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.

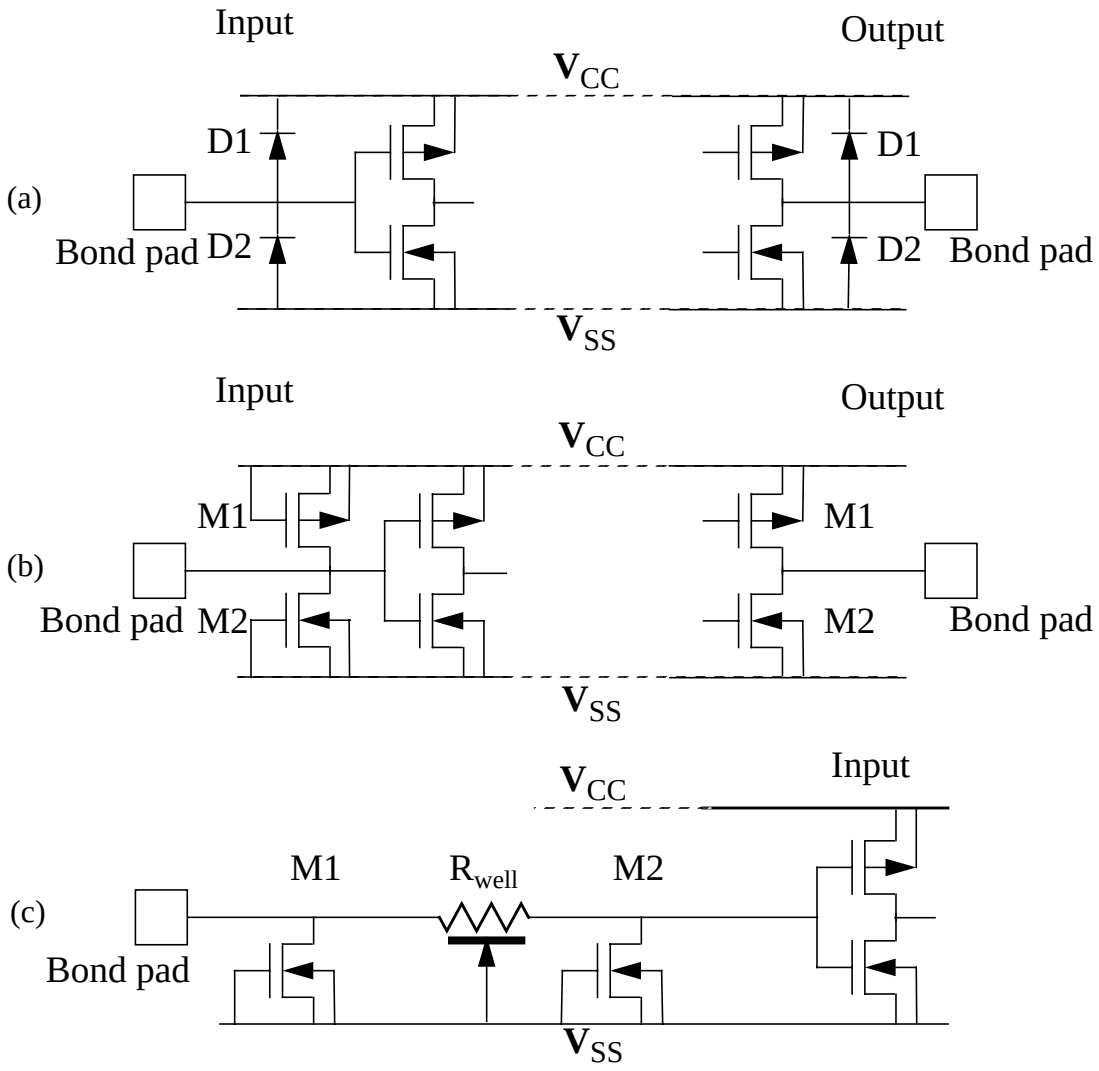

Fig. 1.1 ESD protection circuits in a CMOS technology: (a) diode protection; (b) CMOS transistor protection; (c) combination resistor/transistor protection circuit.

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.

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

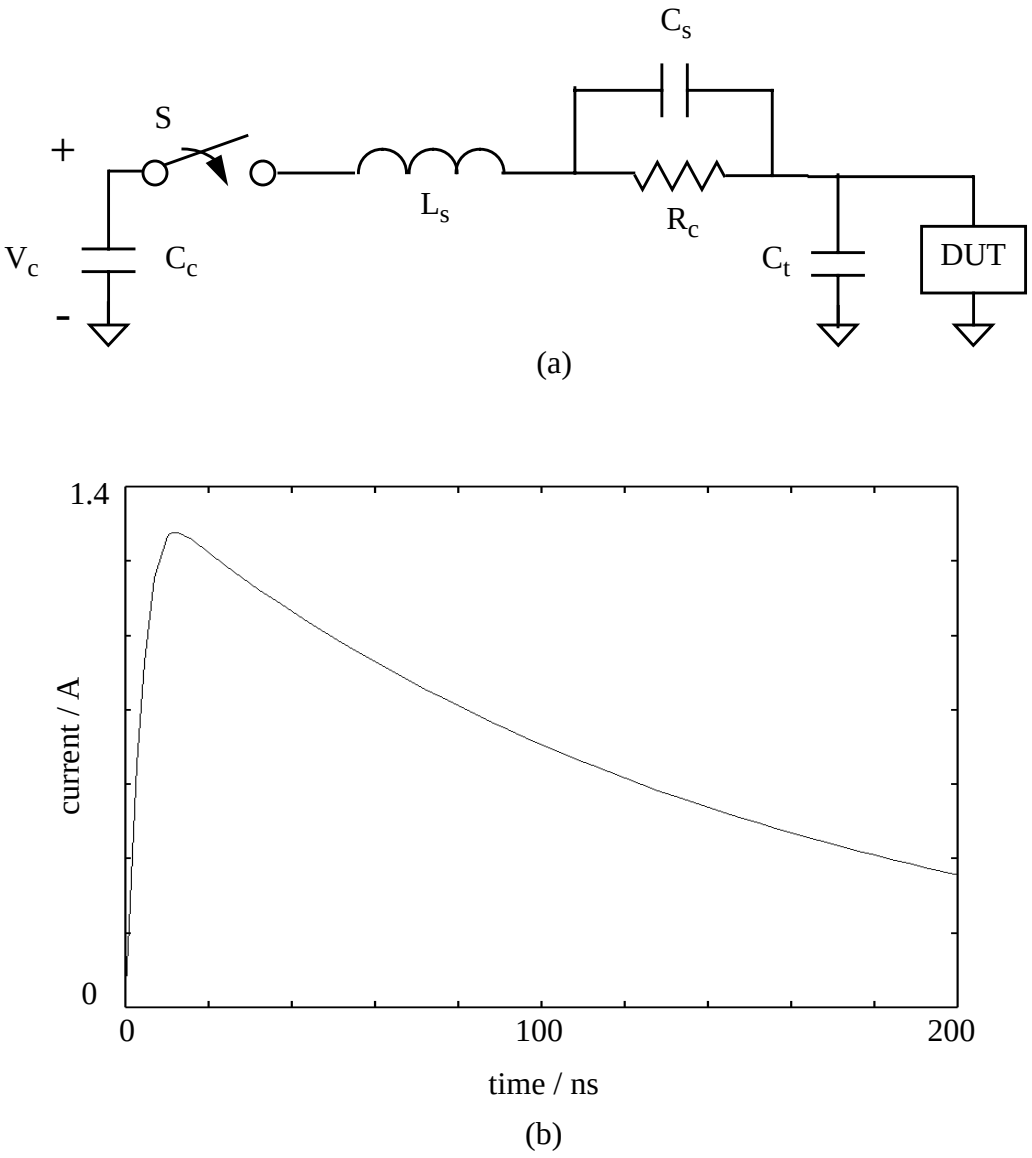

Fig. 2.2 *(a) Circuit model for the HBM and MM. Capacitor C_c is charged to the test voltage, V_c, and then discharged through R_c to the device under test (DUT) by closing switch S. Parasitic circuit elements are represented by series inductance L_s, stray capacitance C_s, and test-board capacitance 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

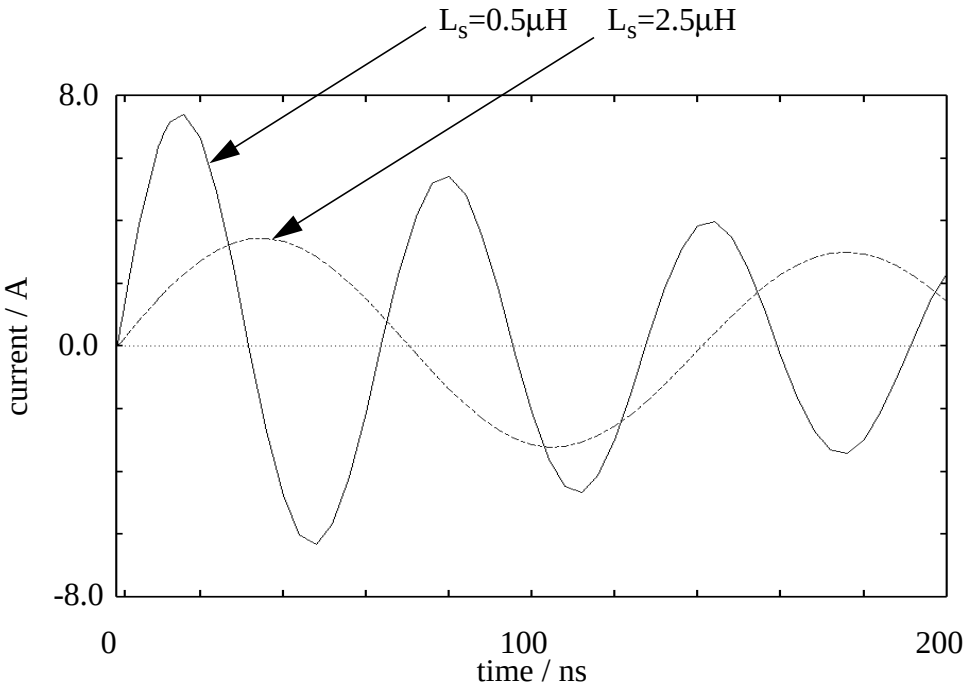

Fig. 2.3 SPICE-generated short-circuit MM output current waveforms for V_c=400V, C_c=200pF, R_c=5Ω, C_s=1pF, and C_t=10pF.

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.

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 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,

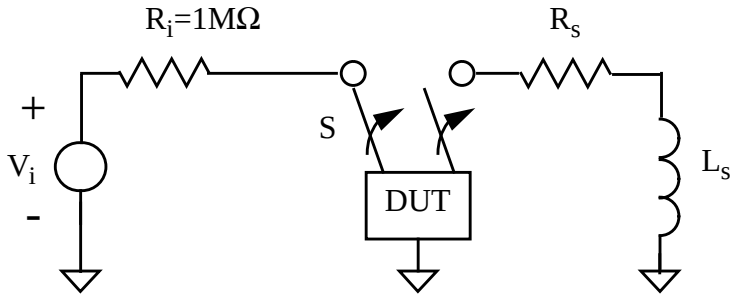

Fig. 2.4 *Circuit model for the CDM. A ground pin of the device under test (DUT) is charged to a voltage V_i, after which the switch S is thrown, removing the voltage source and connecting a different pin of the device to ground.*

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

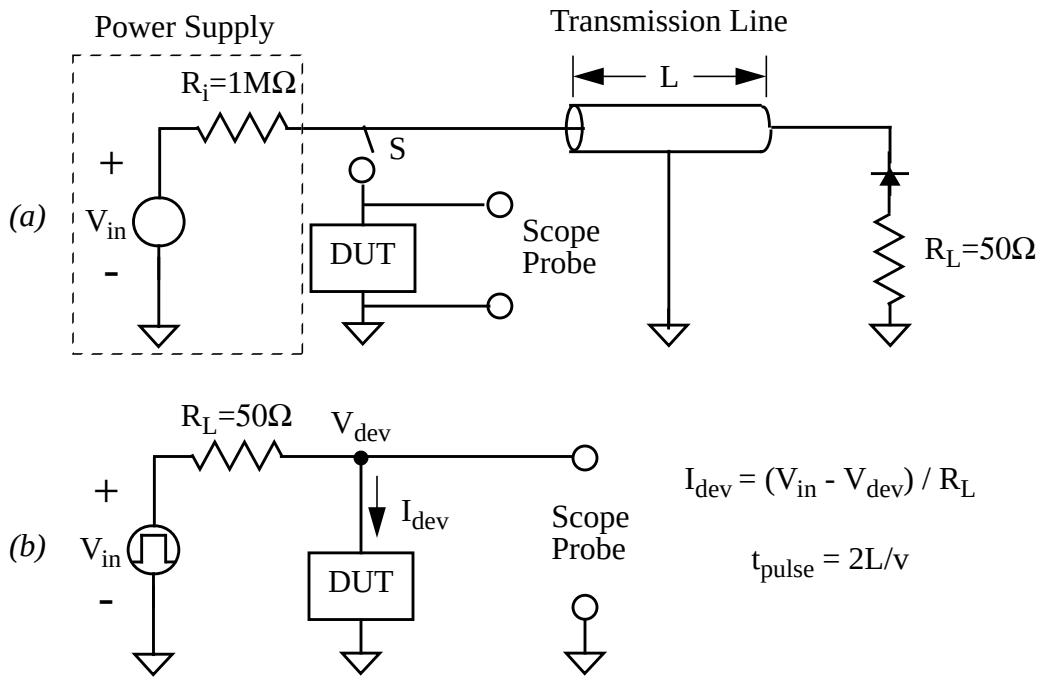

Fig. 2.5 *(a) TLP schematic: a transmission line of length L is charged to voltage V_{in} and then discharged through the device under test (DUT) when switch S closes. An oscilloscope voltage probe across the DUT monitors the circuit response. (b) Equivalent circuit of the TLP setup: the input is a square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

