Simulation of the Gate Current Sensitivity of a 1.5 nm Gate Oxide nMOST

Thomas Feudel [1] and Andreas Schenk

Technical Report 02/99

Abstract

Scaled MOS devices with gate oxides as thin as 1.5 nm are affected by direct-tunneling gate leakage. We demonstrate the sensitivity of the direct-tunneling gate current to variationsof both process parameters and parameters in the model of direct tunneling [3].The zero of the gate current is a probe for the form of the oxide near the drain-side corner of the poly gate. We argue that it can also be used as a monitor of the wear-out process caused by trapping of positive charge in this region.

1. Integrated Systems Engineering AG Zurich

1.0 Introduction

With the continuous downscaling of MOS devices the usage of 1.5 nm equivalent gate oxide layers is expected for devices with a gate length of 100 nm for the year 2006 [1]. The feasibility to manufacture thermal oxides of 1.5 nm thickness has recently been demonstrated by Momose et al. [2]. Unfortunately, such devices are dominated by direct-tunneling gate leakage currents. This study gives an outline of the direct tunneling model, the main model impact parameters and investigates the sensitivity of the gate current on process parameter variations by means of simulation.

2.0 Physical Model of Direct Tunneling

Direct tunneling is the main gate leakage mechanism for oxides thinner than 3 nm. It turns into Fowler-Nordheim (FN) tunneling at oxide fields of approximately 6 MV/cm, independently of the oxide thickness. In our device simulator Dessis direct tunneling is modeled as interface condition based on the barrier penetration description which assumes: single parabolic bands in all materials, plane waves in gate electrode and silicon, no fixed oxide charges, elastic tunneling (no phonon coupling), and relaxation of momentum conservation (no band structure mismatch). The parameter "oxide mass" (m_{ins}) (or "tunnel mass") accounts for the unknown excitation spectrum of extremely thin amorphous oxides and the fact that tunneling is never monoenergetic in device application. This is because the energetic distance between injection energy and local barrier potential can rapidly change across the oxide, and due to the thermal broadening. The commonly accepted range of values for the tunnel mass m_{ins} is 0.39 m_0 - 0.5 m_0. The Dessis default value is 0.42 m_0 which was confirmed by many FN experiments. The other crucial parameter is the barrier height $E_{barrier}$ of the potential barrier induced by the oxide. Here, the default value is 3.15 eV which does not seem to change essentially even for the thinnest oxides possible, as confirmed by a number of experiments. There are a couple of less crucial parameters, e.g. the Fermi energy of the gate electrode ($E_{F,M}$), the effective masses in gate (m_M) and silicon (m_s) which depend on the injection energy, and the permittivity ε_{ins} of the ultra-thin layer. All of them have a non-exponential impact, i.e. they influence the transmission probability by a factor much less than one order of magnitude.

Gate oxides as thin as 1.5 nm require to avoid two approximations which are commonly applied to thicker oxides: the WKB approximation and the neglection of image forces. Both problems were solved for the direct tunnel model in a flexible way. By introducing a trapezoidal pseudo-barrier, the oxide wave functions are mapped to Airy functions with the only numerical complication of an additional inner iteration at three discrete energies per interface point. The non-locality induced by the image potential and the coupling of the tunnel current to the drift-diffusion current naturally lead to convergence problems, in particular far from equilibrium. A practical way to improve convergency and to speed up the simulation is to model the effect of the image potential by an effective barrier lowering of the trapezoidal barrier. This was done in all simulations of this project ($E_{barrier}$ -> 2.8 eV). The effective barrier lowering often can be

determined by a reference calculation or it can be taken as a fit parameter. The image potential is switched off by giving the parameters E_0, E_1, and E_2 the same value.

A full description of the physics and of some applications of the model can be found in [3].

3.0 Impact of Model Parameters and Grid Refinement

1) *Lateral Refinement at the Gate Corner:* An optimized grid (Fig. 1) is the precondition for a successful device simulation. As shown in Fig. 2, the gate tunnel current exhibits a sharp peak near the drain-side gate corner which must be resolved. On the other hand, too many grid points at the Si-SiO_2 interface will slow down the simulation because much CPU time is used up by the direct tunneling model. To find the trade-off, various levels of refinement were tested. Finally, a level was chosen for the further simulations (named “fine” in Fig. 3), where the gate tunnel current showed a negligible deviation from that obtained when using an extremely fine grid (named “finest” in Fig. 3). This optimal grid is shown in Fig. 1 together with the spatial distribution of the y-component of the electron current density. For this, we have used a vertical grid spacing of 0.05 nm for the inversion layer and a spacing of 1.7 nm $\times$ 1.3 nm for the drain corner.

2) *Impact Ionization:* The gate tunnel current could be influenced by impact ionization at the edge of the drain extension. Because of the local carrier heating the position of the quasi Fermi level will change there. This results in a stronger occupation of final states of the tunneling process (electrons tunnel from gate to Si) and hence, in a suppression of direct tunneling. A “Hydrodynamic” simulation would be necessary to study this effect. Unfortunately, no convergency could be obtained in the hydrodynamic mode of Dessis even when trying to ramp the drain voltage first. The difference in the drift-diffusion mode is caused by the small change of the electron density when impact ionization is switched on. The effect is negligible as shown in Fig. 4. At large V_{gs} such that $V_{gs} > V_{ds}$, the tunneling direction becomes reversed near the drain-side gate corner and carrier heating will have some effect. However, under this bias condition the gate tunnel current monotonously decreases from source to drain!

3) *Fermi Energy of the Electrode:* A rigid derivation of the transmission probability proves its dependence on the transverse momentum. In order to avoid an additional numerical integration, this momentum was approximated by the Fermi momentum of the gate electrode. The respective Fermi energy becomes the parameter $E_{F,M}$ of the model. In aluminum gates it has the large value of 11.7 eV, whereas in poly gates it should be replaced by the mean thermal energy. Fig. 5 demonstrates that the gate current is only little affected even if $E_{F,M}$ is varied by two orders of magnitude.

