ELSEVIER

Journal of Electrostatics 36 (1995) 81–92

Journal of
ELECTROSTATICS

Fast turn-on of an NMOS ESD protection transistor: measurements and simulations[☆]

J.R.M. Luchies[a,]*, C.G.C.M. de Kort[b], J.F. Verweij[a]

[a] *MESA Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands*
[b] *Philips Research Laboratories, WA-1.159, Prof. Holstlaan 4, 5656 AA Eindhoven, Netherlands*

Received 26 October 1994; accepted after revision 15 March 1995

Abstract

The transient turn-on of the parasitic bipolar transistor of an NMOS transistor was studied. The voltages appearing at internal nodes of protection and functional circuit after application of 350 ps rise-time pulses have been measured using electro-optic sampling. For very fast transients the triggering of the protection transistor shifts from an avalanche multiplication current towards a displacement current-induced triggering, thereby lowering the trigger voltage. With our circuit simulation mode we are able to predict the outcome of human body model and charged device model testing.

Keywords: ESD; CDM; HBM; Electro-optical sampling; Simulator; Transient modeling

1. Introduction

The lateral bipolar npn transistor formed by the source, drain and substrate of an NMOS transistor is often applied as an electro-static discharge (ESD) protection transistor in CMOS circuits. For very fast pulses, which are typical for charged device model (CDM) tests, failure modes different than those in standard human body model (HBM) and machine model (MM) testing may occur [1]. Typical CDM failures are usually found in the gate oxides of the functional circuit instead of the protection transistor. It is expected that the different turn-on times of the individual elements of the protection circuit become of critical importance. For very fast transients the protection transistors may not trigger into snapback fast enough and gate oxides in the functional circuit can be damaged (Fig. 1).

[☆] © 1994. Reprinted with permission, after revision, from Electrical Overstress/Electrostatic Discharge Symposium Proceedings, EOS-16, Las Vegas, NV, USA, September 27–29, 1994.

* Corresponding author. Tel.: +31 53 894007, Fax. +31 53 341903 (after 10 October 1995 resp. +31 53 4894007 and +31 53 4341903). E-mail: janmarc@ice.el.utwente.nl.

SSDI 0304-3886(95)00023-2

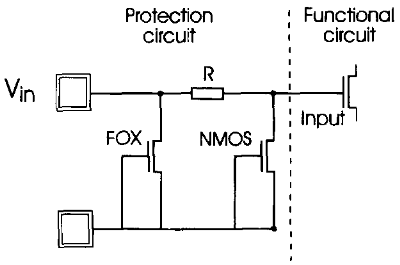

Fig. 1. Example of an input protection circuit and functional circuit, showing the individual elements, a field oxide transistor, FOX, an NMOS transistor and a resistor, R, and some internal nodes of the circuit.

Proper understanding of the protection circuit behaviour requires knowledge of the voltages at the internal nodes during a fast transient. In several publications external measurements of voltages and currents during an ESD event were shown [2,3]. But these measurements give only limited insight into the voltages appearing inside the circuit. With techniques such as electro-optic sampling [4] and e-beam [5] it is possible to determine the voltage waveforms inside the chip under periodic stress conditions.

In this paper we will determine the response of the parasitic bipolar transistor of the NMOS device during fast transient pulses, by measuring the internal node voltages. A circuit model will be used to explain the measurements. 2D-device simulations have been performed to obtain more accurate parameter values for the bipolar transistor model. The transistor model is then used to explain the outcome of ESD stress tests.

In the first section theory of triggering and turn-on of bipolar transistors will be covered. Then the experimental set-up and transistor model will be described. Hereafter, the experimental results will be shown and modelling of the transient behaviour will be discussed. The transient transistor model will then be used to model ESD stress tests. Finally, conclusions are drawn.

