INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book. These are also available as one exposure on a standard 35mm slide or as a 17" x 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600
Electrical and thermal modeling of electrostatic discharge protection structures for submicron VLSI

Stiegler, Harvey J., Ph.D.
Rice University, 1989
RICE UNIVERSITY

ELECTRICAL AND THERMAL MODELING OF ELECTROSTATIC DISCHARGE PROTECTION STRUCTURES FOR SUBMICRON VLSI

by

HARVEY J. STIEGLER

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE

W. L. Wilson, Jr., Professor of Electrical and Computer Engineering, Director

T. A. Rabson, Professor of Electrical and Computer Engineering

S. A. Dodds, Associate Professor of Physics

Houston, Texas

May, 1989
ABSTRACT

ELECTRICAL AND THERMAL MODELING OF ELECTROSTATIC DISCHARGE PROTECTION STRUCTURES FOR SUBMICRON VLSI

by
Harvey J. Stiegler

A modeling technique has been developed which simulates a semiconductor device subjected to electrostatic discharge (ESD) stress according to the human body model (HBM). To accomplish this, a computer program was developed which solves the electron and hole continuity equations, Poisson’s equation, and the heat flow equation in one dimension. The program has been applied to npn structures typical of the parasitic bipolar devices found in MOS output stages.

Profiles from lightly-doped drain (LDD), double-diffused drain (DDD), and graded drain (GD) device structures were investigated. The performance of these various profiles under ESD stress has been compared in order to understand their functioning and to determine the important design parameters. It is found that device heating is reduced for structures in which the doping profile rises steeply to a high concentration in the drain region near the metallurgical junction. The rate of heating is related to reduced carrier saturation velocity due to local heating and its effects on charge distribution, electric field, and total potential drop across the reverse-biased junction.
The modeling technique presented gives results which are in reasonable agreement with measured data. This technique should be a useful tool for evaluating new device structures, fabrication processes, or process changes before committing to the costly and time-consuming process of actual device fabrication.
ACKNOWLEDGEMENTS

Many people have helped to make this work possible. First, I would like to thank my advisor Dr. Bill Wilson for the time and effort he has devoted to this work, for his helpful ideas and suggestions, and for allowing me to pursue a research topic of my own choosing. I would also like to thank the management and staff of Texas Instruments, Inc., who provided support and computer resources which made this work possible. In particular, I would like to thank Bob Rountree of T.I. whose efforts in the field provided the original idea and inspiration for this research. Bob also provided encouragement and many helpful ideas along the way. Dr. Charvaka Duvvury, also of T.I., provided a number of very helpful discussions as well as data on actual devices for validating the simulations. The time devoted by my committee members to reading the manuscript and offering suggestions is also gratefully acknowledged. Finally, I would like to thank my wife Sandra for her patience, understanding, and encouragement during this research work and during the preparation of this manuscript.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Background</td>
<td>7</td>
</tr>
<tr>
<td>2.1. Protection Devices</td>
<td>7</td>
</tr>
<tr>
<td>2.2. Bipolar Breakdown</td>
<td>13</td>
</tr>
<tr>
<td>2.3. Failure Modes</td>
<td>18</td>
</tr>
<tr>
<td>3. Simulation Method</td>
<td>22</td>
</tr>
<tr>
<td>3.1. Basic Semiconductor Equations</td>
<td>22</td>
</tr>
<tr>
<td>3.2. Models</td>
<td>24</td>
</tr>
<tr>
<td>3.3. Numerical Techniques</td>
<td>28</td>
</tr>
<tr>
<td>4. Data and Analysis</td>
<td>42</td>
</tr>
<tr>
<td>4.1. Structure Descriptions</td>
<td>42</td>
</tr>
<tr>
<td>4.2. Simulation Conditions</td>
<td>53</td>
</tr>
<tr>
<td>4.3. General Performance Analysis</td>
<td>53</td>
</tr>
<tr>
<td>4.4. Functional Analysis</td>
<td>65</td>
</tr>
<tr>
<td>4.5. Comments on the Validity of the Simulations</td>
<td>101</td>
</tr>
<tr>
<td>5. Conclusions and Directions for Further Work</td>
<td>104</td>
</tr>
<tr>
<td>Appendix</td>
<td>107</td>
</tr>
<tr>
<td>References</td>
<td>110</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>Human body model and machine model test circuits</td>
</tr>
<tr>
<td>2</td>
<td>Schematic of typical MOS input and output circuits</td>
</tr>
<tr>
<td>3</td>
<td>Circuit and cross section of a typical input protection structure</td>
</tr>
<tr>
<td>4</td>
<td>Cross section of a typical CMOS output structure</td>
</tr>
<tr>
<td>5</td>
<td>Parasitic npn bipolar transistor in breakdown under HBM stress</td>
</tr>
<tr>
<td>6</td>
<td>Doping density, charge, and electric field in the space charge region</td>
</tr>
<tr>
<td>7</td>
<td>The location of the simulation profile in an output structure</td>
</tr>
<tr>
<td>8</td>
<td>The structure of lightly doped drain, double diffused drain, and graded drain devices</td>
</tr>
<tr>
<td>9</td>
<td>Cross-sectional view of Structure 1 (LDD)</td>
</tr>
<tr>
<td>10</td>
<td>Doping Profile 1 (LDD)</td>
</tr>
<tr>
<td>11</td>
<td>Cross-sectional view of Structure 2 (DDD)</td>
</tr>
<tr>
<td>12</td>
<td>Doping Profile 2 (DDD)</td>
</tr>
<tr>
<td>13</td>
<td>Cross-sectional view of Structure 3 (DDD)</td>
</tr>
<tr>
<td>14</td>
<td>Doping Profile 3 (DDD)</td>
</tr>
<tr>
<td>15</td>
<td>Cross-sectional view of Structure 4 (GD)</td>
</tr>
<tr>
<td>16</td>
<td>Doping Profile 4 (GD)</td>
</tr>
<tr>
<td>17</td>
<td>Cross-sectional view of Structure 5 (GD)</td>
</tr>
<tr>
<td>18</td>
<td>Doping Profile 5 (GD)</td>
</tr>
<tr>
<td>19</td>
<td>External circuit for simulations</td>
</tr>
<tr>
<td>20</td>
<td>Temperature distribution for Profile 1 during HBM stress</td>
</tr>
<tr>
<td>21</td>
<td>Temperature distribution for Profile 2 during HBM stress</td>
</tr>
<tr>
<td>22</td>
<td>Temperature distribution for Profile 3 during HBM stress</td>
</tr>
<tr>
<td>23</td>
<td>Temperature distribution for Profile 4 during HBM stress</td>
</tr>
<tr>
<td>24</td>
<td>Temperature distribution for Profile 5 during HBM stress</td>
</tr>
<tr>
<td>25</td>
<td>Potential distribution for Profile 1, t = 0 to 1 ns</td>
</tr>
<tr>
<td>26</td>
<td>Potential distribution for Profile 2, t = 0 to 1 ns</td>
</tr>
<tr>
<td>27</td>
<td>Potential distribution for Profile 3, t = 0 to 1 ns</td>
</tr>
<tr>
<td>28</td>
<td>Potential distribution for Profile 4, t = 0 to 1 ns</td>
</tr>
<tr>
<td>29</td>
<td>Potential distribution for Profile 5, t = 0 to 1 ns</td>
</tr>
<tr>
<td>30</td>
<td>Log maximum temperature versus time for Profiles 1-5</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>31. $\tau_{10}$ versus $V_{ab}$</td>
<td>63</td>
</tr>
<tr>
<td>32. Carrier saturation velocity versus temperature</td>
<td>66</td>
</tr>
<tr>
<td>33. Critical current density versus doping density for various temperatures</td>
<td>66</td>
</tr>
<tr>
<td>34. $J_o$ for Profile 1</td>
<td>68</td>
</tr>
<tr>
<td>35. $J_o$ for Profile 2</td>
<td>68</td>
</tr>
<tr>
<td>36. $J_n$ for Profile 1</td>
<td>69</td>
</tr>
<tr>
<td>37. $J_p$ for Profile 1</td>
<td>69</td>
</tr>
<tr>
<td>38. $J_n$ for Profile 2</td>
<td>70</td>
</tr>
<tr>
<td>39. $J_p$ for Profile 2</td>
<td>70</td>
</tr>
<tr>
<td>40. $J_n/J_o$ for Profile 1, $t = 0$ to 48 ns</td>
<td>72</td>
</tr>
<tr>
<td>41. $J_n/J_o$ for Profile 2, $t = 0$ to 100 ns</td>
<td>72</td>
</tr>
<tr>
<td>42. $J_n/J_o$ for Profile 3, $t = 0$ to 100 ns</td>
<td>73</td>
</tr>
<tr>
<td>43. $J_n/J_o$ for Profile 4, $t = 0$ to 30 ns</td>
<td>73</td>
</tr>
<tr>
<td>44. $J_n/J_o$ for Profile 5, $t = 0$ to 100 ns</td>
<td>74</td>
</tr>
<tr>
<td>45. Charge distribution for Profile 1, $t = 1$ to 20 ns</td>
<td>76</td>
</tr>
<tr>
<td>46. Electric field distribution for Profile 1, $t = 1$ to 20 ns</td>
<td>76</td>
</tr>
<tr>
<td>47. Potential distribution for Profile 1, $t = 1$ to 20 ns</td>
<td>77</td>
</tr>
<tr>
<td>48. Electron density for Profile 1, $t = 1$ to 20 ns</td>
<td>77</td>
</tr>
<tr>
<td>49. Hole density for Profile 1, $t = 1$ to 20 ns</td>
<td>78</td>
</tr>
<tr>
<td>50. Charge distribution for Profile 1, $t = 20$ to 36 ns</td>
<td>80</td>
</tr>
<tr>
<td>51. Electric field distribution for Profile 1, $t = 20$ to 36 ns</td>
<td>80</td>
</tr>
<tr>
<td>52. Potential distribution for Profile 1, $t = 20$ to 36 ns</td>
<td>81</td>
</tr>
<tr>
<td>53. Electron density for Profile 1, $t = 20$ to 36 ns</td>
<td>81</td>
</tr>
<tr>
<td>54. Hole density for Profile 1, $t = 20$ to 36 ns</td>
<td>82</td>
</tr>
<tr>
<td>55. Charge distribution for Profile 1, $t = 36$ to 48 ns</td>
<td>83</td>
</tr>
<tr>
<td>56. Electric field distribution for Profile 1, $t = 36$ to 48 ns</td>
<td>83</td>
</tr>
<tr>
<td>57. Potential distribution for Profile 1, $t = 36$ to 48 ns</td>
<td>84</td>
</tr>
<tr>
<td>58. Electron density for Profile 1, $t = 36$ to 48 ns</td>
<td>84</td>
</tr>
<tr>
<td>59. Hole density for Profile 1, $t = 36$ to 48 ns</td>
<td>85</td>
</tr>
<tr>
<td>60. Charge distribution for Profile 2, $t = 5$ to 100 ns</td>
<td>87</td>
</tr>
<tr>
<td>61. Electric field distribution for Profile 2, $t = 5$ to 100 ns</td>
<td>87</td>
</tr>
<tr>
<td>62. Potential distribution for Profile 2, $t = 5$ to 100 ns</td>
<td>88</td>
</tr>
<tr>
<td>63. Electron density for Profile 2, $t = 5$ to 100 ns</td>
<td>88</td>
</tr>
<tr>
<td>64. Hole density for Profile 2, $t = 5$ to 100 ns</td>
<td>89</td>
</tr>
<tr>
<td>65. Charge distribution for Profile 3, $t = 5$ to 100 ns</td>
<td>90</td>
</tr>
<tr>
<td>66. Electric field distribution for Profile 3, $t = 5$ to 100 ns</td>
<td>90</td>
</tr>
<tr>
<td>67. Potential distribution for Profile 3, $t = 5$ to 100 ns</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Description</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>68.</td>
<td>Electron density for Profile 3, t = 5 to 100 ns</td>
</tr>
<tr>
<td>69.</td>
<td>Hole density for Profile 3, t = 5 to 100 ns</td>
</tr>
<tr>
<td>70.</td>
<td>Charge distribution for Profile 4, t = 2 to 30 ns</td>
</tr>
<tr>
<td>71.</td>
<td>Electric field distribution for Profile 4, t = 2 to 30 ns</td>
</tr>
<tr>
<td>72.</td>
<td>Potential distribution for Profile 4, t = 2 to 30 ns</td>
</tr>
<tr>
<td>73.</td>
<td>Electron density for Profile 4, t = 2 to 30 ns</td>
</tr>
<tr>
<td>74.</td>
<td>Hole density for Profile 4, t = 2 to 30 ns</td>
</tr>
<tr>
<td>75.</td>
<td>Charge distribution for Profile 5, t = 5 to 100 ns</td>
</tr>
<tr>
<td>76.</td>
<td>Electric field distribution for Profile 5, t = 5 to 100 ns</td>
</tr>
<tr>
<td>77.</td>
<td>Potential distribution for Profile 5, t = 5 to 100 ns</td>
</tr>
<tr>
<td>78.</td>
<td>Electron density for Profile 5, t = 5 to 100 ns</td>
</tr>
<tr>
<td>79.</td>
<td>Hole density for Profile 5, t = 5 to 100 ns</td>
</tr>
</tbody>
</table>
List of Tables