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

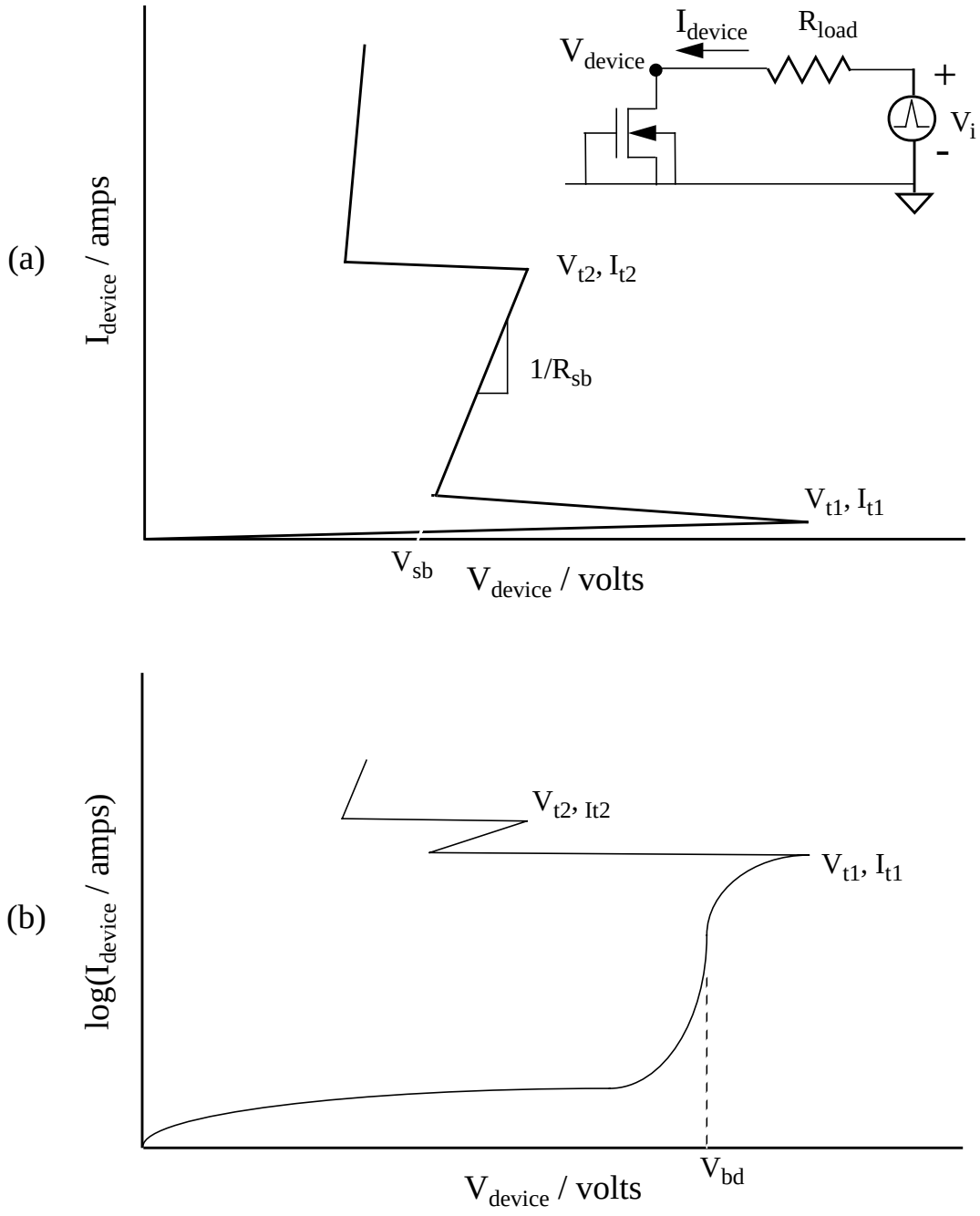

Fig. 2.6 *Qualitative I-V curve for an NMOS transistor subjected to a positive ESD pulse: (a) linear scale shows snapback trigger point (V_{t1}, I_{t1}), snapback voltage (V_{sb}) and resistance (R_{sb}), and second-breakdown trigger point (V_{t2}, I_{t2}), with circuit shown inset; (b) log scale shows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

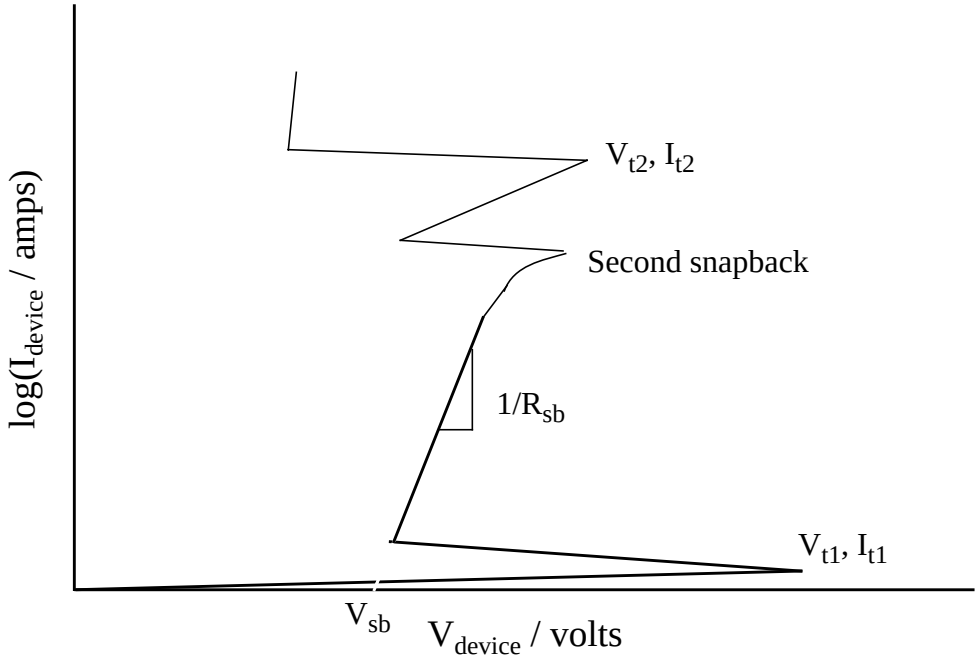

Fig. 2.7 Depiction of second snapback in a qualitative transient I-V curve for an NMOS transistor subjected to a positive ESD pulse.

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

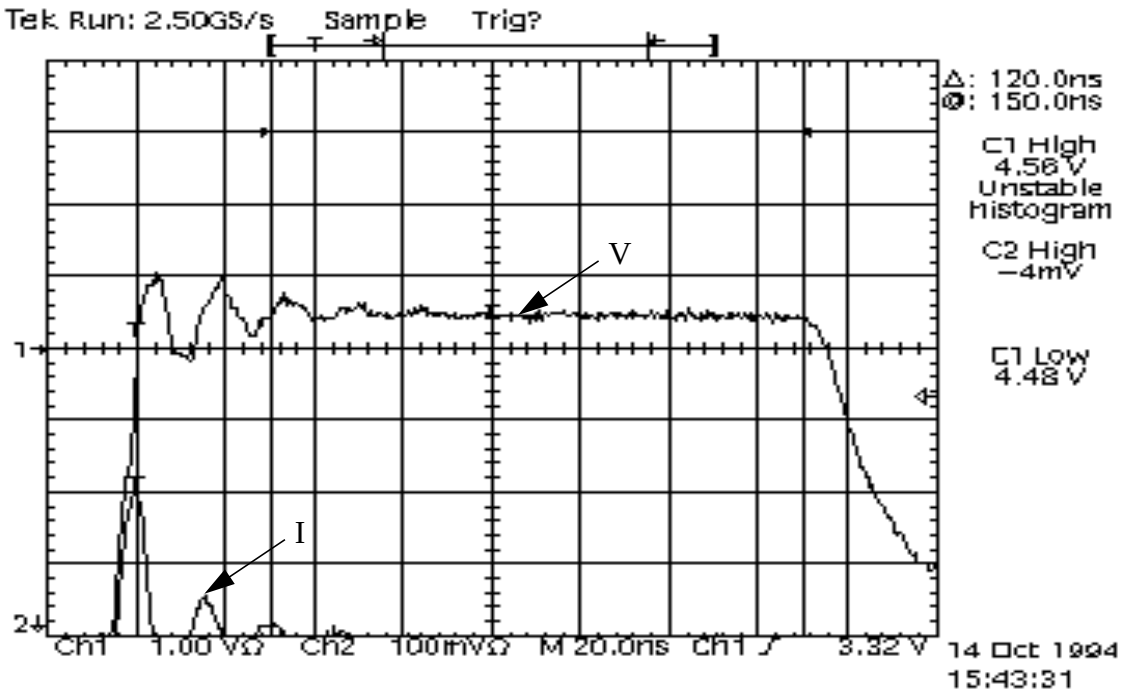

Fig. 2.8 *A screen capture of a Tektronix TDS 684A digitizing oscilloscope displays the device voltage (Ch1) and current (Ch2) response of a 50/0.6μm device to a 4.6V, 150ns input pulse. After some initial ringing, the device voltage settles to a value approximately equal to the input voltage and the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

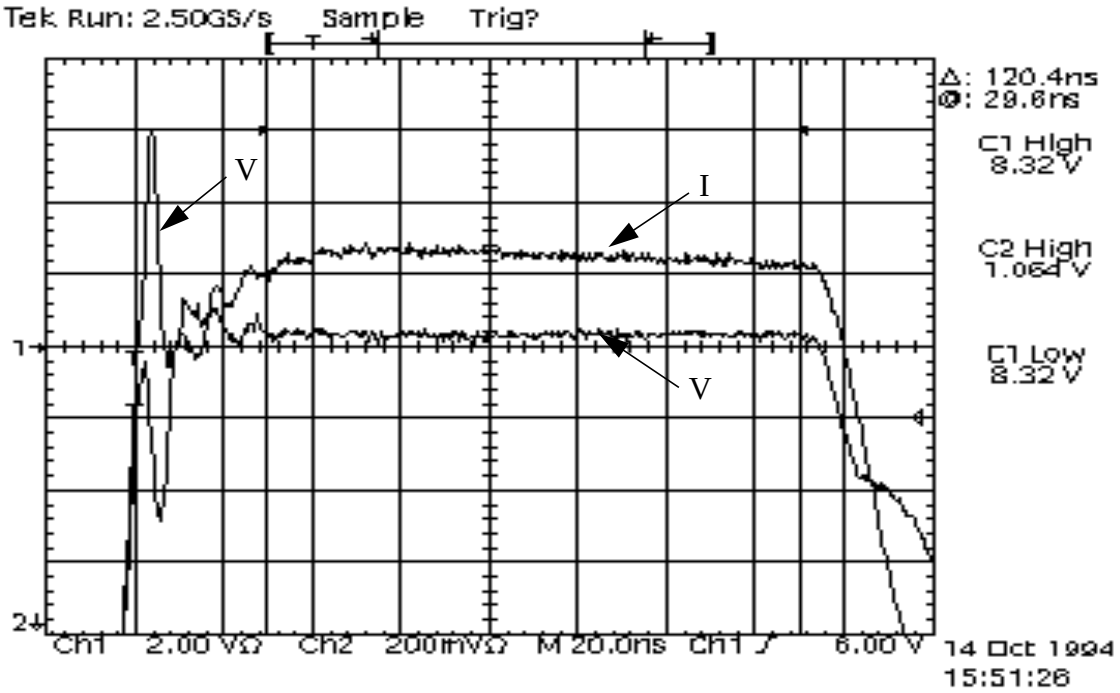

Fig. 2.9 *The screen capture shows the current and voltage response to a 15V, 150ns input pulse. The device voltage breaks down and snaps back in the first few nanoseconds of the pulse.*

