soln   #Vctrl        Ictrl          I2
  1  0.000000e+00  1.903203e-16  -4.790550e-16
  2  2.000000e-01  4.411067e-05  -1.166555e-15
  3  6.000000e-01  1.233240e-04  -2.412571e-15
  4  1.400000e+00  1.985247e-04  -3.623966e-15
  5  3.000000e+00  2.104332e-04  -3.936303e-15
  6  4.600000e+00  2.155307e-04  -4.374722e-13
  7  6.200000e+00  2.189477e-04  -2.059647e-11
  8  7.000000e+00  2.202971e-04  -6.990150e-11
 *9  7.800000e+00  2.214088e-04  -1.707327e-10
 10  8.600000e+00  2.224028e-04  -3.638914e-10
 11  9.400000e+00  2.232137e-04  -6.557824e-10
 12  1.020000e+01  2.238725e-04  -1.043467e-09
 13  1.180000e+01  2.249711e-04  -2.208417e-09
 14  1.227965e+01  2.252584e-04  -2.671811e-09
 15  1.258651e+01  2.254261e-04  -2.980391e-09
 16  1.290104e+01  2.255882e-04  -3.308354e-09
 17  1.354587e+01  2.258872e-04  -3.995203e-09
*18  1.441648e+01  2.262760e-04  -5.076190e-09
 19  1.517061e+01  2.266211e-04  -5.076190e-09
 20  1.818697e+01  2.280737e-04  -1.389891e-08
 21  3.025183e+01  2.340778e-04  -1.841648e-07
 22  3.326644e+01  2.357641e-04  -3.360147e-07
 23  4.526743e+01  2.472668e-04  -2.984657e-06
 24  4.787346e+01  2.524107e-04  -4.783998e-06
 25  5.085805e+01  2.612441e-04  -4.783998e-06
 26  5.436859e+01  2.808594e-04  -1.633647e-05
*27  5.639889e+01  3.073427e-04  -2.612463e-05
 28  5.729060e+01  3.385567e-04  -3.460666e-05
*29  5.719992e+01  3.725458e-04  -3.844234e-05
 30  5.586886e+01  4.090192e-04  -3.611460e-05
 31  5.534786e+01  4.193991e-04  -3.517154e-05
```

Fig. A.72 The output file, mesvg.5.out, for the GaAs MESFET example.

Bibliography

[1] T.J. Green and W.K. Denson, “Review of EOS/ESD field failures in military equipment,” *Proc. 10th EOS/ESD Symp.*, 1988, pp. 7-14.

[2] C. Diaz, C. Duvvury, S.-M. Kang, and L. Wagner, “Electrical overstress (EOS) power profiles: A guideline to qualify EOS hardness of semiconductor devices,” *Proc. 14th EOS/ESD Symp.*, 1992, pp. 88-94.

[3] R. Merril and E. Issaq, “ESD Deign Methodology,” *Proc. 15th EOS/ESD Symp.*, 1993, pp. 233-237.

[4] F. Kuper, J.M. Luchies, and J. Bruines, “Suppression of soft ESD failures in a submicron CMOS process,” *Proc. 15th EOS/ESD Symp.*, 1993, pp. 117-122.

[5] S. Shabde, G. Simmons, A. Baluni, and R. Back, “Snapback-induced gate dielectric breakdown in graded-junction MOS structures,” *Proc. IEEE Int. Reliability Physics Symp.*, 1984, pp. 165-168.

[6] C. Duvvury, R. McPhee, D. Baglee, and R. Rountree, “ESD protection reliability in 1μm CMOS technologies,” *Proc. IEEE Int. Reliability Physics Symp.*, 1986, pp. 199-205.

[7] R. McPhee, C. Duvvury, R. Rountree, and H. Domingos, “Thick oxide device ESD performance under process variations,” *Proc. 8th EOS/ESD Symp.*, 1986, pp. 173-181.

[8] T.L. Polgreen and A. Chatterjee, “Improving the ESD Failure Threshold of Silicided n-MOS Output Transistors by Ensuring Uniform Current Flow,” *IEEE Trans. Elec. Devices*, vol. ED-39, 1992, pp. 379-388.

[9] R. Rountree, “ESD protection for submicron CMOS circuits: issues and solutions,” *IEDM Tech. Dig.*, 1988, pp. 580-583.

[10] I. Morgan, Advanced Micro Devices internal document, 1992.

[11] M. Middendorf and T. Hanksen, “Observed physical defects and failure analysis of EOS/ESD on MOS devices,” *Intl. Symp. for Test & Failure Analysis*, 1984, pp. 205-213.

