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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.1.3 Calibration of the Snapback I-V Curve

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

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

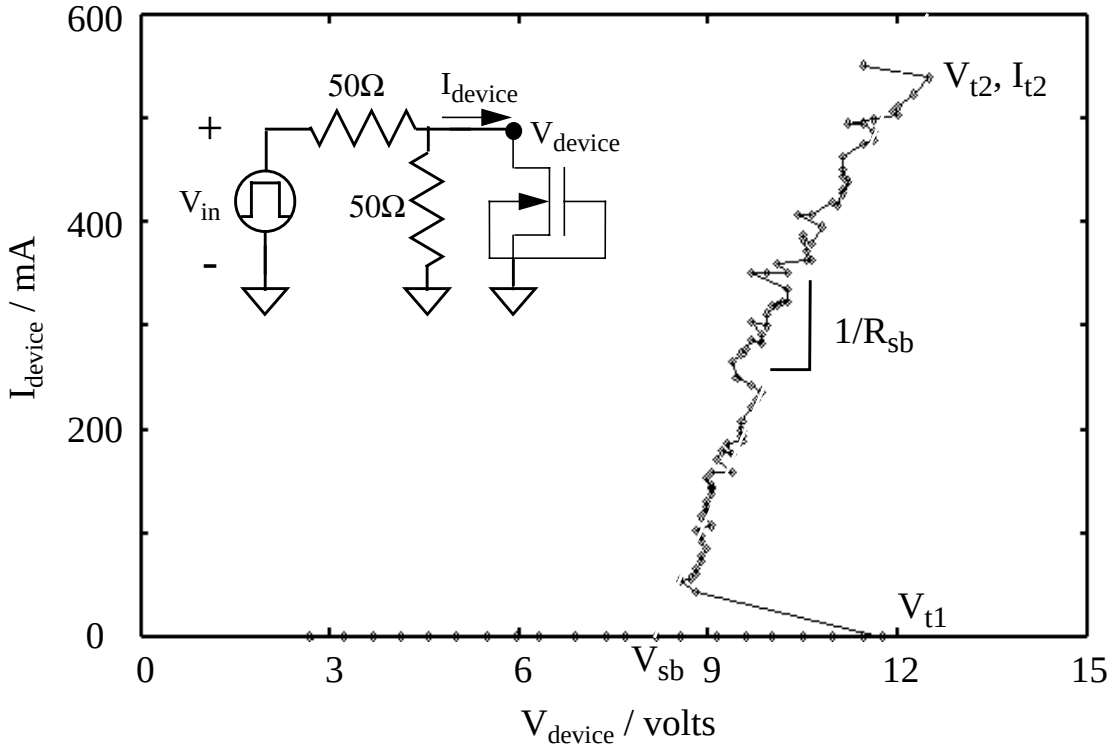

Fig. 4.41 I-V points from the transmission-line pulse sweep of a standard 50/0.75μm test structure (equivalent circuit shown inset). The trigger voltage (V_{t1}), snapback voltage (V_{sb}), snapback resistance (R_{sb}), and second-breakdown point (V_{t2}, I_{t2}) can be extracted from the curve.

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

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

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

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

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

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

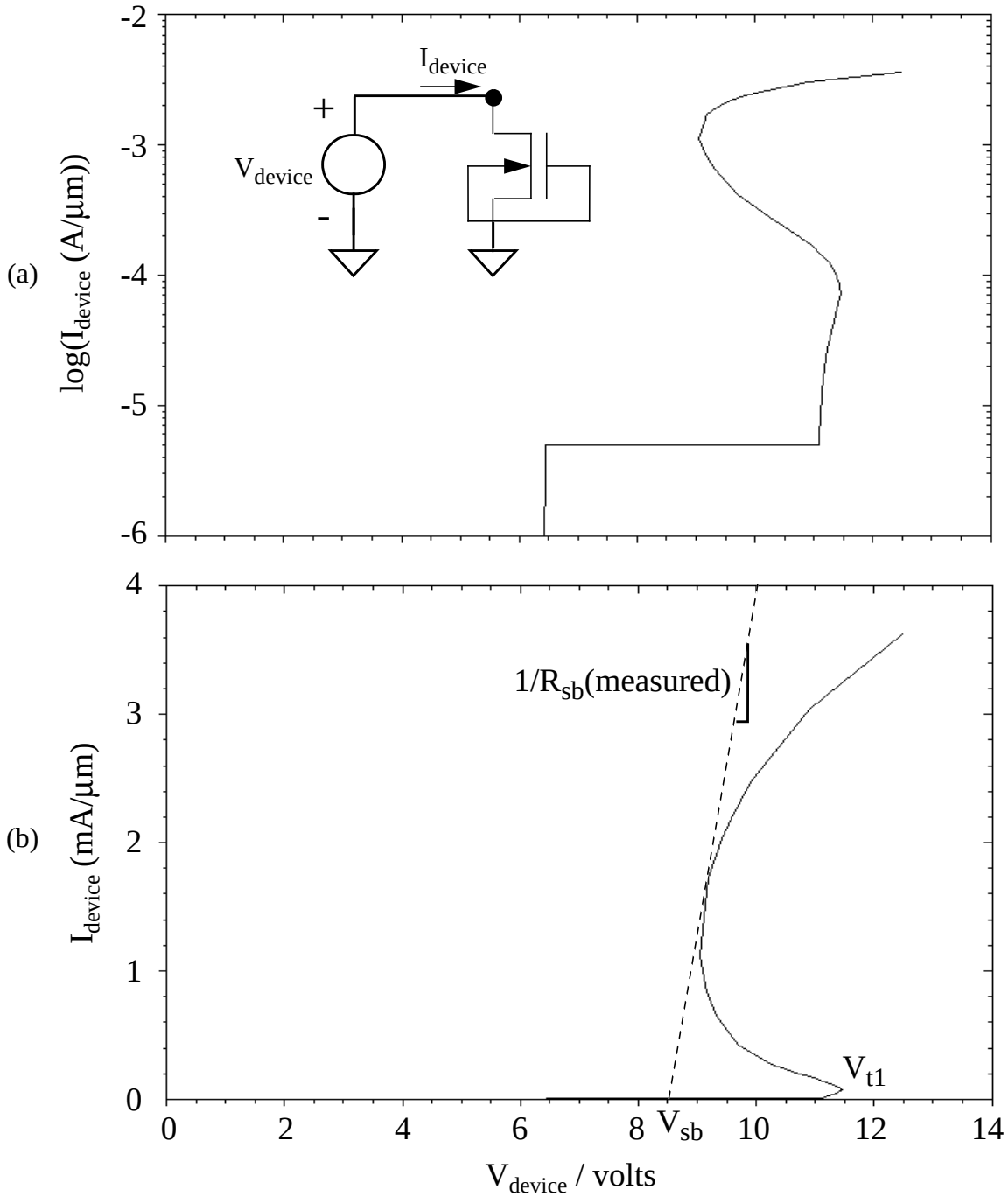

Fig. 4.42 *Device current per width is plotted on a log (a) and linear (b) scale vs. device voltage for a dc-sweep simulation of the standard structure with proper gridding and impact-ionization modeling. The snapback voltage is extracted using a line determined by the measured snapback resistance. To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

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

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

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

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

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

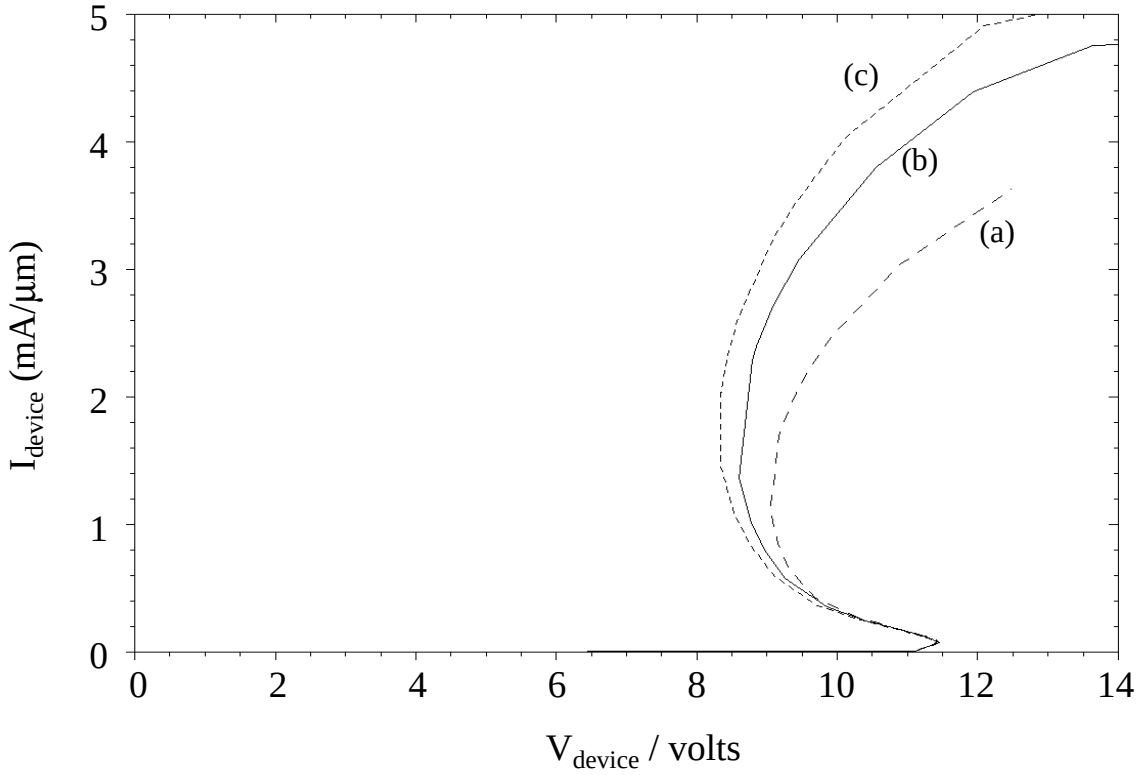

Fig. 4.43 Simulated I-V sweep for T=297K boundary conditions on (a) electrical contacts; (b) perimeter of simulation structure; and (c) perimeter of structure with reduced dependence of impact-ionization rate on temperature.