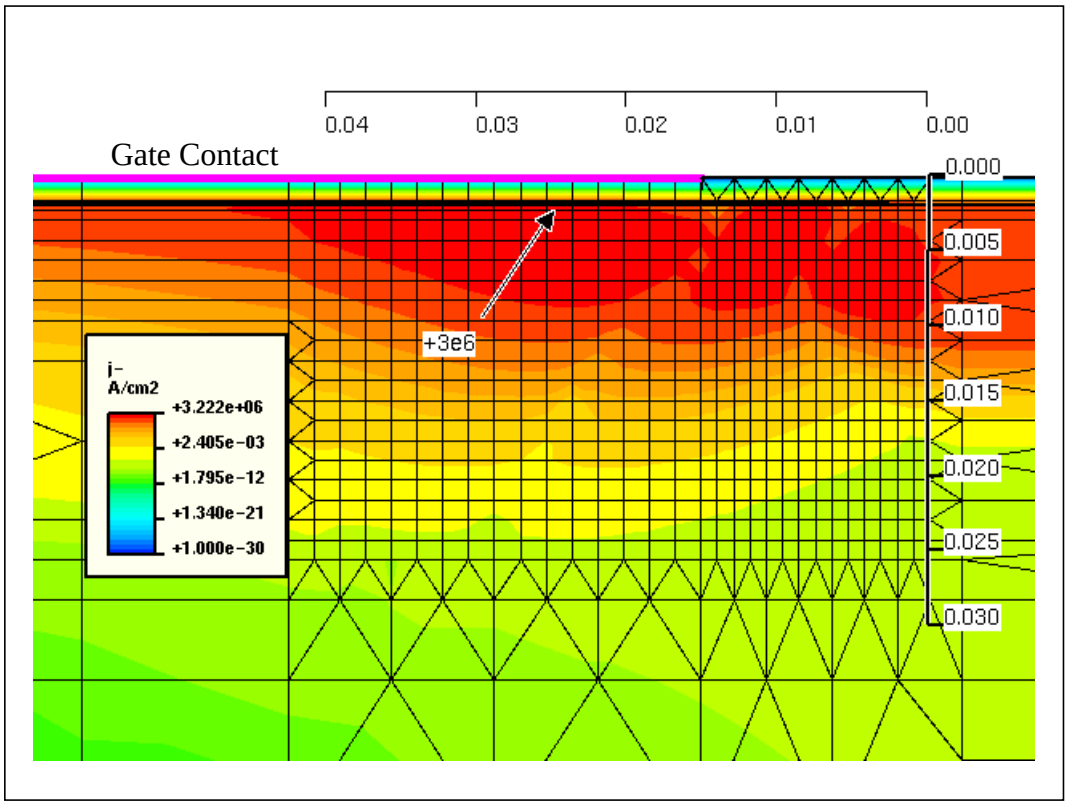

FIGURE 1. Optimal lateral grid refinement at the drain-side gate corner and spatial distribution of the y-component of the electron current density at V_{ds} = 3 V and V_{gs} = 1 V.

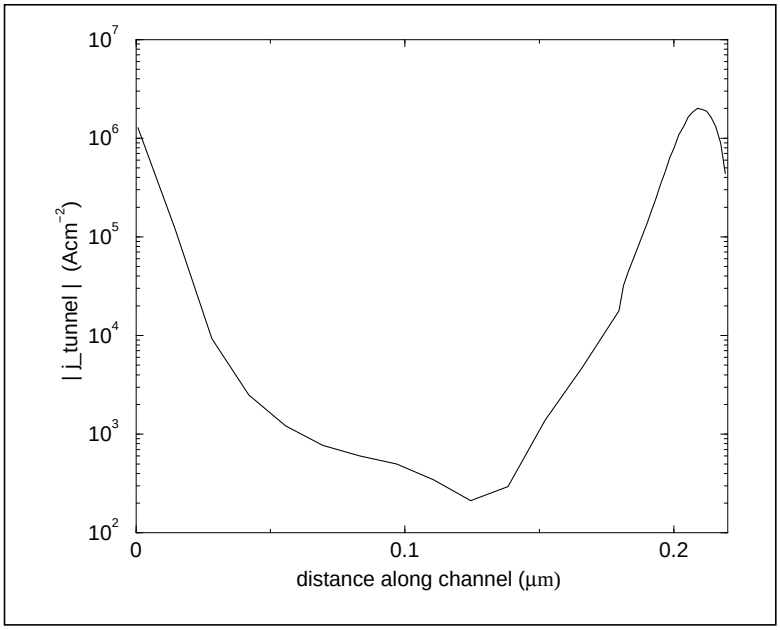

FIGURE 2. Lateral profile of the gate tunnel current along the entire gate oxide at V_{ds} = 3 V and V_{gs} = 1 V. The sign of the current changes at the minimum of the curve.

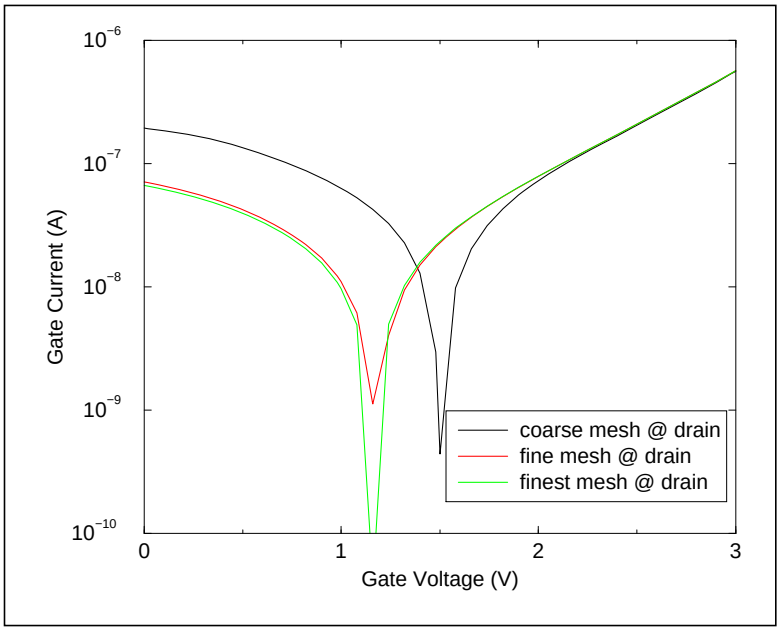

FIGURE 3. Gate tunnel current at V_{ds} = 3 V for different lateral refinements near the drain-side gate corner.