[12] H. Melchior and M.J.O. Strutt, “Secondary Breakdown in Transistors,” *Proc. IEEE*, 1964, pp. 439-440.

[13] K. Mayaram, J.-H. Chern, L. Arledge, and P. Yang, “Electrothermal Simulation Tools for Analysis and Design of ESD Protection Devices,” *IEDM Tech. Dig.*, 1991, pp. 909-912.

[14] J. Colvin, “The identification and analysis of latent ESD damage on CMOS input gates,” *Proc. 15th EOS/ESD Symp.*, 1993, pp. 109-116.

[15] S. Aur, A. Chatterjee, and T. Polgreen, “Hot Electron Reliability and ESD Latent Damage,” *Proc. IEEE Int. Reliability Physics Symp.*, 1988, pp. 15-18.

[16] O.J. McAteer, R.E. Twist, and R.C. Walker, “Latent ESD Failures,” *Proc. 4th EOS/ESD Symp.*, 1982, pp. 41-48.

[17] “MIL STD 883.C/3015.7 notice 8,” *Military Standard for Test Methods and Procedures for Microelectronics: ESD Sensitivity Classification*, March 22, 1989.

[18] A. Amerasekera and J. Verwey, “ESD in Integrated Circuits,” *Quality and Reliability Engineering International*, vol. 8, 1992, pp. 259-272.

[19] K. de Kort, J.M. Luchies, and J.J. Vrehen, “The transient behavior of an input protection,” *4th European Conference on Electron and Optical Beam Testing of Electronic Devices*, 1993, pp. 7-15-7-18.

[20] C. Duvvury, R.N. Rountree, and O. Adams, “Internal chip ESD phenomena beyond the Protection Circuit,” *Proc. IEEE Int. Reliability Physics Symp.*, 1988, pp. 19-25.

[21] N. Khurana, T. Maloney, and W. Yeh, “ESD on CHMOS devices--equivalent circuits, physical models and failure mechanisms,” *Proc. IEEE Int. Reliability Physics Symp.*, 1985, pp. 212-223.

[22] Y. Fong and C. Hu, “High-Current Snapback Characteristics of MOSFETs,” *IEEE Trans. Elec. Dev.*, vol. ED-37, 1990, pp. 2101-2103.

[23] A. Amerasekera, L. van Roozendaal, J. Bruines, and F. Kuper, “Characterization and Modeling of Second Breakdown in NMOSTs for the Extraction of ESD-Related Process and Design Parameters,” *IEEE Trans. Elec. Dev.*, vol. ED-38, 1991, pp. 2161-2168.

[24] C. Diaz, C. Duvvury, and S.M. Kang, “Studies of EOS susceptibility in 0.6 μm nMOS ESD I/O protection structures,” *Proc. 15th EOS/ESD Symp.*, 1993, pp. 83-91.

[25] K.R. Mistry, D.B. Krakauer, and B.S. Doyle, “Impact of Snapback-Induced Hole Injection on Gate Oxide Reliability of N-MOSFETs,” *IEEE Elec. Dev. Letts.*, vol. 11, 1990, pp. 460-462.

[26] C.S. Rafferty, M.R. Pinto, and R.W. Dutton, “Iterative methods in semiconductor device simulation,” *IEEE Trans. Elec. Dev.*, vol. ED-32, 1985, pp. 2018-2027.

[27] M.R. Pinto, C.S. Rafferty, H. Yeager, and R.W. Dutton, “PISCES-IIB,” Technical Report, Department of Electrical Engineering, Stanford University, 1985.

[28] R.J.G. Goossens, S. Beebe, Z. Yu, and R.W. Dutton, “An Automatic Biasing Scheme for Tracing Arbitrarily Shaped I-V Curves,” *IEEE Trans. Computer-Aided Design*, vol. CAD-13, 1994, pp. 310-317.

[29] “MEDICI Two-Dimensional Semiconductor Device Simulation, Version 1.1” Technology Modeling Associates, Inc., Palo Alto, CA, 1993.

[30] "ATLAS 2D Device Simulation Framework User's Manual, Edition 2," Silvaco International, Santa Clara, CA, 1994.

[31] H.S. Carslaw and J.C. Jaeger, Conduction of Heat in Solids, 2nd Ed., Oxford, Clarendon Press, 1959.

[32] A. Amerasekera, A. Chatterjee, and M.-C. Chang, "Prediction of ESD Robustness in a Process Using 2-D Device Simulations," *Proc. IEEE Int. Reliability Physics Symp.*, 1993, pp. 161-167.

