Investigation of Thermal Breakdown Mechanism in 0.13 μm technology ggNMOS under ESD Conditions

Leonardo M. Hillkirk[1,2,3], Jung-Hoon Chun[2] and Robert W. Dutton[2]

[1] KTH, Royal Institute of Technology, Department of Microelectronics and Information Technology,
Electrum 229, SE-164 40 Kista, Sweden

[2] Center for Integrated Systems, Stanford University, Stanford, California 94305-4075, USA

[3] Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL
United Kingdom

Phone: +44-(0)161-275 4119; Fax: +44-(0)161-275 4056; e-mail: leonardo@imit.kth.se

Abstract—**2D Transient device simulations reproducing conditions similar to Electro-Static Discharge (ESD) conditions have been performed for the entire safe operating area (SOA) of a 0.13 μm technology, ground-gated N-channel Metal-Oxide-Semiconductor (ggNMOS) transistor up to 2nd breakdown, using a set of macroscopic physical models related to previous studies [1] implemented in MEDICI. The simulations results indicate the potential influence of a source-end mechanism of destruction, in addition to the previously reported drain-end avalanche generation of electron-hole pairs and subsequent thermal runaway in the proximity of the carriers generation spot as a result of the large carrier density [1, 2, 3]. Under dynamic conditions and with non-zero contact resistance, thermal runway is also observed on the source-side of the device indicating that, for values of the contact resistance on the order of 5.4e^{-6} Ohms-cm^2, substantial damage can occur at the source end. The simulation results are in qualitative agreement with experimental results where it is observed that, after electrical and subsequent thermal runaway, damage is localized not only at the drain region but also at the source region of the device. Thus, the ESD related destruction of a 0.13 μm gate ggNMOS may not be the result of a single destruction mechanism, but the consequence of coupled events, depending on the design characteristics of a particular device.**

Keywords: ESD, ggNMOS, SOA, device simulation, heat generation and transport, device reliability.

I. Introduction

Electro-Static Discharge (ESD) is the process of electric charge transfer between to bodies (i.e. a between a "charged" person and an Integrated Circuit IC) having different electrostatic potential. Depending on the difference in the amount of accumulated electric charge in each of the bodies (which determine the electric potential that can easily amount to several thousand Volts), there is a possibility that a large electric current will circulate when the bodies come in contact or are close enough to each other for dielectric (i.e. air) breakdown to occur. ESD is a very fast physical process that typically occurs in a ns time scale.

The large current densities circulating in the small ULSI devices as a result of ESD events result in very high local temperatures that can easily lead to the destruction of the device. For that reason, ESD is a major cause of concern in the IC industry, as a high percentage of all IC failures are ESD related. Consequently, there is a need for computer simulation tools that can predict how failure occurs and give a guidance on which design parameters can be optimized in order to minimize ESD related failures.

II. Computer Simulations

2D Transient device simulations reproducing ESD conditions have been performed for the entire SOA of a 0.13 μm technology ggNMOS transistor up to 2nd breakdown, using a set of macroscopic physical models related to previous studies [1] implemented in MEDICI. For this purpose, a half sine wave 1 Amp (equivalent to 10 mA per μm of device width), 100 ns long current pulse was applied to the drain, while all the other contacts (source, gate and substrate) where grounded. Further details of the ggNMOS transistor structure used in the simulations presented here can be found in [1].

Heat generation and transport was accounted for by solving the classical lattice heat equation as implemented in MEDICI

$$\rho c_v(T)\frac{\partial T}{\partial t} + \vec{\nabla}\cdot\left(-\kappa(T)\vec{\nabla}T\right) = H$$

where ρ is the density of the material, c_v is the lattice heat capacity at constant volume (which is an increasing function of temperature), T is the temperature in °K, κ is the lattice thermal conductivity (which is a decreasing function of T in the temperature range between room temperature and a few hundred °C), and H is the heat generation term. The nature of the heat generation term in semiconductor devices has been described among others by Lindefelt [4] and Wachutka [5]. Both these theories describe the nature of heat generation based on fundamental physical principles. The theory presented by [4] was derived under rather general conditions, allowing it to be applied to a large range of cases. In this theory, the total heat generation in semiconductors is the result of seven generation terms. However, ref. [6] has shown that, even under extreme but realistic conditions (sub-μs time scale) and in bipolar devices, the classical Joule and carriers recombination heat terms are largely dominant, and found no need to include the remaining five terms to describe heat generation. These two most relevant terms (Joule and carriers recombination heat) are the ones used in this investigation to calculate the temperature distribution inside the device. By doing so, it is implied that the local electron and hole temperatures are assumed to be equal to the lattice temperature at any point in time (i.e. the carriers are assumed