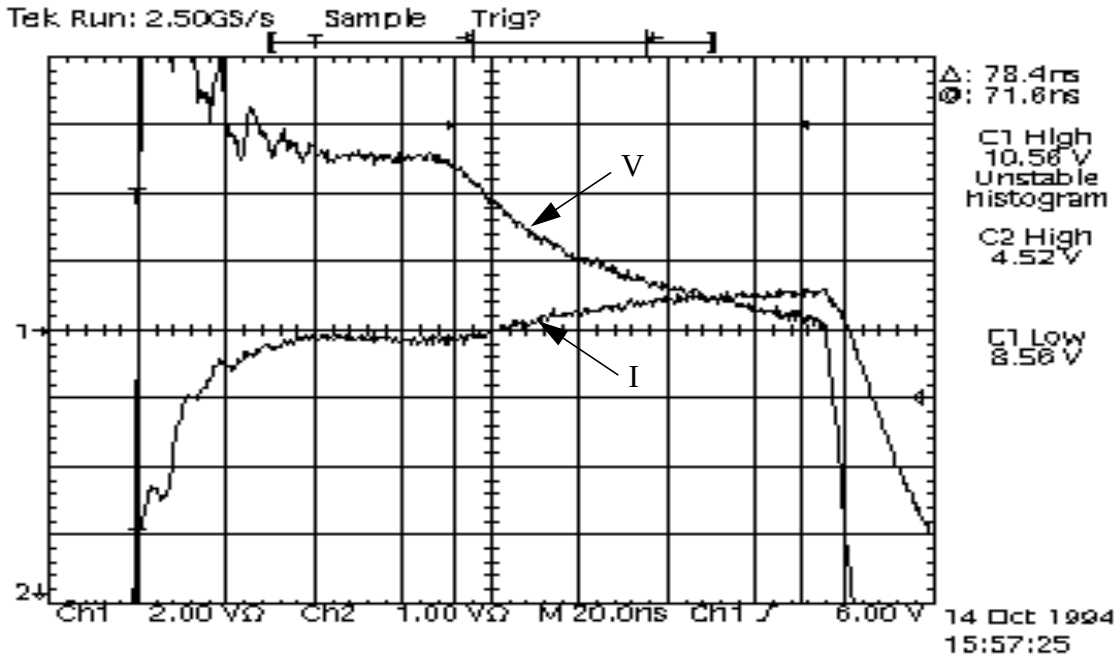

Fig. 2.10 *Second breakdown is observed as a drop in the device voltage and rise in device current about 72ns into the 36V, 150ns input pulse.*

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

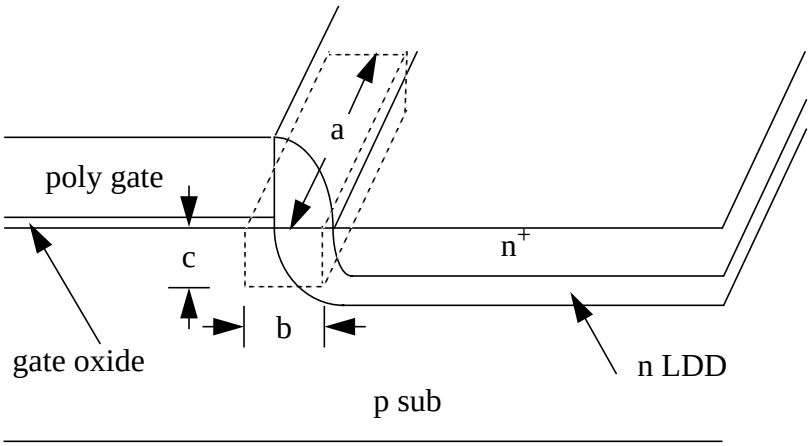

Fig. 2.11 3D thermal box region (dotted lines) of heat dissipation in an NMOS transistor subjected to a positive ESD pulse. The dimension a is equal to the device width, b is related to the gate length, and c is approximately equal to the diffusion depth.

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

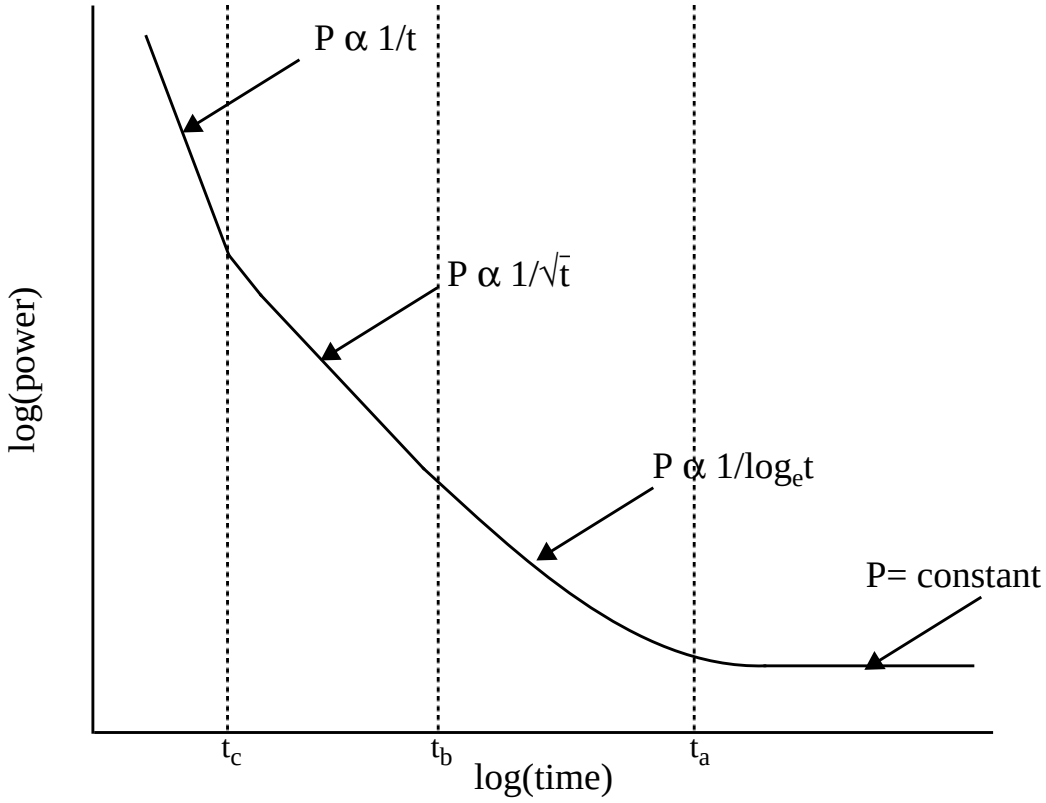

Fig. 2.12 A qualitative schematic of input power-to-failure vs. time-to-failure predicted by an analytical thermal model.

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

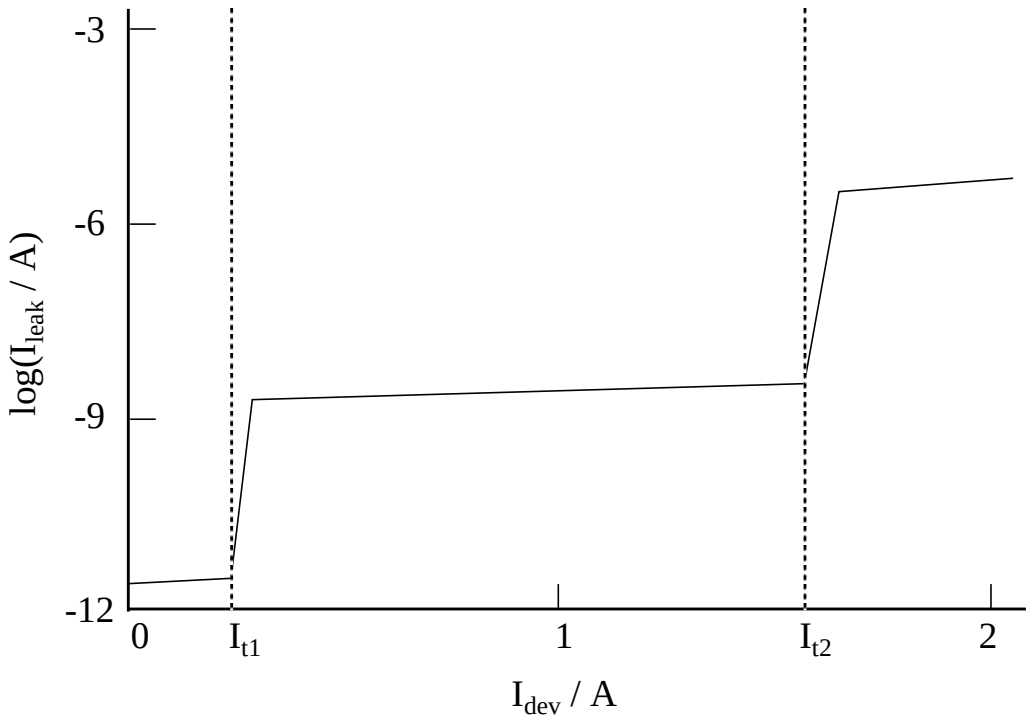

Fig. 2.13 Qualitative plot of device leakage evolution vs. stress-current level of a TLP experiment. Transitions are evident at the snapback and second-breakdown points.

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

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

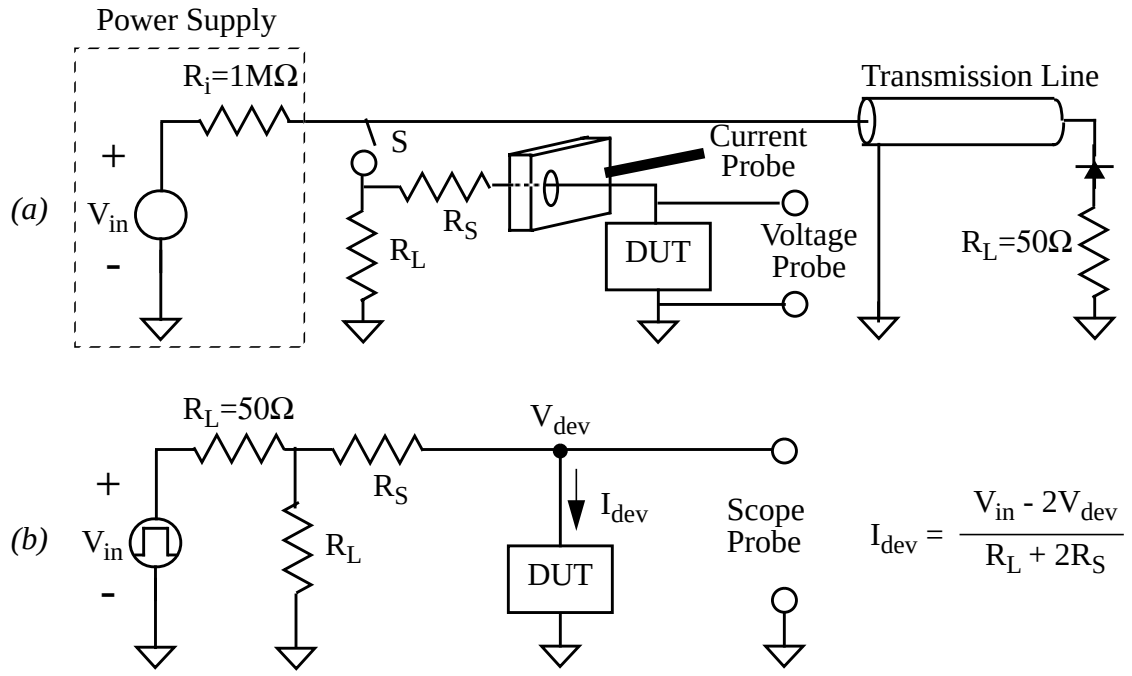

Fig. 2.14 *(a) Advanced TLP schematic: a series-parallel resistor combination has been added to enhance current resolution. A current probe is now used to directly measure the device current on the oscilloscope. (b) Equivalent circuit of the TLP setup.*

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.

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.

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

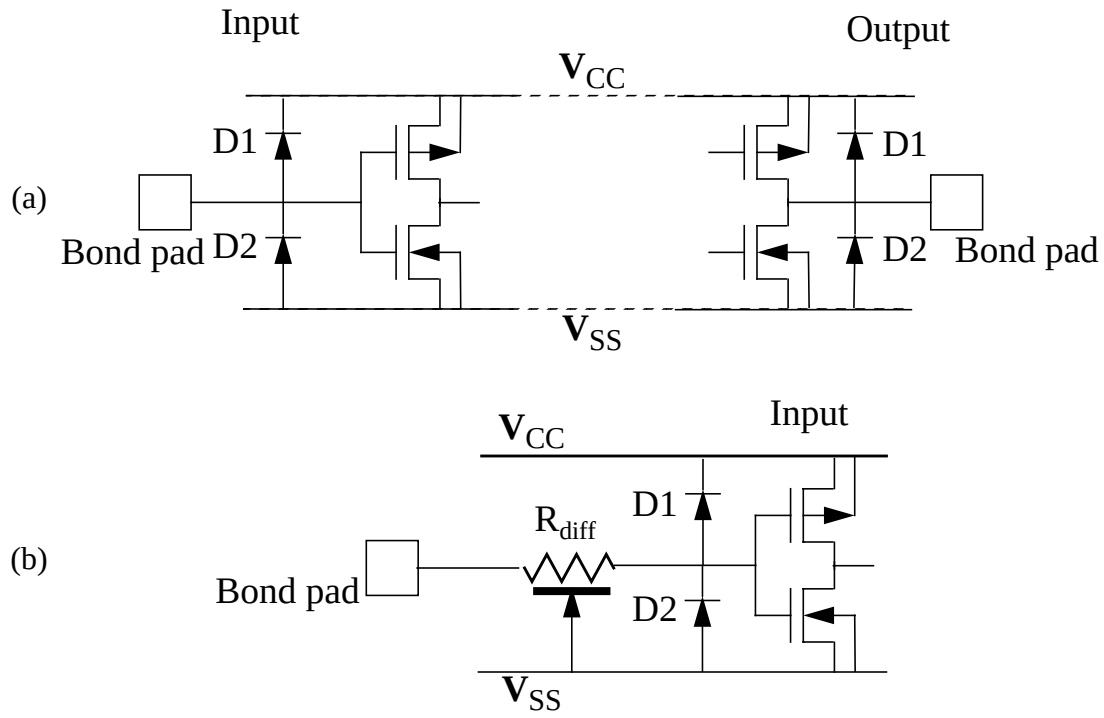

Fig. 2.15 *(a) ESD diode protection circuit in a CMOS technology; (b) use of a series resistor in combination with diode protection. A diffused resistor has a distributed resistance and also forms a distributed diode.*

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.