2. Protection transistor triggering

There are several ways to trigger a bipolar transistor into snapback. The most common is avalanche generation in the reverse biased collector–base depletion region due to a high electric field. In a circuit model for the transistor, a current source for the avalanche current can be added. This type of triggering can be modelled with the multiplication factor, M, using the well-known Miller formula [6]:

$$M = \frac{1}{1 - (V_{\mathrm{cb}}/V_{\mathrm{cbO}})^n}, \tag{1}$$

where V_{cb} is the applied collector–base voltage, V_{cbO} the collector–base breakdown voltage and n an empirical constant. The multiplication factor can also be included in the expressions for the collector and base currents. The base current I_B is then given by (see Appendix A):

$$I_B = I_S(T)\frac{1 - M\alpha_F}{\alpha_F}\left[\exp\left(\frac{-qV_{eb}}{kT}\right) - 1\right], \tag{2}$$

where α_F is the forward base current gain, and V_{eb} the internal emitter–base voltage. When $M\alpha_F \geqslant 1$ the bipolar transistor can provide its own base current and will subsequently go into snapback.

The trigger voltage, V_{t1}, in a quasi-DC situation and without an external base–emitter voltage applied, for which the criterion $M\alpha_F \geqslant 1$ is valid, can be written as

$$V_{t1} \approx V_{cb0} + V_{be,on} = V_{cb0} - I_B R_b, \tag{3}$$

where R_b is the base resistance and $V_{be,on}$ the internal base–emitter voltage. Although it must be stated that the collector–base breakdown voltage is a strong function of the junction curvature. Therefore, in a quasi-DC approach, 3D effects will play an important role and deviations from the trigger voltage as given in Eq. (3) may be found. The sustaining voltage is given by [7]

$$V_s = V_h + I_C(R_c + R_e), \quad \text{where } V_h \approx V_{be,on} + V_{cb0}[\beta_F + 1]^{-1/n}, \tag{4}$$

where V_h is the holding voltage, and β_F the current gain (see Appendix A).

Several ways have been proposed to lower the trigger voltage. One way is to bias the gate voltage of the NMOS transistor [8], another is to bias the substrate (the base of the bipolar transistor) [9]. These methods rely on other ways of providing a base current than the usual avalanche. For transients, the capacitive coupling of the collector and base comes into play. In this case the base–emitter voltage drop needed to trigger the transistor may be provided by the additional collector–base capacitor displacement current. This type of triggering is also known as dV/dt triggering, which has been experimentally identified with e-beam [5, 10] and electro-optic sampling techniques [11]. The current through the collector–base capacitor can be written as

$$I_{C_{cb}} = C_{cb}\frac{dV_{cb}}{dt}. \tag{5}$$

The circuit model has thus to be extended with the junction and depletion capacitors. The complete model is shown in Fig. 2. The capacitors have been added to the model as non-linear elements. For very fast transients, the combination of base resistance and collector–base and base–emitter capacitors will be the determining factor in the triggering behaviour. This combination of capacitors and base resistance will then more or less act as a capacitive voltage divider. A simplified model for transient triggering is shown in Fig. 3.

Still the avalanche multiplication current has to become high enough to sustain the snapback mode, because for snapback the criterion $M\alpha_F \geqslant 1$ still holds. At the moment of triggering, the base–emitter voltage has become 0.6 V, then the injection of electrons from the emitter into the base starts. The electrons may then travel to the

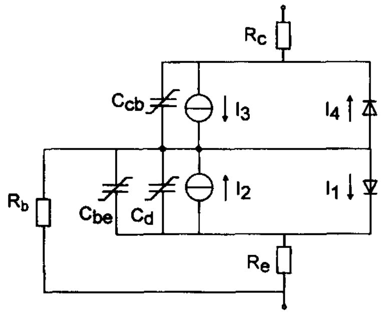

Fig. 2. Lumped element model of the protection transistor with the capacitors for transient simulations. The currents I_1 to I_4 are explained in Appendix A.

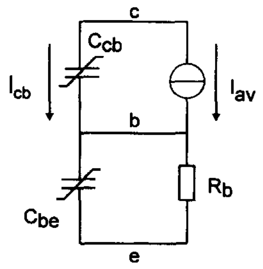

Fig. 3. Simplified lumped element model of the protection transistor for triggering both by the avalanche multiplication current as well as by the displacement current of the collector–base capacitor.

collector or recombine with the excess majority carriers in the base. The collector current will increase, but the actual turn-on will be determined by time needed to charge the base with excess minority carriers.