1. Coefficients For the Mobility Model Equations ........................................ 25
2. Coefficients For the Impact Ionization Model ........................................ 26
3. Values of Some Constants for Silicon .................................................. 28
4. Summary of Simulation Profiles ............................................................ 53
5. Summary of Snap Back Voltage and Temperature Performance .............. 62
6. Profile Heating Time Constants and Snap Back Voltages ....................... 64
7. Snap Back Potential Comparison by Structure Type .................................. 102
8. Performance Properties of Structures 3 through 5 .................................. 103
CHAPTER 1

Introduction

With the maturation of integrated circuit technology in the VLSI era, two requirements have been demanded of those who design and fabricate modern circuits. The first is ever more aggressive technology, driven by the requirement to put ever faster and more complex circuits on silicon. The second is increased reliability. Today’s integrated circuits are expected to withstand handling and operation in harsh environments and to operate for many years without degradation.

One environmental hazard to which integrated circuits are exposed which has received increasing attention in recent years is that of exposure to electrostatic discharge (ESD). A packaged integrated circuit (IC) can be exposed to the buildup and discharge of high electrostatic potentials during several phases of its life, including handling by humans, shipping, and handling by automated testing or assembly equipment. Since metal-oxide-semiconductor (MOS) IC’s are fabricated with silicon dioxide dielectrics and semiconductor junctions with breakdown voltages on the order of tens of volts, it is clear that exposure to even a minimal discharge of static electricity could have serious consequences to internal structures.

An early impetus to improve ESD protection was given by the inclusion of a specification for protection levels in a U.S. military standard, MIL-STD-883. This
standard attempted to divide "sensitive" from "nonsensitive" devices by specifying a testing method and a protection level to be achieved. The level of protection required was 2 kV, and the testing method has become widely used and is known as the "human body model," or HBM, because it attempts to duplicate the conditions experienced in human handling of sensitive components. The human body model consists of a 100 pF capacitor, charged to the test voltage, which can be connected to the device under test through a 1.5 kΩ series resistor and a switch. This circuit is illustrated in Figure 1(a).

The human body model circuit, however, has not proved entirely satisfactory for evaluating exposure to ESD in all circumstances. In particular, discharges experienced in automatic handling equipment, whether for device testing or end-equipment assembly, are unlikely to include a large resistance in the discharge path. This has led to the development of an alternative testing circuit, the machine model (MM), which is illustrated in Figure 1(b). The machine model consists of a 200 pF charged capacitor which is applied directly to the device under test. This testing method attempts to reduce series resistance to the minimum possible value and due to the lack of an external current limiting resistor typically results in much lower measured failure voltages than does the HBM. In fact, the difference in the nature of the testing method considered may place different demands on the design of the internal protection structure; and designing a structure to perform well on both tests is a formidable challenge.
Figure 1(a). Human body model test circuit.

Figure 1(b). Machine model test circuit.
Regardless of whether the HBM or MM is used, testing proceeds in a similar manner. The external capacitor is charged to the test voltage, the switch is closed to apply stress to the device under test, and the part is examined for evidence of failure. Since the purpose of the test is usually to determine the failure threshold of a given device, the test begins with a low voltage which is gradually increased for each succeeding pulse until evidence of device degradation or failure is observed. Device failure is generally indicated when excess leakage current or an internal short is detected for a given device terminal. Because this is a destructive test, it is normally employed only to qualify a device or process and is not intended for production use. Testing is often done by applying stress between a device input or output pin and a power supply pin, usually ground. However, it is also valid to test a device by applying stress between combinations of pins, neither of which is a power supply. In such cases, complex discharge paths may result which are beyond the scope of this work. Here it will be assumed that the stress is applied between a single input or output terminal and ground.

As experience has increased and expertise in the design of ESD protection circuits has improved, the level of discharge to which an IC can be safely exposed has steadily increased. The designer's task has been complicated, however, by the advance of MOS technology. The pursuit of higher performance has brought structures fabricated with thinner dielectrics and smaller feature sizes. At the same time, expectations of protection levels have not decreased but in fact have continued to increase. Scaling of MOS technology to the level of one-micron and submicron
active devices has brought further process innovations designed to counter certain performance and reliability problems. Diffusion areas clad with refractory metal silicides have been introduced to reduce sheet resistances and improve the speed performance of circuits by reducing RC delay. Another feature, known as the graded drain (GD) or lightly-doped drain (LDD), has also been introduced to counter the increased electric field within active MOS devices brought on by reduced dimensions and the subsequent increase in device degradation due to the generation of hot carriers. Unfortunately, both of these process innovations have been found to degrade the ESD performance of devices, especially those employed in output structures, although the reasons for this have not been well-understood [1, 2, 3].

Even though much progress has been made in improving the performance of ESD protection structures, understanding of the physical principles involved in device failure has not advanced rapidly and most progress has been made by empirical design. Typically, improvement in protection structures has been made either by iterative trial and error design or by producing specialized test vehicles with a large array of varied structures. Testing and failure analysis have been used to empirically determine the best structures from those tested, and proposed improvements based on experience have been incorporated in later designs for further evaluation. There are two drawbacks to this method. First, it leads to the accumulation of large amounts of empirical data but relatively modest physical understanding of the failures. Second, the cycle time required for an iteration in the design loop can range from a few weeks to many months.
The goal of this work is to provide a means for evaluating the performance of ESD protection structures and understanding their operation and failure mechanisms by simulation. In order to accomplish this, a computer program was designed which performs a transient solution of the semiconductor equations, including the heat-flow equation, in one dimension. The program has been applied to doping profiles which have been extracted from the structures of actual test devices. It was thus possible to compare the simulated results with actual measured data in order to check the validity of the analysis.

Chapter Two of this work discusses types of protection devices, their characteristics, and operating modes. Some background on bipolar avalanche breakdown, second breakdown, and failure modes is also included in Chapter Two. Chapter Three discusses the solution of the semiconductor equations, the models used for the various material parameters, and the numerical methods used. Chapter Four contains the actual simulation data, analysis of the data, and discussion of the validity of the simulations. Finally, conclusions and suggestions for further work are presented in Chapter Five.
CHAPTER 2

Background

2.1. Protection Devices

Within modern VLSI circuits there exist several device structures which may be subject to damage from externally applied electrostatic discharge. However, the most sensitive of these is generally the gate oxide of MOS devices. Other components which may be damaged by exposure to ESD include semiconductor junctions; polysilicon, aluminum, or other metal interconnect; and contacts between metallization and the underlying silicon layers. The primary purpose of protection structures is to prevent the rupture of the MOS gate dielectric due to excessive electric field and to protect other elements of the circuitry from damage due to high potentials or heating effects.

Within the package of an integrated circuit, the package pins are connected to bonding pads on the surface of the chip via small bonding wires. Although some pins may perform a double function, the external connections of an IC can generally be classified as either inputs or outputs. The characteristics of input and output pins are quite different, and the problem of providing adequate ESD protection is quite different between the two. This derives from the differences in circuit connection as illustrated in Figure 2. In an input circuit (Figure 2(a)), the bonding pad is connected
Figure 2(a). A typical MOS input circuit.

Figure 2(b). A typical MOS output circuit.
to the insulated gate of an MOS device. In the absence of a protection circuit there is no conducting path to dissipate charge short of a breakdown in the MOSFET's gate dielectric. However, the task of protecting the input is made easier by the fact that the input device is usually small, does not source or sink DC current, and the introduction of parasitic protection elements usually does not produce noticeable performance degradation. The output circuit (Figure 2(b)), on the other hand, connects the source and drain of large driver devices to the bonding pad. Since the output must drive large capacitive loads through rapid voltage swings, it is particularly sensitive to protection circuits which introduce parasitic resistance into the circuit. Because output devices must be very tightly coupled to the pad, they are particularly difficult to protect.

A wide variety of devices has been examined for use as protection structures. In NMOS processes, n⁺p diodes and lateral npn transistors have been widely used. The n⁺p diode provides good protection against negative input stress, shunting the charge to substrate, but is not effective against positive input stress. Lateral npn transistors can be used in several forms including thick-oxide devices, formed from the parasitic bipolar under the isolation field oxide; thin oxide devices, formed from the parasitic bipolar under an n-channel MOSFET; and field-plate diodes, similar to the thin oxide device but with the gate permanently grounded [3, 4, 5]. With the emergence of CMOS as the prevalent technology, p⁺n diodes, lateral and vertical pnp transistors, and SCRs have become available for use. In conventional p-substrate n-well CMOS technology, however, structures such as the p⁺n diode which rely on conduction
through the n-well may not prove particularly effective due to the high resistance of the n-well [3, 6].

Input structures typically have achieved good protection for a number of years with thick-oxide-type lateral npn protection devices which are usually followed by a resistance and a field plate diode secondary protection structure [4, 5]. The circuit diagram and cross section of such a device is illustrated in Figure 3. Such devices are particularly effective in conjunction with abrupt-junction fabrication processes. More recently, with the emergence of CMOS process variations which taper the junction profile, the lateral SCR has found increasing use as an input protection device [7, 8].

Output structures, on the other hand, are unable to take advantage of the thick-field or lateral SCR protection devices because of the more rapid turn on and lower breakdown voltage of the parasitic npn associated with the n-channel device. In CMOS output buffers, the p⁺n diode between the output and the n-well may provide some protection by conducting charge to the positive power supply node during positive stress; but its usefulness is usually limited by the series resistance of the n-well, as mentioned previously, and the finite capacitance from the positive power supply to ground which provides a limited charge sink [6, 9]. For these reasons, the output transistors are, in effect, their own protection structure and the central element is the n-channel device. Empirical study has yielded layout guidelines for achieving optimum results with a given technology, and a typical structure is shown in cross section in Figure 4. However, the fundamental limitation to output ESD protection is the functioning of the n-channel device and its parasitic npn bipolar transistor under
Figure 3(a). Circuit diagram of a typical input protection structure.