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

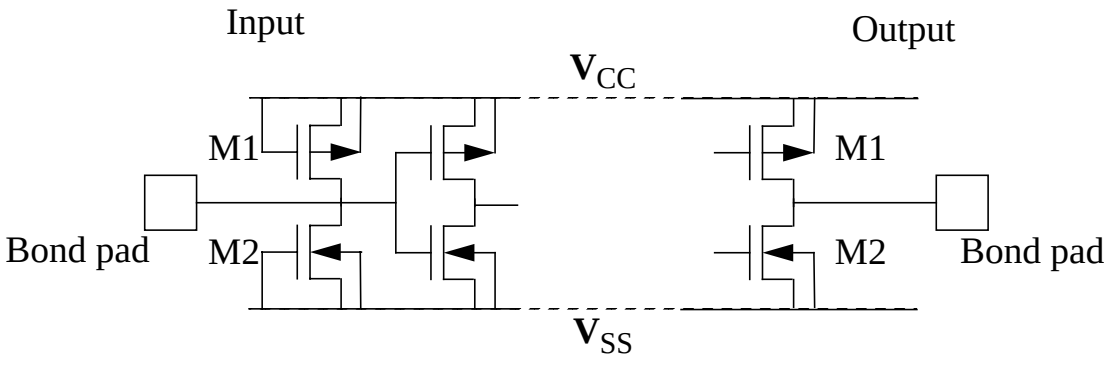

Fig. 2.16 CMOS input and output protection. The input protection transistors protect the thin gates of the input buffer, while the output protection transistors double as the output buffer.

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.

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. 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

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

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.

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

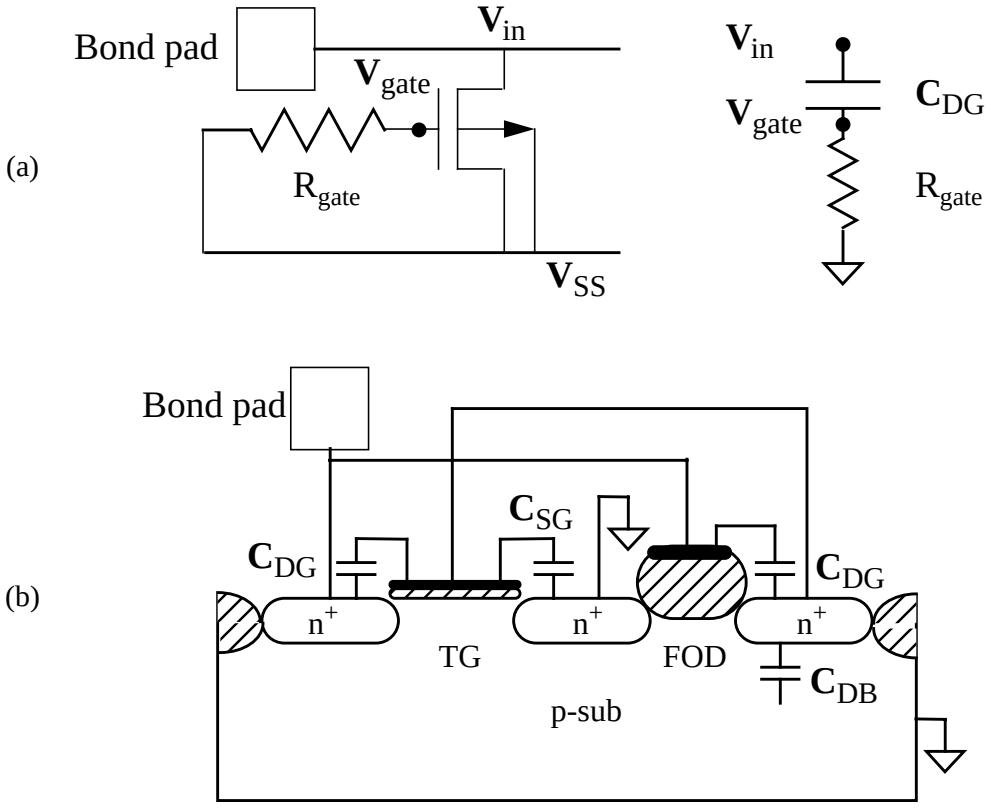

Fig. 2.17 Gate-bouncing techniques: (a) employment of a gate-bounce resistor and equivalent circuit; (b) dynamic gate coupling method.

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.

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

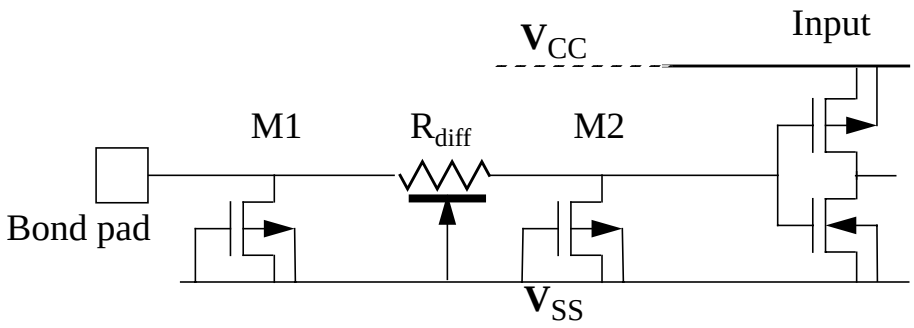

Fig. 2.18 Combination resistor/transistor ESD input protection circuit, featuring a diffused resistor (R_{diff}) between wide (M1) and narrow (M2) NMOS transistors. Resistance from gate to ground is not shown.

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

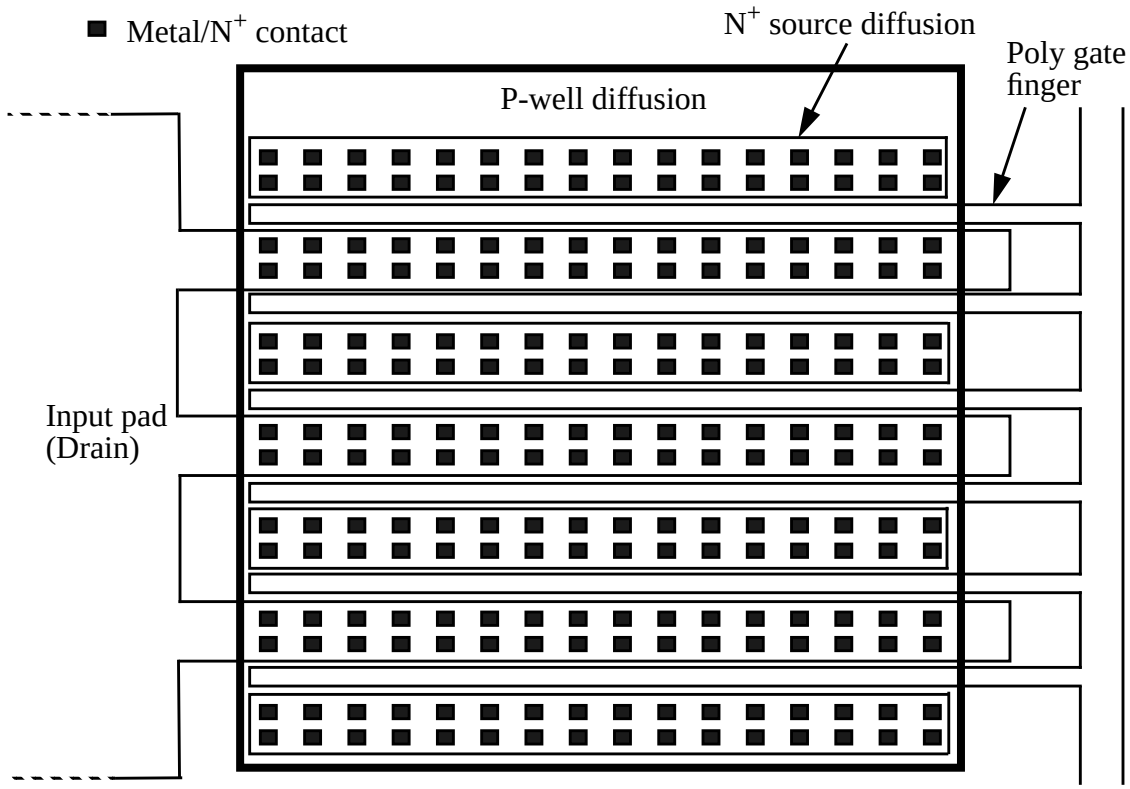

Fig. 2.19 Layout of a multiple-finger NMOS transistor between input and V_{SS}. There are three drain fingers and six poly gates. Another finger (not shown) branches off of the drain pad to the input buffer of the IC.

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

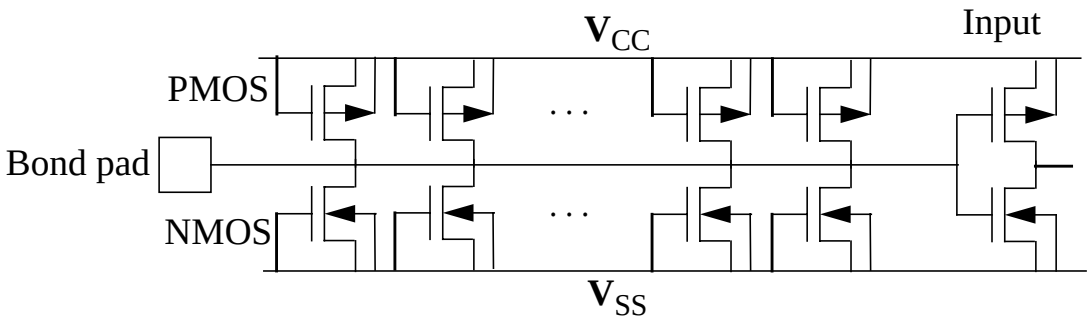

Fig. 2.20 Circuit diagram of CMOS input protection using multifinger structures.

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.

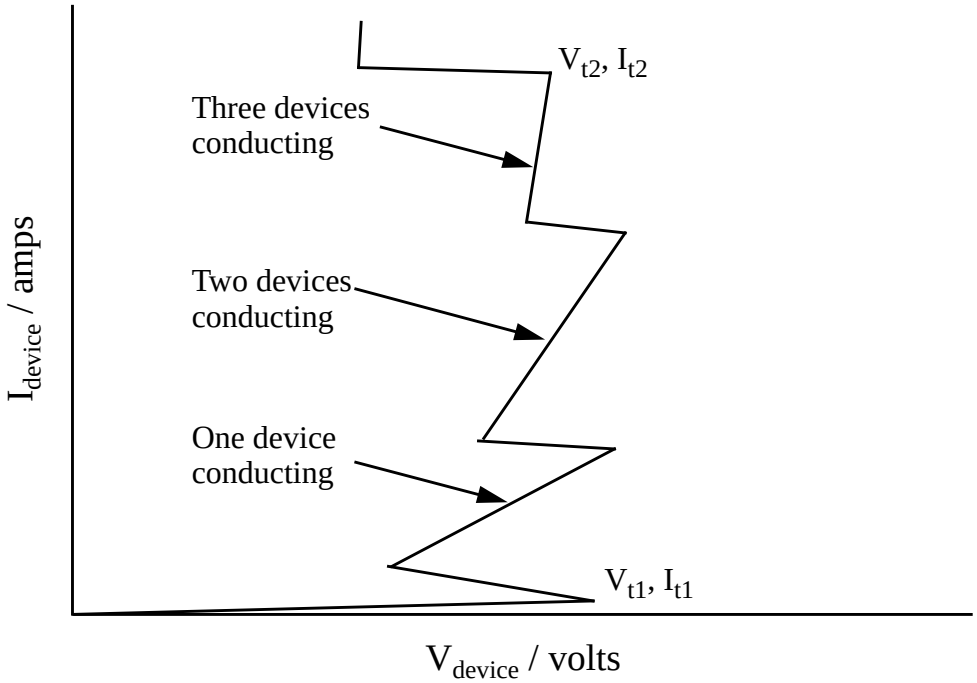

Fig. 2.21 Qualitative TLP I-V curve for an NMOS multifinger structure subjected to a positive ESD pulse (cf. Fig. 2.6). Each snapback indicates another device turning on.