The turn-on time of a bipolar transistor can be modelled with the diffusion capacitance C_D. This capacitance models the excess minority carrier charge stored in the base. For the standard case, the charge in the base is provided by the base current, and subsequently the charge stored on the diffusion capacitance is modelled by

$$Q(C_D) = \tau_{BF}(I_3 - I_4), \qquad (6)$$

where τ_{BF} is the forward charge-control variable. The base charge can only be provided by the portion of the collector current which is not contributing to the negative base current.

3. Experimental set-up

We have applied the electro-optic sampling technique [12] to measure the internal node voltages with high accuracy (time resolution: 50 ps, voltage accuracy: 2 mV). The technique makes use of an electro-optic crystal, which is placed in the vicinity of a metal line (Fig. 4). A small probe-tip is positioned at the metal line. Optical pulses generated by a diode laser are then used to 'scan' the electric signal inside the e-o crystal. This signal is converted into an electric signal which is fed back to a reference electrode. In this way it is possible to keep the detected signal zero and then the signal amplitude at the reference electrode equals the signal amplitude on the probe-tip (the metal line) at the time of sampling. The measurement technique can also be used at high impedance nodes (input gates), since the capacitive load is smaller than 30 f F [4]. The technique is capable of probing at 1 μm wide metal lines. Furthermore, very high amplitudes may be measured with this technique.

We used an HP 8116A pulse generator, to generate the 4 ns rise-time pulses. Since the technique is a sampling technique, the applied pulses were kept low in amplitude and width to prevent sample heating. After each measurement sequence the leakage current was measured to see whether the sample was damaged. For the very fast rise-time pulses we used an Avtech pulse generator to generate 350 ps rise-time pulses.

Samples with two different layouts for a 1 μm CMOS technology were available. The protection transistor was a 100/1 NMOS transistor. The samples differed in the resistor between the input-pad and the first gate oxide. Sample one had only a few Ohm resistance, whereas sample two had 2 kΩ resistance. The voltage waveforms were measured at four nodes inside the chip (Fig. 5). The voltages at the drain (1) and source (2) of the protection transistor could be measured. Furthermore, the voltage across the gate (3-2) could be measured. By measuring the voltage across the ground track (2-4), the current through the protection device can be extracted.

A compact circuit model including avalanche multiplication of the bipolar transistor in the protection circuit has been developed. The circuit simulations were carried

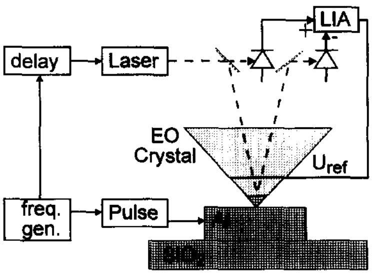

Fig. 4. Schematic representation of the electro-optic sampling set-up, showing a frequency generator for the sampling frequency, a delay to 'scan' the pulse from the pulse generator. A second output of the frequency generator is used to trigger the diode laser, which emits optical pulses of 50 ps duration. The lock-in amplifier determines the detected difference in signal.

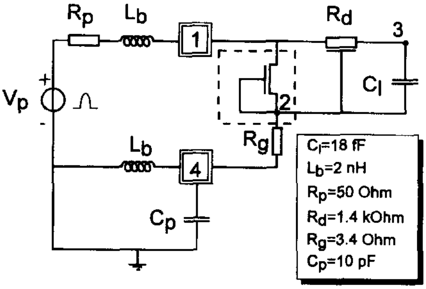

Fig. 5. Lumped model of the protection circuit including the model for the parasitic bipolar transistor. The abbreviations in the figure are: the resistance of the ground track, R_g, the inductance of the bondwires, L_b, a diffusion resistance, R_d, parasitic capacitance, C_p, the pulse generator, V_p, and its internal resistance, R_p. The input gate is modelled with a capacitor, C_l.

out using Pstar as a simulator [13]. The complete equivalent circuit model of the experimental set-up is shown in Fig. 5. The model accounts for the pulse generator including its 50 Ω resistance and all parasitic inductors, which have been shown to play an important role in the measurements. Most of the parameter values were extracted with DC measurements. The parasitic inductors and capacitors were fitted, and have reasonable values. The parameters for the junction capacitors were measured. For the depletion capacitance a standard model has been implemented. To obtain accurate model parameter values for the protection transistor (such as for the base resistance), 2D-device simulations have been carried out with MEDICI. The base resistance was found to be 130 Ω. The collector resistance was 5 Ω.