0-7803-7826-1/03/$17.00 © 2003 IEEE

to be always in thermal equilibrium with the lattice) throughout the entire device. By analytical calculations, ref. [3] has shown that, under ESD operating conditions and in ULSI MOS transistors, neglecting the carriers heat capacity and thermal conductivity and thus assuming overall thermal equilibrium results in an approximately 6% overestimation of the calculated lattice peak temperature due to an underestimation of the rate of heat removal away from the heat source. This error was found to be acceptable in view of other uncertainties involved in some of the other models used in the simulations. Furthermore, there is experimental evidence suggesting that at very large current densities and in the µs to ms time scale, Joule heat generation largely dominates over carrier-recombination heat generation, even in the case of wide bandgap materials such as SiC [7, 8, 9].

The values of the specific contact resistances at the metal-semiconductor junctions have been set constant to $5.4e^{-6}$ Ohms-cm^2 for both the source and drain contacts. This is not strictly correct as the metal-semiconductor contact resistances are predicted to be a monotonously decreasing function of temperature [1, 10]. While a new temperature dependent contact resistance model has recently been demonstrated [10], it is not currently available for simulation studies. Hence, the constant contact resistance leads to an overestimation of the generated Joule heat at the metal-semiconductor junctions. Nevertheless, the results presented in this paper show a good correlation with experimental results obtained under ESD conditions, where the presence of thermal damage at the source-end of the device in addition to the damage seen at the drain-end appears to be the consequence of Joule heating at the source metal-semiconductor contact (compare figs. 2, 6 and 7 below).

III. Results and Discussion

Fig. 1 shows the dynamic IV characteristics up to the point of 2nd breakdown resulting from the application to the drain of a half sine wave 1 Amp (equivalent to 10 mA/µm for the 100 µm wide device presented here) 100 ns long current pulse. Initially, the positive voltage applied to the heavily doped drain results in the extraction of electrons increasing in this way the volume of the fixed positive charge region on the drain side of the drain-substrate junction. At the same time, holes from the substrate drift toward the bulk contact, increasing the volume of the fixed negative charge region on the substrate side of the drain-substrate junction. This results in a reverse biased drain-substrate junction that sustains most of the applied voltage and blocks any further diffusion current transport, and in a forward biased junction at the source-substrate junction. At this point (from zero bias up to first break down), the small current flow is due to the thermally generated leakage current originated at the drain-substrate reverse biased junction. If the path of the holes across the bulk is highly resistive (something that happens soon as the current density increases if the bulk doping concentration is low), the preferred path of the holes will be thorough the forward biased source-substrate junction, resulting in the injection of most of the holes into the source, while simultaneously electrons from the source are injected into the substrate. This event marks the turn-on of the parasitic bipolar transistor. When this occurs, the drain of the original NMOS becomes the collector of the bipolar, while the source becomes the emitter and the substrate and bulk the base. This transition from NMOS to bipolar operating mode occurs soon after first breakdown. At the point of first breakdown, the critical voltage for the generation of electron-hole pairs by impact ionization is reached, and the drain-substrate junction can no longer sustain any voltage. This results in an abrupt fall of the drain-source voltage drop. At the same time, due to the large generation of carriers as a result of impact ionization, the current begins to increase sharply. As the current density increases, the drain-source voltage drop continues to increase, but it not longer localized at the drain-substrate junction, but closer to the impact ionization generation point, and across the substrate. In the case of silicided contact devices, as the one presented in this study, a comparatively large contact resistance also exists at the metal-semiconductor interface at the source and drain contacts. In these type of devices, a large voltage drop will also occur at the metal-semiconductor junctions, leading to a substantial heating of these regions (see figs. 2, 6 and 7 below).

