Edited by H. Ryssel, P. Pichler - September 1995

Scaling of Conventional MOSFET's to the 0.1-μm Regime

M.J. van Dort, J.W. Slotboom, and P.H. Woerlee

Philips Research Laboratories,
Prof. Holstlaan 4,
5656 AA Eindhoven, The Netherlands

Abstract

As fundamental limits of MOSFET's are being explored, new device structures have been proposed in order to maintain good short-channel behaviour in the deep submicron regime. These advanced transistors usually require complex channel and source/drain engineering, and will probably not be excepted by industry if the conventional way of scaling is still feasible. The conventional MOSFET is the benchmark for semiconductor industries. This paper addresses some of the issues which are important when conventional MOSFET's are scaled down to the deep submicron regime.

1. Introduction

In the past decades CMOS processes have been successfully scaled down to submicron dimensions. Intensive research efforts have been put into the issue of how to best scale the devices from one generation to the next one. Various scaling laws have been proposed (see e.g. [1, 2, 3]). Each of these scaling scenario's successfully miniaturize the CMOS processes, but they have different philosophies regarding issues like for instance the internal electric fields, the current density and the power consumption. In particular, an important boundary condition for the scaling factor for voltage was set by the immunity against hot-carrier degradation. Straightforward down scaling turned out to be impossible, and different drain structures have been proposed and implemented in order to meet the 10-year lifetime criterion against hot-carrier degradation. The best-known of the changes in the drain configuration are the introduction of the LDD and the LATID implantation, but also more advanced structures have been investigated. All the structures that modify the gate/drain overlap have in common that they improve the lifetime of the transistor at the expense of the DC or AC performance of the devices.

The aim of this paper is to discuss two important boundary conditions for device scaling in the deep-submicron regime, the hot-carrier degradation and the threshold voltage. Non-local carrier heating is very important in the sub 0.1-micron regime, making it possible to abandon the LDD structure for the low power supply voltages (V_{dd}) needed for these devices and return to the conventional S/D structure. This issue will be addressed in Sec. 2. Of course, the threshold voltage V_{t} has to be scaled in accordance with the decrease of V_{dd}. The low V_{t}'s required for proper circuit operation make it difficult to turn off the transistor completely. The reason is that it is not possible to scale the subthreshold swing accordingly. This puts a severe demand on λ_V, the scaling factor for voltages. The alternative is to relax λ_V. This will however have a severe consequence for the performance of the future CMOS generations, as it will increase the power dissipation and the current density to unacceptable high values [2]. Optimization in the deep-submicron regime is therefore likely to be dominated by clever power management, a proper choice for the scaling factor for the (threshold) voltage and the inevitable poor off-state performance of these MOSFET's. A thorough understanding of the mechanisms determining the V_t in the deep submicron regime is thus extremely important. This issue will be addressed in Sec. 3. Short-channel related issues are finally discussed in Sec. 4.

2. Hot-carrier immunity

For design rules smaller than 2.0 μm hot-carrier degradation is a severe problem. The lateral electric fields at the drain are high and some of the electrons gain energies high enough to surpass the Si-SiO$_2$ barrier. Subsequent trapping of these hot carriers in the gate oxide degrades the MOSFET's to an extent where they cannot be used any more.

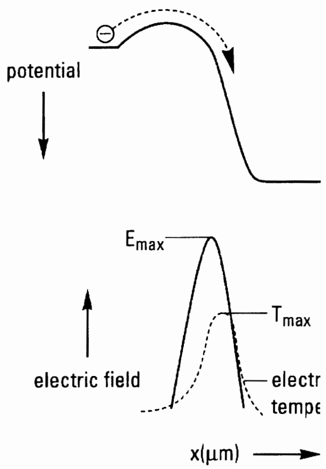


Figure 1: *Non-local carrier heating. The electron temperature is lagging behind the electric field due to the finite energy relaxation length. The energy of the electrons can be lowered by reducing the width of the electric field as well as by a reduction of $E_{\max}$.*