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

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

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

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

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

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

4.1.4 Calibration of Thermal Failure

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

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

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

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

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

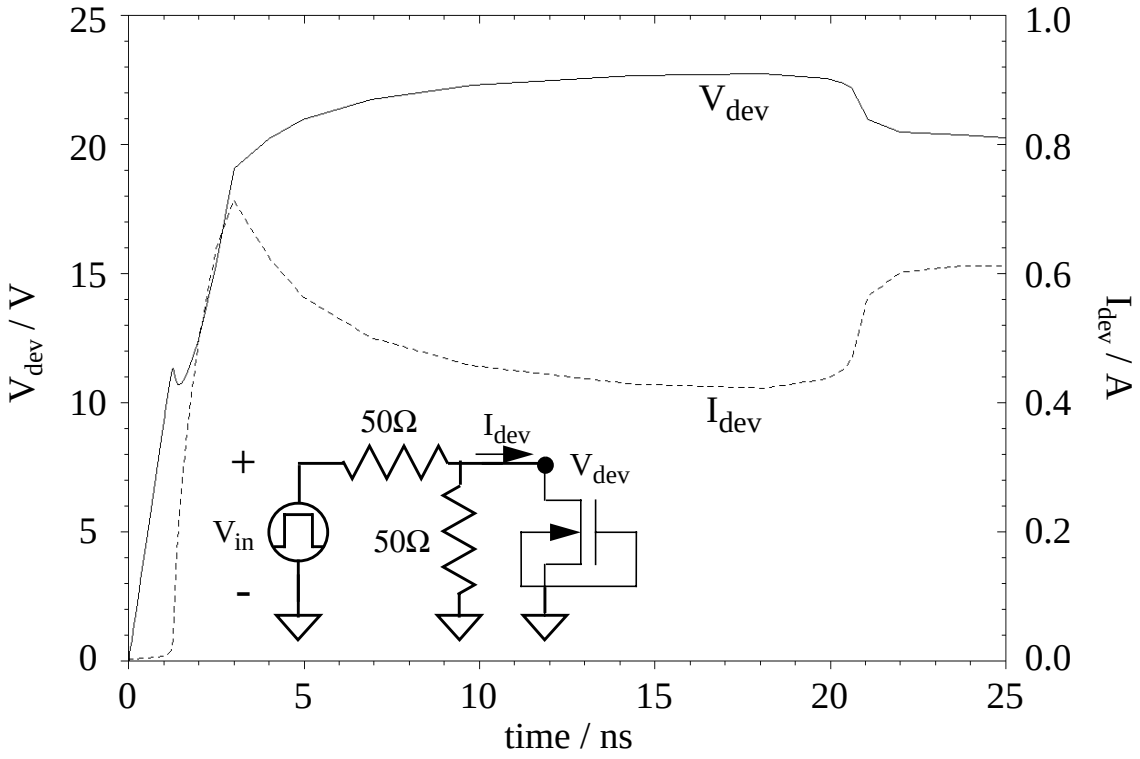

Fig. 4.44 Device voltage and current vs. time for a transient simulation of the 100/0.75μm structure (cf. Fig. 2.10). Second breakdown is observed at 21ns, corresponding to a peak temperature in the device of 1510K. The simulation circuit is shown inset.

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

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

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

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

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

4.2 MOSFET Snapback I-V Results

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

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

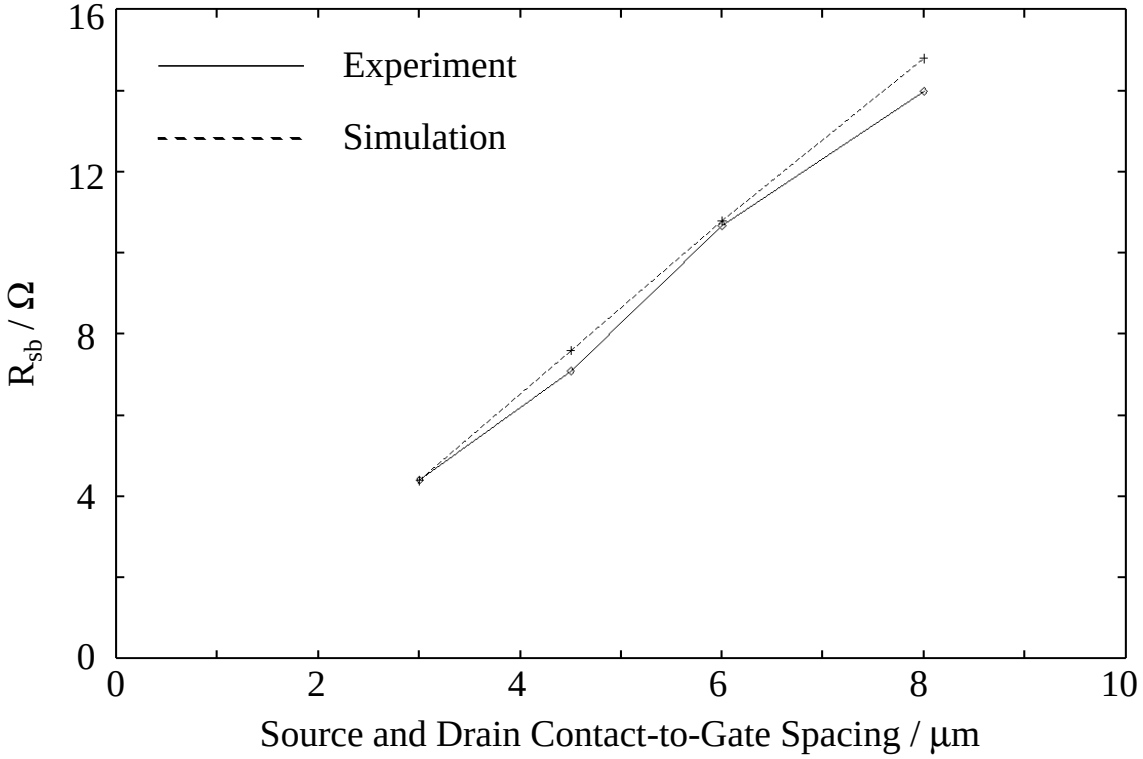

Fig. 4.45 Experimental and simulated snapback resistance, R_{sb}, vs. contact-to-gate spacing for a 50/0.75μm MOSFET test structure. The contact-to-gate spacing refers to the distance from the drain contacts to the gate edge and the source contacts to the gate edge.

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

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

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

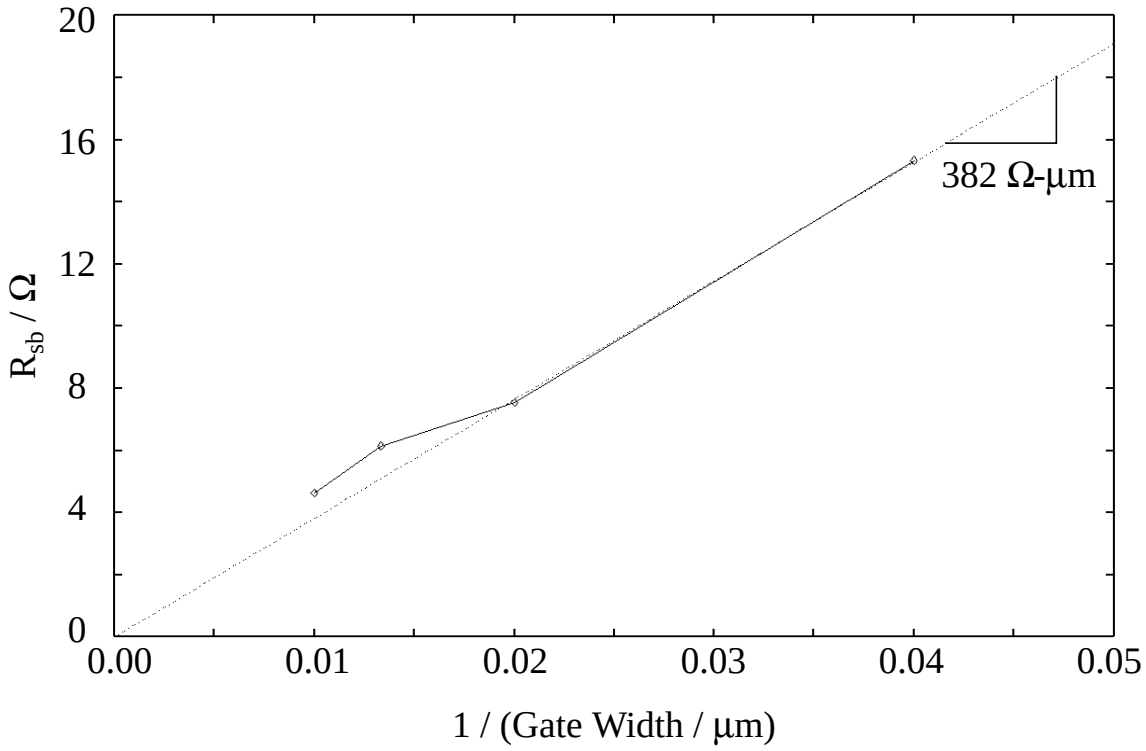

Fig. 4.46 Experimental snapback resistance, R_{sb}, (connected points) vs. inverse gate width, W, for 0.75μm test structures. The dashed line indicates that $R_{sb} \times W = 382$Ω-μm for gate widths less than 50μm.