[33] A. Chatterjee, T. Polgreen, and A. Amerasekera, "Design and Simulation of a 4 kV ESD Protection Circuit for a 0.8μm BiCMOS Process," *IEDM Tech. Dig.*, 1991, pp. 913-916.

[34] O. J. McAteer, Electrostatic Discharge Control, McGraw-Hill, New York, 1990.

[35] H. Hyatt, H. Calvin, an H. Mellberg, "A Closer Look at the Human ESD Event," *Proc. 3rd EOS/ESD Symp.*, 1981, pp. 1-8.

[36] O.J. McAteer, "Electrostatic Damage in Hybrid Assemblies," *Annual Reliability and Maintainability Symposium Proceedings*, 1978, pp. 434-442.

[37] Z. Yu, D. Chen, R.J.G. Goossens, and R.W. Dutton, "Accurate Modeling and Numerical Techniques in Simulation of Impact-Ionization Effects on BJT Characteristics," *IEDM Tech. Dig.*, 1991, pp. 901-904.

[38] D.C. Wunsch and R.R. Bell, "Determination of Threshold Failure Levels of Semiconductor Diodes and Transistors due to Pulse Voltages," *IEEE Trans. Nucl. Sci.*, vol. NS-15, Dec. 1968, pp. 244-259.

[39] V.M. Dwyer, A.J. Franklin, and D.S. Campbell, "Thermal Failure in Semiconductor Devices," *Solid-State Electronics*, vol. 33, 1990, pp. 553-560.

[40] D.L. Lin, "ESD Sensitivity and VLSI Technology Trends: Thermal Breakdown and Dielectric Breakdown," *Proc. 15th EOS/ESD Symp.*, 1993, pp. 73-81.

[41] C. Duvvury and C. Diaz, "Dynamic Gate Coupling of NMOS for Efficient Output ESD Protection," *Proc. IEEE Int. Reliability Physics Symp.*, 1992, pp. 141-150.

[42] S.M. Sze, Physics of Semiconductor Devices, 2nd Ed., John Wiley, New York, 1981.

[43] C. Duvvury, C. Diaz, and T. Haddock, "Achieving Uniform nMOS Device Power Distribution for Sub-micron ESD Reliability," *IEDM Tech. Dig.*, 1992, pp. 131-134.

[44] Z. Yu, D. Chen, L. So, and R.W. Dutton, "PISCES-2ET Two Dimensional Device Simulation for Silicon and Heterostructures," Technical Report, Integrated Circuits Laboratory, Stanford University, 1994.

[45] Z. Yu and R.W. Dutton, "SEDAN III - A generalized electronic material device analysis program," Technical Report, Stanford University, 1985.

[46] S. Selberher, Analysis and Simulation of Semiconductor Devices, Springer-Verlag, New York, 1984.

[47] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, "A Physically Based Mobility Model for Numerical Simulation of Nonplanar Devices," *IEEE Trans. Computer-Aided Design*, vol. CAD-7, 1988, pp. 1164-1171.

[48] D.M. Caughey and R.E. Thomas, "Carrier mobilities in silicon empirically related to doping and field," *Proc. IEEE*, vol. 55, 1967, pp. 2192-2193.

[49] J.W. Slotboom, G. Streutker, G.J.T. Davids, and P.B. Hartog, "Surface Impact Ionization in Silicon Devices," *IEDM Tech. Dig.*, 1987, pp. 494-497.

[50] J.G. Rollins and J. Choma, Jr., "Mixed-Mode PISCES-SPICE Coupled Circuit and Device Solver," *IEEE Trans. Computer-Aided Design*, vol. CAD-7, 1988, pp. 862-867.

[51] Z. Yu and R.W. Dutton, "A Modularized, Mixed IC Device/Circuit Simulation System," *Proc. Synthesis and Simulation Meeting and International Interchange*, 1992, pp. 444-448.

[52] S. Ohtani, M. Yoshida, N. Kitagawa, and T. Saitoh, "Model of leakage current in LDD output MOSFET due to low-level ESD stress," *Proc. 12th EOS/ESD Symp.*, 1990, pp. 177-181.

[53] S. Tam, P.-K. Ko, and C. Hu, "Lucky-Electron Model of Channel Hot-Electron Injection in MOSFETs," *IEEE Trans. Elec. Dev.*, vol. ED-31, 1984, pp. 1116-1125.

[54] B.S. Doyle, D.B. Krakauer, and K.R. Mistry, "Examination of Oxide Damage During High-Current Stress of n-MOS Transistors," *IEEE Trans. Elec. Dev.*, vol. ED-40, 1993, pp. 980-985.