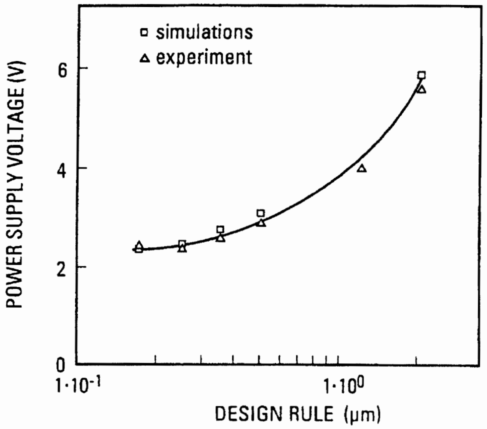


Figure 2: *Measured and simulated maximum power supply voltage $V_{\mathrm{DD,MAX}}$ as a function of the design rule of the process. The leveling-off of $V_{\mathrm{DD,MAX}}$ is caused by non-local carrier heating [7].*

The energy distribution of the electron population depends on the magnitude of the electric field and on the shape of the electric field peak. If the width of the field peak is large ($\geq$ 100 nm), the energy distribution of the electrons stays approximately in equilibrium with the local electric field. In this situation the degradation of the MOSFET can be avoided by a reduction of the maximum field at the drain (figure 1). This is accomplished with the introduction of the LDD or LATID implantations. These drain structures are used for process generation with design rules down to 0.35 μm. For generation with minimum feature lengths of 0.25 μm or less, the width of the electric field peak is so narrow that we begin to notice the effect of the finite energy relaxation length. In this situation, the electrons never reach the energies 'belonging to' the maximum electric field (see figure 1). We are then in the regime of non-local carrier heating: a reduction of the width of the electric field causes the average energy of the electrons to drop. We can benefit from this physical effect by making the S/D profile as steep as possible [4].

Another effect which helps to reduce the energy of the electrons is the presence of the Si-SiO$_2$ interface. In a properly scaled device the shallow source/drain junctions are used and the current path of the electrons is very close to the interface. The electrons notice the presence of this interface as an extra scattering mechanism. This in itself makes it more difficult for the electrons to gain enough energy to cause damage [5].

The energy distribution, or the electron temperature, can be modeled by solving the hydro-dynamical equation in addition to the normal drift-diffusion equations. A fully self-consistent solution is CPU intensive and often not necessary. An efficient post-processing method to calculate the substrate currents incorporating both the surface impact ionization as well as the non-local carrier heating has been presented in [6].

Figure 2 depicts the maximum power supply voltage $V_{\mathrm{DD,MAX}}$ as a function of the design rules for devices scaled according to the quasi-constant-voltage scaling laws [7]. Significant scattering of $V_{\mathrm{DD,MAX}}$ has been reported in the literature for 0.1-μm devices, but all values for $V_{\mathrm{DD,MAX}}$ are well above 1.5 V. The maximum allowed power supply voltage is therefore likely to be above the value required according to the a realistic scaling scenario.

3. Threshold voltage

In the previous section it was demonstrated that hot-electron degradation is not an important issue for devices in the deep-submicron regime with low V_{dd}. The scaling of the threshold voltage V_{t} on the other hand is less straightforward. The problem we encounter is that the subthreshold slope can not be scaled. Too low a V_{t} will imply that the transistor can not be completely turned off: a significant leakage current will flow with 0 Volts applied to the gate. This is for most applications not acceptable and this will probably lead to a different scaling factor for the threshold voltage than the one used for the power supply voltage. Low values of V_{t} are acceptable when the subthreshold slope is as steep as possible and the spread in V_{t} is kept to an absolute minimum value. In this section the threshold voltage is discussed in more detail.

CMOS processes scaled down to the 0.1-μm regime need high substrate doping levels in order to maintain good short-channel behaviour. The use of high concentrations of channel doping will introduce new effects on the threshold voltage. Firstly, the quantization of the inversion layer becomes noticeable when the electrons are exposed to high normal electric fields. Secondly, the statistical distribution of the dopant atoms in the channel wil affect the mean value as well as the spread of the threshold voltage.