4. Experimental results

In Fig. 6 the measured and simulated pulses for 4 ns rise-time pulses are depicted. The model can accurately describe the transient behaviour for these pulses: when the input voltage exceeds the trigger voltage $V_{t1} \approx 14$ V, the protection transistor goes into snapback and the voltage drops to the sustaining voltage $V_s \approx 11$ V. In this way, the protection transistor can adequately protect the 20 nm input gate oxide. Avalanche is in this case the dominant factor in the triggering behaviour of the parasitic bipolar transistor. The time needed from triggering towards snapback is approximately 0.9 ns. This time is approximately equal to the turn-on time of the lateral bipolar transistor which should be in the order of 1 ns [14].

For a short rise-time pulse ($t_{rise} = 350$ ps), the voltage pulse is depicted in Fig. 7 for the measurement and the simulation. The trigger voltage is approximately 12 V, which is lower than for the 4 ns rise-time pulse. This means that the avalanche multiplication current is not the only collector current in the turn-on behaviour. The simulations agree quite satisfactorily with the measurements; the same trigger and sustaining voltages are found. The transition from triggering towards snapback seems

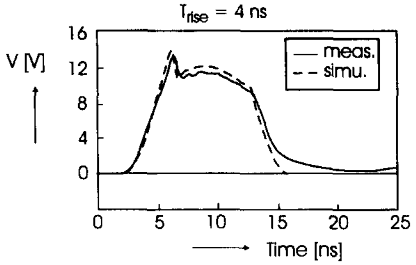

Fig. 6. Voltage appearing across the grounded-gate NMOS transistor after application of a 4 ns rise-time pulse. After triggering at 14 V, the voltage is clamped to the sustaining voltage of 11 V.

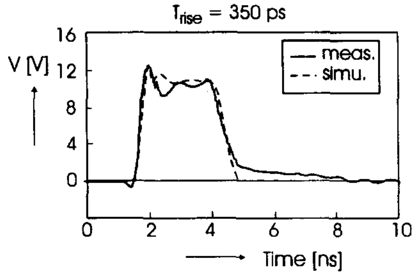

Fig. 7. Voltage appearing across the grounded-gate NMOS transistor after application of a 350 ps rise-time pulse. After triggering at 12 V, the voltage is clamped to a safe value (10 V) for the input gate.

to take an estimated 0.35 ns, which is much shorter than for the former pulse. The reason for this difference is however unclear. The simulations predict a somewhat different turn-on behaviour. A more detailed model of the turn-on time seems to be appropriate.

The current flowing through the protection transistor was determined by measuring the voltage across the ground track of the circuit. The resistance of the ground track was determined with DC measurements ($R_g = 3.4\ \Omega$). Both measurement and simulation are shown in Fig. 8. The measurements indicate that the protection is conducting some 250 mA during the fast pulse. The simulation predicts however a lower value. Also large oscillations are found in the measurements, which are not seen in the simulations. This may be due to the subtraction of the two voltages measured at each side of the ground track. Furthermore, the differences between the simulations and the measurements may be explained by the inaccuracies in the modelling of some parasitic elements, such as the bondwire to ground. The simulations showed that the increase in temperature for these currents (and duty-cycles) is negligible.

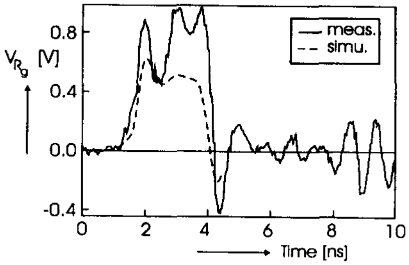

Fig. 8. Voltage appearing across the ground track of the protection structure after application of the 350 ps rise-time pulse. Herewith it is possible to determine the current through the protection transistor.

5. ESD-test measurements and simulations

Since our model seems accurate enough to describe fast transient measurements, we have applied the model to explain peculiar results in standard ESD tests. The measurements were carried out on two types of samples, which gave the same HBM results but differed for the CDM test. The fail voltages are shown in Table 1. All simulations were performed for HBM and CDM fail voltages to investigate the mechanisms that caused the failures.