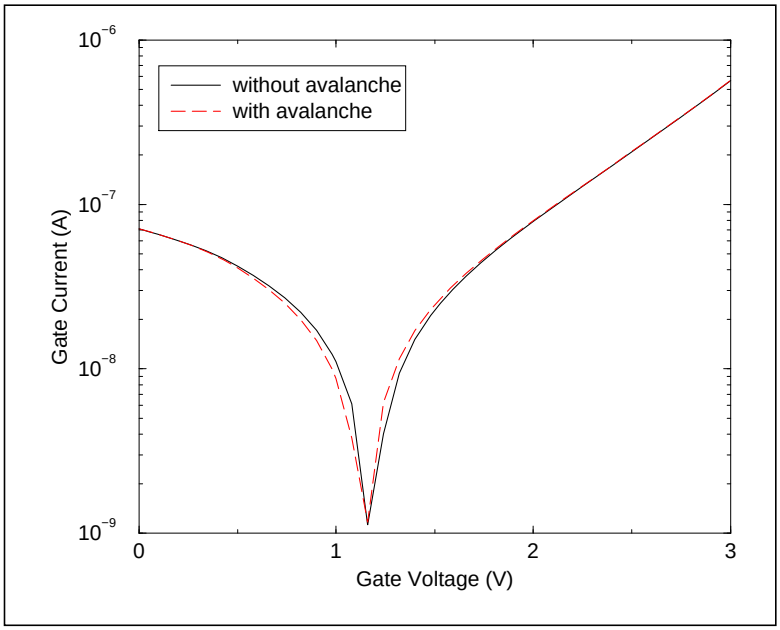

FIGURE 4. Drift-diffusion simulation of the gate tunnel current at V_{ds} = 3 V with and without impact ionization, respectively.

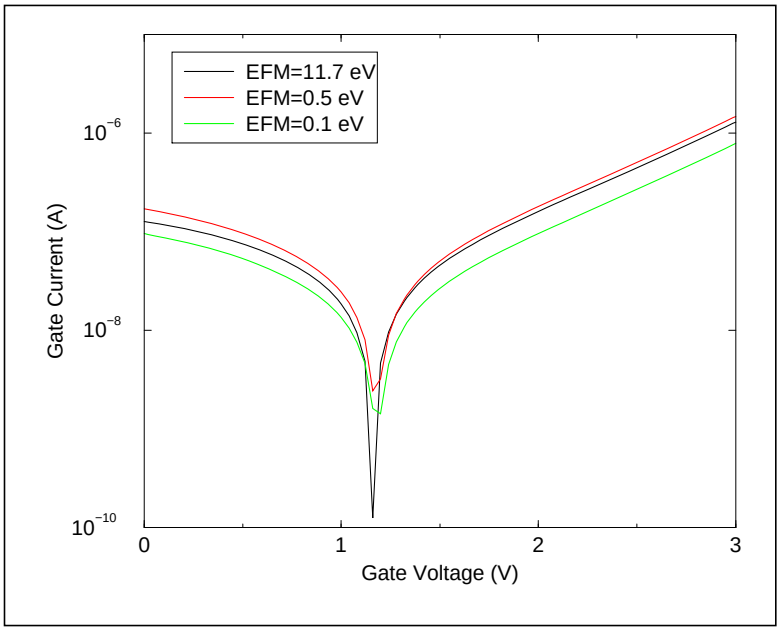

FIGURE 5. Gate tunnel current at V_{ds} = 3 V for different values of the Fermi energy in the gate electrode.

4.0 Process Simulation and Process Variations

The process simulation has been performed with our 1D simulator Tesim (until gate patterning) and the 2D simulator Dios according to the specifications given in [2]. After a boron channel implantation at 50 keV, 5×10^{13} cm^{-2}, 1.5 nm of thermal oxide have been grown in a rapid thermal oxidation (RTO) step at 800 °C (10 sec., dry O_2 ambient). The gate electrode has been formed of in-situ phosphorus-doped polysilicon. To form ultra-shallow junctions, the source-drain (S/D) extensions have been doped by outdiffusion from PSG spacers during the S/D anneal step. The phosphorus concentration in the spacer oxide was 3×10^{21} cm^{-3}. Finally, the S/D regions have been implanted with arsenic at 30 keV, 3×10^{15} cm^{-2}. The back-end processing has been neglected, because we were focussing only on a comparison between different process variants.

For our simulations, we have concentrated on a transistor with 0.22 µm gate length and 5 µm gate width. Since tilted implantations did not influence the device behavior, the process simulation area could be restricted to a half-cell. Fig. 6 shows the net doping profile, the polysilicon gate, the oxide spacer and the silicon region.

1) Gate Oxide Thickness: The thickness has been varied between 1.5 and 4 nm. We assumed a homogeneous thickness for the entire gate region. The thickness measured with TEM on a 150 mm wafer was in the range between 1.3 and 1.7 nm with a 3σ value of 0.39 nm [2]. Our larger values have been used only to demonstrate the application range of the model.

2) Gate Corner: The highest value of the electric field occurs at the drain corner because there is a large voltage drop between the gate (at about 0 V) and the drain (e.g. at 3 V) electrode. While ramping the gate voltage, this voltage drop changes and we have first a strong electron injection into the gate over a short distance, whereas later on the current flows from the gate to the source contact over nearly the full gate length. The shape of the lower gate corner should therefore strongly influence the gate current. It is typically influenced during processing by the polysilicon etching or underetching and polysilicon reoxidation steps. In our simulation we have modified the geometry of the gate corner by linearly increasing the oxide thickness for a certain length L and height H (Fig. 7).

3) Gate Overlap: As an equivalent to the lateral diffusion of the S/D extension profile, we have modified the position of the gate corner with respect to the position of the p-n junction. This has a strong impact on the drain resistance and the distribution of the electric field. The value has been varied by ±5 nm per side.

4) Extension Profile: In addition to the previous impact parameters, the shape of the extension profile not only changes the lateral underdiffusion but also the peak concentration. The tunneling probability is higher for increasing concentrations. In our simulations we kept the oxide phosphorus concentration constant and modified the phosphorus segregation coefficient SEG at the SiO_2/Si interface to control the penetration depth. Fig. 8 shows the phosphorus profiles for different segregation coefficients for a cut line 0.2 µm distant from the center of the channel region (L_{gate}=0.22 µm). A penetration depth of 30 nm [2] can be achieved with a segregation coefficient of about 3.