3.1. Quantization effects

An accurate modeling of the threshold voltage is essential for the comparison of the different scaling scenarios. A systematic deviation has been found between the observed V_t and the V_t simulated with a conventional device simulator (figure 3). The simulated V_t is always lower and this difference increases with increasing substrate doping level.

This deviation has been attributed to the quantum-mechanical splitting of the energy level in the conduction band. At the onset of strong inversion an analytical solution of the Schrödinger equation is available [8] and it is quite easy to calculate the threshold voltage with inclusion of the quantization effects. A simple model that accounts for this QM effect has been presented in [9] and compared with the selfconsistent solution [10]. The result is

$$V_t^{\mathrm{QM}} \approx V_t^{\mathrm{CLAS}} + \Delta\Psi_{\mathrm{S}}\left(1 + \frac{1}{2C_{\mathrm{ox}}}\sqrt{\frac{\varepsilon_{\mathrm{Si}} q N_A}{\phi_B}}\right), \tag{1}$$

with

$$q\Delta\Psi_{\mathrm{S}} \approx \frac{13}{9}\beta\left(\frac{\epsilon_S}{4qu_T}\right)^{1/3} E_y(0)^{2/3}. \tag{2}$$

In this formula, $E_y(0)$ is the perpendicular electric field at the Si-SiO$_2$ interface, u_T the thermal voltage and $\beta = 4.3 \times 10^{-8}$ eVcm. The perpendicular electric field at the onset of strong inversion is given by $E_y(0) \approx qN_AW_m/\epsilon_S$.

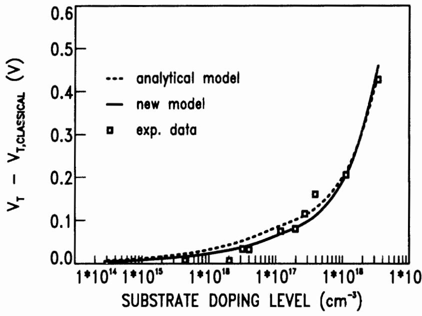


Figure 3: *Deviation between the classically-simulated and the measured* <u>*long-channel*</u> V_t *as a function of the doping concentration* (t_{ox}= *14 nm) [9].*

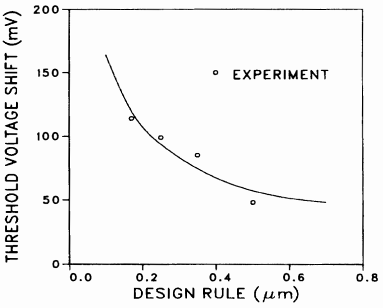


Figure 4: *Deviation between the classically-simulated and the measured* <u>*long-channel*</u> V_t *as a function of the design rule. The measurements have been done on devices which were scaled according to the QCV scaling rules. Solid line is the theoretical curve for this scaling scenario [9].*

This model has been implemented in the device simulator MINIMOS, and is used throughout this paper to calculate the threshold voltage. Figure 3 shows the V_t– shift as a function of the doping concentration at fixed oxide thickness t_{ox}. In a properly scaled device, the oxide thickness decreases as a function of the design rule. Figure 4 shows the measured and simulated V_t shift for devices scaled according to the quasi-constant-voltage scaling rules. From this figure we observe that the quantization effect can not be neglected in the deep-submicron regime.

3.2. Statistical variations of the channel dopants

Another problem we encounter when conventional MOSFET's are scaled to deep submicron dimensions is of a more statistical nature. If the active area ($W \times L$) of a MOSFET is small, the depletion layer charge consists of relatively few dopant atoms. The number of atoms building up this depletion layer will fluctuate. In addition, these charges will not be distributed uniformly. These two effects have an effect on the threshold voltage. It has been shown that the average value of the threshold voltage, $< V_t >$, drops due to the microscopic distribution of dopant atoms [11]. More importantly, it has a significant effect on the spread σ_{V_t} of the V_t. Significant fluctuations of the threshold voltage have recently been reported [12].