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

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

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]

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

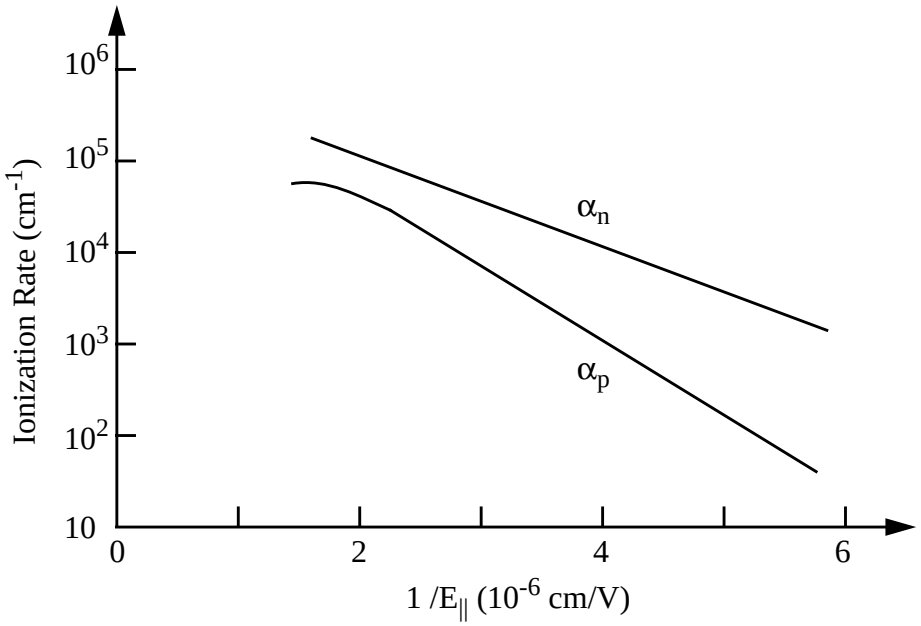

Fig. 3.22 Qualitative plot of impact-ionization rates for electrons (α_n) and holes (α_p) as a function of electric field for silicon at 300K.

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

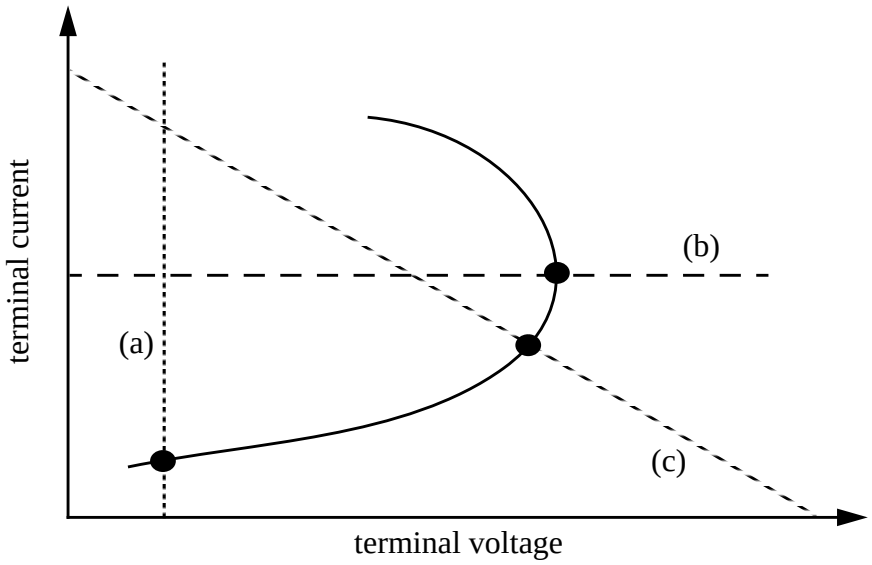

Fig. 3.23 Schematic representation of various types of bias specification: (a) voltage control, (b) current control, (c) load-line control. The simulator will converge to the intersection of the (dashed) constraint line and the I-V curve.

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

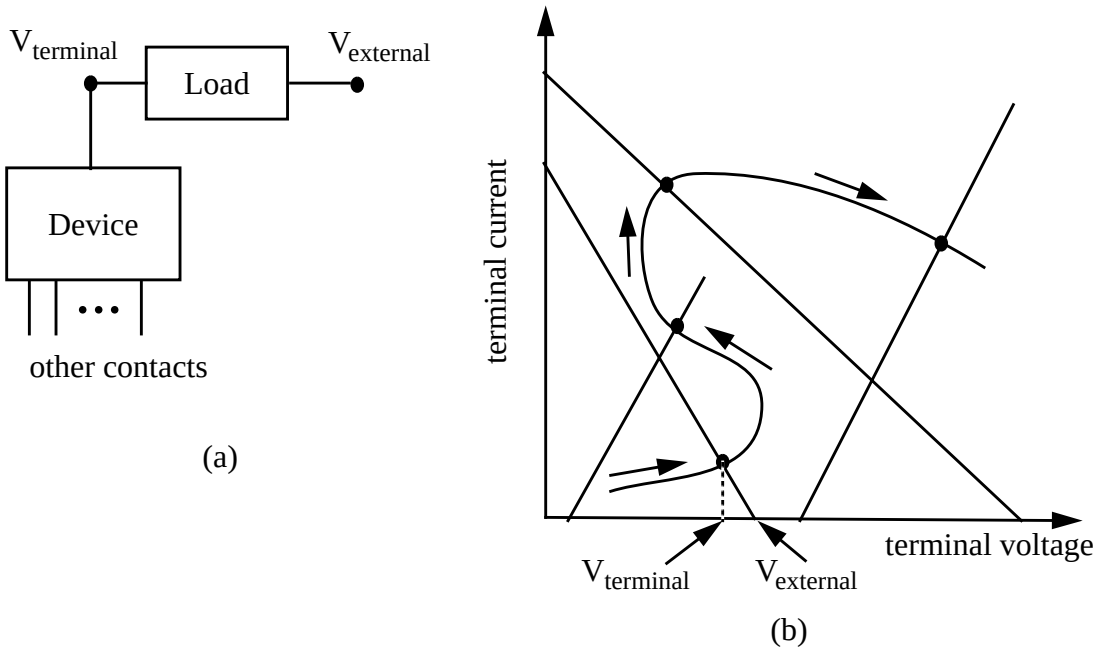

Fig. 3.24 *(a) Schematic of a general device with external load and voltage; (b) adapting the load line (solid lines) along the I-V curve allows for optimal convergence at each operating point.*

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

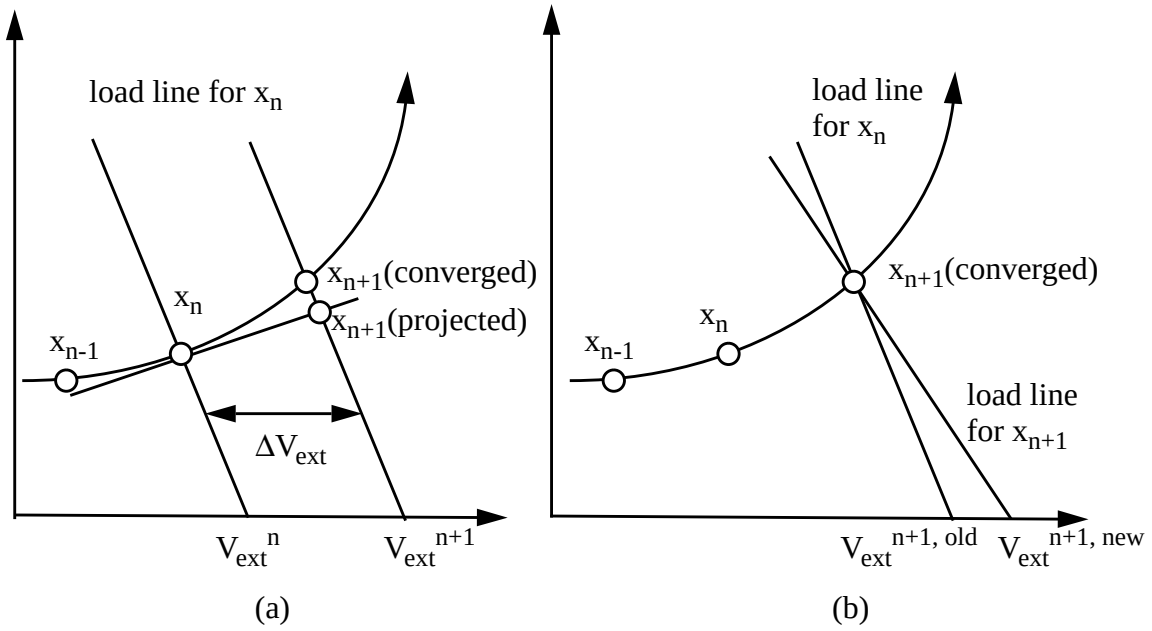

Fig. 3.25 *(a) Projection: the solution is advanced by incrementing the external voltage from V_{ext}^{n} to V_{ext}^{n+1} while the load resistance is held constant; (b) Recalibration: the load resistance is changed so that the load line is perpendicular to the trace at the new solution point. This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

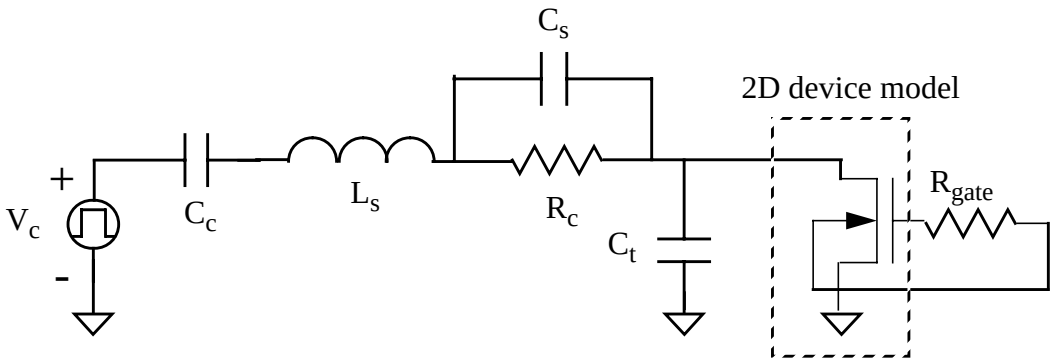

Fig. 3.26 Mixed-mode circuit model for an NMOS transistor subjected to the human-body model. The voltage source and all circuit elements are defined with SPICE models, except for the transistor, which is defined by the 2D device simulator.

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.

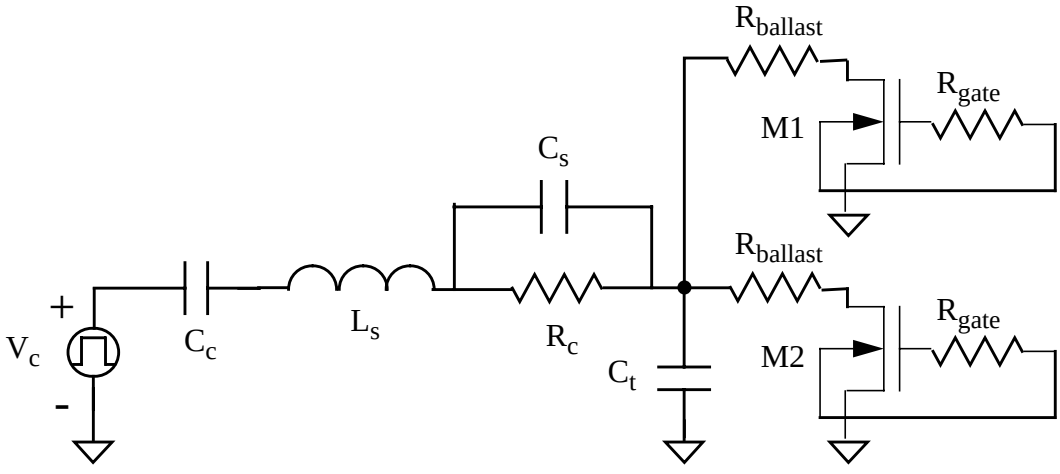

Fig. 3.27 Mixed-mode circuit model for a multiple-finger ESD NMOS structure subjected to the human-body model. Ballast resistors are placed between the output of the HBM circuit and each drain to facilitate uniform turn-on of transistors M1 and M2.