5) Channel Profile: The channel profile has been modified in order to change the gradient of the p-n junction. Three different profiles with a constant dose of 5×10^{13} cm^{-2} have been simulated (Fig. 9). The 'default' profile is the result of a default process simulation where the shape of the as-implanted boron profile is only changed by a minor degree for a high-temperature treatment at 800 °C for 10 seconds. An almost flat profile can be achieved by using an artificially increased point defect concentration. We have used an interstitial scaling factor of 10 for our modified +1 model for the channel implant step. A boron surface peak, which corresponds to depth of the extension region, is the result of a modified boron segregation coefficient. We have used a value of 1 instead of using the default value of 0.35 for the final anneal step at 1000 °C. Figure 10 shows the lateral net doping profile in a depth of 3 nm below the silicon surface, clearly indicating the channel, the extension and the heavily doped drain region.

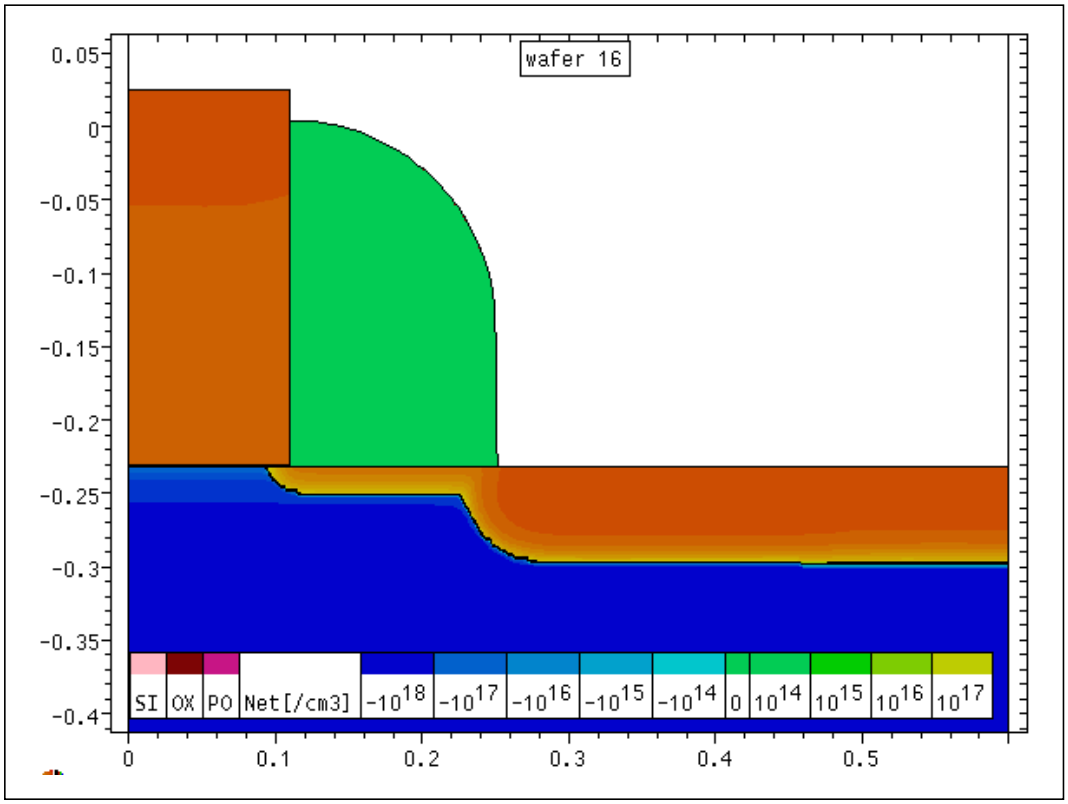

FIGURE 6. 2D cross section of the simulated structure showing the net doping profile.

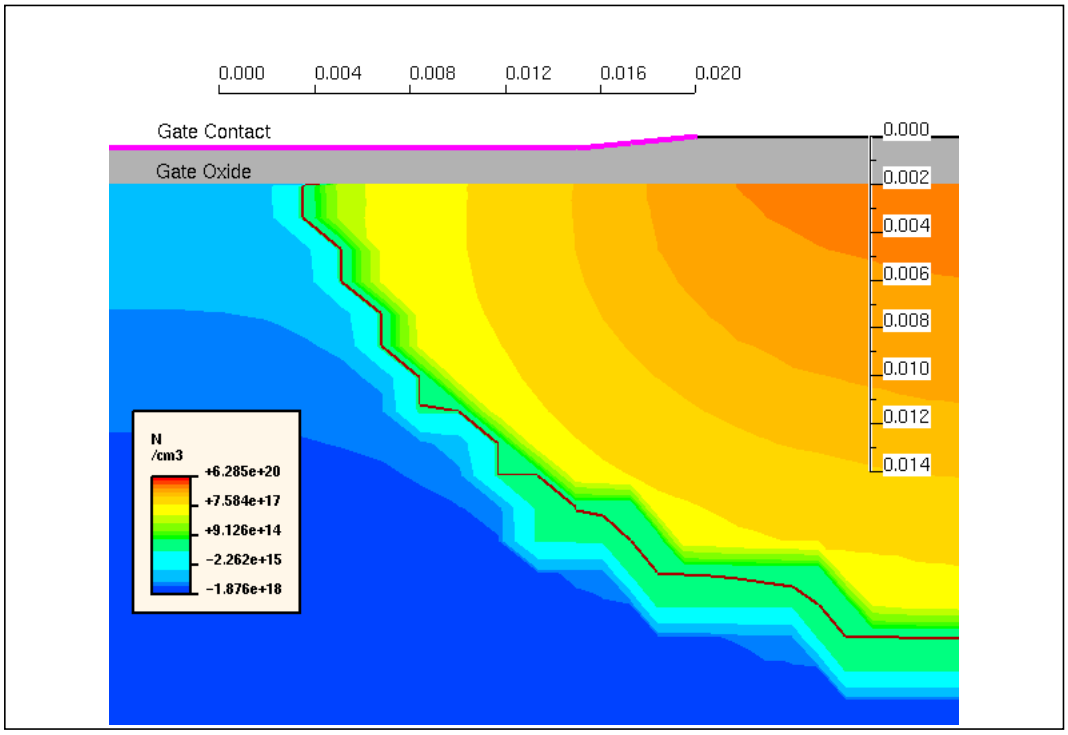

FIGURE 7. Zoom into the 2D structure at the drain corner. The width of the gate oxide is linearly increased over a length of L=5 nm for a height of H=0.5 nm.