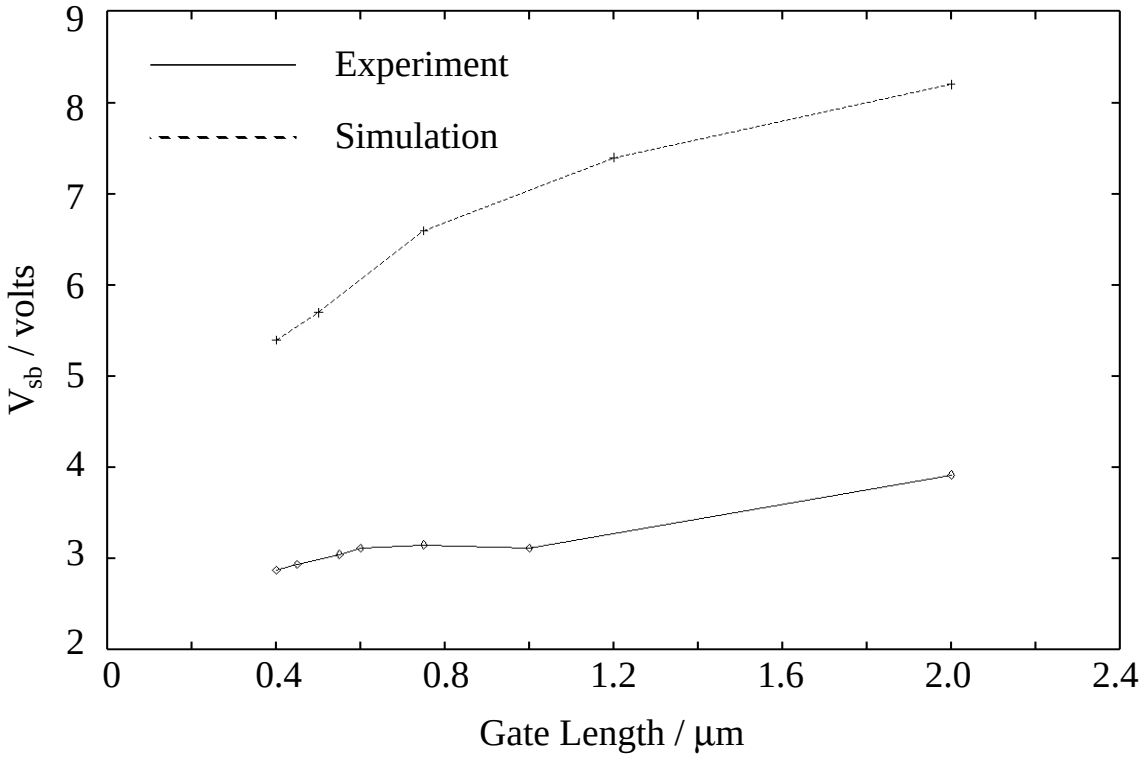

Fig. 4.47 Experimental and simulated snapback voltage, V_{sb}, vs. gate length for 20μm-wide test structures. The experimental results are for fully salicided structures, while the simulation results are for structures with 1.0μm CGS.

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

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

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

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

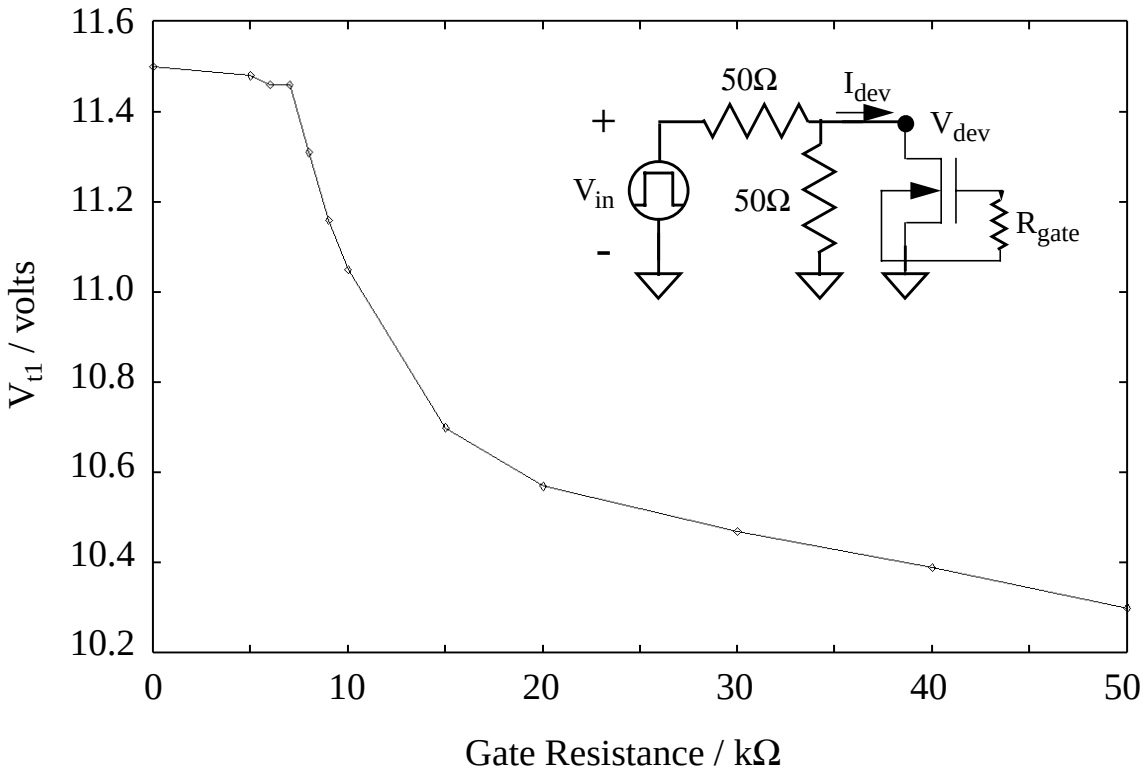

Fig. 4.48 *Simulated trigger voltage, V_{t1}, vs. gate resistance, R_{gate}, for the 50/0.75μm test structure. Simulations predict that R_{gate} does not have a significant effect until it reaches a value of about 10kΩ.*