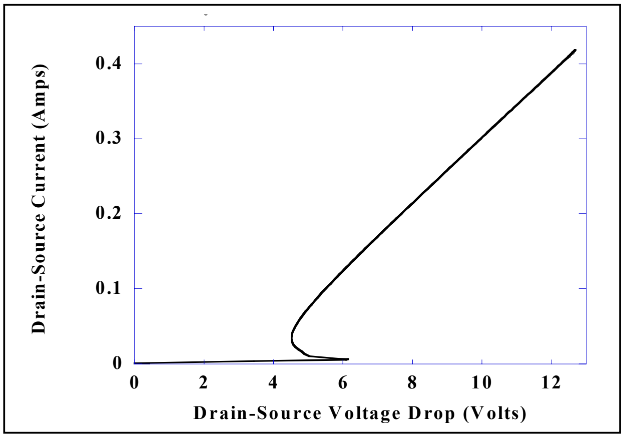

Figure 1. Dynamic IV characteristics up to the point of 2nd breakdown resulting from the application to the drain of a 10 mA/µm half sine wave current pulse. All the other contacts were grounded. 1st breakdown and the subsequent snapback can be seen at the bottom of the figure.

Fig. 2 shows the time evolution of the relative temperature distribution inside the ggNMOS after the onset of 1st breakdown. The figure displays the sequence of temperature distributions 0.1, 0.5, 1, 2, 4 and 13.3 ns after the beginning of the current pulse. In the beginning of the series, the highest temperature ("hot spot" in light colors) is located around the impact ionization carrier generation point situated under the gate at the edge of the drain-substrate junction. This is so because of the high current density in the impact ionization region (the generation process itself is endothermic, but due to the increased carrier concentration, a large amount of generated Joule heat causes an overall exothermic spot). As time goes, a second "hot spot" under the metal contact of the source region starts to develop. This hot spot is a consequence of Joule heating due to the voltage drop at the metal-semiconductor interface resulting from the contact resistance between the metal contact and the semiconductor material, and does not appear if the contact resistances in the simulation are set to zero (ideal Ohmic contacts), in concordance with previous related studies [1, 2, 3]. The lack of symmetry between the heat generation under the drain and source contacts follows the asymmetry of the current density distribution (fig. 3), and the position of the impact ionization and recombination regions (figs. 4 and 5). The heat absorbed by the generation process at the impact ionization carrier generation point might be responsible for slowing down the heating of the metal-semiconductor contact