Figure 3(b). Cross section of a typical input protection structure.
Figure 4. Cross section of a typical CMOS output structure.
stress; and these basic characteristics are set by the fabrication process used. The parasitic npn bipolar transistor associated with n-channel output structures is the primary device examined in this work.

The failure threshold of well-designed protection structures has been found to scale directly with device width in a very repeatable fashion, indicating that uniform conduction is the normal operating mode for stresses which the structure is able to tolerate. For structures involving either thick- or thin-oxide n-channel devices, the operation mode is acknowledged to be that of bipolar breakdown of the parasitic lateral npn transistor. As the applied voltage is increased, the collector-base junction of the npn transistor enters avalanche breakdown (figure 5), leading to the phenomenon known as "snap back" in which the terminal voltage is rapidly reduced as current conduction increases. If the applied stress is not excessive, this conduction will dissipate the necessary charge. However, if the stress is increased, the device may enter the operating region known as "second breakdown," which is usually followed by device degradation or catastrophic failure.

2.2. Bipolar Breakdown

The operation of parasitic bipolar devices in ESD protection structures is equivalent to the CEO mode of operation, that is, collector-emitter breakdown with the base terminal open-circuited. This is a well-understood phenomenon, and a description can be found in many standard references [10, 11, 12]. When a positive potential is applied to the collector with respect to the emitter, most of the potential
Figure 5. Parasitic npn bipolar transistor in breakdown under HBM stress.
will appear across the reverse-biased collector-base junction; but the base-emitter junction will be slightly forward biased. The collector current will be the sum of the collector-base junction generation current, \( I_{CBO} \), and the injected current from the emitter which transits the base (\( \alpha I_E \)). Now, with \( I_B = 0 \), \( I = I_E = I_C = (\alpha I_E + I_{CBO}) \).

If the voltage is increased, avalanche multiplication will take place at the collector-base junction. The avalanche multiplication factor \( M \) is usually taken to be of the form

\[
M = \frac{1}{1 - \left( \frac{V}{BV_{CBO}} \right) \eta}
\]  

(1)

where \( BV_{CBO} \) is the breakdown voltage of the collector-base junction and \( \eta \) is a constant usually between 3 and 6. From equation (1) it is apparent that as \( V \) approaches \( BV_{CBO} \) the multiplication factor \( M \) becomes very large. Now, the current flowing in the device will be multiplied by \( M \) at the collector-base junction, which gives:

\[
M (\alpha I + I_{CBO}) = I
\]

or

\[
I = \frac{MI_{CBO}}{1 - \alpha M}
\]  

(2)

Thus, when \( \alpha M = 1 \), the current flowing in the device is unconstrained except by the external circuitry, resulting in a lower terminal voltage due to increased current conduction and a subsequent voltage drop across the internal impedance of the source.
The distribution of charge, electric field, and potential within a bipolar device experiencing breakdown will vary depending upon the magnitude of the current density being forced through the device. Dunn and Nuttall have provided an explanation of this operation within an ideal epitaxial npn bipolar transistor [13]. Figure 6(a) shows the doping profile of such an idealized device. This device consists of an \( n^+ \) collector contact, an \( n^- \) collector region, and a p-type base region. Not shown is the \( n^+ \) emitter region. When the transistor is unbiased or is biased such that only a small current flows, the situation is as depicted in Figure 6(b), where the donors and acceptors exposed by the absence of carriers in the depletion region gives rise to an electric field which peaks at the metallurgical junction. If the current density is increased until \( J = J_o = qN_{dc}v_s \), where \( N_{dc} \) is the collector doping density in the \( n^- \) region and \( v_s \) is the carrier saturation velocity, then the ionized donors in the formerly depleted region are completely neutralized as shown in Figure 6(c). At this point, the peak electric field has reached its maximum spatial extension; hence the potential drop across the device is at its maximum. If the current density continues to increase, additional electrons are required in the collector region as shown in Figure 6(d) leading to a gradient in the magnitude of the electric field and a reduction in the total potential across the device. Continued increase in current density leads to the situation shown in Figure 6(e) where the quasi-neutral base has been extended into the collector region and the potential drop has been further reduced. This analysis will be important later in this work.
Figure 6. Doping density, charge, and electric field in the space charge region (after Dunn and Nuttall) [13].
Unlike collector-emitter breakdown due to avalanche multiplication, bipolar second breakdown is a less well-understood phenomenon despite being the subject of many studies since it was first reported by Thornton and Simmons in 1958 [14, 15, 16, 17, 18]. A precise definition of second breakdown is difficult to give, but the term is generally applied to a further drop in collector-emitter potential, occurring after the initial snap back, which is accompanied by increased current conduction. Another difference between simple avalanche breakdown and second breakdown is that the latter is usually accompanied by device degradation or catastrophic failure. Some characteristics which have been identified with second breakdown are a delay time between initial snap back and the onset of second breakdown, a minimum triggering energy required to initiate second breakdown, and an association with current constriction suggestive of thermal runaway.

2.3. Failure Modes

In modern VLSI protection structures dielectric breakdown is rarely seen and electrothermal migration is the primary failure mode [3]. The initial snap back which occurs during bipolar breakdown provides a reduction in terminal voltage for the device under stress and, therefore, a reduction in the power being delivered to the device. Although the device potential is clamped to a relatively low voltage during the initial stages of the ESD event, eventually enough power will be delivered to the device to cause heating and the device may subsequently fail due to thermal runaway. Even though the active area may be considered a small region within a large thermal mass,
the ESD event is quite rapid with respect to thermal events and heat flow may not be a large factor in determining the response of a structure to stress. Nevertheless, it is expected that the maximum rate of temperature increase will likely occur near the silicon surface due to the substantially lower thermal conductivity of silicon dioxide compared to silicon [5].

In considering the thermal response of semiconductor devices to stress induced by electromagnetic pulse (EMP), Wunsch and Bell proposed a model based on localized heating at a junction which has proven useful for estimating failure thresholds in some applications [19]. Assuming localized heating at the junction, they derived the following expression for the relationship between power and temperature rise in a simple structure:

\[
\frac{P}{A} = \sqrt{\frac{\pi \kappa \rho_m c}{T_m - T_i}} t^{-\frac{1}{2}}
\]

(3)

where \( P \) is the power delivered to the junction, \( A \) is the junction area, \( \kappa \) is the thermal conductivity, \( \rho_m \) is the mass density, \( c \) is the specific heat, \( T_m \) is the final temperature, and \( T_i \) is the initial temperature. They were then able to use this expression to plot power versus expected time to failure for three different cases: (1) heating from ambient to 675 °C; (2) heating from ambient to 1415 °C; and (3) heating from ambient to 1415 °C over one-tenth of the junction area. Case one considers that failure may occur at 675 °C, the melting point of aluminum, a material widely used as metallization in semiconductor circuits; case two considers that failure will occur when the melting point of silicon is reached; and case three considers that failure may
occur due to current constriction causing localized heating beyond the melting point of silicon over only a portion of the device area.

Modern CMOS processes have introduced features designed to improve performance and increase reliability due to such effects as hot carrier stress. Among these features are the cladding of diffused areas with refractory metal silicide and the fabrication of MOS transistors with lightly doped drain or graded drain diffusions. These features have had a pronounced impact on the ability to achieve adequate ESD protection and have been the subject of some study [1, 2].

The cladding of diffusion areas with refractory metal silicide is designed to reduce resistance and increase circuit performance. Unfortunately, improved conductivity decreases the effects of emitter ballasting by diffusion resistance in protection structures which is known to be an effective countermeasure against nonuniform current density. Other effects, such as current crowding in the low resistance silicide, may also be present, accounting for degradation brought about by silicide. Although important, the effects of silicided junctions will not be considered further in this work.

Junction grading, intended to reduce electric field and hence hot carrier generation, also has been shown to impact ESD protection. Shabde et al reported that the failure incidence in graded drain devices increases with the amount of junction grading [1]. This has been characterized by McPhee et al for thick oxide protection devices, but the correlation is rather weak for these devices [2]. However, for thin oxide devices the correlation is much clearer [9]. Chen has found that snap back voltage,
which generally increases with increased junction grading, correlates inversely with ESD pass voltage, except that a steeper slope occurs at higher snap back voltages where the data was taken from LDD devices [3]. This suggests that ESD failure threshold may be adversely impacted by junction grading and that LDD-type devices may introduce additional failure mechanisms. These points are addressed later in this work.
CHAPTER 3

Simulation Method

This chapter presents the methods used to solve the semiconductor equations in order to simulate a one-dimensional device subjected to ESD stress. Numerical solution of the semiconductor equations is a challenging area which has attracted many workers over the past two decades or more and remains a challenging area today. Many difficulties arise, including ill-conditioning of the equations, nonlinearities, and instability of solutions on a spatial grid; and many strategies have been tried to overcome these difficulties. Since the purpose of this work is to understand physical phenomena, rather than to develop new and novel numerical techniques, relatively straightforward and conventional techniques have been used.

3.1. Basic Semiconductor Equations

The operation of semiconductor devices is described by the following four basic equations:

\[
\frac{\partial n}{\partial t} = \frac{1}{q} \nabla \cdot j_n + G_n - U_n \quad (4)
\]

\[
\frac{\partial p}{\partial t} = -\frac{1}{q} \nabla \cdot j_p + G_p - U_p \quad (5)
\]

\[
\nabla \cdot (\varepsilon \nabla \psi) = -q (p - n + N_D^+ - N_A^-) \quad (6)
\]
\[
\frac{\partial T}{\partial t} = \frac{1}{\rho_m c} (\nabla \cdot [\kappa(T)\nabla T] + H)
\] 

(7)

These equations represent the continuity equations for electrons and holes (4 and 5), Poisson’s equation (6), and the heat flow equation (7). In the above, n and p are the electron and hole concentrations, respectively, \( \Psi \) is potential, T temperature, t time, \( \mathcal{J}_n \) and \( \mathcal{J}_p \) the electron and hole current densities, G and U the carrier generation and recombination rates, \( \varepsilon \) the permittivity of the material, q the electronic charge, \( \kappa \) the thermal conductivity, and H the heat generation term which is given by \( \mathcal{J}_n \mathcal{E} \). Also, \( N_D^+ \) and \( N_A^- \) are the ionized donor and acceptor doping densities, \( \rho_m \) is the mass density, and \( c \) is the specific heat.

The expressions for electron and hole current density are given by equations (8) and (9):

\[
\mathcal{J}_n = q(n\mu_n \mathcal{E}_n + D_n \nabla n)
\] 

(8)

\[
\mathcal{J}_p = q(p\mu_p \mathcal{E}_p - D_p \nabla p)
\] 

(9)

Here, \( \mu_n \) and \( \mu_p \) are the electron and hole mobilities, \( D_n \) and \( D_p \) the electron and hole diffusivities, and \( \mathcal{E}_n \) and \( \mathcal{E}_p \) the "effective" electric field for electrons and holes. This work follows common practice and sets

\[
\mathcal{E}_n = \mathcal{E}_p = \mathcal{E} = -\nabla \Psi
\]

which ignores some second order effects and also assumes Boltzmann carrier statistics [20]. The latter introduces a significant approximation, particularly at higher temperatures. The above expressions for current density also should include a third
term which represents current that arises from a gradient in temperature [21]. This term has not been implemented in the present work.