4.3 Device Failure Results

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

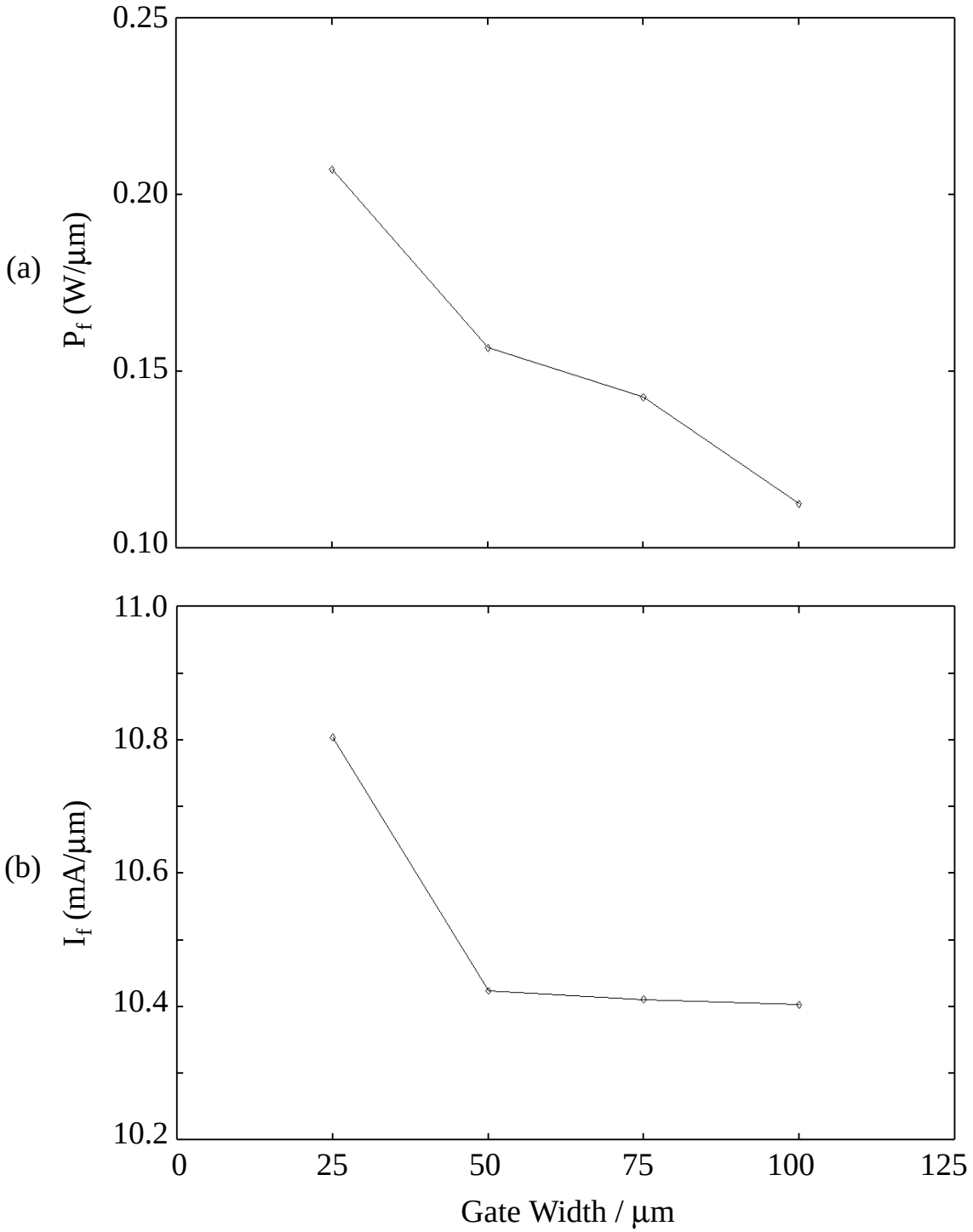

Fig. 4.49 Power to failure, P_f, (a) and current to failure, I_f, (b) vs. device width for 0.75μm test structures subjected to stepped 200ns transmission-line pulses. Each value is divided by the device width to normalize the results.

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

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

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

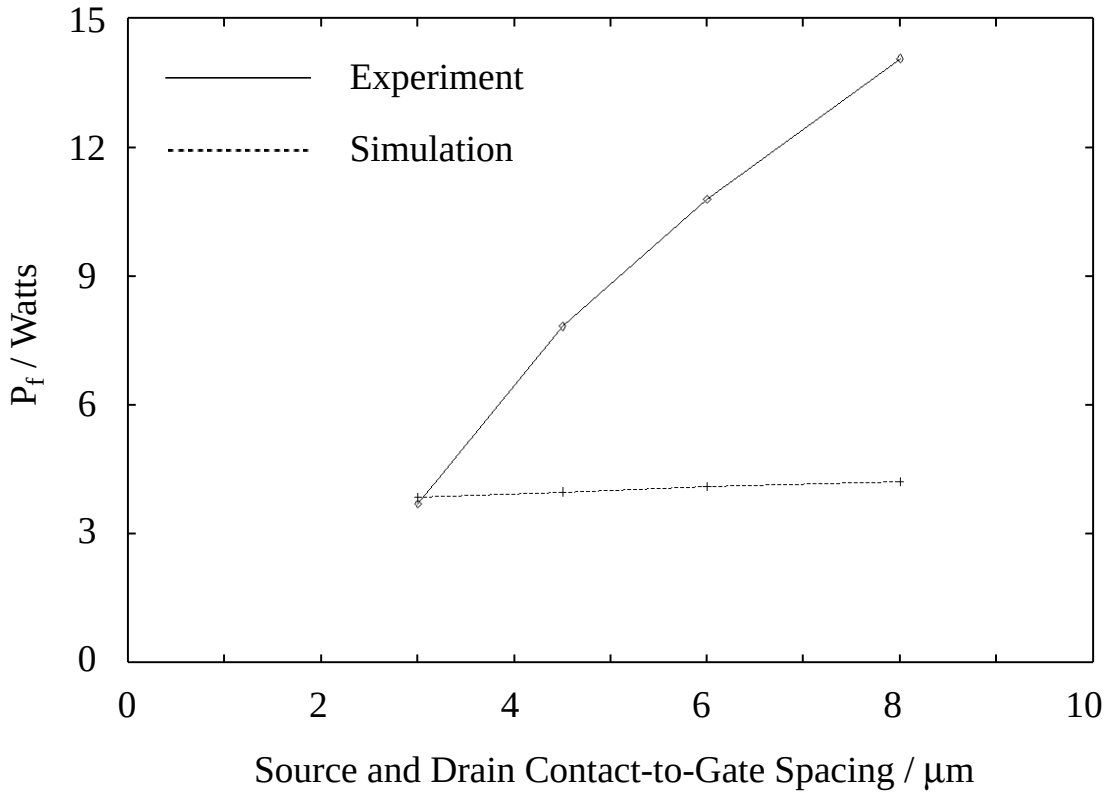

Fig. 4.50 Simulated and experimental power-to-failure, P_f, vs. contact-to-gate spacing for 50/0.75μm test structures subjected to 200ns TLP pulses.

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

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

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

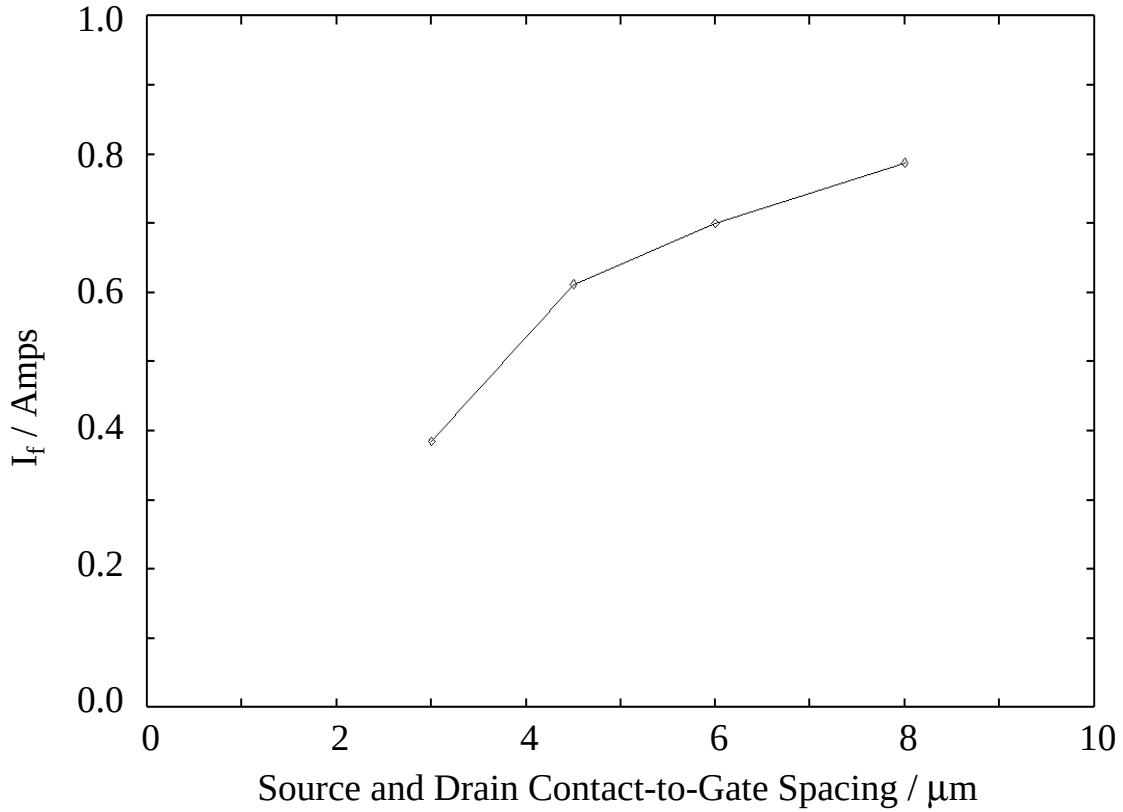

Fig. 4.51 Experimental current-to-failure, I_f, vs. contact-to-gate spacing for 50/0.75μm test structures subjected to 200ns TLP pulses.