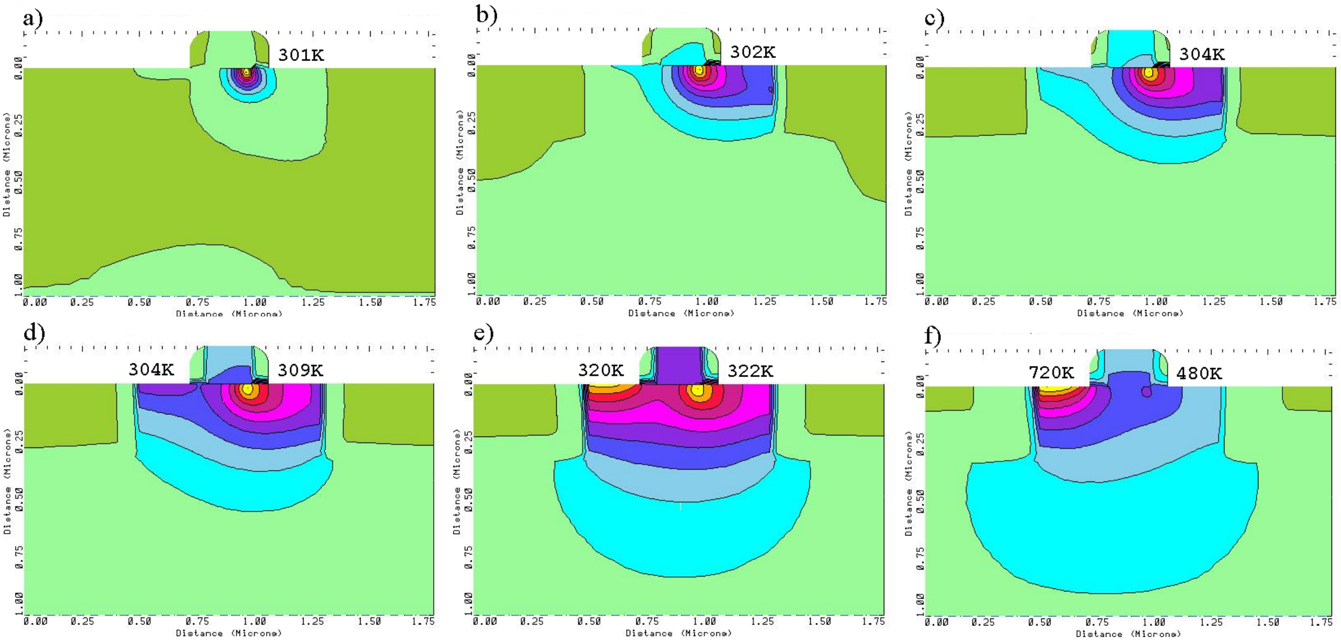

Figure 2. a) to f) Time evolution of the relative temperature distribution inside the 0.13 µm ggNMOS during the application of a 1 Amp half sine wave 100 ns long current pulse, after the onset of 1st breakdown. Time evolves from left to right, top to bottom. The figures display the relative temperature distribution 0.1, 0.5, 1, 2, 4 and 13.3 ns after the beginning of the current pulse. The temperature of the device at the beginning of the current pulse is 300K. The peak temperature of the "hot spots" is indicated inside each frame. The temperature at the source-side hot spot just before 2nd breakdown is approx. 720K. For more details, see the main body of the text above.

at the drain, while the heat emitted by the recombining carriers at the carriers recombination point located at the source-substrate junction under the gate oxide (fig. 5) generates additional heat that might help to accelerate the heating of the metal-semiconductor contact at the source. Just before 2nd breakdown, a large number of intrinsic carriers are thermally generated at the source at the metal-semiconductor interface. The largest concentration of intrinsic carriers is found in the region where the temperature is the highest and indicates where thermal runaway will eventually occur (fig. 6). The presence of hot spots on both sides of the gate correlates with experimental results where device failure is observed on both the drain and source regions (fig. 7).

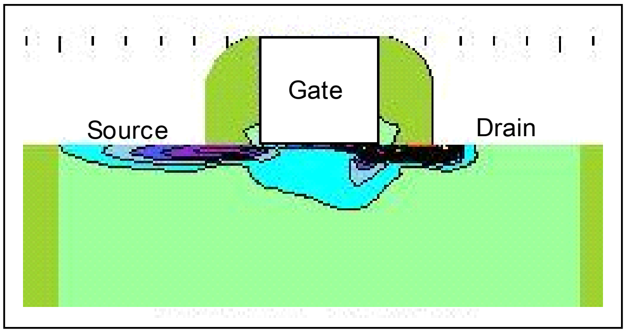

Figure 3. Current density distribution inside the 0.13 µm ggNMOS 3 ns after the beginning of the current pulse.

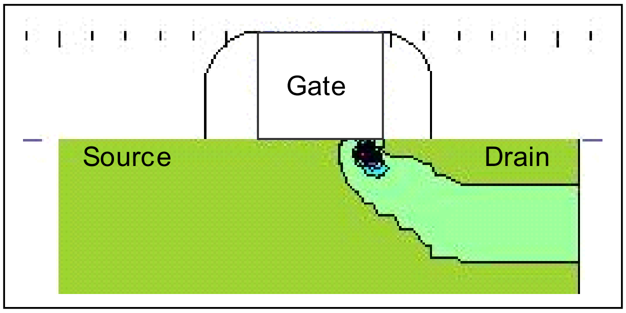

Figure 4. Total generation rate in the 0.13 µm ggNMOS 1 ns after the beginning of the current pulse. The position of the impact ionization spot at the drain-substrate PN junction remains unchanged from 1st and up to 2nd breakdown.