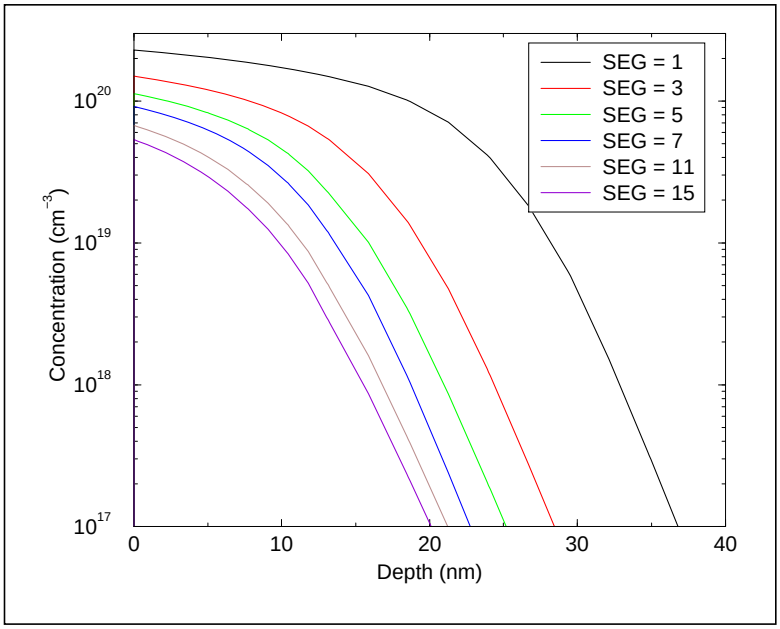

FIGURE 8. Shape of the phosphorus extension profile for a constant oxide concentration of 3×10^{21} cm^{-3} and different segregation coefficients.

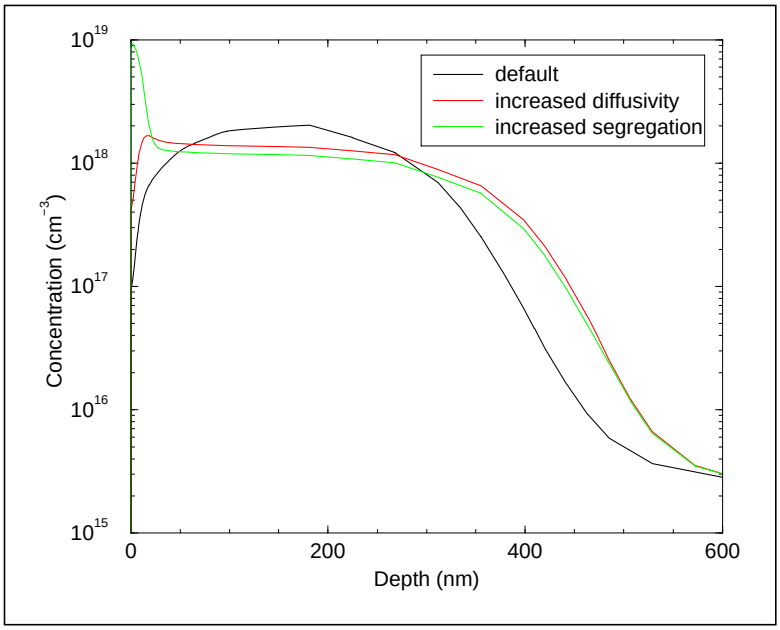

FIGURE 9. Simulation and modification of the boron channel profile.

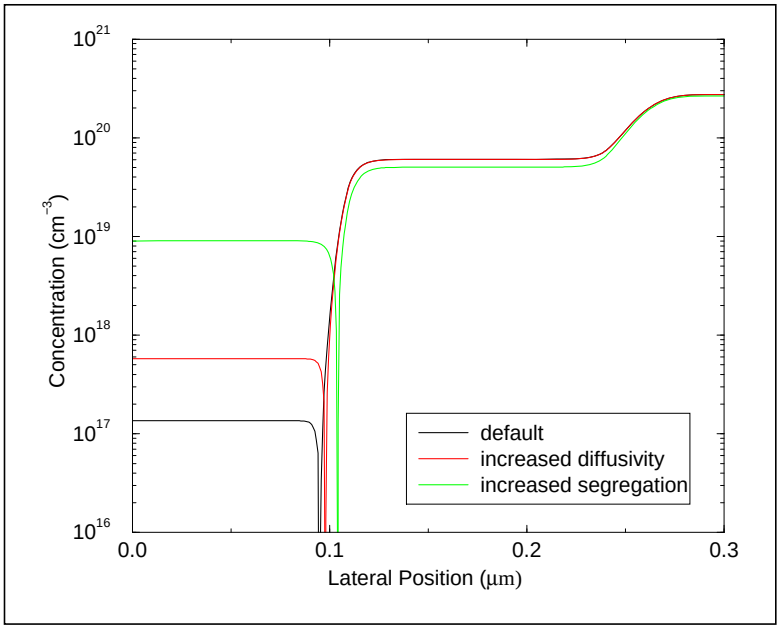

FIGURE 10. Lateral net doping profile in a depth of 3 nm below the silicon surface.

5.0 Results and Discussion

Figures 11 to 18 show the influence of the process parameter as they have been discussed in section 4.0. The drain voltage was 3 V for all simulations.

A variation of the homogeneous gate oxide thickness (Fig. 11) causes a linear shift of the I_g-V_{gs} characteristic. Already a variation by 0.4 nm changes the current by two orders of magnitude. The gate voltage at which the current minimum occurs is not influenced by the gate oxide thickness. For oxide thicknesses higher than 3 nm the effect of direct tunneling becomes negligible.