3.2. Models

In order to accomplish semiconductor device simulation, models are required for the various parameters which appear in the above equations. Typically, these parameters may be dependent upon a number of variables such as electric field, doping density, and temperature. Temperature dependence is a particular problem in this work since it is desired to simulate devices over a range from room temperature to near the melting point of silicon. Unfortunately, published experimental data is generally restricted to the lower portion of this range; therefore, the available models are not suitable for use over the entire temperature range desired. In order to accomplish the desired simulation, extrapolation of the available data is used; therefore, results in the upper temperature regimes should be considered as trends and not as numerically accurate.

3.2.1. Mobility and Diffusivity

The electron and hole mobilities, $\mu_n$ and $\mu_p$ respectively, are calculated from expressions given by Orvis and Yee [22]:

$$\mu = \frac{\mu_0 (T/T_0)^{-\gamma}}{[1 + (|E|/E_0)^\beta]^{1/\beta}}$$  \hspace{1cm} (10)

where
\[ \mu_0 = \frac{\mu_{\text{max}} - \mu_{\text{min}}}{1 + (N/N_r)^\alpha} + \mu_{\text{min}} \]

and

\[ E_c = \frac{2.319 v_m}{[1 + 0.8\exp(T/600)]\mu_0(T/T_0)^{-\gamma}} \]

and where \( T_0 = 300 \, ^\circ \text{K} \) and the various coefficients are given in Table 1.

<table>
<thead>
<tr>
<th>( \mu_{\text{max}} ) (cm(^2)/V-s)</th>
<th>( \mu_{\text{min}} ) (cm(^2)/V-s)</th>
<th>( N_r ) (cm(^{-3}))</th>
<th>( v_m ) (cm/s)</th>
<th>( \alpha )</th>
<th>( \gamma )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_n )</td>
<td>1330</td>
<td>65</td>
<td>8.5\times10^{16}</td>
<td>1.1\times10^{7}</td>
<td>0.72</td>
<td>2.42</td>
</tr>
<tr>
<td>( \mu_p )</td>
<td>495</td>
<td>47.7</td>
<td>6.3\times10^{16}</td>
<td>9.5\times10^{6}</td>
<td>0.76</td>
<td>2.20</td>
</tr>
</tbody>
</table>

These expressions, in turn, are based on the earlier work of Caughey and Thomas [23] and Jacoboni et al [24]. The reported experimental data used for curve fitting ranges up to an electric field of 1–3\times10^5 V/cm and a temperature of 430 \(^\circ\text{K}\). These are less than ideal for the present work, but the curve fits are quite good throughout the range of the data and the models are expected to be useful for some extrapolated range above the measured data. The diffusion coefficients are calculated by assuming the validity of the Einstein relation:

\[ D = \left[ \frac{kT}{q} \right] \mu \]  (11)
3.2.2. Generation Rate (Impact Ionization)

The generation rate of electrons and holes is computed according to the following expression:

\[ G = \frac{1}{q} (\alpha |\vec{j}_n| + \beta |\vec{j}_p|) \]  

(12)

where

\[ \alpha = a_1 \exp \left\{ -\frac{a_2 + a_3(T - 300)}{|E|} \right\} \]  

(13)

\[ \beta = a_1 \exp \left\{ -\frac{a_2 + a_3(T - 300)}{|E|} \right\} \]  

(14)

and the coefficients are given in Table 2.

<table>
<thead>
<tr>
<th>E (V/cm)</th>
<th>a_1 (cm(^{-1}))</th>
<th>a_2 (V/cm)</th>
<th>a_3 (V/cm(^{-2})°K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electrons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E &lt; 2.4\times10^5</td>
<td>2.6\times10^6</td>
<td>1.43\times10^6</td>
<td>1.3\times10^3</td>
</tr>
<tr>
<td>2.4\times10^5 &lt; E &lt; 4.2\times10^5</td>
<td>6.2\times10^5</td>
<td>1.08\times10^6</td>
<td>1.3\times10^3</td>
</tr>
<tr>
<td>E &gt; 4.2\times10^5</td>
<td>5.0\times10^5</td>
<td>9.90\times10^5</td>
<td>1.3\times10^3</td>
</tr>
<tr>
<td>Holes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E &lt; 5.1\times10^5</td>
<td>2.0\times10^6</td>
<td>1.97\times10^6</td>
<td>1.1\times10^3</td>
</tr>
<tr>
<td>E &gt; 5.1\times10^5</td>
<td>5.6\times10^5</td>
<td>1.32\times10^6</td>
<td>1.1\times10^3</td>
</tr>
</tbody>
</table>

The data for these coefficients was taken over range of 22 °C to 150 °C [25], so the same cautions given above for mobility also apply here. The present form of the model was reported by Orvis and Yee and was modified slightly by them from the original data to provide continuous functions for improved numerical solution [22].
3.2.3. Recombination Rate

The recombination rate $U$ is computed according to the Shockley-Read-Hall model [26, 27]:

$$U = \frac{pn - n_i^2}{\tau_p(n + n_i) + \tau_n(p + n_i)}$$  \hspace{1cm} (15)

where $\tau_n$ and $\tau_p$ are the minority carrier lifetimes. For this work $\tau_n = \tau_p = 0.1$ $\mu$s.

3.2.4. Intrinsic Carrier Concentration and Energy Gap

The intrinsic carrier concentration is modeled as

$$n_i^2 = 6.059463 \times 10^{31} T^3 \exp \left( - \frac{qE_g}{kT} \right)$$  \hspace{1cm} (16)

where $E_g$ is the energy gap for silicon given by [12]:

$$E_g = q \left( 1.17 - \frac{4.73 \times 10^{-4} T^2}{T + 636} \right)$$  \hspace{1cm} (17)

3.2.5. Thermal Conductivity

Heasell has shown that thermal conductivity can be modeled by the following simple expression [28]:

$$\kappa = \frac{350}{T - 68} \text{ W/cm}^{-1} \text{K}$$  \hspace{1cm} (18)

This equation is based on a curve fit of measured data at temperatures up to 700 $^\circ$K.
3.2.6. Silicon Constants

The values of some physical properties of silicon used in this work are given in Table 3.

<table>
<thead>
<tr>
<th>parameter</th>
<th>symbol</th>
<th>value</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>permittivity</td>
<td>ε</td>
<td>$1.054 \times 10^{-12}$</td>
<td>F/cm</td>
</tr>
<tr>
<td>mass density</td>
<td>ρ_m</td>
<td>2.33</td>
<td>gm/cm$^3$</td>
</tr>
<tr>
<td>heat capacity</td>
<td>c</td>
<td>0.7</td>
<td>J/gm$^{-oK}$</td>
</tr>
<tr>
<td>melting point</td>
<td></td>
<td>1688</td>
<td>$^{oK}$</td>
</tr>
</tbody>
</table>

3.3. Numerical Techniques

The semiconductor equations are solved by applying the Crank-Nicolson and finite difference methods. Current density expressions after Scharfetter and Gummel are employed to avoid numerical instability on the discretization grid and the variables are scaled to improve conditioning. The resulting nonlinear equations are solved by Newton’s method. The solution of the heat equation is separated from the other equations for simplification and reduced computational effort.

3.3.1. Temporal Differencing

Temporal differencing of the time-dependent equations is achieved in the following form [29, 30]:

$$f(t + \Delta t) = f(t) + (1 - \beta) \Delta t f'(t) + \beta \Delta t f'(t + \Delta t)$$  \hspace{1cm} (19)

In this format, $\beta = 0$ gives the explicit or Euler formulation, $\beta = 1$ gives the fully implicit or backward Euler formulation, and $\beta = 0.5$ gives the Crank-Nicolson
method. In this work the Crank-Nicolson method is used because of its properties of guaranteed convergence and unconditional stability.

3.3.2. Spatial Differencing

The semiconductor equations are discretized on a one-dimensional finite difference mesh where \( i \) is the mesh index. This is accomplished as follows:

First, the following quantities are defined:

\[
\Delta x^+(i) = x(i+1) - x(i)
\]

\[
\Delta x^-(i) = x(i) - x(i-1)
\]

\[
\Delta x_A(i) = \frac{\Delta x^+(i) + \Delta x^-(i)}{2}
\]

\[
E^+(i) = \frac{\Psi(i) - \Psi(i+1)}{\Delta x^+}
\]

\[
E^-(i) = \frac{\Psi(i-1) - \Psi(i)}{\Delta x^-}
\]

\[
E_A(i) = \frac{E^+ + E^-}{2}
\]

Then, the electron continuity equation is discretized as:

\[
\frac{\partial n}{\partial t} = \frac{1}{q\Delta x_A} \left( J_n^+ - J_n^- \right) + (G_n - U_n)
\]

(20)

and similarly the hole continuity equation becomes:
\[
\frac{\partial p}{\partial t} = -\frac{1}{q\Delta x_A} (J^+_p - J^-_p) + (G_p - U_p)
\]

where \(J^+_n\) and \(J^+_p\) are the electron and hole current densities between mesh points \(i\) and \((i+1)\), and \(J^-_n\) and \(J^-_p\) are the current densities between mesh points \((i-1)\) and \(i\). Similarly, Poisson's equation becomes:

\[
\frac{(E^+ - E^-)}{\Delta x_A} = \frac{q}{\epsilon} (p - n + N^+_D - N^-_A)
\]

3.3.3. Current Density Expressions

It is well-known that numerical instabilities can occur for the continuity equations if the potential difference between two adjacent mesh points exceeds \(2kT/q\) [29]. Scharfetter and Gummel introduced a formulation of the current density expressions which overcomes this difficulty and which has become widely accepted [31]. Following this formulation, the magnitude of the current densities in one dimension is given by the following expressions:

\[
J_n = q\mu_nE \left\{ \frac{n(i+1)}{1 - \exp \left( -\frac{\mu_nE}{D_n} \right) \Delta x} + \frac{n(i)}{1 - \exp \left( \frac{\mu_nE}{D_n} \right) \Delta x} \right\}
\]

\[
J_p = q\mu_pE \left\{ \frac{p(i+1)}{1 - \exp \left( \frac{\mu_pE}{D_p} \right) \Delta x} + \frac{p(i)}{1 - \exp \left( -\frac{\mu_pE}{D_p} \right) \Delta x} \right\}
\]

These expressions are derived by assuming constant electron and hole current densi-
ties between adjacent mesh points. In addition to providing numerical stability, these expressions have the property that they reduce to the conventional central difference expression for diffusion current density in the limit of small electric field and the upwind difference approximation for drift current in the limit of large electric field [29].