[55] H. Haddara and S. Cristoloveanu, "Two-dimensional modeling of locally damaged short-channel MOSFETs operating in the linear region," *IEEE Trans. Elec. Dev.*, vol. ED-34, 1987, pp. 378-385.

[56] A. Schwerin, W. Hansch, and W. Weber, "The relationship between oxide charge and device degradation: A comparative study of n- and p-channel MOSFETs," *IEEE Trans. Elec. Dev.*, vol. ED-34, 1987, pp. 2493-2500.

[57] S.H. Voldman and V.P. Gross, "Scaling, Optimization and Design Considerations of Electrostatic Discharge Protection Circuits in CMOS Technology," *Proc. 15th EOS/ESD Symp.*, 1993, pp. 251-260.

[58] G. Kreiger, "ESD in Integrated Circuits--General Introduction," *ESD in Integrated Circuits Short Course Proceedings*, sponsored by University of California, Berkeley, 1992.

[59] M.E. Law, C.S. Rafferty, and R.W. Dutton, "SUPREM-IV Users Manual," Integrated Circuits Laboratory, Stanford University, 1988.

[60] C.D. Thurmond, "The Standard Thermodynamic Function of the Formation of Electrons and Holes in Ge, Si, GaAs and GaP," *J. Electrochem. Soc.*, vol. 122, 1975, pp. 1133-1141.

[61] R.S. Muller and T.I. Kamins, Device Electronics for Integrated Circuits, 2nd Ed., John Wiley, New York, 1986.

[62] C. Jacoboni, C. Canali, G. Ottaviani, and A.A. Quaranta, "A Review of Some Charge Transport Properties of Silicon," *Solid-State Electronics*, vol. 20, 1977, pp. 77-89.

[63] R. van Overstraeten and H. DeMan, “Measurement of the Ionization Rates in Diffused Silicon p-n Junctions,” Solid-State Electron., vol. 13, 1970, pp. 583-608.

[64] S.M. Sze, VLSI Technology, 2nd Ed., McGraw-Hill, New York, 1988, p. 118.

[65] S. Daniel and G. Krieger, “Process and Design Optimization for Advanced CMOS I/O ESD Protection Devices,” *Proc. 12th EOS/ESD Symp.*, 1990, pp. 206-213.

[66] A. Stricker, D. Gloor, and W. Fichtner, “Layout Optimization of an ESD-Protection n-MOSFET by Simulation and Measurement,” *Proc. 17th EOS/ESD Symp.*, 1995, pp. 205-211.

[67] S.G. Beebe, “Methodology for Layout Design and Optimization of ESD Protection Transistors,” *Proc. 18th EOS/ESD Symp.*, 1996, pp.265-275.

[68] C. Duvvury and A. Amerasekera, “Advanced CMOS Protection Device Trigger Mechanisms During CDM,” *Proc. 17th EOS/ESD Symp.*, 1995, pp.162-174.

[69] S. Voldman, S. Furkay, and J. Slinkman, “Three-Dimensional Transient Electrothermal Simulation of Electrostatic Discharge Protection Circuits,” *Proc. 16th EOS/ESD Symp.*, 1994, pp. 246-256.

[70] BBN/Catalyst Handbook, Bolt Beranek and Newman Inc., 1992.

[71] K. Verhaege et al., “Analysis of HBM ESD Testers and Specifications Using a 4th Order Lumped Element Model,” *Proc. 15th EOS/ESD Symp.*, 1993, pp. 129-137.

[72] H. Gieser and M. Haunschild, “Very-Fast Transmission Line Pulsing of Integrated Structures and the Charged Device Model,” *Proc. 18th EOS/ESD Symp.*, 1996, pp. 85-94.

[73] C. Diaz, S.M. Kang, and C. Duvvury, “Circuit-Level Electrothermal Simulation of Electrical Overstress Failures in Advanced MOS I/O Protection Devices,” *IEEE Trans. Computer-Aided Design*, vol. CAD-13, 1994, pp. 482-493.

[74] S. Ramaswamy, E. Li, E. Rosenbaum, and S.-M. Kang, “Circuit-Level Simulation of CDM-ESD and EOS in Submicron MOS Devices,” *Proc. 18th EOS/ESD Symp.*, 1996, pp. 316-321.

[75] S.L. Lim, X.Y. Zhang, S. Beebe, and R.W. Dutton, “A Computationally Stable Quasi-Empirical Compact Model for the Simulation of MOS Breakdown in ESD Protection Circuit Design,” *Proc. Intl. Conf. on Simulation of Semiconductor Processes and Devices*, 1997, pp. 161-164.