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

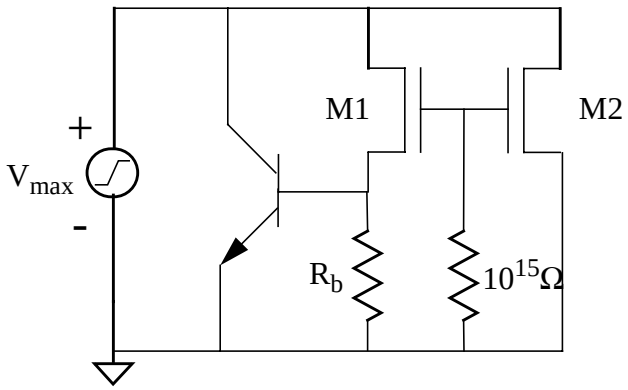

Fig. 3.28 ESD protection circuit used for SPICE simulations by Chatterjee et al. [33]. M1 is an NMOS transistor designed to facilitate the turn-on of the npn transistor during ESD. M2 represents the output NMOS transistor being protected.

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.

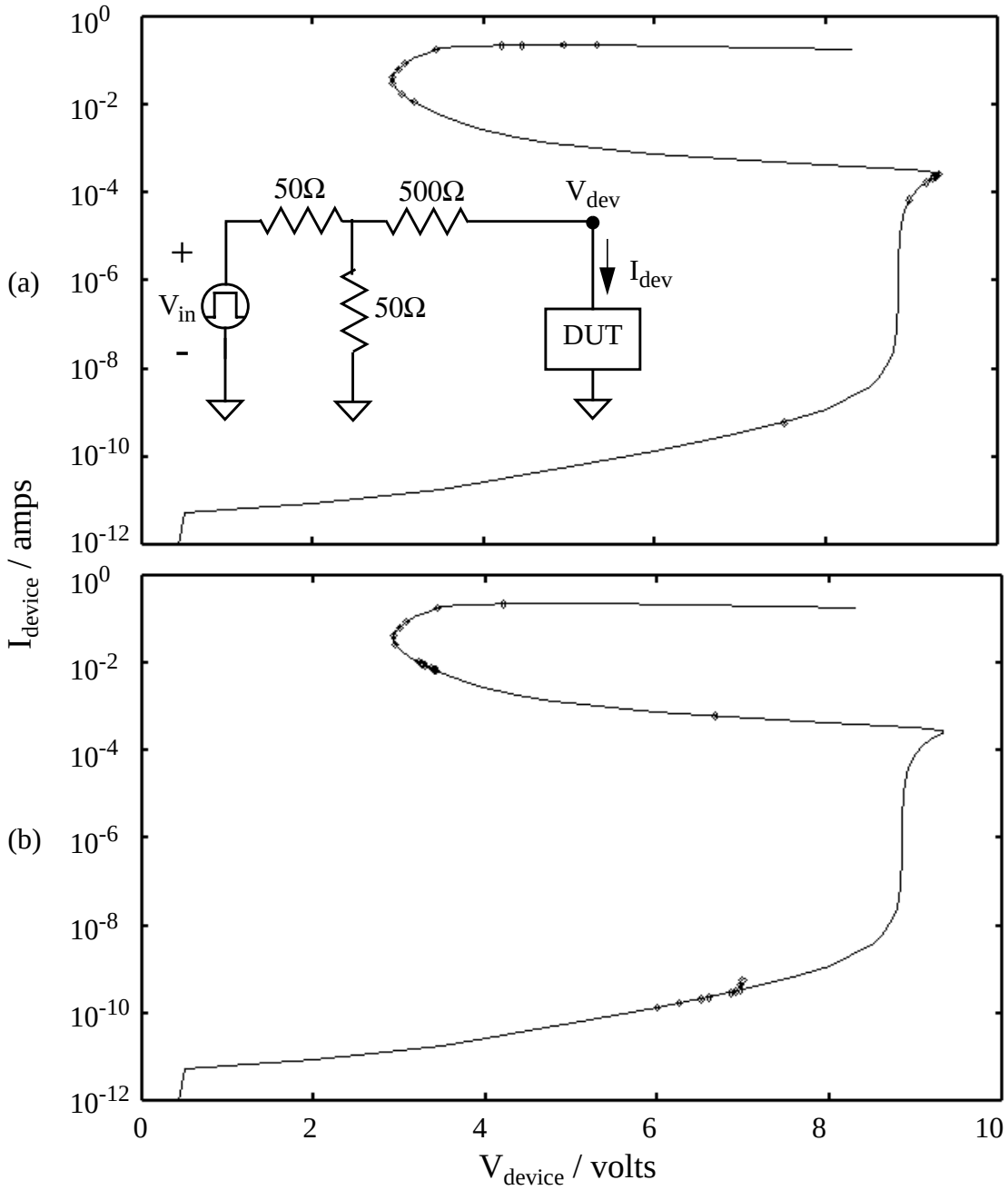

Fig. 3.29 *I-V curves for curve-tracing (solid line) and TLP (points) simulations for a 20/0.5μm MOSFET with (a) a 2000Ω gate resistor and (b) an 8000Ω gate resistor, with the TLP circuit shown inset. Each point represents one non-catastrophic (maximum temperature < 1688K) 100ns TLP simulation with a 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.*

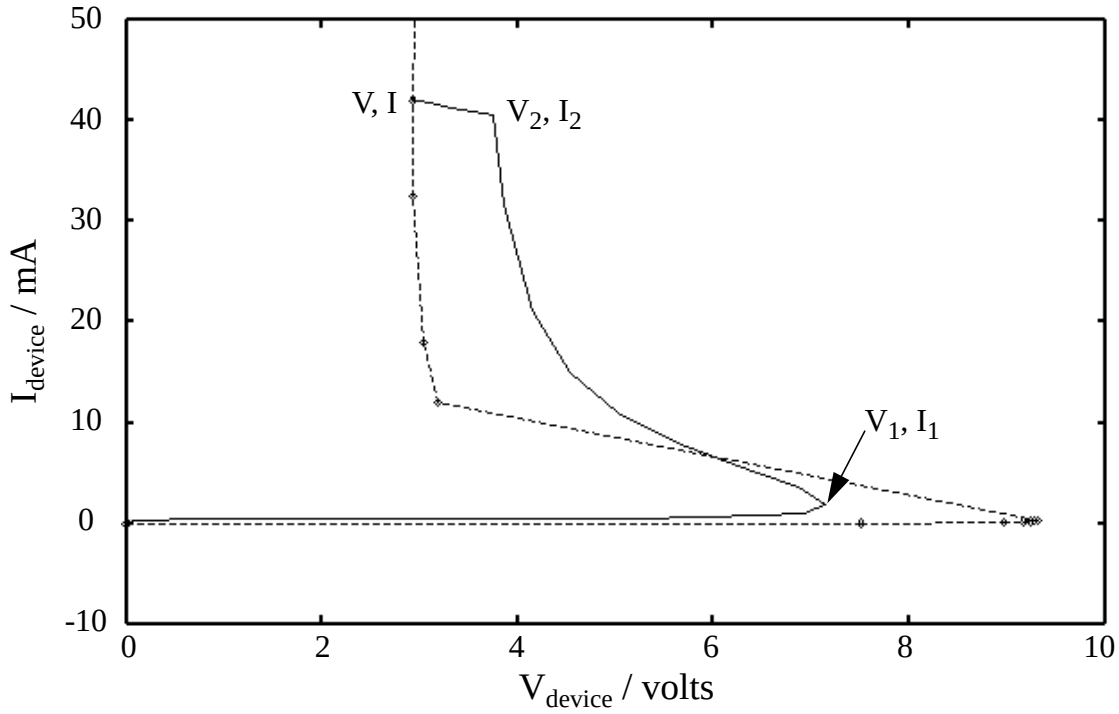

Fig. 3.30 *I-V curves of a single TLP simulation (solid line, V_{in}=50V in the circuit of Fig. 3.27) and of points resulting from a group of TLP simulations (dashed line). Point (V_1, I_1) corresponds to the turn-on of the parasitic bipolar transistor. V_2 and I_2 are the device voltage and current 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

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

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

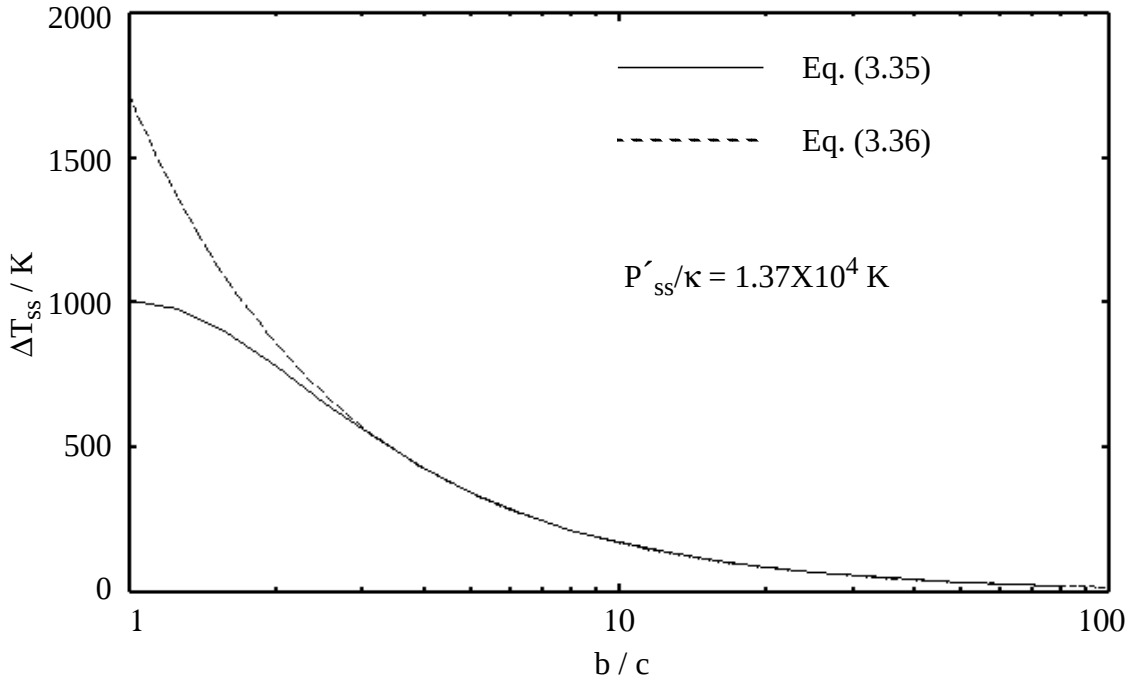

Fig. 3.31 The dependence of the steady-state change in peak temperature, ΔT_{ss}, on b/c *(log scale) as described by Eq. (3.35) is approximated by Eq. (3.36).*

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 $c = 2.0\mu m$),

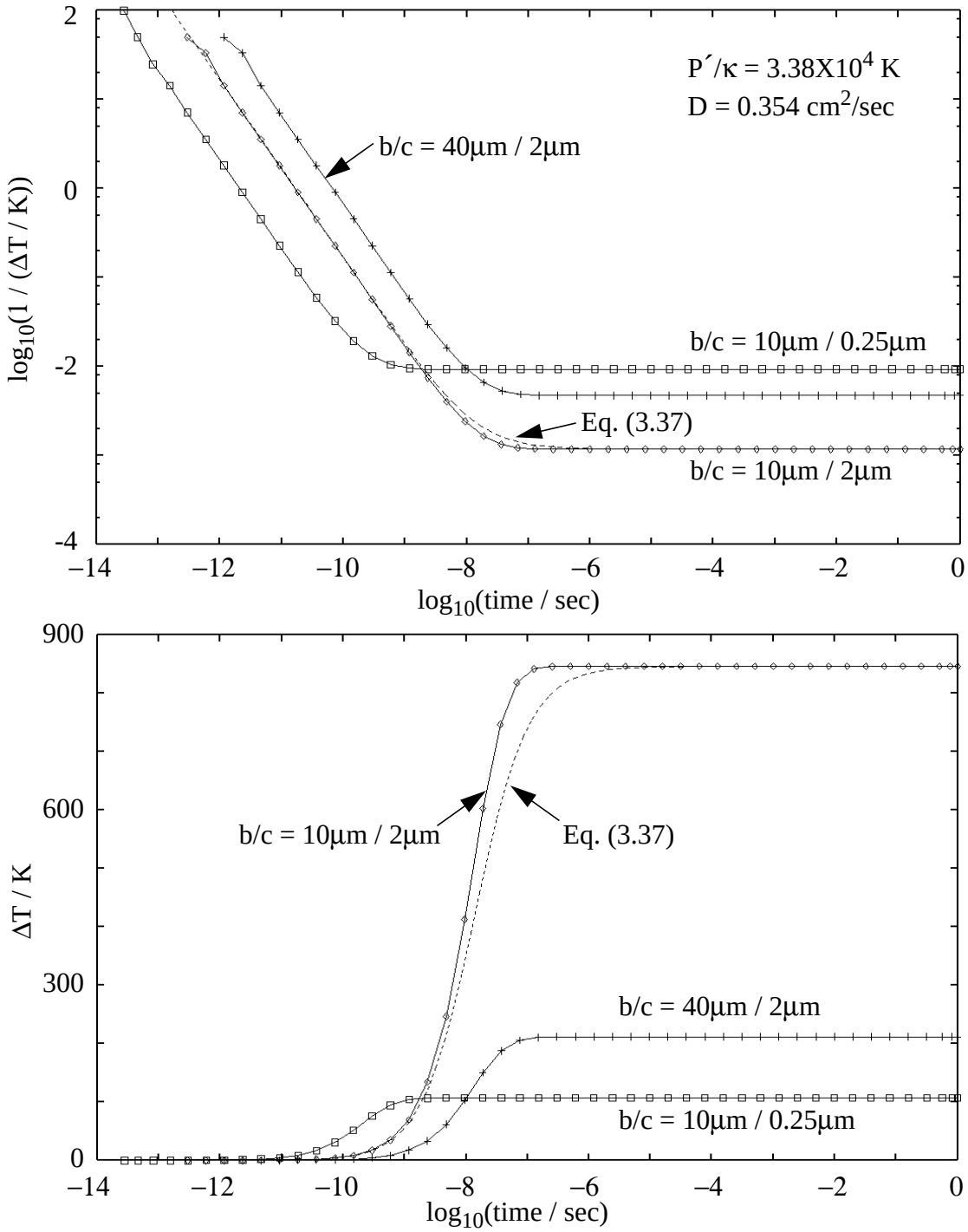

Fig. 3.32 Simulated 1/ΔT vs. time (a) and ΔT vs. time (b) curves for various length/ width ratios in a uniformly doped semiconductor region with a constant applied power. The temperature on the perimeter of the rectangular device is fixed at 300K. Eq. (3.37), the analytic approximation, is plotted for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