3.3.4. Scaling

Because the semiconductor equations deal with quantities that vary over many orders of magnitude, scaling them for best numerical results can be quite difficult. Several schemes have been used by different workers; some of these can be found in [21, 32, 33, 34]. Here, a method similar to that of Kreskovsky and Grubin [29] has been used. The following reference quantities are defined:

\( N_r \): the maximum doping density

\( \Psi_r \): the maximum expected potential

\( X_r \): the maximum device dimension

\( V_r \): the maximum carrier velocity

\( \mu_r \): the maximum carrier mobility

\( D_r \): the maximum carrier diffusivity

\( \varepsilon_r \): the maximum structure permittivity
\[ t_r = \frac{X_r}{V_r} \text{: the computed time scaling factor.} \]

The carrier and doping densities are scaled by \( N_r \):

\[ n = \frac{n}{N_r} \]

\[ p = \frac{p}{N_r} \]

\[ N_D^+ = \frac{N_D^+}{N_r} \]

\[ N_A^- = \frac{N_A^-}{N_r} \]

Potential is scaled by \( \Psi_r \):

\[ \Psi = \frac{\Psi}{\Psi_r} \]

Length is scaled by \( X_r \):

\[ x = \frac{x}{X_r} \]

Time is scaled by \( t_r \):

\[ t = \frac{t}{t_r} \]

Mobility is scaled by \( \mu_r \):
\[ \tilde{\mu} = \frac{\mu}{\mu_r} \]

Diffusivity is scaled by \( D_r \):

\[ \tilde{D} = \frac{D}{D_r} \]

Permittivity is scaled by \( \varepsilon_r \):

\[ \tilde{\varepsilon} = \frac{\varepsilon}{\varepsilon_r} \]

Applying these gives the following expressions:

\[
\left( \frac{N_r}{t_r} \right) \frac{\partial \tilde{n}}{\partial t} = \frac{1}{qX_r} \nabla \tilde{\psi}_n + (G_n - U_n)
\]

\[
\frac{\partial \tilde{n}}{\partial t} = \left( \frac{t_r}{qN_r X_r} \right) \nabla \tilde{\psi}_n + \left( \frac{t_r}{N_r} \right) (G_n - U_n)
\]  

(25)

Likewise,

\[
\frac{\partial \tilde{p}}{\partial t} = - \left( \frac{t_r}{qN_r X_r} \right) \nabla \tilde{\psi}_n + \left( \frac{t_r}{N_r} \right) (G_n - U_n)
\]  

(26)

In the above, \( \tilde{\psi}_n \), \( \tilde{\psi}_p \), \( G_n \), \( G_p \), \( U_n \), and \( U_p \) are in real (unscaled) units. Poisson's equation is scaled as follows:

\[
\left( \frac{\varepsilon_r X_r}{X_r^2} \right) \nabla^2 (\varepsilon \tilde{\psi}) = - qN_r (\tilde{\phi} - \tilde{n} + \tilde{N}_D^+ - \tilde{N}_A^-)
\]
\[ \nabla \cdot (\varepsilon \nabla \psi) = - \frac{qN_r X_r^2}{\varepsilon_r \Psi_r} (\tilde{n} - \tilde{\tilde{n}} + \tilde{N}_D^+ - \tilde{N}_A^-) \] (27)

The current density expressions are scaled as follows:

\[ J_n = q \left\{ \frac{N_r \mu_r \Psi_r}{X_r} \tilde{\mu}_n \tilde{E} \right\} \frac{\tilde{n}(i+1)}{1 - \exp \left[ - \left( \frac{\mu_r \Psi_r}{D_r} \right) \left( \frac{\tilde{\mu}_n \tilde{E}}{\tilde{D}_n} \right) \Delta \tilde{\tilde{n}} \right]} + \frac{\tilde{n}(i)}{1 - \exp \left[ - \left( \frac{\mu_r \Psi_r}{D_r} \right) \left( \frac{\tilde{\mu}_n \tilde{E}}{\tilde{D}_n} \right) \Delta \tilde{\tilde{n}} \right]} \] (28)

\[ J_p = q \left\{ \frac{N_r \mu_r \Psi_r}{X_r} \tilde{\mu}_p \tilde{E} \right\} \frac{\tilde{p}(i+1)}{1 - \exp \left[ - \left( \frac{\mu_r \Psi_r}{D_r} \right) \left( \frac{\tilde{\mu}_p \tilde{E}}{\tilde{D}_p} \right) \Delta \tilde{\tilde{p}} \right]} + \frac{\tilde{p}(i)}{1 - \exp \left[ - \left( \frac{\mu_r \Psi_r}{D_r} \right) \left( \frac{\tilde{\mu}_p \tilde{E}}{\tilde{D}_p} \right) \Delta \tilde{\tilde{p}} \right]} \] (29)

The tilde, indicating variables in scaled, dimensionless terms, will be dropped from the remainder of this discussion. Where appropriate in the discussion of numerical methods, the variables are assumed to be scaled as described; otherwise, variables are assumed to take on their unscaled values.
In this work, \( N_r \) and \( X_r \) are taken from the device size and doping profile, \( \Psi_r = 1 \, \text{V}, \quad V_r = 8.5 \times 10^6 \, \text{cm/sec}, \quad \mu_r = 1500 \, \text{cm}^2/\text{V-sec}, \quad D_r = \mu_r(300 \, k)/q, \) and \( \varepsilon_r = 1.054 \times 10^{-12} \, \text{F/cm}. \)

### 3.3.5. Newton's Method

Temporal and spatial differencing of the semiconductor equations produces a set of algebraic equations which must be solved at each time step of the transient analysis. However, the equations are nonlinear and cannot be solved by direct methods, so the iterative technique known as Newton's method is applied [21, 35]. This is done by constructing functions for each of the primary variables (\( n, p, \) and \( \Psi \)) at each grid point \( i \) \((1 \leq i \leq i_{\text{max}})\) that should equal zero at each new time step:

\[
f_n(i) = n(t) + (1 - \beta) \Delta t \left. \frac{\partial n}{\partial t} \right|_t + \beta \Delta t \left. \frac{\partial n}{\partial t} \right|_{t+\Delta t} - n(t + \Delta t) \quad (30)
\]

\[
f_p(i) = p(t) + (1 - \beta) \Delta t \left. \frac{\partial p}{\partial t} \right|_t + \beta \Delta t \left. \frac{\partial p}{\partial t} \right|_{t+\Delta t} - p(t + \Delta t) \quad (31)
\]

\[
f_{\Psi}(i) = \nabla \cdot (\varepsilon \nabla \Psi) + q \left( p - n + N_D^+ - N_A^- \right) \quad (32)
\]

Also, the solution vector \( X \) and function matrix \( F \) are defined:
Now, given $X(t)$ at any current time $t$, a solution $X(t + \Delta t)$ is sought which minimizes $F$. Since the system is nonlinear, Newton's method is applied:

$$F'(X^n)\Delta X^{n+1} = -F(X^n)$$

(33)

where the superscript $n$ indicates the iteration number. Due to the fact that each of the individual component functions is, in general, a function of $n$, $p$, and $\Psi$ at the points $(i-1)$ and $(i+1)$, as well as at point $i$, the matrix $F$ takes on a block tridiagonal structure which can easily be solved by standard routines such as the band-matrix solver of LINPACK [36]. The solution continues until a stopping criteria is met or until the maximum number of iterations is exceeded, whichever occurs first. For the present work, the stopping criteria was

$$\frac{1}{\|X^n\|_2 - \|X^{n-1}\|_2} \leq 10^{-6}$$
and the maximum number of iterations allowed per time point was 50.

3.3.6. Thermal Solution

Temporal and spatial differencing of the heat equation follows a very similar approach to that discussed above for the electrical equations. In this work, the solution of the heat equation is separated from the solution of the continuity equations and Poisson's equation. Although it is possible to solve all four equations simultaneously, separation of the heat equation reduces the dimension of the problem substantially, thus saving computation time. Here an alternating solution technique is used, with several electrical solutions being made between successive solutions of the heat equation. Although a formal investigation of the accuracy lost due to nonsimultaneous solution of the heat equation has not been made, this approach is justified by the relatively slow evolution of thermal phenomena when compared to electrical phenomena and by the relatively small time steps used in this work. The maximum electrical time step for these simulations was 10 ps and the maximum thermal time step was 100 ps. There is also precedent for this approach in the literature contributed by other workers in the field [22]. Here a heat solution is performed at least every ten electrical steps. With the separation of the heat equation from the electrical equations and making the approximation that thermal conductivity is constant throughout a heat solution time interval, it is possible to use a single, direct solution. The matrix is of tridiagonal form and the LINPACK tridiagonal solution routines are used.
3.3.7. Boundary Conditions

Boundary conditions for the carrier continuity equations are given by assuming charge neutrality and infinite recombination velocity at the contacts. That is, the following conditions are enforced at the contacts:

\[ p_n = n_i^2 \]  \hspace{1cm} (34)

and

\[ p - n + N_D^+ - N_A^- = 0 \]  \hspace{1cm} (35)

The boundary conditions for potential are given by assuming that the contact at \( i_{\text{max}} \) is at ground potential (i.e. 0 volts) and the potential on the contact at \( i = 1 \) is given by solving an external circuit corresponding to the Human Body Model. For these simulations, the Human Body Model has been modified slightly by adding a small capacitor at the device terminal to simulate parasitic capacitance due to packaging and interconnect which is normally encountered in practice; the exact circuit is illustrated in Chapter Four. For these simulations also, no work function difference at the contacts was implemented nor was any provision made for potential drop due to contact resistance.

Boundary conditions for solution of the heat equation are thermally insulating on all sides. This is well-justified for the top surface since the thermal conductivity of silicon dioxide is much smaller than that of silicon. It is also reasonable on the ends since the devices simulated often are part of a "ladder" structure and hence have mirror symmetry. On the bottom, however, it should be expected that some heat loss
will occur into the silicon substrate. Still, it can be considered that negligible heat loss occurs in the vertical direction over the time duration of these simulations if it is considered that the simulation region is a narrow strip near the surface of a larger conducting region. The validity of this assumption is borne out by the correlation of the simulations to actual measured failure results.

3.3.8. Initial Transient Solution

The initial starting solution for the simulation is very important since the most rapid changes take place at the beginning of the simulation. In principle, it is possible to choose a starting guess and let the solution algorithm run until steady to obtain a steady-state starting solution; however, in practice this may require very small time steps and a sophisticated time-step control algorithm to obtain usable results. Here a rough estimate of the solution is made, followed by solution of Poisson's equation using the same Newton's method routine used for the main algorithm. It is possible to obtain a complete steady-state solution from Poisson's equation alone by replacing p and n in the \( f_\psi \) functions with appropriate expressions involving the Fermi level related to the local potential [37]. The resulting system of equations converges easily and more rapidly, and the resulting solution is presented to the main algorithm which is allowed to run until steady. This combination rarely fails to produce a usable initial solution for the transient analysis.
3.3.9. Simulation Sequence

The simulation proceeds in the following sequence. First, the structure is defined, variables are scaled appropriately, and a very crude initial guess is made as to carrier and potential distribution. The structure is initially at a uniform 300°K. Next, an initial solution is obtained as described above with the external circuit disconnected. Then the actual transient analysis is begun by changing electrical boundary conditions to connect the charged external circuit to the device under test. For the structures considered in this work, initially a potential difference will build up across the reverse-biased junction. When the electric field reaches sufficient magnitude, impact ionization will begin to take place and avalanche breakdown will occur, causing the device potential to snap back. Some potential drop remains, however, and as the device begins to conduct heating occurs due to the $J^2$ product. As the analysis proceeds, the solutions are recorded for later analysis.