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

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

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

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

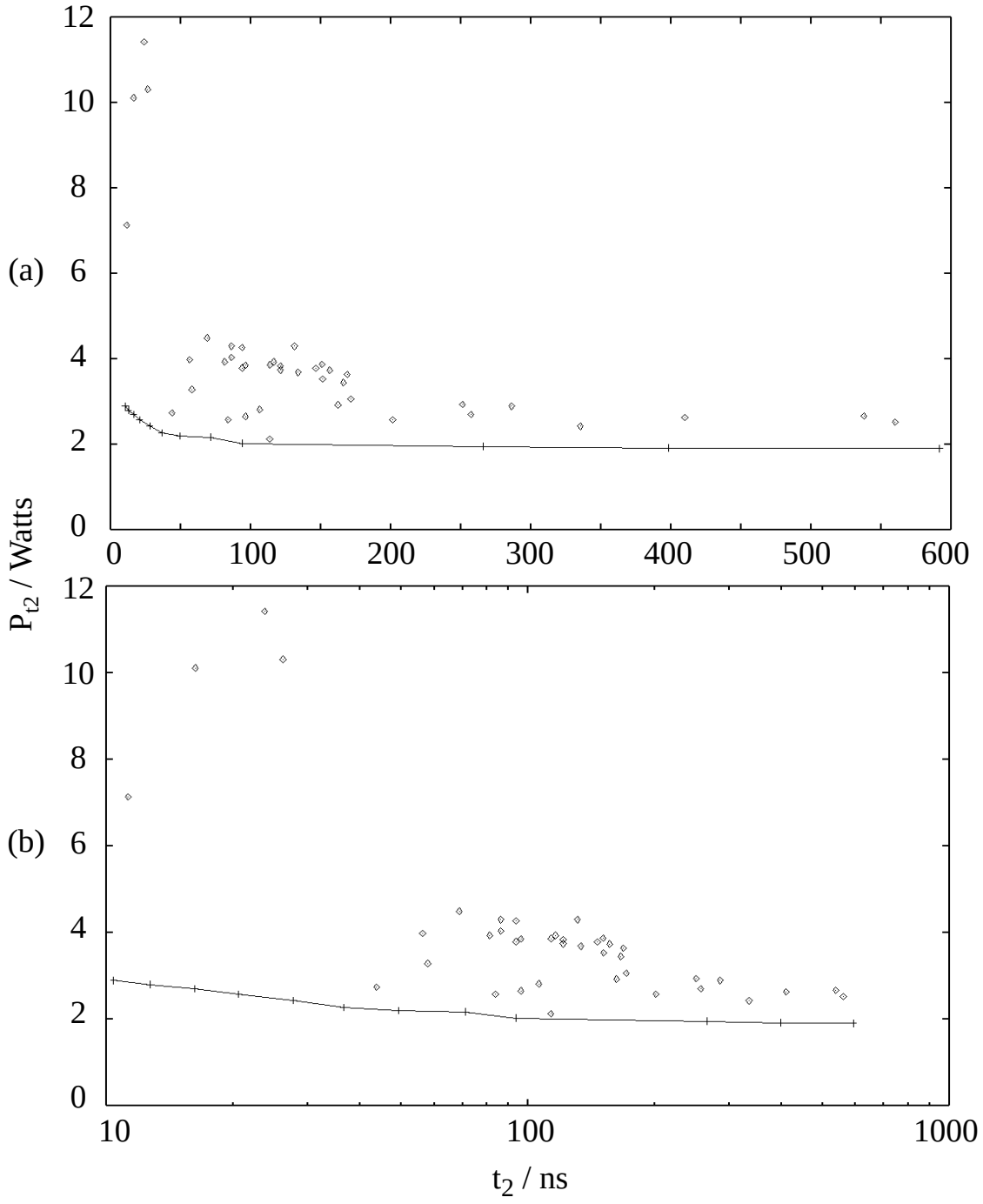

Fig. 4.52 *Power at second breakdown, P_{t2}, vs. time to breakdown, t_2, for a 25/0.75μm structure plotted on linear (a) and semi-log (b) scales. Experimental results (points) are extracted from TLP experiments using various pulse lengths, while simulation points (line) are taken from simulations with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

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

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

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

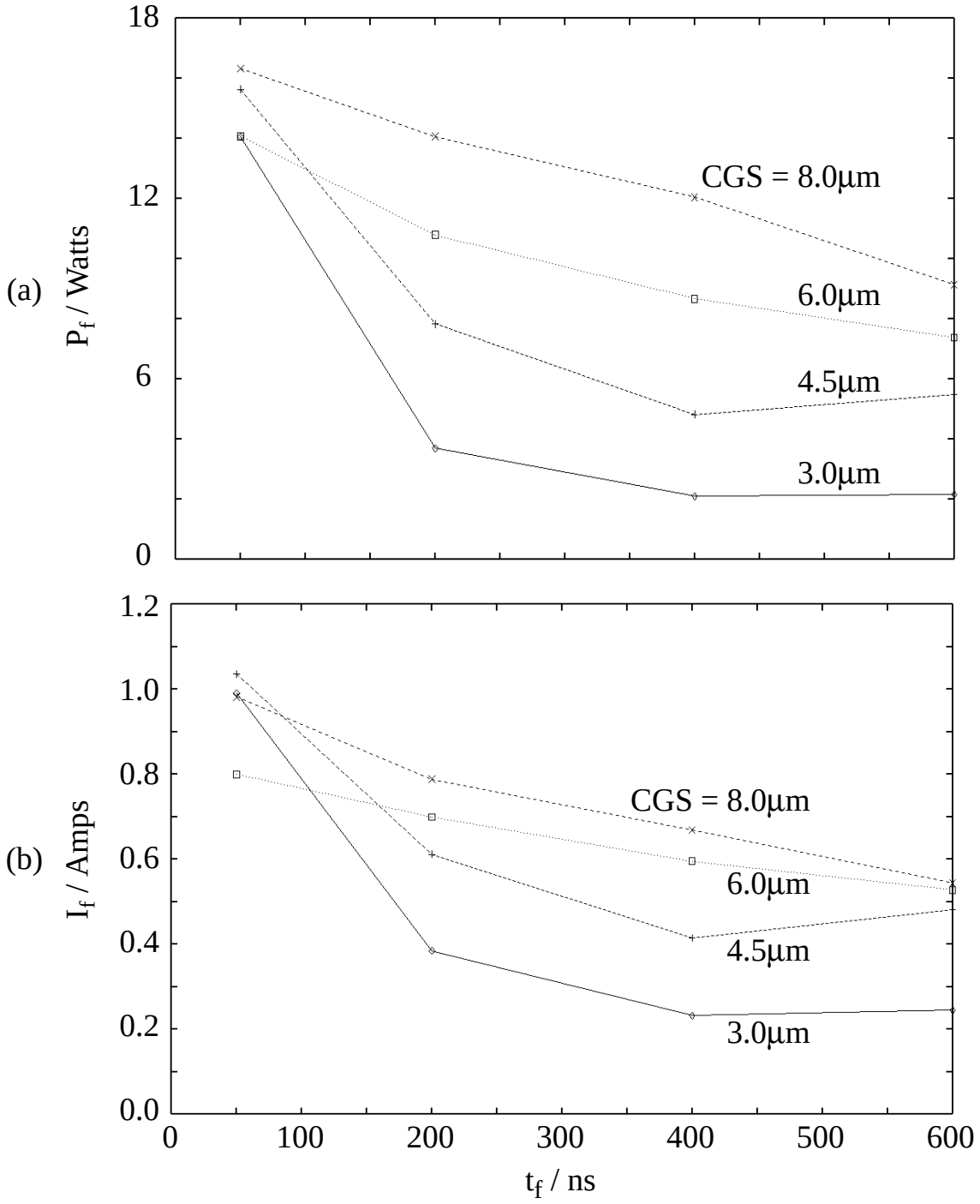

Fig. 4.53 Experimental power-to-failure (a) and current-to-failure (b) vs. time-to-failure, t_f, for 50/0.75μm test structures with varying contact-to-gate spacings (CGS). In these plots, the time to failure is equal to the TLP pulse width and the failure condition is defined as 1μA leakage.

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

4.4 Design Example

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 5

Design and Optimization of ESD Protection Transistor Layout

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

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

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

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

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

5.1 Methodology

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

5.1.1 Characterization of Test Structures

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

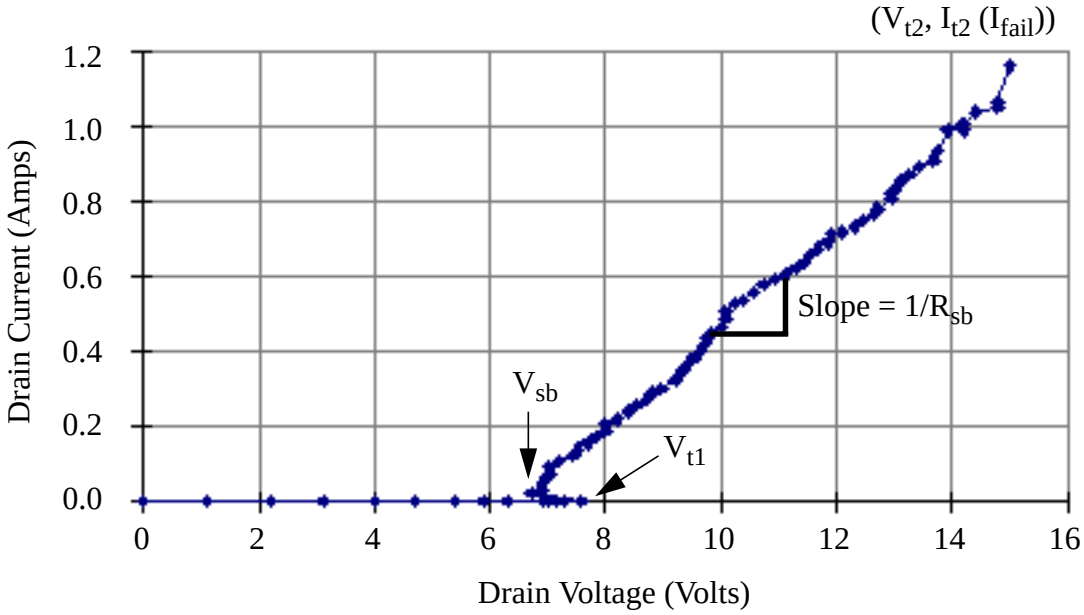

Fig. 5.54 Snapback I-V curve for a 50/0.6μm NMOS transistor generated by TLP. Critical I-V parameters are the trigger voltage (V_{t1}), snapback voltage (V_{sb}), snapback resistance (R_{sb}), and thermal-runaway or second-breakdown point (V_{t2}, I_{t2} (I_{fail})).