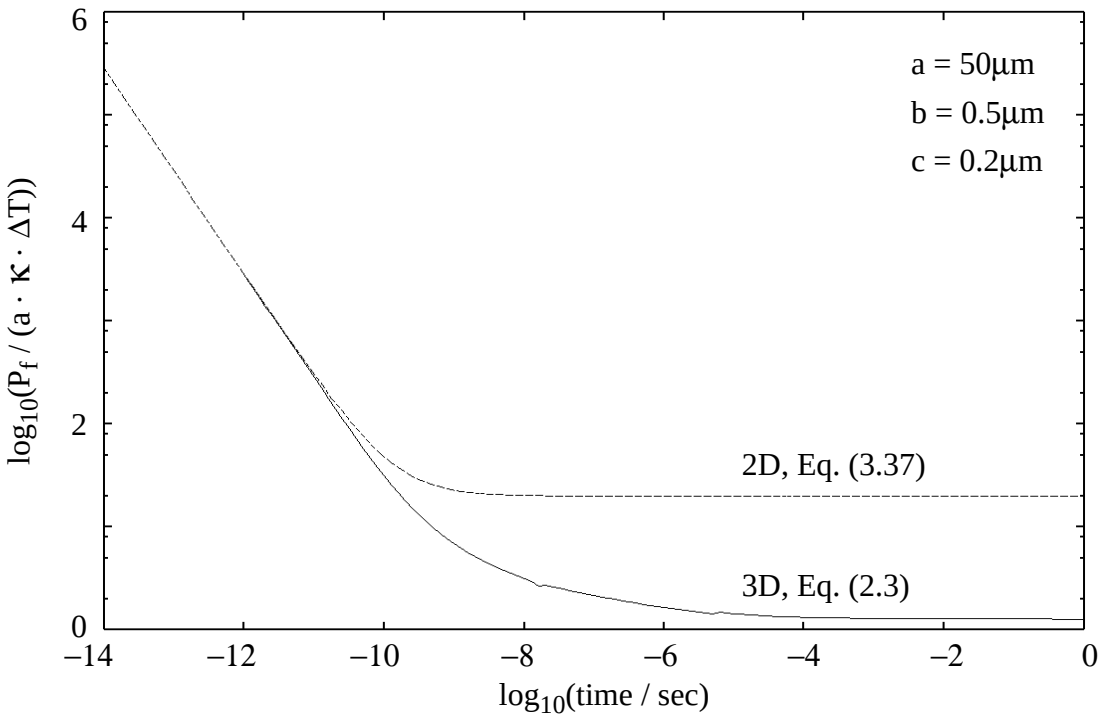

Fig. 3.33 *Power to failure, normalized by a, κ, and ΔT, is plotted vs. time to failure for the 2D and 3D implementations of the thermal box model. The time constants for the given box dimensions are t_a = 5.6μs, t_b = 560ps, and t_c = 90ps.*

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:

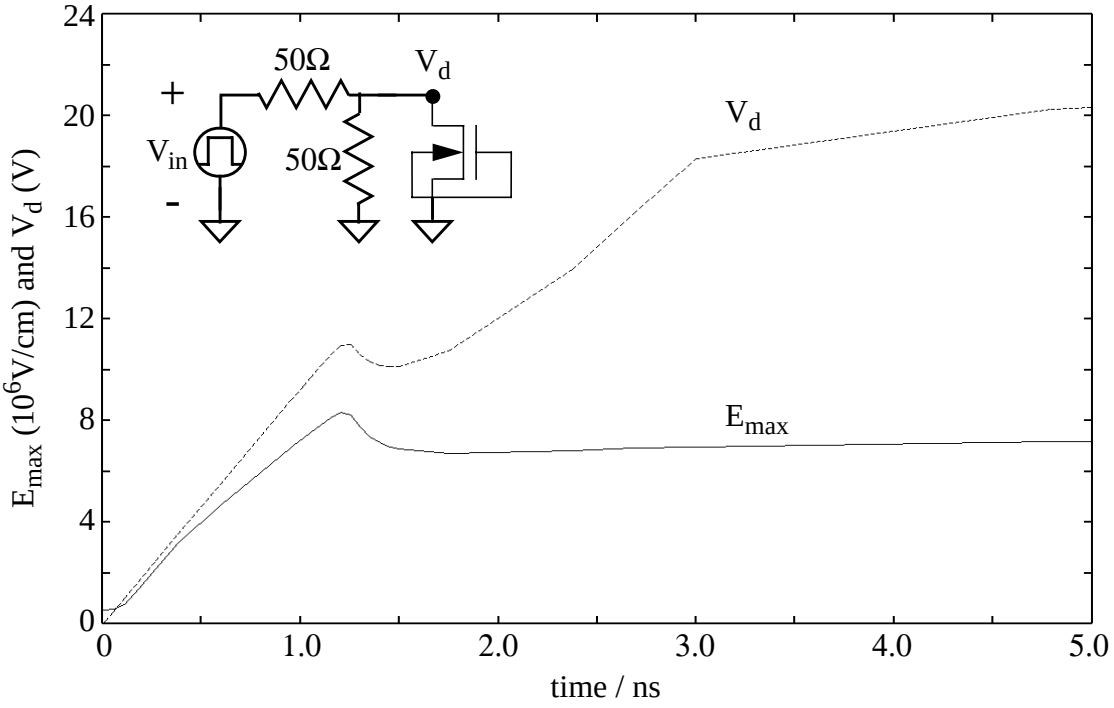

Fig. 3.34 The maximum electric field in the gate oxide (E_{max}, in MV/cm) of an ESD-protection MOSFET subjected to a square pulse with a 3ns rise time is plotted vs. time. As seen from the plot of the input voltage at the drain of the device, V_d, the reduction in E_{max} is due to the device snapping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

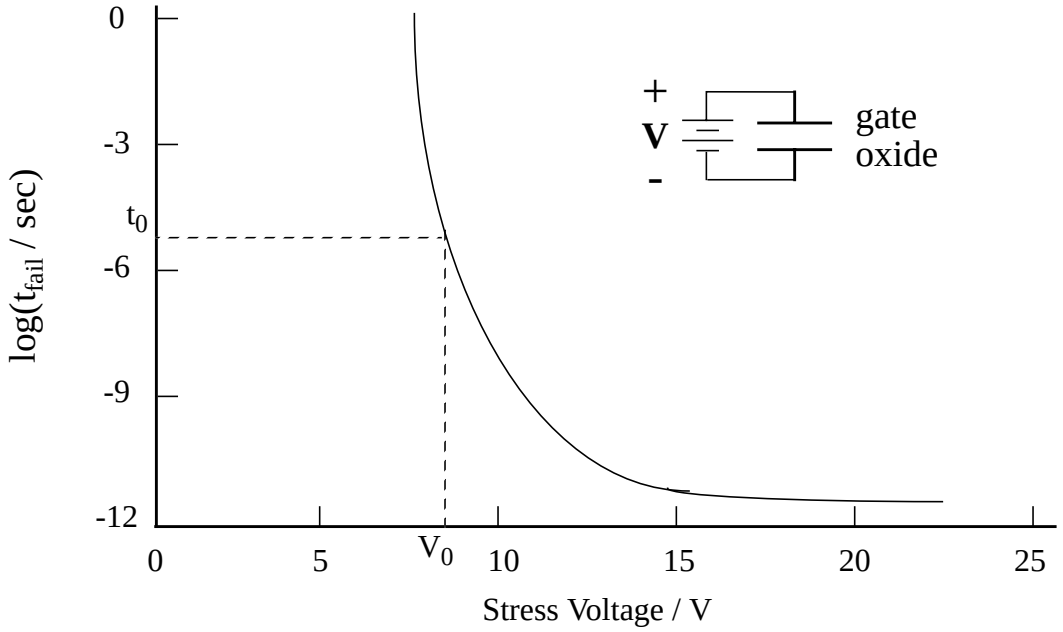

Fig. 3.35 A qualitative plot of time-to-failure vs. stress voltage reveals the time-dependent dielectric breakdown behavior of a gate oxide. If the voltage V_0 is applied across the oxide for a time greater than t_0, the oxide will rupture.

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

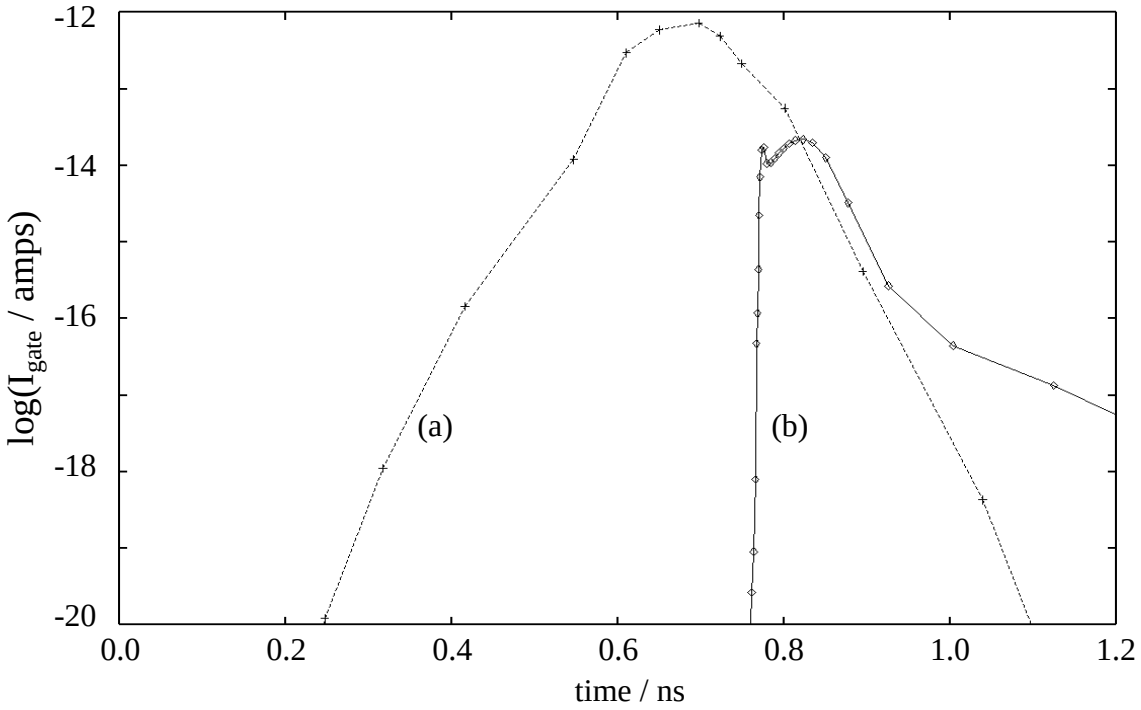

Fig. 3.36 *Gate current vs. time for 50/0.75μm MOSFETs with (a) 10KΩ gate resistor and (b) grounded gate. The drain is subjected to a square-wave pulse with a rise time of 3ns as depicted in the inset of Fig. 3.34. For both structures, the peak in gate current coincides with the time snapback occurs.*