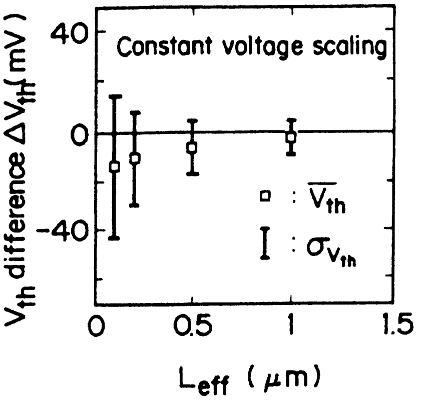


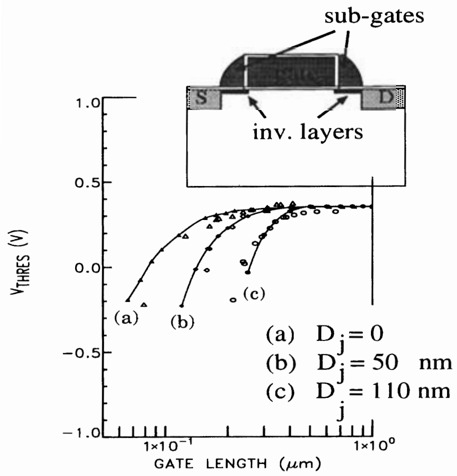
Figure 5: V_t *distribution due to the* N_A *distribution. Constant voltage scaling laws are used. Figure taken from ref. [11].*

Figure 5 shows the V_t distribution as a function of the design rule for devices scaled according to the constant-voltage scaling laws. Miniaturization of the spread is mandatory for circuit operation when the devices are operating at low values of V_{dd} and V_t. This effect could well be a limitation for ultimate conventional MOSFET's and might have a severe impact on the way the devices are scaled.

4. Short-channel devices

The limits of the conventional MOSFET can further be explored by examining the control of the short-channel behaviour, or the V_t roll off. A key parameter in the various scaling scenario's is the junction depth D_j. It determines the drain-induced barrier lowering of the short-channel MOSFET's. A minimum junction depth has been realized in experiments by Noda et al.[13] using two subgates to induce inversion layers acting as drain extensions (see inset of figure 6). These two inversion layers mimick infinitesimal shallow junctions (inversion layers are typically 30 Å thick). The larger junction depths used in figure 6 are oridinairy S/D constructions. Accurate simulations of the D_j dependence is important. Figure 7 shows the original data as well as the MINIMOS simulations that we have performed on these data. The original data can be reproduced using device simulations and the model for V_t described in Sec. 3.

Now that we have verified the simulation tools, we can estimate the limits of the conventional MOS scaling. The long-channel threshold voltage $V_{\mathrm{t,L}}$ has been varied and the minimum effective channel length for which we still get good short-channel behaviour has been determined. As a criterion for good short-channel behaviour we have assumed that the V_{t} of the MOSFET with the minimum gate length is at least $0.75 \times V_{\mathrm{t,L}}$, when the voltage applied to the drain is to $V_D = 1.5$ V. This means that we test the punch-through behaviour.



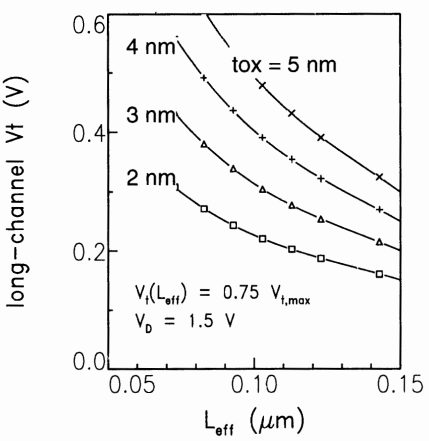


Figure 6: *Threshold voltage V_{t} versus the gate length for different junction depths D_{j}. Open symbols are the data from [13], solid symbols are the results of our MINIMOS simulations.*

Figure 7: *Long-channel V_{t} versus the mir imum effective channel with good shor channel behaviour.*

We have implicitly assumed that the power supply voltage of deep-submicron generation will be 1.5 V. This is a reasonable assumption and has been used in many experimental studies in the 0.1 μm regime. The results of the simulations are displayed in figure 7.