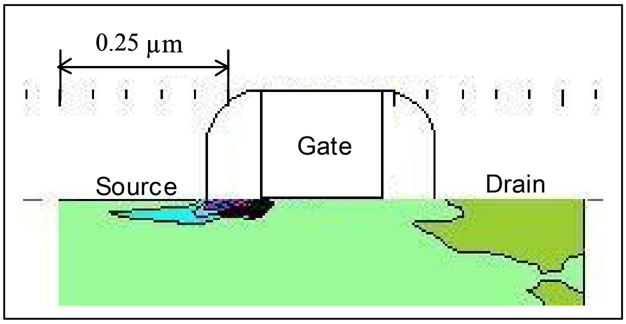

Figure 5. Total recombination rate in the 0.13 µm ggNMOS 7 ns after the beginning of the current pulse. As the device approaches 2nd breakdown, the volume where recombination occurs shrinks towards the edge of the source-substrate PN junction right under the gate.

Only a qualitative agreement with experimental results is claimed here. Several theoretical issues must be reviewed in order to achieve a more physically stringent description of ESD phenomena in sub-micron semiconductor devices.

These issues include:

1) As the characteristic length of ULSI MOSFET devices continues to shrink, device size approaches the scale of quantum transport processes. The use of the macroscopic heat equation to simulate these small structures may result in an underestimation of the amount of generated heat [11]

2) The metal-semiconductor contact resistances are predicted to be a monotonously decreasing function of temperature [1, 10]. This phenomenon should be considered in the simulation studies.

3) Calibration of the surface and bulk mobility models at very high current densities and temperatures should be accomplished.

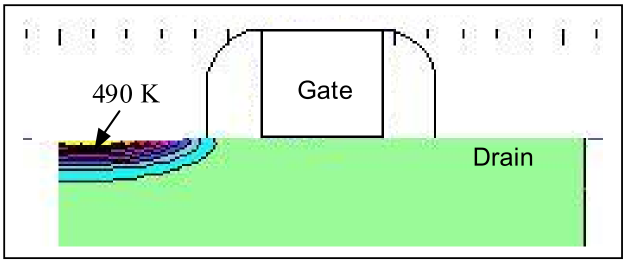

Figure 6. Intrinsic carrier concentration in the 0.13 µm ggNMOS 10 ns after the beginning of the current pulse, approaching the point of 2nd breakdown. Observe the increasing concentration of thermally generated carries in the source region at the metal-semiconductor interface. The temperature of the "hot spot " is indicated.

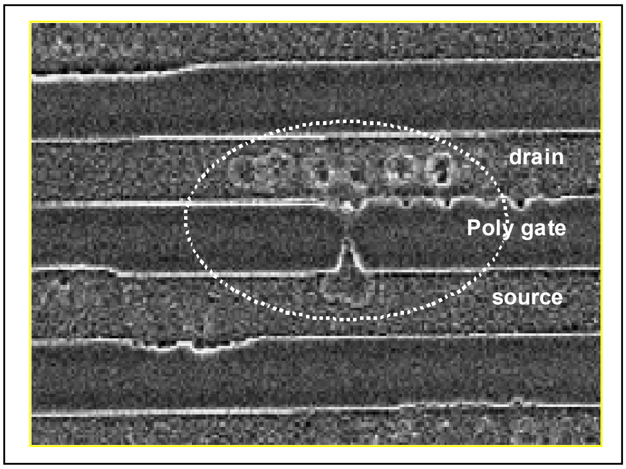

Figure 7. Photograph showing the results of ESD failure on a ggNMOS on both the drain (several rounded damage spots) and source (displaying a drop shaped molten area) regions [10]. This sort of ESD damage has also been reported by other groups.

IV. SUMMARY OF WORK AND CONCLUSIONS

As the size of CMOS transistors continues to scale down in the quest for ever faster ICs, the relative contribution of the contact resistances to the device voltage drop increases, since contact resistance is an intrinsic material property that does not scale down as device size decreases.