Apparently, the failure modes induced by the HBM and CDM are different. Failure analysis revealed that the protection transistor was damaged for HBM stress, whereas the gate oxide of the functional circuit was ruptured in case of CDM stress. We have therefore compared the voltage appearing across the gate oxide of the functional circuit for both samples using circuit simulation. The simulations were carried out using the protection circuit model as described before and with HBM and CDM circuit equivalents. The simulations were done for the fail voltages (Table 1) and the results are shown in Fig. 9.

For relative slow pulses (like in HBM; rise-time ≈ 5 ns) the voltages appearing at the gate oxides of the functional circuit are approximately equal for both samples. This voltage is too low to cause gate oxide breakdown for this time regime. The temperature increase inside the protection transistor is approximately 550 K for both samples, which should lead to second breakdown.

The voltages across the gate oxide is, however, different in case the same CDM voltage is simulated for the two samples. The maximum voltage across the gate for both samples is equal when different CDM stress (CDM; rise-time ≈ 0.4 ns) voltages are applied. This simulations show that the peak voltage across the gate oxide is approximately equal when the fail CDM stress voltages are simulated. The simulated voltage across the gate oxide is found to be 45 V, for these fast transients. This voltage drop across the gate oxide could result in gate oxide breakdown. Such a high breakdown voltage was not yet reported in the literature but it may be extrapolated from the experiments of Fong et al. [15]. The temperature increases are much lower

Table 1
Human body model and charged device model fail voltages for samples 1 and 2

Fail voltages	HBM (V)	CDM (V)
Sample #1	5000	450
Sample #2	5000	1200

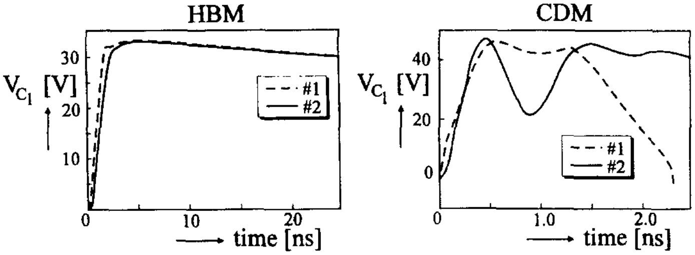

Fig. 9. Simulation results for 5 kV HBM test and for 450, 1200 V CDM test. In case of HBM stress, the voltage appearing at the gate oxide did not exceed 32 V. In case of CDM stress, the peak voltages appearing at the gate was approximately equal for 450 and 1200 V CDM stress-voltage (fail voltages for the tests) on, respectively, samples #1 and #2.

than in HBM. The temperature increase in the protection transistor is 30 and 180 K for samples #1 and #2, respectively.

Since the transistor is switching fast enough, we find that the limiting factor for ESD protection of fast transients seems not to be the turn-on speed of the lateral bipolar transistor of the NMOS transistor, but is very likely due to the combined RC-products of the protection and functional circuit. The layout of the protection-/functional circuit combination then determines the ESD performance. Much care should be taken with the use of metal lines and placement of resistors. Also the resistance of the drain diffusions (the spreading resistance) in protection transistors can give rise to a significant voltage drop, which can lead to the destruction of gate oxides in the functional circuit.

6. Conclusions

Fast transient measurements were carried out on ESD protection circuits using the electro-optic sampling technique. The electro-optic sampling technique allows the determination of high amplitude voltage measurements with high accuracy. With this technique we could determine the internal node voltages inside a protection circuit during fast transient signals.

Simulations were performed to verify the measurements. This, however, requires accurate knowledge of model parameters, which are hard to extract and strongly non-linear. Therefore, we used 2D-device simulations, both DC and transient, to get improved model parameters. With the improved parameters we could predict the fast transient triggering and turn-on of the lateral bipolar transistor. Furthermore, with our model we have been able to predict the observed differences between the outcome of human body model and charged device model testing.