For instance, if we aim a MOS generation with $L_{\mathrm{eff,min}}$= 0.10 μm and $V_{\mathrm{t,L}}$= 0.35 V, we see from figure 7 that the oxide thickness should be 3.5 nm or less. This value for the oxide thickness is expected to be the lower limit for conventional $\mathrm{SiO_2}$. Below 3.5 nm direct tunneling through the oxide becomes too important.

4.1. Reverse short channel effect (RSCE)

The MOSFET is an intrinsic 2D device, and an accurate simulation of the 2D doping profile is essential. One of the most important phenomena for short-channel devices is RSCE (figure 8), which is caused by anomalous diffusion effects near the edge of the gate. RSCE can be caused by 2D oxidation-enhanced diffusion due to the gate reoxidation [14], or 2D transient-enhanced diffusion due to the implantation of the LDD or the source and drain [15].

Although these effects can in principle be simulated with a 2D process simulator, an accurate prediction of RSCE is still difficult. This is mainly caused by the complex nature of the poin-defect dynamics, especially for TED, and the inability to directly measure the 2D doping profile. The diffusion coefficient of boron can be approximated by $D_B \approx D_B^* \times C_I/C_\mathrm{I}^*$ where $*$ denotes the equilibrium value and C_I the 2D interstitial profile. Simulation of the RSCE thus requires the interstitial distribution.

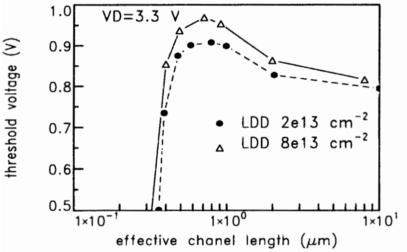


Figure 8: *Reverse short channel effect. This figure illustrates the RSCE due to 2D TED. The LDD dose was varied in this experiment.*

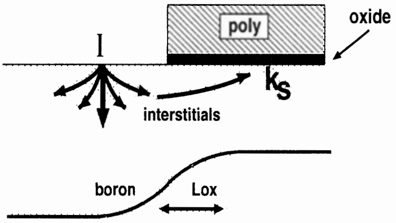


Figure 9: *Simulation of RSCE requires the simulation of C_I back to thermal equilibrium. Excess interstitials are injected (rate I) during oxidation or implantation and absorbed at the interface with rate k_S.*

Key parameters for the simulations of the evolution of the interstitial profile are models for the injection I and k_S/D_I, the ratio of the recombination rate at the interface and the diffusion coefficient of interstitials (figure 9). Injection of point defects is in this case caused by oxidation or by implantation.

Special test structures have been designed to investigate the 2D boron profile after 2D TED, showing an important role in the 2D point-defect dynamics for the extended defects formed during armorphizing implantations [16]. The modeling of the initial conditions for TED and an efficient inclusion of the extended defects in the process simulations is still a challenging problem.

5. Conclusions

Some of the issues concerning the scaling of MOSFET's have been discussed. From the point of view of device operation, there seems to be no fundamental limit. Quasi-conventional devices an effective channel length as small as 0.05 μm have been fabricated [17]. In the deep-submicron regime there will be less emphasis on the hot-carrier performance. The scaling of the V_t and V_dd is more important. For MOSFET's in the 0.1-μm regime with low V_t's, it is absolutely mandatory to minimize the spread in the transistor parameters, especially for the V_t. Techniques to identify the process parameters responsible for the fluctuations in the V_t are available. This issue will be discussed elsewhere [18]. It is expected that quantization effects and the statistical distribution of the impurities will become more important. It is still an open question of how to implement physical models to account for these effects efficiently in the device simulators. For MOSFET's in the 0.1 -μm regime, it is further expected that anomalous diffusion effects will dominate the process simulations. Although progress has been made in this field in the last couple of years, it is still an important research topic. Conventionally scaled MOSFET are expected to be operational for process generations designed for effective channel lengths below 0.1 μm.