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

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

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

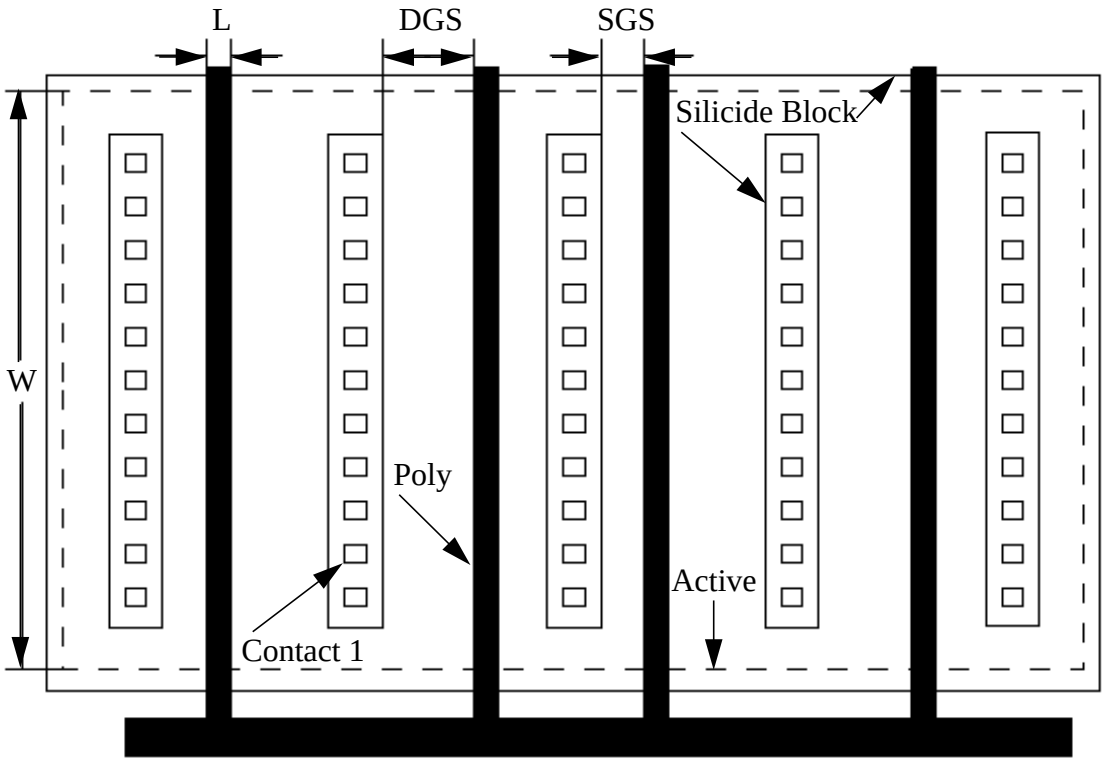

Fig. 5.55 *Layout of a four-fingered ESD structure showing finger width (W), gate length (L), and source (SGS) and drain (DGS) contact-to-gate spacing (actually silicide-to-gate spacing).*

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

5.1.2 Correlation of TLP to the Human Body Model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

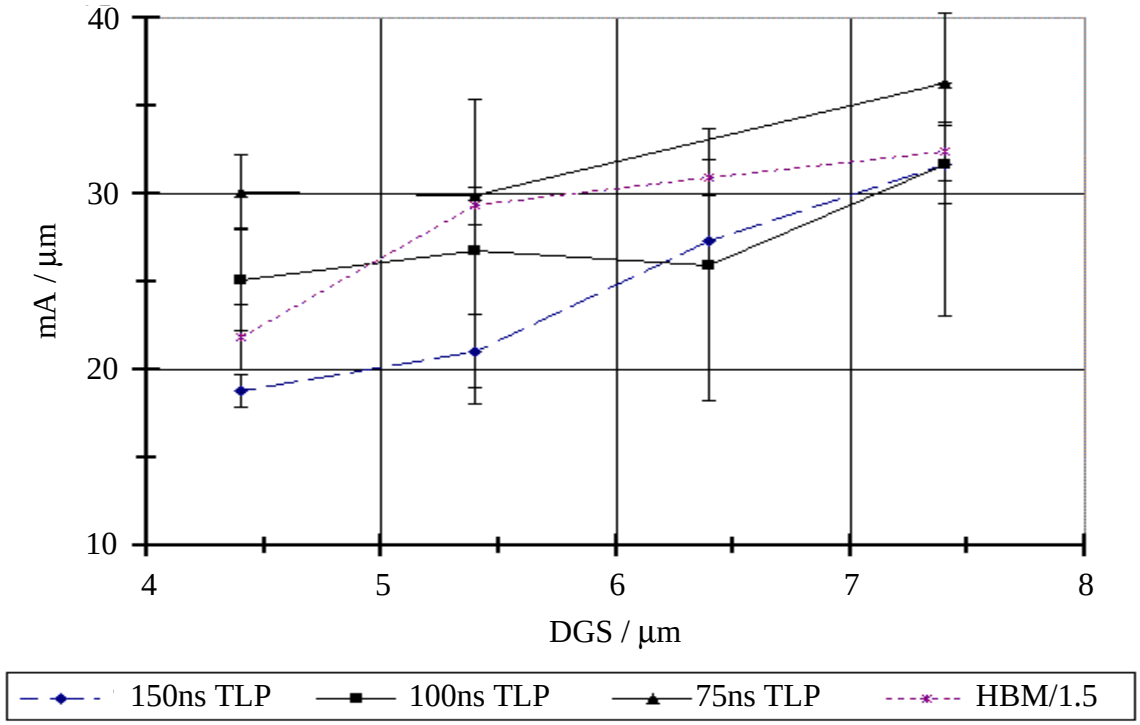

Fig. 5.56 Normalized (divided by width) withstand current vs. drain-side CGS for HBM stressing and 75, 100, and 150ns TLP stressing of 50/0.6µm single-finger transistors. For HBM, the withstand voltage is converted to mA by dividing by 1.5. Error bars represent 95% confidence intervals.

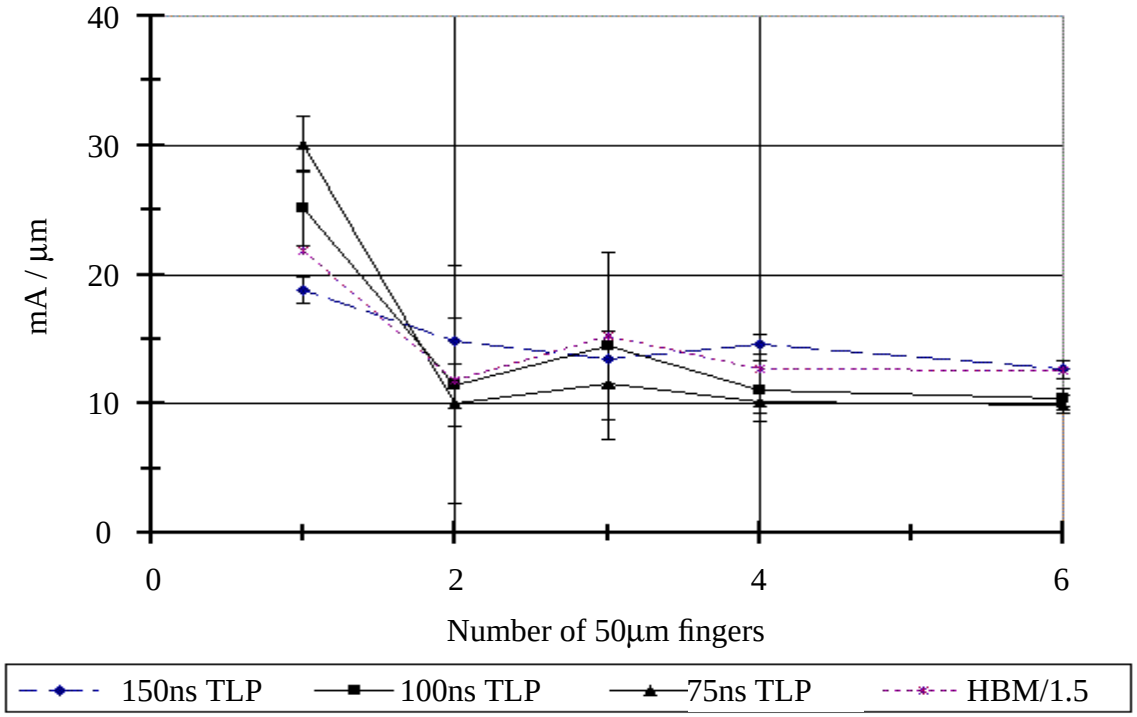

Fig. 5.57 Normalized (divided by width) withstand current vs. number of 50/0.6µm fingers for HBM stressing and 75, 100, and 150ns TLP stressing. Error bars represent 95% confidence intervals.