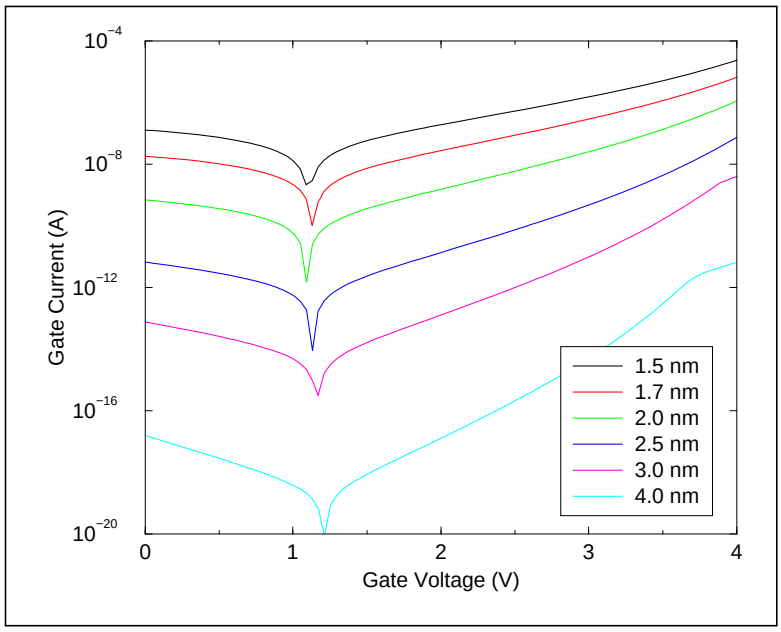

FIGURE 11. Gate current in dependency on the gate voltage for different gate oxide thicknesses.

The zero of the gate current at the gate voltage $V_{gs} = V_{g0}$ is determined by the balance between in-tunneling from the gate at the drain side of the oxide and out-tunneling into the gate at the source side. Since the in-tunneling current peaks sharply a few nanometers away from the drain-side gate corner, the actual shape of the oxide in this region has a dramatic impact on the position of V_{g0}. That V_{g0} can be used as a probe for the oxide shape is demonstrated in Fig. 12. The reference is a planar oxide of 1.5 nm thickness. If the corner of the poly gate would be rounded off with a bending radius of approximately 1nm, only a slight reduction of the in-tunneling occurs which results in a minor decrease of V_{g0}. This is because the peak of the gate current does not match with the modified gate corner in this case. If the thickness variation of the oxide near the gate corner covers the whole current peak, a large effect on the gate current is observed (Fig. 12). If the oxide thickness is increased by 0.5 nm over a distance of only 5 nm, the gate current is reduced by one order of magnitude.

The shift of V_{g0} is due to the reduction of the in-tunneling current at the drain-side gate corner. Therefore, the total gate current at $V_{gs} = 0$ V rapidly decreases with rising slope. From the measurement both the value of V_{g0} **and** the current level at $V_{gs} = 0$ V

are well defined. To a certain extent this allows to fit L and H independently! Hence, the gate tunnel current is indeed a useful tool to obtain information about the form of the oxide near the corner of the poly gate.

The threshold voltage and the drive current are not changed at all since there is sufficient gate-to-drain overlap. On the other hand, an increase of the gate oxide thickness by 0.5 nm would increase the threshold voltage by 18% and decrease the drive current by 17%. This means that a small gate reoxidation would improve the device performance by a strong degree.

It should also be mentioned that local variations of the oxide thickness at the drain corner have a very similar influence. The typical variation of the oxide thickness in a 35 nm TEM sample was 0.15 nm [2]. According to our simulations, this changes already the gate current by about one order of magnitude.

The simulation also shows, that only the oxide shape at the drain corner is important. Changing the shape both at the source and the drain side has only a minor influence on the gate current, even for a high gate bias.

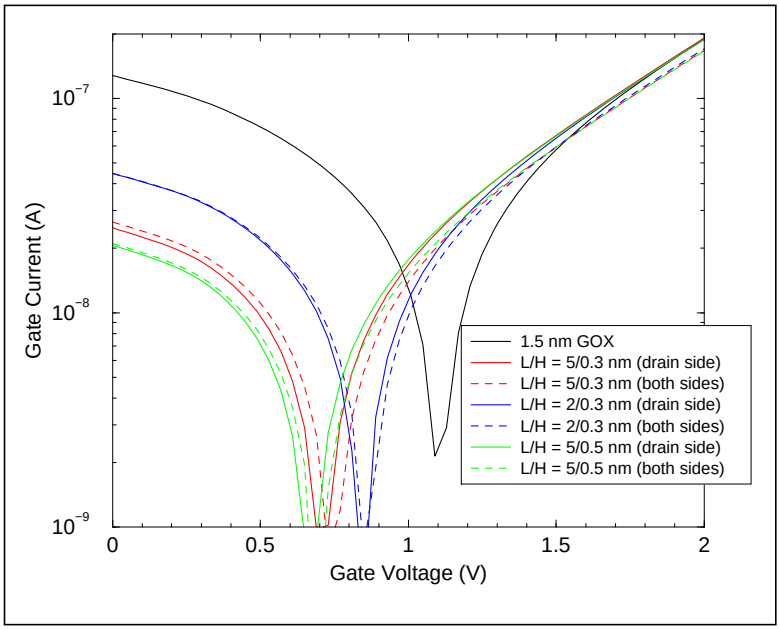

FIGURE 12. Gate current in dependency on the gate voltage for different shapes of the lower gate corner. The gate oxide thickness has been linearly increased either at the drain side only or for both sides.

Reducing the gate length and the gate-to-drain overlap (while keeping the effective channel length) reduces the parasitic gate current (Fig. 13) since the electric field is reduced. The threshold voltage and the drive current are not influenced (Fig. 14). From a technology point of view, also a gate reoxidation step in front of the spacer oxide deposition would help to reduce the gate current.