For very fast transients the triggering of the protection transistor shifts from an avalanche multiplication current towards a displacement current-induced triggering. In this way the device is triggered very fast into snapback. But before the actual turn-on time is reached already large currents will be supported, which could be enough for some CDM cases. For our protection circuit the limiting factor in CDM protection is therefore not the turn-on time of the protection transistor, but the layout of resistors and parasitics.

Acknowledgements

These investigations in the program of the Foundation for Fundamental Research on Matter (FOM) have been supported (in part) by the Netherlands Technology Foundation (STW). We are grateful to Joris Vrehen for doing the measurements with the electro-optic sampling set-up. It is also a pleasure to thank Fred Kuper for many fruitful and helpful discussions on the subject of ESD protections. Without his stimulation and kindly supplied samples this project would not have been possible.

Appendix A

The transistor model used in the circuit simulations is shown in Fig. 10. The model is basically an Ebers–Moll model with a number of extensions describing mechanisms

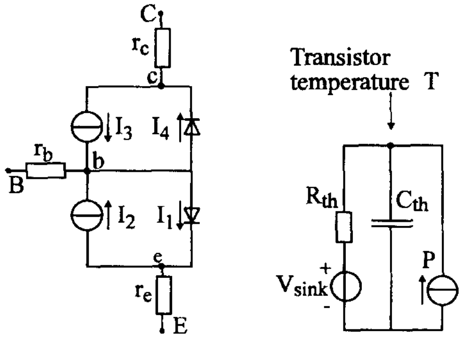

Fig. 10. Basic transistor model including the thermal behaviour.

playing an important role in breakdown. For evaluating the main mechanisms involved in breakdown of a parasitic bipolar transistors, a coupled electrical–thermal transistor model is used. The electrical model is based on the one used by Latif and Bryant [16], which includes avalanche and high injection.

Avalanche is often considered to be the initiating mechanism for snapback, which has been included in the lateral transistor model through the Miller formula. High injection in the base is included, modelling the transition in the voltage–current relationship. An empirical relation describing the current dependency of the current gain is also included. The collector resistance, R_c, is an equivalent spreading resistance of the drain. The base resistance is formed by the substrate resistance. Additionally, the depletion capacitors of the base–emitter, base–collector junctions as well as the diffusion capacitance are added to the model to simulate large signal transients. The equations for the currents in a bipolar transistor including the influence of avalanche multiplication are the following:

$$\begin{aligned} I_E = &- I_{ES}(T)\left[\exp\left(\frac{-qV_{eb}}{kT}\right) - 1\right] \\ &+ \alpha_R I_{CS}(T)\left[\exp\left(\frac{-qV_{cb}}{kT}\right) - 1\right] = -I_1 + I_2, \end{aligned} \tag{A.1}$$

$$\begin{aligned} I_C = &M\alpha_F I_{ES}(T)\left[\exp\left(\frac{-qV_{eb}}{kT}\right) - 1\right] \\ &- MI_{CS}(T)\left[\exp\left(\frac{-qV_{cb}}{kT}\right) - 1\right] = I_3 - I_4, \end{aligned} \tag{A.2}$$

$I_{ES}(T)$ and $I_{CS}(T)$ are defined as

$$I_S(T) = \alpha_F I_{ES}(T) = \alpha_R I_{CS}(T) = I_S(T_0)\left(\frac{T}{T_0}\right)^3 \frac{\exp\left[-\frac{E_G}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right]}{\left(1 + \theta\exp\left[\frac{-qV_{eb}}{2kT}\right]\right)}. \tag{A.3}$$

The factor M represents the avalanche multiplication factor given by the Miller formula:

$$M = \frac{1}{1 - (V_{cb}/V_{cbO})^n} \tag{A.4}$$

where n is a parameter with a value in the range of about 1–10 with a typical value of 3. The base current I_B is equal to $I_1 + I_4 - I_3 - I_2$:

$$\begin{aligned} I_B = I_S(T)\Bigg\{&\frac{1 - M\alpha_F}{\alpha_F}\left[\exp\left(\frac{-qV_{eb}}{kT}\right) - 1\right] \\ &+ \frac{M - \alpha_R}{\alpha_R}\left[\exp\left(\frac{-qV_{cb}}{kT}\right) - 1\right]\Bigg\}. \end{aligned} \tag{A.5}$$