3.3.10. Algorithm Performance

Although not extremely robust in its present implementation, the algorithm used produces reasonably stable results for the types of profiles being investigated. Scaled as indicated above, the system of equations produces a condition number of about $10^{10}$ to $10^{11}$ which is certainly not ideal, but combined with double precision arithmetic on the IBM mainframe used produces several significant figures of usable results [36]. Fixed time step control is used beginning with very fine time steps of 0.001 ps and increasing rapidly to a maximum time step of 10 ps for the electrical
solutions; heat solutions are performed about every ten time steps. Such fine initial
time steps are, perhaps, overkill; however, since adequate performance was obtained
with this scheme, it was not deemed worthwhile to invest a large effort in more
sophisticated time step control.
CHAPTER 4

Data and Analysis

4.1. Structure Descriptions

Five thin-oxide MOSFET structures of three basic types were investigated in this work. Structure 1 is similar to a device structure that is used widely in modern CMOS integrated circuits [3]. Structure 2 is a hypothetical structure that is identical to the first except for the omission of a single implant step. Structures 3, 4, and 5 represent a set of test devices that were fabricated in a "split lot;" that is, a wafer lot was fabricated in which all wafers were processed identically except at a small number of steps where a split occurred and certain experimental parameters were varied. These structures are representative of the parasitic npn bipolar transistors associated with n-channel output devices. Figure 7 illustrates a portion of the output structure shown earlier along with the location of the simulation profile.

The three basic types of structures represented here, in the terminology of Chen [3], are classified as graded drain (GD), double diffused drain (DDD), and lightly doped drain (LDD). These different structures are illustrated schematically in Figure 8. The basic differences between these types of structures are generally confined to the processing steps immediately surrounding the Source-Drain (SD) implant(s) and their associated high-temperature annealing or diffusion operations and are described
Figure 7. The location of the simulation profile in an output structure.
Figure 8(a). The structure of a graded drain device.

Figure 8(b). The structure of a double diffused drain device.

Figure 8(c). The structure of a lightly doped drain device.
The Graded Drain device, illustrated in figure 8(a), is fabricated by implanting a single species, here phosphorus, after the MOSFET gate structure has been patterned and etched. A sidewall spacer of silicon dioxide may be added to the edge of the polysilicon gate before the implant to block the implanted species from the region immediately adjacent to the gate. This allows the phosphorus to diffuse laterally during subsequent high temperature steps without creating excessive gate-diffusion overlap. Space is thus provided in which the phosphorus concentration can become graded in the vicinity of the junction for the purpose of reducing the $n^+$ concentration near the junction, contributing to the reduction of the electric field and consequently reducing the quantity of hot carriers generated during normal operation.

The Double Diffused Drain device, pictured in figure 8(b), is fabricated similarly to the Graded Drain device, except that two species, usually phosphorus and arsenic, are implanted at the same time. The phosphorus, which diffuses faster during high temperature steps, forms the actual junction while the arsenic lags behind creating a more heavily doped region away from the junction. This allows a lighter phosphorus dose to be used, reducing the concentration near the junction without compromising the sheet resistance of the diffused areas away from active devices.

The Lightly Doped Drain device, shown in figure 8(c), is a further extension of the DDD structure. Here a light phosphorus "reach-through" implant is performed prior to the deposition of the sidewall oxide spacer. This is then followed by the double SD implant. Since the reach-through implant is usually done at low energy, this
results in a shallow n− region extending from the heavy n+ drain region to a point under the gate electrode. With three implants in the SD process, this structure is the most complicated of the three.

The doping profiles for this work were created by simulating the device process flows with the SUPRA two-dimensional process simulation program [38]. Since the device structures are symmetrical, it was only necessary to simulate half of the structure in SUPRA; hence, the structure and doping profile plots which follow depict only the left half of the device. The structures to be simulated were constructed by flipping the structure about the right-hand (x = 2.0 μm) vertical axis. The one-dimensional doping profile required for the electrical and thermal simulation was then extracted along a horizontal line at the y = 0.09 μm coordinate, which is approximately 400 to 500 angstroms below the silicon-silicon dioxide interface in these structures.

Structure 1 is of the LDD type. A sidewall spacer of 3000 angstroms was used between the n− reach-through implant and the double SD implants. Structures 2 and 3 are of the DDD type. Structure 2 is identical to the first except for the omission of the reach-through implant. Structure 3 was processed through a different process flow and used a sidewall oxide spacer of 2500 angstroms. Structures 4 and 5 are of the GD type and were processed through the same flow as Structure 3 except for the omission of the arsenic SD implant and variations in the phosphorus implant dose. The resulting structures and their associated one-dimensional doping profiles are shown in Figures 9 through 18. The cross-sectional figures show the polysilicon gate
Figure 9. Cross-sectional view of Structure 1 (LDD).

Figure 10. Doping Profile 1 (LDD).
Figure 11. Cross-sectional view of Structure 2 (DDD).

Figure 12. Doping Profile 2 (DDD).
Figure 13. Cross-sectional view of Structure 3 (DDD).

Figure 14. Doping Profile 3 (DDD).
Figure 15. Cross-sectional view of Structure 4 (GD).

Figure 16. Doping Profile 4 (GD).
Figure 17. Cross-sectional view of Structure 5 (GD).

Figure 18. Doping Profile 5 (GD).
on top of gate oxide on top of the silicon substrate. The metallurgical junction is shown by a dashed line. To the upper left of the junction is the n⁺ drain region. Below and to the right of the junction is the p-type substrate. The indicated doping contours are at decade intervals beginning with $10^{16}$ cm⁻³ near the junction.

The one-dimensional doping profiles are shown with their associated structure cross section. The most notable feature is the n⁻ plateau in Profile 1 produced by the reach-through implant. The background doping concentration of the p region (indicated by $N_A$) is similar for Profiles 1 and 2 and is also similar for Profiles 3, 4, and 5, indicating the two families to which these structures belong. The peak n⁺ concentrations (given by $N_D$), however, vary substantially due to the various implant splits. More subtle differences exist also which are extremely important. A careful comparison of the various profiles will reveal differences in the slope of the net doping concentration, especially in the n-type region immediately adjacent to the metallurgical junction.

Table 4 summarizes the properties of these five profiles. $N_{D_{\text{max}}}$ is the maximum net doping concentration in the n⁺ SD region, $N_{A_{\text{max}}}$ is the maximum net doping concentration in the p region, and $x_{\text{jct}}$ is the coordinate of the metallurgical junction for later reference.
Table 4. Summary of Simulation Profiles.

<table>
<thead>
<tr>
<th>Profile</th>
<th>Type</th>
<th>( N_{\text{D}_{\text{max}}} )</th>
<th>( x_{\text{cl}} )</th>
<th>( N_{A_{\text{max}}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LDD</td>
<td>( 1.70 \times 10^{20} )</td>
<td>1.10( \mu \text{m} )</td>
<td>( 7.52 \times 10^{16} )</td>
</tr>
<tr>
<td>2</td>
<td>DDD</td>
<td>( 1.69 \times 10^{20} )</td>
<td>0.91( \mu \text{m} )</td>
<td>( 7.54 \times 10^{16} )</td>
</tr>
<tr>
<td>3</td>
<td>DDD</td>
<td>( 1.72 \times 10^{20} )</td>
<td>0.97( \mu \text{m} )</td>
<td>( 1.70 \times 10^{16} )</td>
</tr>
<tr>
<td>4</td>
<td>GD</td>
<td>( 1.60 \times 10^{19} )</td>
<td>0.99( \mu \text{m} )</td>
<td>( 1.98 \times 10^{16} )</td>
</tr>
<tr>
<td>5</td>
<td>GD</td>
<td>( 4.55 \times 10^{18} )</td>
<td>0.97( \mu \text{m} )</td>
<td>( 2.01 \times 10^{16} )</td>
</tr>
</tbody>
</table>

4.2. Simulation Conditions

All electrical and thermal simulations in this work were conducted under the same external conditions. The HBM external circuit was used but was modified with the addition of a second capacitor to represent parasitic package capacitance (see Figure 19). The main external capacitor was 100 pF and was charged to 1.5 kV. A 1.5 kΩ series resistor was used, and the parasitic device package capacitance was 10 pF. The assumed device cross-sectional area was 100 μm². Simulations were run until a simulation time of 100 ns or until the maximum structure temperature exceeded the melting point of silicon, whichever occurred first.

4.3. General Performance Analysis

A measure of the comparative overall performance of the structures can be seen by examining the evolution of the device temperature with time. Temperature as a function of space and time for Profiles 1 through 5 is depicted in Figures 20 through 24 and summarized in Table 5. From these results it is easy to rank Profiles 1 through 4 in terms of relative performance. Profile 5, on the other hand, exhibits some unusual behavior, rising to a temperature of 1622 °K at the drain contact in
Figure 19. External circuit for simulations.
Figure 20. Temperature distribution for Profile 1 during HBM stress.

Figure 21. Temperature distribution for Profile 2 during HBM stress.
Figure 22. Temperature distribution for Profile 3 during HBM stress.

Figure 23. Temperature distribution for Profile 4 during HBM stress.
Figure 24. Temperature distribution for Profile 5 during HBM stress.
34.4 ns, and then cooling down with time. This unusual functionality will be discussed further in a later section.

Since the HBM external circuit appears almost like a current source (with exponential decay time constant of 150 ns), each of the structures is subjected to almost identical total current density. Therefore, a measure of the power being delivered to the device during the early heating stages of the ESD event is given by the terminal voltage immediately after snap back. In this work, snap back voltage is designated $V_{sb}$ and is defined as the device terminal potential measured after snap back at $t = 1$ ns. This can be measured in Figures 25 through 29 and is also summarized in Table 5. From the table it can be seen that among Profiles 1 through 3 a higher $V_{sb}$ implies more rapid heating and consequently a lower ESD failure threshold. On the other hand, Profile 4 does not quite fit this trend since it heats too rapidly relative to its value of $V_{sb}$; a possible reason for this is discussed later. The unusual behavior of Profile 5 has already been noted. A tentative conclusion can be drawn then that ESD HBM performance then is substantially dependent upon potential drop and the power delivered to the junction, although exceptions exist which require further explanation.
Figure 25. Potential distribution for Profile 1, t = 0 to 1 ns.

Figure 26. Potential distribution for Profile 2, t = 0 to 1 ns.
Figure 27. Potential distribution for Profile 3, t = 0 to 1 ns.

Figure 28. Potential distribution for Profile 4, t = 0 to 1 ns.
Figure 29. Potential distribution for Profile 5, t = 0 to 1 ns.
Table 5. Summary of Snap Back Voltage and Temperature Performance.

<table>
<thead>
<tr>
<th>Profile</th>
<th>V&lt;sub&gt;sb&lt;/sub&gt; (Volts)</th>
<th>T&lt;sub&gt;max&lt;/sub&gt; (°K)</th>
<th>t(T&lt;sub&gt;max&lt;/sub&gt;) (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.074</td>
<td>1666</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td>4.545</td>
<td>1149</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>3.854</td>
<td>841</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>5.995</td>
<td>1586</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>7.454</td>
<td>1622</td>
<td>34.4</td>
</tr>
</tbody>
</table>

Another way of analyzing the general performance of these structures is to plot maximum temperature versus time. Figure 30 shows a plot of the logarithm of maximum temperature versus time for Profiles 1 through 5. Here the data for Profile 5 has been truncated at t = 20 ns in order to suppress data resulting from changes in functionality and to focus on the initial heating phase of that profile. It can be seen that the maximum temperature follows almost a straight line with time on the semilogarithmic plot. It can also be noted that the curves for Profiles 2 and 3 exhibit a slight downward curvature, indicating for longer analyses either that some reduction in maximum temperature is experienced. This may be due to heat flow effects or to the reduced rate of heating as the potential in the external capacitor, and consequently the structure current density, decays. However, these effects are not significant in the present analysis. Applying a straight-line least-squares fit to the data of Figure 30 gives a time constant, τ<sub>10</sub>, for each profile which is representative of the time required for a decade increase in maximum structure temperature. These values are given in Table 6, along with the values of V<sub>sb</sub> already presented. This data has also been plotted in Figure 31. A good correlation is seen between the logarithm of the time constant τ<sub>10</sub> and the snap-back potential V<sub>sb</sub> except again for that given by Profile 4.
Figure 30. Log maximum temperature versus time for Profiles 1-5.