In this work, 2D Transient device simulations reproducing conditions similar to ESD conditions have been performed for the entire SOA of a 0.13 µm technology ggNMOS transistor up to 2nd breakdown using MEDICI. The simulation results, which are based on macroscopic physical models related to previous studies [1] show that, under dynamic conditions and for the representative metal-semiconductor contact resistance value of 5.4e^{-6} Ohms-cm^2 set at both the source and drain contacts, a source-end mechanism of destruction at the metal-semiconductor interface contributing to the failure of the 0.13 µm gate ggNMOS exits, in addition to the previously reported drain-end destruction mechanism due to avalanche generation of electron-hole pairs and subsequent thermal runaway in the proximity of the carriers generation spot as a result of the large carrier density [1, 2, 3]. The simulations show that under the described conditions thermal runway occurs firstly at the source-end of the device, indicating that substantial damage may also take place in that region. The simulation results are in qualitative agreement with experimental results where it is observed that, after electrical and subsequent thermal runaway, damage is localized not only at the drain region but also at the source region of the device. Thus, the ESD related destruction of a 0.13 µm gate ggNMOS may not be the result of a single destruction mechanism, but the consequence of coupled events, depending on the design characteristics of a particular device.

ACKNOWLEDGMENT

L. M. Hillkirk would like to thank the John and Karin Engbloms Stipendiefond for providing the funding in support of this collaborate project at Stanford University.

REFERENCES

[1] K. H. Oh, "Investigation of ESD performance in advanced CMOS technology", PhD Thesis, CIS, Center for Integrated Systems, Stanford University, Stanford, California, USA, 2002.

[2] K. Esmark, "Device Simulation of ESD Protection Elements", PhD Thesis, ETH, Swiss Federal Institute of Technology, Zurich, Switzerland, 2002.

[3] S. Beebe, "Characterization, modeling and design of ESD protection circuits", PhD Thesis, CIS, Center for Integrated Systems, Stanford University, Stanford, California, USA, 1998.

[4] U. Lindefelt, "Heat generation in semiconductor devices", *J. Appl. Phys.*, Vol. 75, No. 2, pp. 942-957 (1994).

[5] G. Wachutka, "Rigorous Thermodynamic Treatment of Heat Generation and Conduction in Semiconductor Device Modeling", *IEEE Trans. Computer-Aided Design*, Vol. 9, No. 11, pp. 1141-1149 (1990).

[6] O. Tornblad, U. Lindefelt and B. Breitholtz, "Heat Generation in Si Bipolar Power Devices: The Relative Importance of Various Contributions", *Solid State Electronics*, Vol. 39, No. 10, pp. 1463-1471 (1996).

[7] L. M. Hillkirk, "Infrared emission properties and dynamic measurements of device surface temperature in epitaxial *SiC* P*i*N power diodes, and comparison to *Si*", *Proc. of the International Conference on SiC and Related Materials 2003* (ICSCRM 2003).

[8] L. M. Hillkirk, B. Breitholtz and M. Domeij, "Space and time resolved surface temperature distributions in Si power diodes operating under self-heating conditions", *Solid State Electronics*, Vol. 45, No. 12, pp. 2057-2067 (2001).

[9] L. M. Hillkirk, "Dynamic surface temperature measurements in SiC epitaxial power diodes performed under single-pulse self-heating conditions", submitted to *Solid State Electronics* (2003).

[10] K. H. Oh, J. H. Chun, K. Banerjee, C. Duvvury and R. W. Dutton, "Modeling of Temperature Dependent Contact Resistance for Analysis of ESD Reliability", *Proc. of the International Reliability Physics Symposium*, pp. 249-255, (2003).

[11] E. Pop, K. Banerjee, P. Sverdrup, R. W. Dutton and K. E. Goodson, "Localized Heating Effects and Scaling of Sub-0.18 Micron CMOS Devices", in *Tech. Dig.. IEDM*, pp. 677-680, (2001).

0-7803-7826-1/03/$17.00 © 2003 IEEE