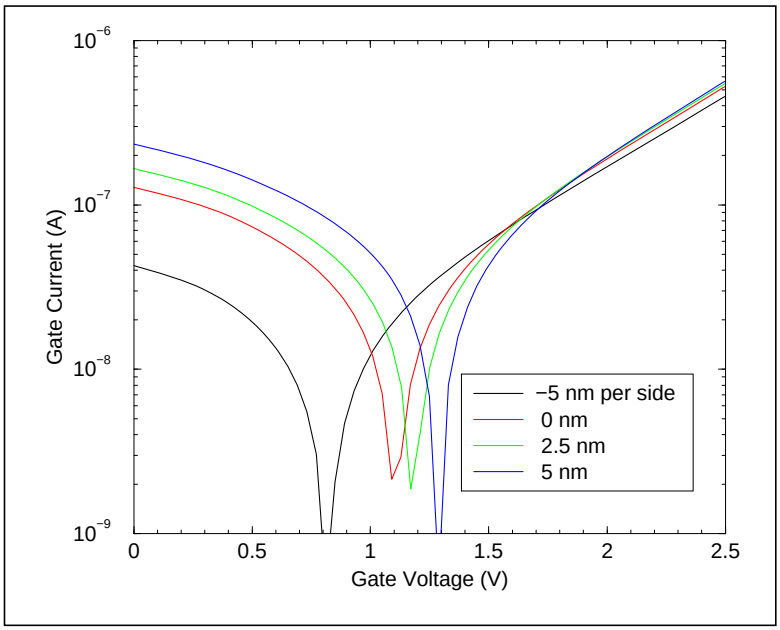

FIGURE 13. Gate current in dependency on the gate voltage for different gate-to-drain overlap. A negative value of -5 nm means a reduction of the gate length from 0.22 μm to 0.21 μm.

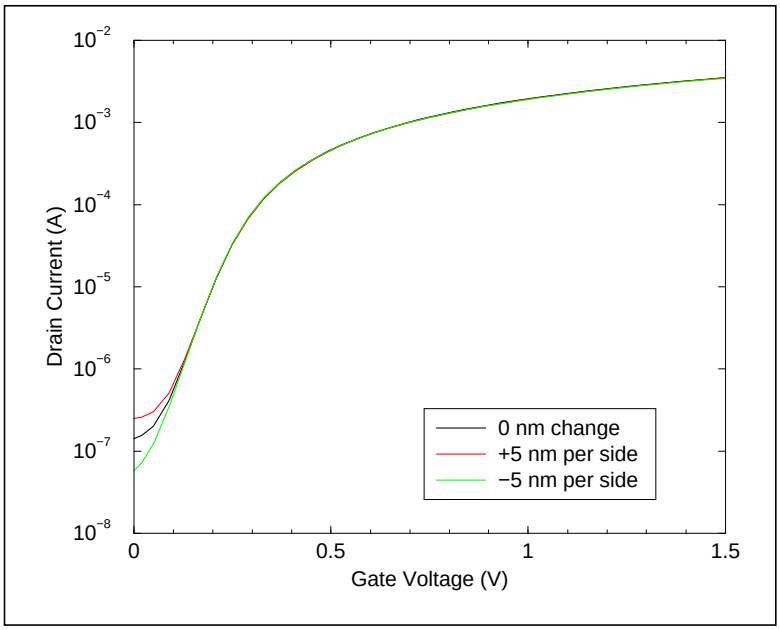

FIGURE 14. Drain current in dependency on the gate voltage for different gate-to-drain overlaps.

The shape of the phosphorus extension profile influences both the absolute value of the gate current and the gate voltage at which the current has its minimum. High concentration levels cause an increased electric field and therefore a large gate current. On the other hand, at high concentration levels also the lateral penetration depth is increased.

This increases the gate-to-drain overlap. The lateral position of the p-n junction was 24 nm distant from the corner of the gate contact for a phosphorus segregation coefficient of 1 (19 nm for SEG=3, 17 nm for SEG=5 and 14 nm for SEG=15). This means that the change of the gate-to-drain overlap is the dominant effect. Unfortunately, high drive currents are also related to high concentration levels since the effective channel length and the drain resistance are reduced (Fig. 16). The threshold voltage is influenced only to a minor degree.

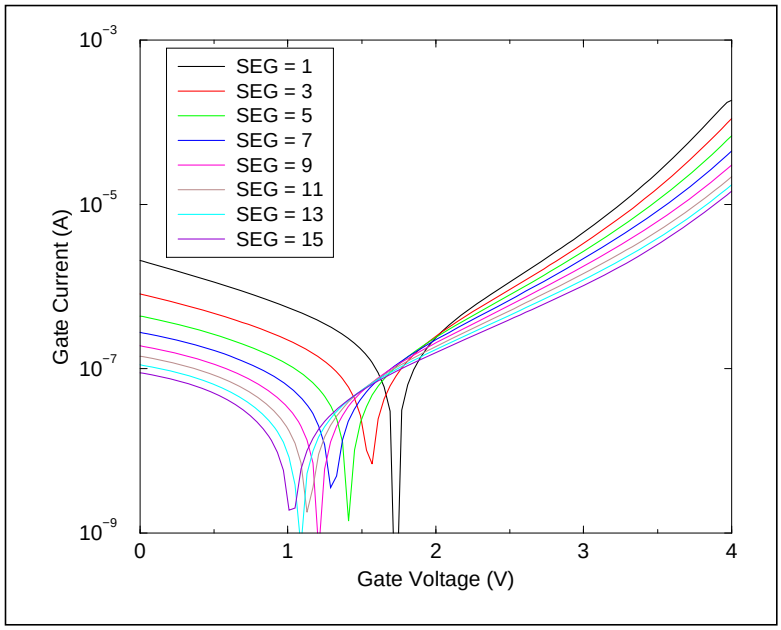

FIGURE 15. Gate current in dependency on the gate voltage for different phosphorus extension profiles.

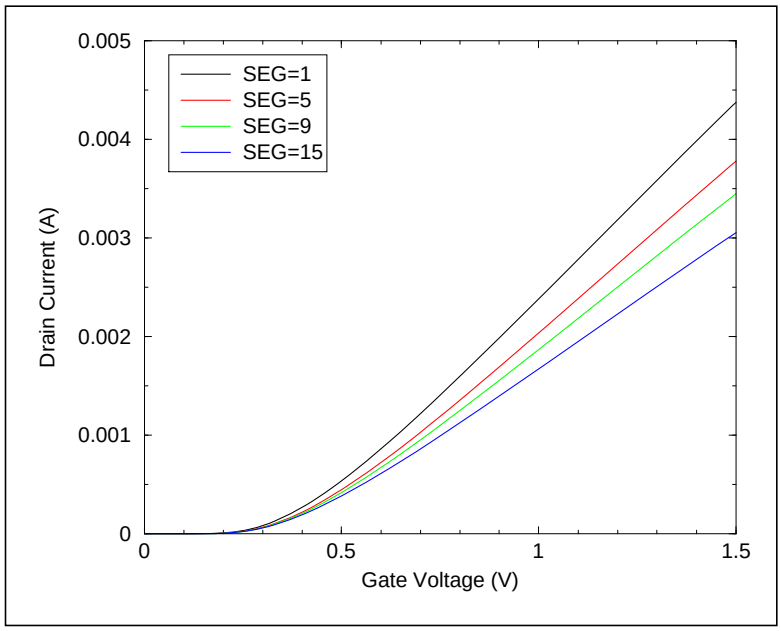

FIGURE 16. Drain current in dependency on the gate voltage for different phosphorus extension profiles.