Figure 31. \( \tau_{10} \) versus \( V_{ab} \).
Two comments are in order at this point. First, the faster-heating of these analyses have been allowed to run until the maximum temperature reached near the melting point of silicon. However, this is a one-dimensional analysis which represents uniform current flow throughout a structure. In actual devices, it is likely that at some temperature lower than silicon melting, structural nonuniformities will lead to uneven heating, resulting in current crowding such as is experienced in bipolar second breakdown and subsequent destruction of the device. Second, up to this point the snap-back potential $V_{sb}$ has been used as an indicator of likely device heating. If conditions remained constant after snap back, the power delivered to the device would remain constant and it would be easy to apply the Wunsch-Bell model to obtain a good idea of the likely device performance. But as will be seen shortly, the potential drop across the structure, and therefore the power delivered to heat the device, is dynamic during the ESD event, so analyses based upon initial conditions after snap back can only be considered a rough guide to likely performance.
4.4. Functional Analysis

In Chapter 2, the work of Dunn and Nuttall was reviewed. The operation of these devices can be understood in the light of their analysis. However, Dunn and Nuttall considered an idealized epitaxial bipolar transistor structure with clear, sharp boundaries between constant doping levels. Also, they did not give consideration to variation in temperature and its effect on the carrier saturation velocity $v_s$. The structures considered in this work, which are more representative of practical devices, feature contoured doping profiles; and the localized heating experienced during breakdown gives rise to an uneven distribution of the saturation velocity throughout the structure. The result is that the critical current density $J_o = qN_Dv_s$, which is the maximum current density that can be carried by the density of electrons from the ionized donor doping atoms, is not constant throughout any area of the n-type drain region.

The expression for saturation velocity given by Jacoboni et al is [24]:

$$v_s = \frac{2.4 \times 10^7}{1 + 0.8 \exp(T/600)}$$

where $T$ is given in °K and the result is given in cm/sec. A plot of this function is shown in Figure 32. This is the underlying expression for $v_s$ used in the mobility model. However, since the mobility expression is also dependent upon doping density, the results obtained by numerical simulation may differ slightly from the accompanying analysis based on saturation velocity $v_s$ and critical current density $J_o$. Still, the results and analysis are in general agreement.
Figure 32. Carrier saturation velocity versus temperature.

Figure 33. Critical current density versus doping density for various temperatures.
Based on the above expression for \( v_s \), the critical current density \( J_o = qN_Dv_s \) can be calculated for various values of the doping density \( N_D \) and various temperatures. This is shown in Figure 33 and is useful for reference in the following analysis.

Since the following analysis makes use of the critical current density \( J_o \), it is important to establish that the value of \( J_o \) is in the neighborhood of the actual device current density during the analysis for some region of interest. This is indeed the case. Figures 34 and 35, which are representative examples, show the calculated value of \( J_o \) for Profiles 1 and 2 during the transient analysis. The actual value of current density in these structures is determined by the external circuit and is initially \( 10^6 \) A/cm\(^2\) with a decay time constant of 150 ns. From the figures, the calculated \( J_o \) in the drain region to the left of the metallurgical junction is clearly smaller than this value for some distance and then makes a transition to a region where it is larger than this value. Not only does \( J_o \) vary throughout the structure, but it will also decrease with time at any point where heating occurs. It is apparent from the figures that the evolution of \( J_o \) with time is much more pronounced in Profile 1 than in Profile 2. This will be examined in more detail later. These plots have been clipped at a maximum of \( 2.0 \times 10^6 \) A/cm\(^2\) to emphasize the area of interest. Of course, given the above definition and the emphasis on conduction by electrons, values for \( J_o \) in the p-type region do not make sense and should be disregarded in the plots.

The following analysis also relies on the ratio of the electron current density \( J_n \) to \( J_o \). Therefore, it is necessary to establish that a substantial majority of the current is carried by electron flow. This can be seen in Figures 36 through 39, which depict
Figure 34. $J_0$ for Profile 1.

Figure 35. $J_0$ for Profile 2.
Figure 36. $J_n$ for Profile 1.

Figure 37. $J_p$ for Profile 1.
Figure 38. $J_n$ for Profile 2.

Figure 39. $J_p$ for Profile 2.
the electron and hole current density for Profiles 1 and 2. For these representative examples, $J_n$ is seen to be almost an order of magnitude greater than $J_p$ throughout the structure, even in the vicinity of the junctions. Therefore, it is reasonable to emphasize the electron current density in the analysis.

In order to visualize the effect of the variation in $J_o$ on the different structures, the ratio of $J_n$ to $J_o$ has been plotted in Figures 40 through 44 for Profiles 1 through 5. These plots have been clipped at 1.0 to emphasize the region where $J_n > J_o$. The distance from the metallurgical junction to the point where $J_n = J_o$ is analogous to the collector width, that is the distance from the junction to the collector contact in the analysis of Dunn and Nuttall, except that here the net charge is not constant due to the complex doping profile. Nevertheless, it is expected that the wider this region, the larger will be the integrated charge and the higher the peak electric field.

Examining the plot of $J_n/J_o$ for Profile 1 the widening of the region left of the junction where $J_n > J_o$ is clearly seen. On the other hand, for Profile 2 such broadening is almost nonexistent. In Profile 3, the slowest heating of the profiles, the decay in total current density outweighs the decrease in $J_o$ due to heating effects, so the region where $J_n > J_o$ actually shrinks back to the right toward the junction. Profile 4 again shows some broadening, although not as much as exists for Profile 1. This is unexpected in the sense that since Profile 4 heats more rapidly than Profile 1, more broadening should be expected; but it is consistent with the fact, already noted, that Profile 4 heats more rapidly than might otherwise be expected. Profile 5, of course, exhibits characteristics due to its already noted unusual behavior in that broadening is
Figure 40. $J_n/J_o$ for Profile 1, $t = 0$ to 48 ns.

Figure 41. $J_n/J_o$ for Profile 2, $t = 0$ to 100 ns.
Figure 42. $J_n/J_o$ for Profile 3, $t = 0$ to 100 ns.

Figure 43. $J_n/J_o$ for Profile 4, $t = 0$ to 30 ns.
Figure 44. $J_n/J_0$ for Profile 5, $t = 0$ to 100 ns.
noted during the early heating phase; but the region narrows again after the peak temperature is reached and cooling begins. The implications of these different behaviors will be explored as each profile is examined in more detail.

4.4.1. Profile 1

Profile 1 is of the LDD type which includes a shallow, lightly doped region near the junction due to the "reach-through" implant. The operation of this structure during the HBM stress is the most complex and interesting "normal" operation of those investigated and provides the most insight into the operation of all of the structures, so it will be presented in more detail than the others. The analysis of Profile 1 has been broken up into three time intervals: 1 ns to 20 ns, 20 ns to 36 ns, and 36 ns to 48 ns. For each time interval plots are presented of net charge, electric field, potential, electron concentration, and hole concentration.

In the time interval from 1 to 20 ns, the maximum temperature rises from 370.5 to 686.8 °K. The boundary between net positive charge and net negative charge is migrating only slightly to the left from \( x = 1.03 \mu m \) to \( x = 1.01 \mu m \), indicating only a slight broadening in the region of negative charge while the peak positive charge also broadens slightly. The resulting electric field makes a minimum at \( 3.4 \times 10^5 \) V/cm at 7 ns, then increases gently but without much broadening. The potential drop across the region of interest also rises after the first few nanoseconds to slightly over 7 V at 20 ns. The electron concentration is seen to be consistent with the minimum required for conduction of \( 10^6 \) A/cm\(^2\) in the 370 to 690 °K range, and the hole concentration is seen to be rising in the \( n^+ \) drain region, indicating the rising intrinsic carrier
Figure 45. Charge distribution for Profile 1, t = 1 to 20 ns.

Figure 46. Electric field distribution for Profile 1, t = 1 to 20 ns.
Figure 47. Potential distribution for Profile 1, t = 1 to 20 ns.

Figure 48. Electron density for Profile 1, t = 1 to 20 ns.
Figure 49. Hole density for Profile 1, $t = 1$ to 20 ns.
concentration n₁ and pn product as a result of rising temperature.

In the time interval from 20 to 36 ns, the maximum temperature increases from 686.8 to 1177 °K. Significant broadening of the negative charge region into the n⁺ drain region begins to occur as evidenced by the migration of the neutral-charge boundary from x = 1.01 μm to 0.94 μm. This is in accordance with the broadening of the area where Jₙ > J₀ seen previously and with the increasing electron concentration required to carry the current density as vₛ decreases. As a result, the peak electric field rises to greater than 5×10⁵ V/cm, broadens substantially, and is accompanied by a significant increase in potential drop to approximately 12 V. The minimum electron concentration continues to increase as required by decreasing vₛ, and the hole concentration continues to rise with increasing temperature.

In the time interval from 36 to 48 ns, the maximum temperature increases further from 1177 to 1666 °K, which is nearing the melting point of silicon. The boundary of the region of net negative charge, continuing to move away from the junction, moves from x = 0.94 μm to x = 0.85μm; and the net positive charge peak increases substantially. This, again, is in accordance with the broadening of the region in which Jₙ > J₀ with increasing temperature. The electric field increases to a peak of about 8.8×10⁵ V/cm at the edge of the broadened area away from the junction yet does not fall off much at the edge nearest the junction, thus contributing to a further increase in potential drop to more than 14 V.
Figure 50. Charge distribution for Profile 1, t = 20 to 36 ns.

Figure 51. Electric field distribution for Profile 1, t = 20 to 36 ns.
Figure 52. Potential distribution for Profile 1, t = 20 to 36 ns.

Figure 53. Electron density for Profile 1, t = 20 to 36 ns.
Figure 54. Hole density for Profile 1, $t = 20$ to $36$ ns.
Figure 55. Charge distribution for Profile 1, \( t = 36 \) to \( 48 \) ns.

Figure 56. Electric field distribution for Profile 1, \( t = 36 \) to \( 48 \) ns.
Figure 57. Potential distribution for Profile 1, t = 36 to 48 ns.

Figure 58. Electron density for Profile 1, t = 36 to 48 ns.
Figure 59. Hole density for Profile 1, $t = 36$ to 48 ns.
4.4.2. Profile 2

In contrast to Profile 1, Profile 2 exhibits relatively well-controlled behavior as temperature increases with time. Due to the steepness of the doping profile for this structure, the region of negative charge broadens only slightly away from the metallurgical junction as indicated by the fact that the charge neutral boundary moves less than 0.01 μm during the analysis interval. There is some deepening of the negative charge region and some increase in the peak positive charge, but the regions remain very narrow when compared to Profile 1. This is consistent with the lack of spreading in the $J_n > J_o$ region as seen previously. As a result, the peak value of the electric field reached is smaller than that for Profile 1; and there is much less broadening of the high-field region. The effect of this is to restrain the maximum potential drop to less than 7 V during the analysis interval. Because this device stays cooler and the carrier saturation velocity is not reduced as much as in the first structure, the minimum electron concentration does not increase as much, although the hole concentration is still rising due to the increase in intrinsic carrier concentration with temperature.