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

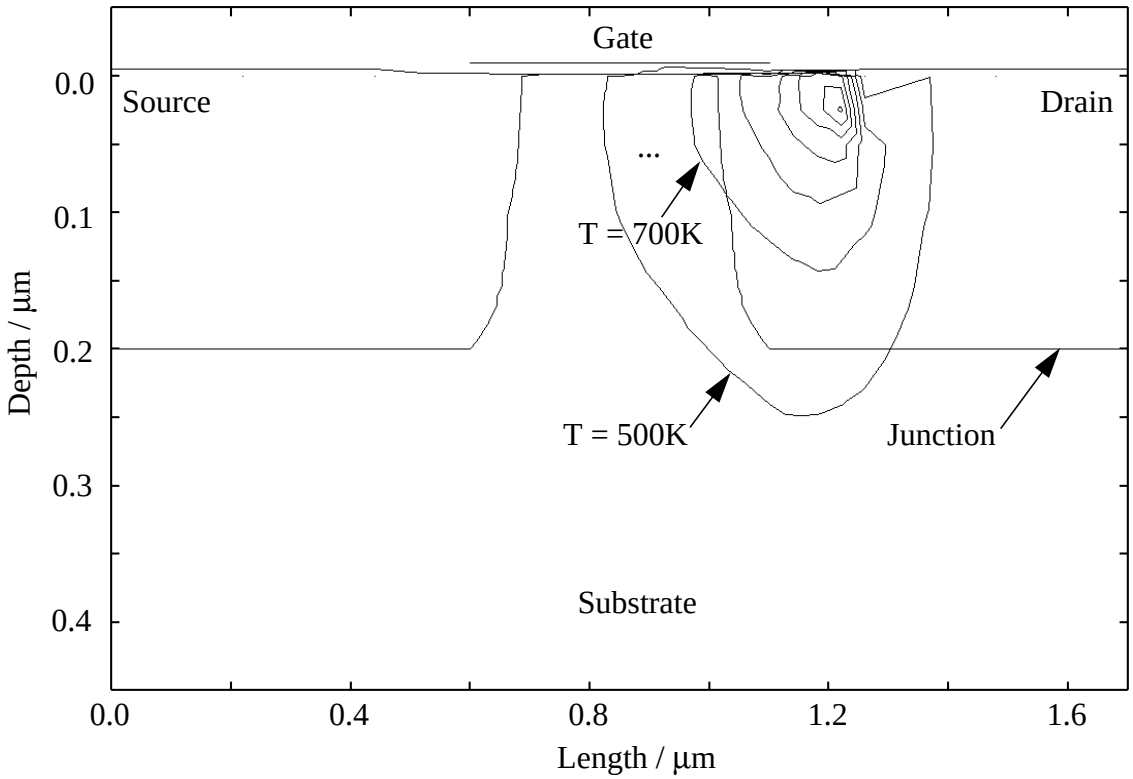

Fig. 3.37 *A constant-temperature contour is plotted for every 200K increment in temperature for a simulation structure at the time of peak ESD stress. Lines are also drawn marking the source and drain junctions of the structure, which is not plotted to scale.*

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

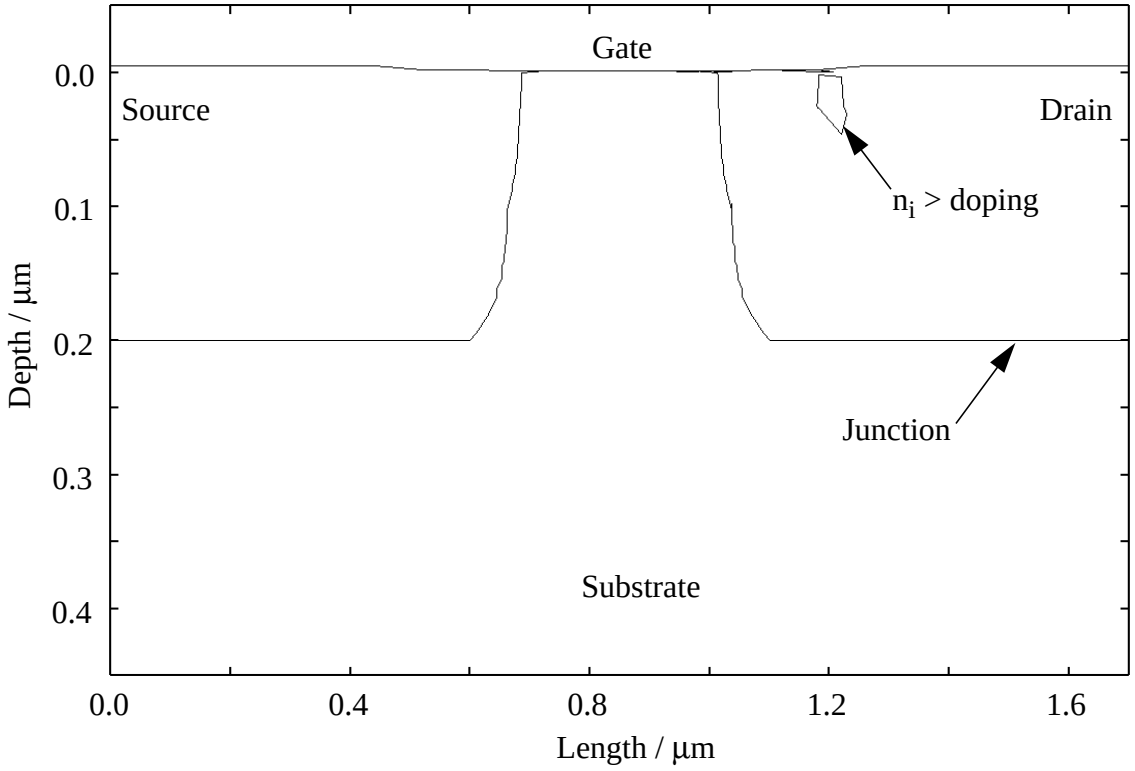

Fig. 3.38 A contour within which the intrinsic carrier concentration, n_i, is greater than the background doping level is drawn for a simulation structure at the time of peak ESD stress. Lines are also drawn marking the source and drain junctions of the structure, which is not plotted to scale.

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

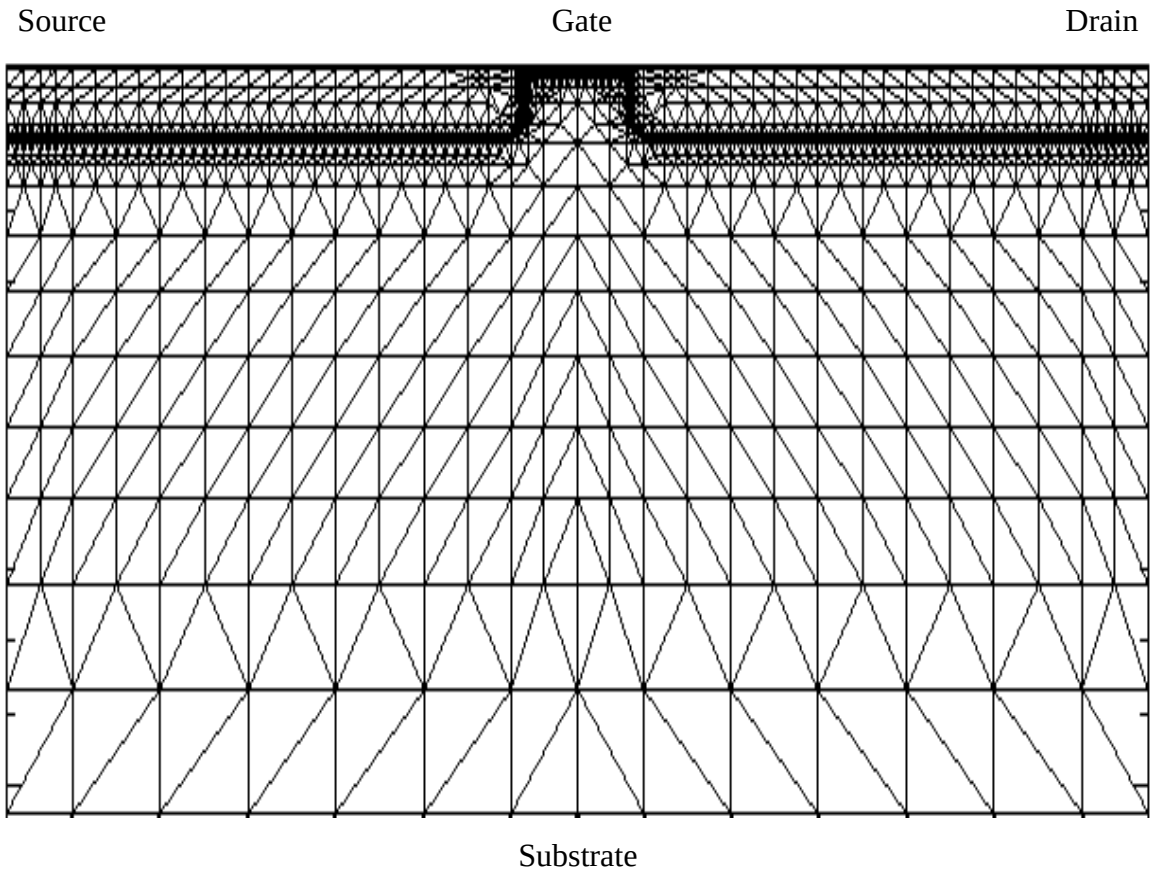

Fig. 4.39 *This example of a MEDICI-generated grid shows the concentration of grid points in the channel, LDD, and junction regions. Several microns of the substrate portion of the simulated structure have been omitted in order to display the grid approximately to scale.*

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.

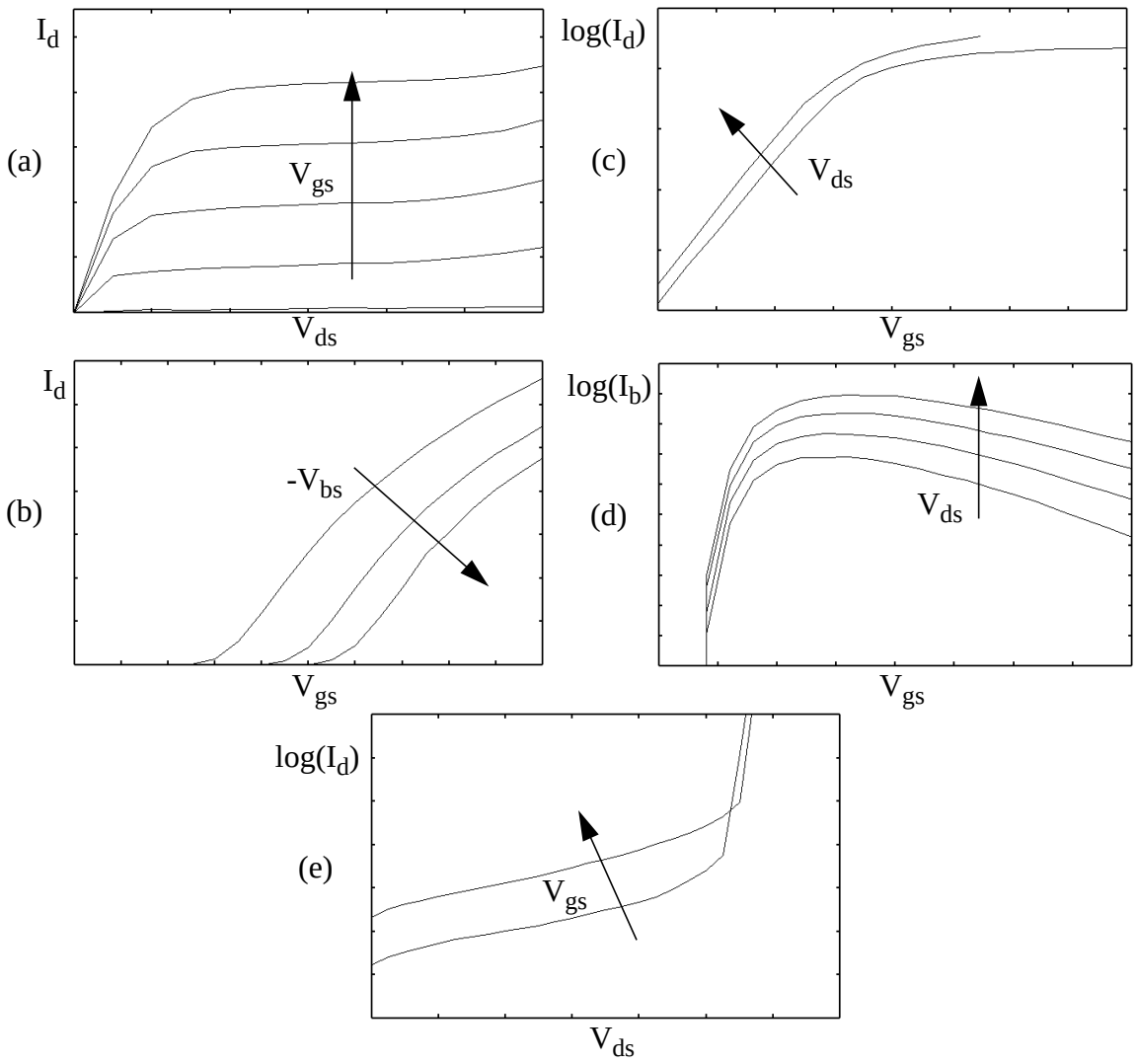

Fig. 4.40 Qualitative depiction of I-V curves used for MOSFET calibration: (a) drain: I_d vs. V_{ds} for stepped V_{gs}; (b) gate: I_d vs. V_{gs} for stepped V_{bs}; (c) sub-threshold: log(I_d) vs. V_{gs} for stepped V_{ds}; (d) substrate: log(I_b) vs. V_{gs} for stepped V_{ds}; (e) breakdown: log(I_d) 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.