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

5.1.3 Development of Second-Order Linear Model

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

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

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

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

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

R = Response

F_i = Factor

a_i, a_j = Model Coefficient

$a_{ij}F_iF_j$ = Model Term

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

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

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

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

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

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

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

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

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

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

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

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

5.1.4 Identification of Critical Current Paths

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

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

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

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

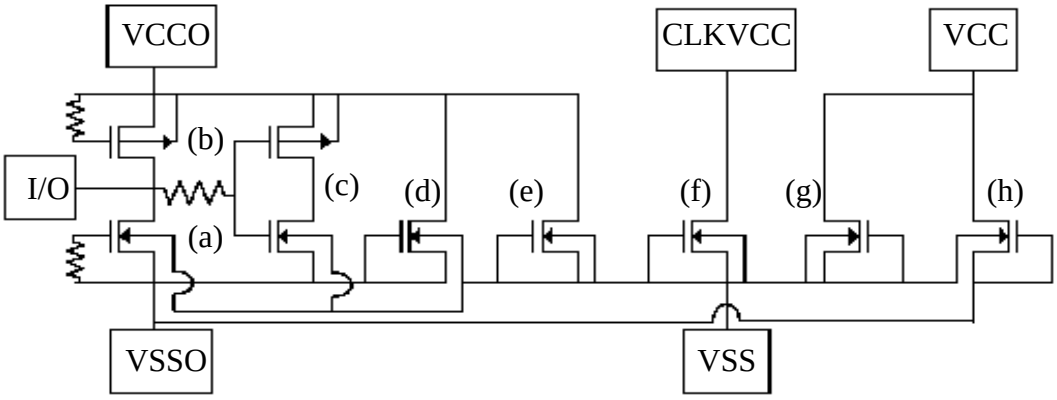

Fig. 5.59 Schematic of critical ESD protection circuits in a chip with split power supplies (VCCO/VSSO and VCC/VSS) and separate clock supply (CLKVCC): (a) n-channel pull-down, (b) p-channel pull-up, (c) CMOS pair representing internal circuitry, (d)-(h) n-channel clamps between various supplies (clamps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

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

5.2 Application

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

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

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

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

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

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

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

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

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

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

Table 5.2 Experimental and modeled SRAM HBM withstand voltages.

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

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

5.3 Analysis

5.3.1 Model Terms

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

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

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

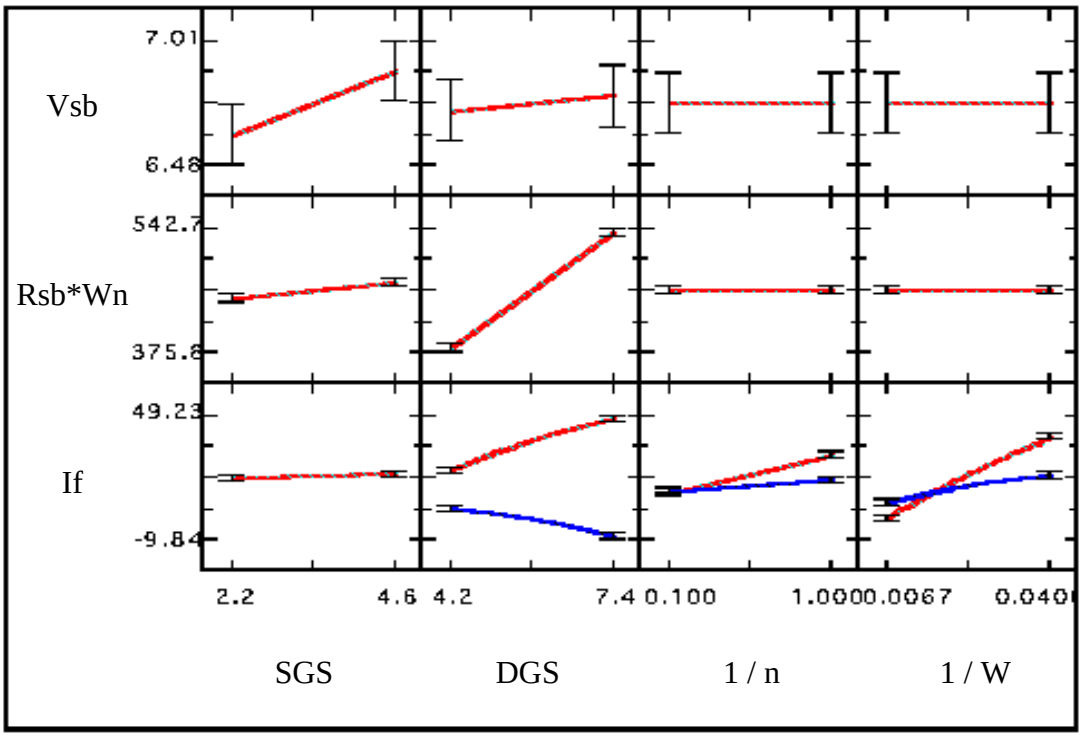

Fig. 5.60 Catalyst model graph for Lot 1 V_{sb}, R_{sb} (multiplied by structure width), and normalized $I_{TLP,ws}$ (If) as a function of SGS, DGS, 1/n, and 1/W.

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

5.3.2 SRAM Model Prediction

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

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

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

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

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

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

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

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

5.4 Optimization

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

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

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

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

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

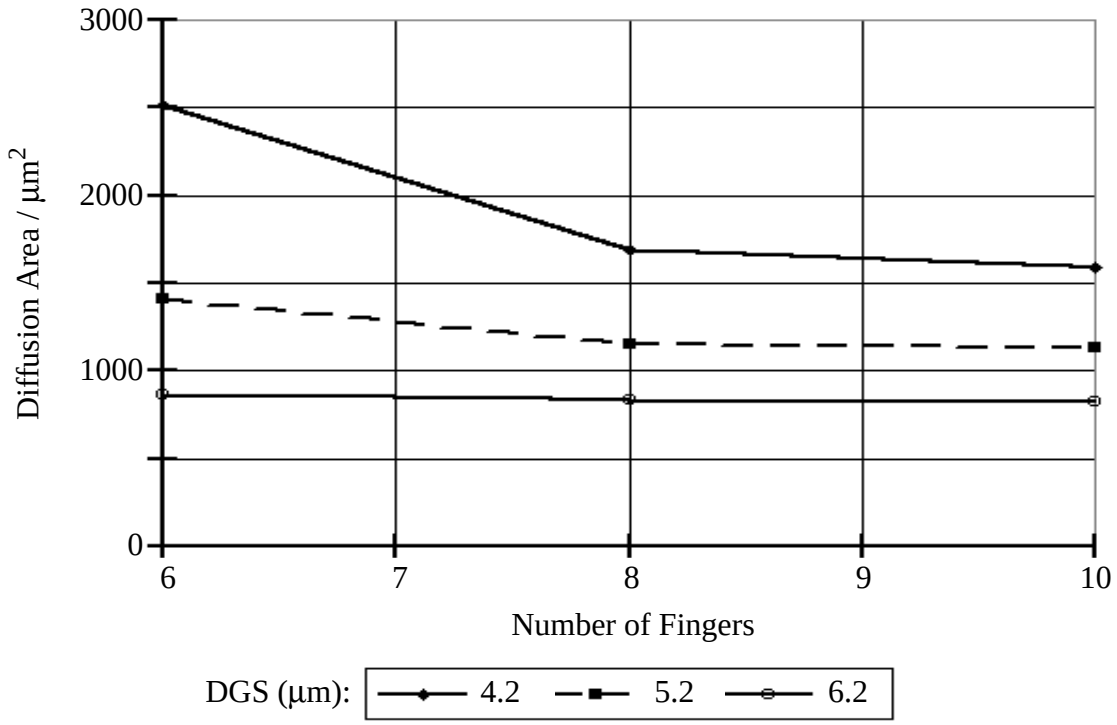

Fig. 5.61 Calculated minimum area of transistor source/drain diffusion needed for 5kV HBM protection for various DGS and number of fingers for Lot 1 with a SGS of 2.2µm.

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

5.5 Summary of Design Methodology

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

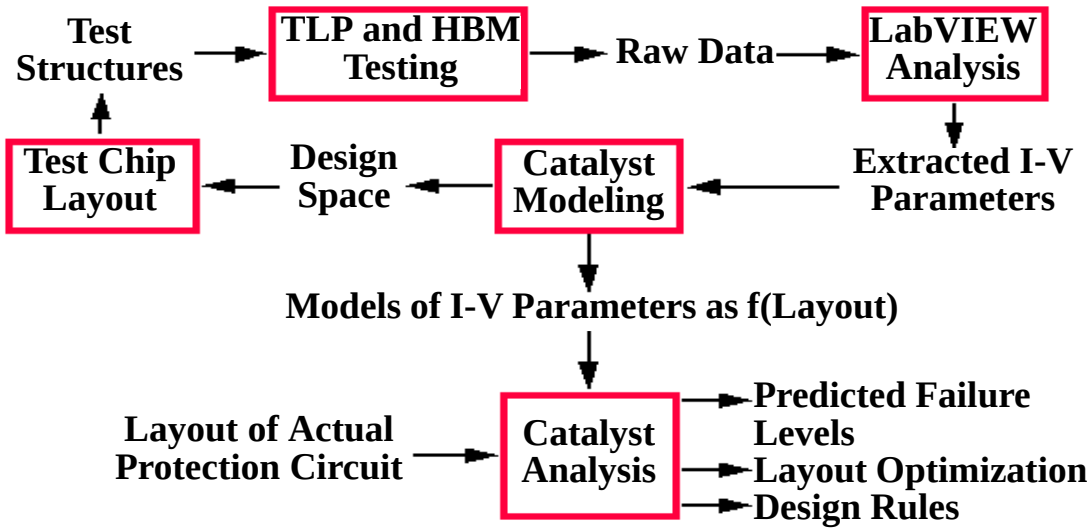

Fig. 5.62 Block diagram of ESD circuit design methodology.