4.4.3. Profile 3

The operation of Profile 3 is quite similar to that of Profile 2, but even better resistance to heating is obtained. Due to additional steepness in the doping profile, no migration of the negative-charge region can be seen in the figure, and the change in charge distribution is limited to a slight deepening in the negative-charge region and a corresponding increase in the positive-charge peak. The electric field peak is
Figure 60. Charge distribution for Profile 2, t = 5 to 100 ns.

Figure 61. Electric field distribution for Profile 2, t = 5 to 100 ns.
Figure 62. Potential distribution for Profile 2, $t = 5$ to 100 ns.

Figure 63. Electron density for Profile 2, $t = 5$ to 100 ns.
Figure 64. Hole density for Profile 2, $t = 5$ to $100$ ns.
Figure 65. Charge distribution for Profile 3, t = 5 to 100 ns.

Figure 66. Electric field distribution for Profile 3, t = 5 to 100 ns.
Figure 67. Potential distribution for Profile 3, \( t = 5 \) to 100 ns.

Figure 68. Electron density for Profile 3, \( t = 5 \) to 100 ns.
Figure 69. Hole density for Profile 3, t = 5 to 100 ns.
further limited in this profile and does not rise above $5.0 \times 10^5$ V/cm during the analysis period. The resulting potential drop rises to only a little above 4 V, so the heating is well-controlled with this profile.

4.4.4. Profile 4

Profile 4 is the first profile examined in which the arsenic source-drain implant was omitted from the process flow. A comparison of the doping profile for this structure with Profile 3 shows the dramatic difference in slope in the region of the arsenic implant. While the slope of Profile 4 is initially higher near the junction than that of Profile 3 due to the higher phosphorus implant dose, the Profiles cross at about $7 \times 10^{17}$ cm$^{-3}$ and then Profile 4 tails off badly. The crossing level is insufficient to support a current density of $10^6$ A/cm$^2$ at much above room temperature, so substantially more broadening should be expected with this profile. This is seen in the plot of net charge. As the region of negative charge broadens and becomes deeper, the region of positive charge also broadens and develops a slightly higher peak. Closer examination of the region of positive charge shows that, while it tails off to the left of the peak, significant net positive charge exists well to the left of the peak. This is due to the lower $N_{D_{\text{max}}}$ and the severe curvature of this profile and to the constraints imposed by the continuity equations. As a result, it is not possible for all of the compensating positive charge to be immediately adjacent to the region of negative charge; therefore, a very wide tail develops in which the electric field is significantly greater than zero. Thus, a potential drop develops well outside the area which was considered the "area of interest" in previous profiles. This extended area of potential
Figure 70. Charge distribution for Profile 4, $t = 2$ to $30$ ns.

Figure 71. Electric field distribution for Profile 4, $t = 2$ to $30$ ns.
Figure 72. Potential distribution for Profile 4, $t = 2$ to $30$ ns.

Figure 73. Electron density for Profile 4, $t = 2$ to $30$ ns.
Figure 74. Hole density for Profile 4, t = 2 to 30 ns.
drop in the $n^+$ region leads to an extended area of heating, and this is likely the reason that Profile 4 does not correlate with the other profiles as indicated earlier and instead heats faster and fails sooner than otherwise would be expected.

4.4.5. Profile 5

A comparison of Profile 5 with Profile 4 shows an even lower $N_{D_{\text{max}}}$ and even shallower slope in doping density. This profile is responsible for the unusual behavior of this structure. During the first 20 ns this structure heats in a similar fashion to the other profiles, except faster. The charge-neutral boundary is seen to be in the vicinity of $x = 0.8 \, \mu m$ and moving further to the left. This location is well up on the slope of the doping profile near the top and approaching $N_{D_{\text{max}}}$. During this early interval a very wide region of high electric field develops, along with a corresponding high potential drop, resulting in the rapid heating. Referring to the previous chart of $J_o$, above 1500 $^o$K at least $3 \times 10^{18} \, \text{cm}^{-3}$ electrons are required to carry $10^6 \, \text{A/cm}^2$, but $N_{D_{\text{max}}}$ is only $4.55 \times 10^{18} \, \text{cm}^{-3}$. Thus, it becomes impossible to sustain the wide region of negative charge required and its compensating region of positive charge without violating the carrier continuity equations. When this occurs, the electric field and its corresponding potential drop collapse and avalanche generation ceases. The structure then enters a period of conduction by space-charge, but resistive heating is drastically reduced due to the much lower potential drop and the temperature profile begins to flatten out. This latter mode may not necessarily be representative of the operation of actual devices, however, because nonuniformities in real devices are likely to induce current crowding and its attendant bipolar second
Figure 75. Charge distribution for Profile 5, t = 5 to 100 ns.

Figure 76. Electric field distribution for Profile 5, t = 5 to 100 ns.
Figure 77. Potential distribution for Profile 5, $t = 5$ to 100 ns.

Figure 78. Electron density for Profile 5, $t = 5$ to 100 ns.
Figure 79. Hole density for Profile 5, t = 5 to 100 ns.
breakdown at a temperature lower than 1600 °K.

4.5. Comments on the Validity of the Simulations

Having considered the operation of several structures by numerical methods, it is important to verify that simulation of a one-dimensional profile provides a meaningful representation of actual structures. First, a comparison of snap-back voltage with measured data is considered. Both Chen and Duvvury et al have given data on snap-back voltage for structures similar to those investigated in this work [3, 9]. However, care should be exercised in making this comparison since the snap-back potential $V_{sp}$ as measured by these workers is defined differently than $V_{sb}$ given earlier in this work. $V_{sp}$ is an extrapolated value calculated by measuring the device under several values of forced current pulses and then fitting a straight line to the measured data. Also, the values measured from real devices are not likely to correlate exactly with those obtained from a one-dimensional simulation due to geometry considerations in the real device; however, if a matching trend can be established, it is an indication that there may be validity to the one-dimensional simulation. Another consideration in comparing the present work with that of Chen is that the profiles are of the same general type but are not necessarily representative of identical process flows. Nevertheless, the correlation is not too bad and is summarized in Table 7.
Table 7. Snap Back Potential Comparison by Structure Type.

<table>
<thead>
<tr>
<th>Profile</th>
<th>$V_{sp}$ (Volts)</th>
<th>$V_{sb}$ (Volts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDD</td>
<td>≈9.2</td>
<td>6.074</td>
</tr>
<tr>
<td>GD</td>
<td>≈8.0</td>
<td>5.995 (Profile 4)</td>
</tr>
<tr>
<td>DDD</td>
<td>≈7.2</td>
<td>4.545 (Profile 2)</td>
</tr>
</tbody>
</table>

A better comparison of snap back voltage can be obtained by comparing with the work of Duvvury et al who measured actual structures corresponding to Profiles 3 through 5 of this work [9]. These values are given in Table 8. Again, although the absolute values are displaced somewhat, probably due to geometry effects in the actual devices, the trend is correct within a reasonable degree of accuracy.

The best indication of the validity of these simulations though is given by comparing the simulated heating rate, given by $\tau_{10}$ calculated earlier, with the measured ESD HBM failure thresholds from actual device structures. The failure threshold was measured by Duvvury on structures of varying width, and a straight-line fit was done to obtain the failure threshold in volts per micron of device width [39]. This data is also given in Table 8 and clearly establishes that the trend in failure threshold tracks the simulated heating rate with reasonable correlation.

Based on the above data, it is believed that the one-dimensional simulation method presented in this work represents a reasonable method of predicting the performance of a class of structures exposed to ESD stress imposed by the HBM method of testing.
<table>
<thead>
<tr>
<th>Profile</th>
<th>$V_{sp}$ (Volts)</th>
<th>$V_{ab}$ (Volts)</th>
<th>$\tau_{10}$ (ns)</th>
<th>Meas. Failure Threshold (V/\mu m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10.0</td>
<td>3.854</td>
<td>272.2</td>
<td>15.22</td>
</tr>
<tr>
<td>4</td>
<td>10.8</td>
<td>5.995</td>
<td>46.9</td>
<td>13.02</td>
</tr>
<tr>
<td>5</td>
<td>12.8</td>
<td>7.454</td>
<td>29.9</td>
<td>9.88</td>
</tr>
</tbody>
</table>
CHAPTER 5

Conclusions and Directions for Further Work

A one-dimensional numerical modeling technique has been presented which gives reasonable predictions of the ESD performance of structures tested according to the HBM method. Evidence that the simulations faithfully represent actual devices is reasonable correlation between simulated and actual snap back voltages and good correlation between simulated heating rate and measured ESD failure threshold. This simulation method is also useful for understanding the mechanisms associated with junction heating and subsequent failure under ESD stress.

Investigation has shown that junction heating results in a local decrease in the carrier saturation velocity and the necessity for increased electron concentration in the space charge region in order to carry the current density imposed by the external HBM circuit. The shape of the doping profile in the $n^+$ region adjacent to the metallurgical junction determines the extent of the region of negative net charge and the shape of the compensating region of positive net charge, and consequently the shape and magnitude of the electric field, amount of potential drop, and device heating rate. Profiles with very steep slopes tend to restrain the advance of the negative charge region and thus control device heating to lower levels. Conversely, profiles with shallow slopes tend to accelerate the advance of the region of negative net charge leading to higher potential drops and accelerated device heating.
The modeling program developed in this work should also prove useful in evaluating the performance of future structures. As new generations of IC processes are developed, the ability to model structures subjected to ESD stress will allow timely process development decisions to be made prior to the fabrication of actual devices. By avoiding a time-consuming iterative development cycle, considerable time and resources can be saved. In addition, new structure designs can be evaluated with this tool provided that they can be modeled in one dimension. For example, the program can also be applied to three-junction pnpn-type devices. This represents a substantial improvement over the empirical design techniques previously in use.

While it is believed that the one-dimensional representation given in this work gives reasonable information about active devices, it would clearly be desirable for future work to extend this simulation to two or perhaps three dimensions. In particular, current constriction, which is believed to be triggered at some temperature below the melting point of silicon and to contribute to the ultimate failure of devices in second breakdown, cannot be investigated with a one-dimensional simulation. A two-dimensional simulation would allow investigation of any current constriction in the vertical direction, as well as a determination of the depth and structure of the conduction layer. A three-dimensional simulation might add information regarding current constriction in the width direction of protection devices, a phenomenon which is almost certainly present in actual structures. In addition, certain refinements should be made in any two- or three-dimensional modeling effort and could improve the present one-dimensional model. These include more flexible boundary conditions
and inclusion of heating effects due to contact resistance. Finally, improved model data at high temperatures would help improve the accuracy of the simulations.
APPENDIX A

Program Flow Chart

1. INITIALIZE
2. READ IN PROFILE
3. SCALE PROBLEM
4. INITIAL GUESS
5. INITIAL SOLN (POISSON'S EQUATION)
6. STEADY STATE SOLN (MAIN ALGORITHM)

A
MAIN ELECTRICAL SOLUTION ROUTINE

CALCULATE BNDRY COND

EVALUATE FUNCTIONS

CALCULATE F' MATRIX

SOLVE FOR ΔX

UPDATE SOLUTION VECTOR

CONVERGED?

> MAX ITER?

NO

NO

YES

YES

ADJUST EXT CIRCUIT FOR NEXT STEP

RETURN

Loop for n time steps
References