The shape of the boron channel profile has a comparably small influence on the gate current (Fig. 17). Only for very high surface concentrations the gate current is reduced, mainly because the gate-to-drain overlap and the net extension concentration are reduced. But this leads also to unreasonably high threshold voltages due to the large charge in the depletion zone (Fig. 18).

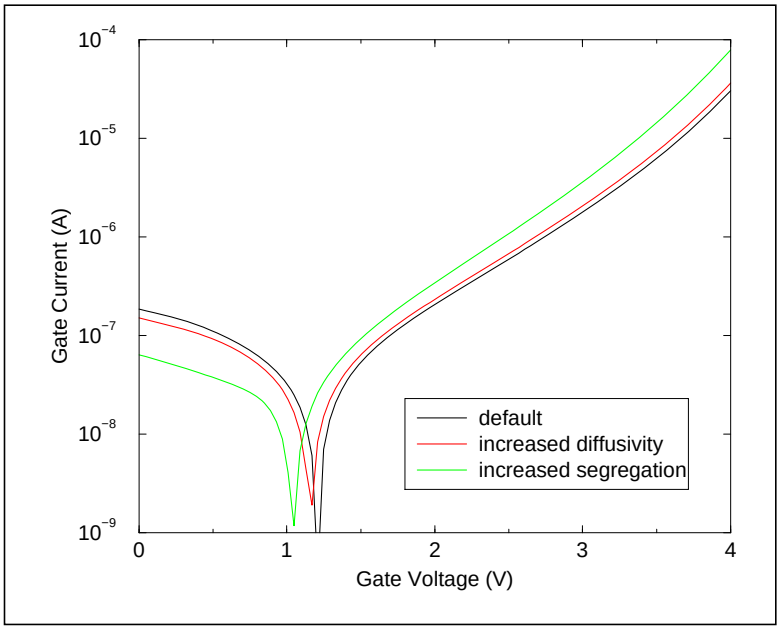

FIGURE 17. Gate current in dependency on the gate voltage for different channel profiles.

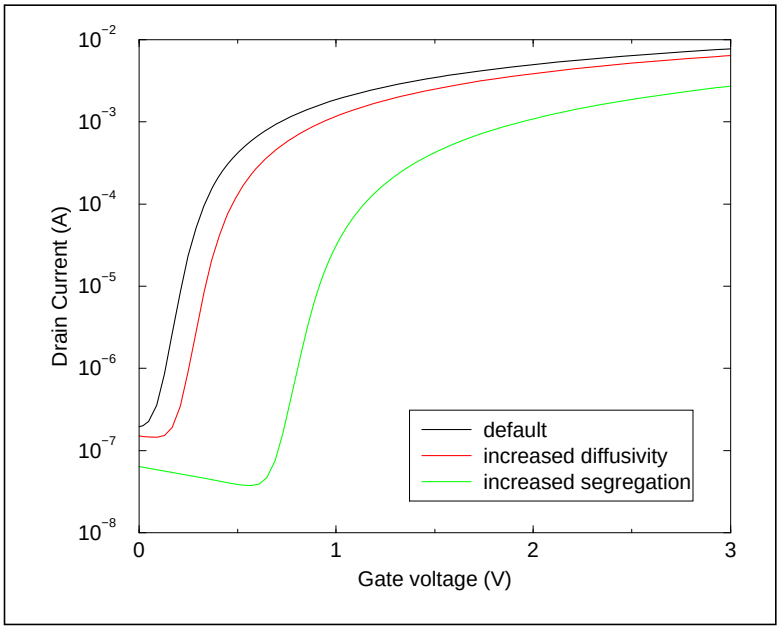

FIGURE 18. Drain current in dependency on the gate voltage for different boron channel profiles, which have been implanted with the same dose.

We have shown that the gate tunnel current is extremely sensitive to the actual shape of the oxide near the corner of the poly gate. In a band diagram this means that the thickness of the potential barrier increases there. The same sensitivity is to be expected with respect to a change of the barrier height. Such a change is caused on a long-term time scale by the trapping of holes which are generated near the interface. These positive charges are trapped in interface states and/or defect states within the oxide layer which results in a local reduction of the barrier potential and, therefore, in a strong local enhancement of the direct tunnel current. The increase of the gate current in turn will speed up the hole creation and trapping. However, the whole process is also influenced by the local decrease of the resistivity of the barrier, i.e. the decreasing potential drop in disfavor of the tunnel current. The final effect is the well-known wear-out of the oxide where the gate current is determined by the small Ohmic resistance of the affected region. Provided a good model for the accumulated positive charge as function of the tunnel-injected charge or current density, the wear-out process can be monitored by the changing gate current.

6.0 Conclusions

We have investigated the influence of essential process parameters on the gate leakage current by using a calibrated direct-tunneling model. It has been found that the gate current is very sensitive to small gate oxide thickness variations at the drain corner of the transistor. Because the doping of the transistor can be controlled very well within the process, the gate current can be regarded as an excellent monitor for small geometry variations or thickness non-uniformities. Within a given process, the gate current can be reduced with an optimized gate reoxidation step. This step increases the local oxide thickness at the drain corner of the gate contact and reduces the gate-to-drain overlap.

7.0 References

[1] National Technology Roadmap for Semiconductors (SIA, San Jose, 1997).

[2] H.S.Momose, S.Nakamura, T.Ohguro, T.Yoshitomi, E.Morifuji, T.Morimoto, Y.Katsumata, and H.Iwai, “Study of the Manufacturing feasibility of 1.5-nm direct-Tunneling Gate Oxide MOSFET’s: Uniformity, Reliability, and Dopant Penetration of the Gate Oxide”, IEEE Trans. Electron Devices, vol. 45, pp. 691-699, March 1998.

[3] A.Schenk and G.Heiser, “Modeling and Simulation of Tunneling through Ultra-Thin Gate Dielectrics”, J.Appl.Phys. vol.81(12),p.7900, 1997.