In all these formulas, the voltages V_{eb} and V_{cb} refer to internal junction voltages. For high V_{cb} this reduces to

$$I_{\mathrm{B}} = I_{\mathrm{S}}(T)\frac{1 - M\alpha_{\mathrm{F}}}{\alpha_{\mathrm{F}}}\left[\exp\left(\frac{-qV_{\mathrm{eb}}}{kT}\right) - 1\right]. \tag{A.6}$$

The current gain β_{F} depends on the current with the fit parameters $C_1 - C_3$;

$$\beta_{\mathrm{F}}^{-1} = C_1 + C_2 I_{\mathrm{C}}^{-0.5} + C_3 I_{\mathrm{C}} = \frac{\alpha_{\mathrm{F}}}{1 - \alpha_{\mathrm{F}}}. \tag{A.7}$$

The power dissipation, P, in the transistor is modeled with a current source in a thermal circuit. The temperature T in the transistor (represented by the voltage across the heat current source) is related to the power dissipation in the transistor by means of the thermal resistance R_{th} and the thermal capacitance C_{th}:

$$T = T_0 + PR_{\mathrm{th}} - R_{\mathrm{th}}C_{\mathrm{th}}\frac{\mathrm{d}T}{\mathrm{d}t}. \tag{A.8}$$

References

[1] T. Maloney, Designing MOS inputs and outputs to avoid oxide failure in the charged device mod· EOS/ESD Symp. Proc., 1988, pp. 220–227.

[2] Y. Fong and C. Hu, Internal ESD transients in input protection circuits, Proc. of IRPS, 1989, pp. 77–81.

[3] D. Krakauer and K. Mistry, ESD protection in a 3.3 V sub-micron salicided technology, EOS/ESD Symp. Proc., 1992, pp. 250–257.

[4] C.G.C.M. de Kort, J.R.M. Luchies and J. Vrehen, The transient behaviour of an ESD input protection, Proc. EOBT Symp., 1993, pp. 7.15–7.18.

[5] R. Kropf, C. Russ, R. Kolbinger, H. Gieser and S. Irl, Zeitaufgelöste Untersuchungen des Snapback-verhaltens eines ESD-Schutzstransistors, Tagungsband 3, ESD Forum, Grainau (Germany), VP Verlags GmbH Herrenberg, 1993, pp. 19–26.

[6] S.L. Miller, Phys. Rev., 105 (1957) 1246.

[7] M. Reisch, On bistable behavior and open-base breakdown of bipolar transistors in the avalanche regime-modeling and applications, IEEE Trans. Elec. Dev., 39 (1992) 1398.

[8] J. Abderhalden, Untersuchungen zur Optimierungen von Schutzstrukturen gegen elektrostatische Entladungen in integrierten CMOS-Schaltungen, Ph.D. Thesis, ETH Zürich.

[9] T. Polgreen and A. Chatterjee, Improving the ESD failure threshold of salicided NMOS output transistors by ensuring uniform current flow, EOS/ESD Symp. Proc., 1989, pp. 167–174.

[10] C. Russ et al., Electro-thermal circuit simulation, one task summary of the ESPRIT-project ESD-protection for sub-micron technologies, 1993.

[11] J.R.M. Luchies, C.G.C.M. de Kort and J.F. Verweij, Bipolar transient turn-on of an ESD protection circuit, Proc. ProRISC IEEE Workshop, 1994, pp. 151–155.

[12] J.A. Valdemaris, G. Mourou and C.W. Gabel, Appl. Phys. Lett. 41(3) (1992) 211.

[13] Pstar, Philips Electronic Design & Tools (ED&T), 1993.

[14] G. Krieger, The dynamics of electrostatic discharge prior to bipolar action related snapback, EOS/ESD, Symp. Proc., 1989, pp. 136–144.

[15] Y. Fong and C. Hu, The effect of high electric field transients on thin gate oxide MOSFETs, EOS/ESD Symp. Proc., 1987, pp. 252–257.

[16] M. Latif and P.R. Bryant, Multiple equilibrium points and their significance in the secondary breakdown of bipolar transistors, IEEE J. Sol. St. Circ., 16 (1981) 8–15.