References

[1] P. Chatterjee, W. Hunter, T. Holloway, and Y. Lin, *IEEE Electron Dev. Lett.*, vol 10, p. 220 (1980).

[2] G. Baccarani, M. Wordeman, R. Dennard, "Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design", *IEEE Trans. Electron Devices*, vol 31, p. 452 (1984).

[3] H. Hu, J. Jacobs, L. Su, and A. Antoniadis, "A Study of Deep-Submicron MOSFET Scaling Based on Experiment and Simulation", *IEEE Trans. Electron Devices*, vol. 42, p. 669 (1995).

[4] J. Slotboom, et al., "Non-Local Impact Ionization in Silicon Devices", *Tech. Digest IEDM*, p. 127 (1991).

[5] J. Slotboom, G. Streutker, G. Davids, and P. Hartog, "Surface Impact Ionization in Silicon Devices", *Tech. Digest IEDM*, p. 494 (1987).

[6] M. van Dort, J. Slotboom, G. Streutker, and P. Woerlee, "Lifetime Calculations of MOSFET's using Depth-Dependent Non-Local Impact Ionization", *Microelectronics Journal* vol. 26, p. 301 (1995).

[7] P. Woerlee et al., "The Impact on Hot-Carrier Degradation and Supply Voltage of Deep-Submicron NMOS Transistors", *Tech. Digest IEDM*, p. 537 (1991).

[8] F. Stern, "Quantum properties of surface space-charge layers", *CRC Crit. Rev. Solid State Sci.* , p. 499, 1974.

[9] M. van Dort et al., "Quantum-Mechanical Threshold Voltage Shifts of MOSFET's Caused by High Levels of Channel Doping", *Tech. Digest IEDM*, p. 495 (1991).

[10] M. van Dort et al., 'A Simple Model for Quantisation Effects in Heavily-Doped Silicon MOSFET's at Inversion Conditions.' *Solid-State Electronics*, Vol.37, p. 411 (1994).

[11] K. Nishinohara, N. Shigyo, and T. Wada, "Effects of Microscopic Fluctuations in Dopant Distributions on MOSFET Threshold Voltage", *IEEE Trans. Electron Devices* vol 39, p. 634 (1992).

[12] T. Mizuno, J Okamura, and A. Toriumi ,"Experimental Study of Threshold Voltage Fluctuation Due to Statistical Variation of Channel Dopant Number in MOSFET's", *IEEE Trans. Electron Devices* vol 41, p. 2216 (1994).

[13] H. Noda, F. Murai, and S. Kimura, "Threshold Voltage Controlled 0.1-μm MOSFET Utilizing Inversion Layer as Extreme Shallow Source/Drain", *Tech. Digest IEDM*, p. 123 (1993).

[14] M. Orlowski, C. Mazuré and F. Lau, "Submicron Short Channel Effects due to Gate Reoxidation Induced Lateral Interstitial Diffusion", *Techn. Digest IEDM*, p. 632 (1987).

[15] C. Rafferty et al., "Explanation of Reverse Short Channel Effect by Defect Gradients", *Tech.Dig. IEDM*, p. 311, 1993.

[16] M van Dort et al., "Two-Dimensional Transient-Enhanced Diffusion and Its Impact on Bipolar Transistors", *Tech.Dig. IEDM*, p 865, 1994.

[17] A. Hori et al., "A 0.05-μm with Ultra Shallow S/D Junctions Fabricated by 5 keV Ion Implantation an Rapid Thermal Annealing", *Tech.Dig. IEDM*, p. 485 (1994).

[18] M. van Dort and D. Klaassen, "Sensitivity Analysis of an Industrial CMOS Process using RSM Techniques", *Proc. SISPAD*, 1995.