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

Chapter 6
Conclusion

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

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

6.1 Contributions

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

6.1.1 Transmission Line Pulsing

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

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

6.1.2 Numerical Device Simulation

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

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

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

6.1.3 Design Methodology

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

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

6.2 Future Work

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

6.2.1 Characterization

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

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

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

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

6.2.2 Modeling

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

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

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

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

6.2.3 Design

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

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

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

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

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

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

Appendix A
Tracer User's Manual

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

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

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

A.1 Command Line

Usage: `tracer inputfile tracefile [outputfile]`

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

A.2 Trace File

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

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

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

A.3 CONTROL Card

A.3.1 Description

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

A.3.2 Syntax

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

A.3.3 Parameters

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

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

A.3.4 Examples

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

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

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

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

A.4 FIXED Card

A.4.1 Description

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

A.4.2 Syntax

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

A.4.3 Parameters

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

A.4.4 Examples

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

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

A.5 OPTION Card

A.5.1 Description

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

A.5.2 Syntax

Simulations with one or two open contacts:

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

Smoothness and step-size control:

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

Control of output files:

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

A.5.3 Parameters

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

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

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

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

A.5.4 Examples

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

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

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

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

A.6 SOLVE Card

A.6.1 Description

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

A.6.2 Syntax

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

A.6.3 Parameters

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

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

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

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

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

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

A.6.4 Examples

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

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

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

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

A.7 Input Deck Specifications

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

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

A.7.1 Load and Solve Cards

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

A.7.2 Contact Card

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

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

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

A.7.3 Method Card

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

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

A.7.4 Options Card

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

A.8 Data Format in Output Files

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

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

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

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

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

A.9 Examples

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

A.9.1 BV_{CEO}

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

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

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

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

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

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

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

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

solve
end
```

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

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

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

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

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

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

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

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

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

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

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

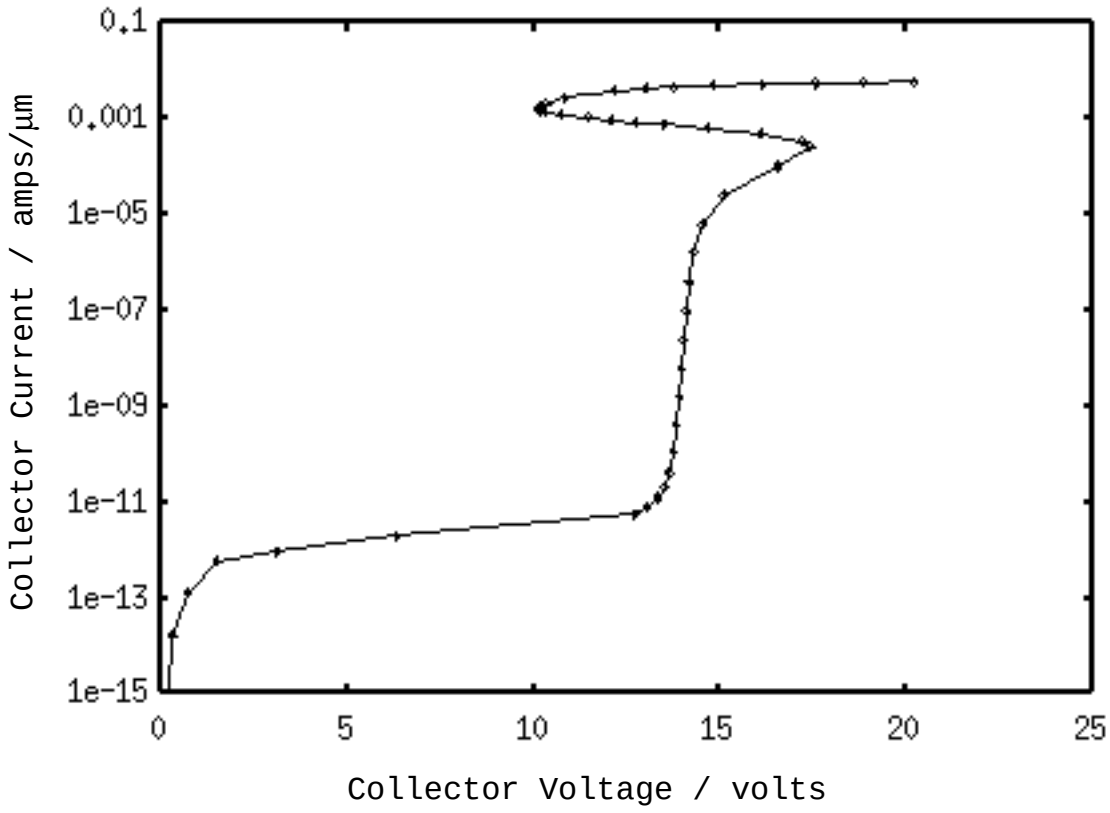

Fig. A.66 Collector current vs. collector voltage for the BV_{CEO} example.

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

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

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

A.9.2 GaAs MESFET

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

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

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

```
title mes.pis

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

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

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

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

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

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

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

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

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

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

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

solve ini

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

end
```

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

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

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

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

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

```
title mesvg.5.pis

option nowarn curvetrace

mesh inf=mes.mesh

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

contact num=2 alu workf=4.84 surf

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

load infil=mesvg.5.ini
solve

end
```

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

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

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

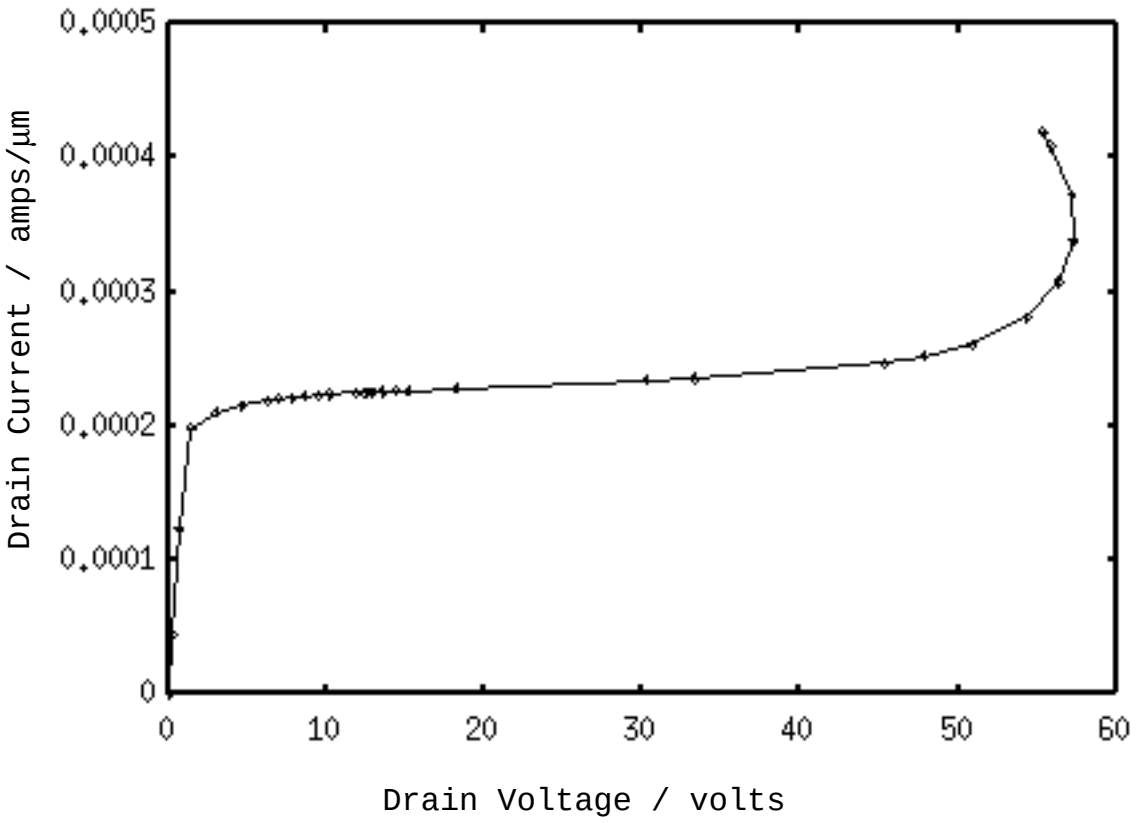

Fig. A.71 Drain current vs. drain voltage for the GaAs MESFET example.

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

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[74] S. Ramaswamy, E. Li, E. Rosenbaum, and S.-M. Kang, “Circuit-Level Simulation of CDM-ESD and EOS in Submicron MOS Devices,” *Proc. 18th EOS/ESD Symp.*, 1996, pp. 316-321.

[75] S.L. Lim, X.Y. Zhang, S. Beebe, and R.W. Dutton, “A Computationally Stable Quasi-Empirical Compact Model for the Simulation of MOS Breakdown in ESD Protection Circuit Design,” *Proc. Intl. Conf. on Simulation of Semiconductor Processes and Devices*, 1997, pp. 161-164.