The advent of Very-Large-Scale-Integration (VLSI) of MOS circuits quickly revealed the need for computer-aided simulation in modern MOS transistor design, especially since experimental investigations are all too often expensive, time consuming and even ineffective. The drawback of all published analytic models lies in the fact that their assumptions and regional approximations are so restrictive that MOS performance can only be predicted in a very limited range. The miniaturisation of the structures accelerated the collapse of these simple models. To make the characterisation of modern devices realistic and reasonable again, the designer is forced to use numerical models of higher orders.

Recently several attempts have been made to develop computer simulation programs based on two dimensional models without restrictive assumptions due to physical constraints (3), (8), (10), (12), (14), (19), (21), (22), (25). But in almost every case, a wider application of these programs was restricted either by limited performance (due to the lack of numerical stability) or by the need for high computer operating costs. Our activity was primarily concerned with the development of a fast and highly user-orientated software tool for practical applications and not for pure academic purposes.

It is beyond the scope of this feature to give a detailed explanation of the operation of a two dimensional simulation program and the physical background. Thus, just a few sentences are given to an outline of this nevertheless important point: In order to accurately
analyse an arbitrary semiconductor structure under all kinds of operating conditions, the basic carrier transport equations, as first given in the classical case by Van Roosbroeck (24), have to be solved. These equations represent a boundary value problem with three nonlinear coupled partial differential equations. These analytic equations have to be transformed into a system of discrete equations and linearised to obey the laws of modern numerical mathematics. Gummel (7) was the first to accomplish this vital task. At the centre of our attention, however, are additional assumptions and simplifications which speed up a computer program and the modelling of physical parameters (e.g. carrier mobility) and these have to be undertaken extremely carefully to obtain quantitatively accurate results. A detailed explanation of how all these difficulties are bypassed in the case of the MOS transistor can be found in (15)-(18).

A didactic example

It is difficult to provide an interesting example for the experienced reader, whilst making it easy to understand for readers with a general interest in modelling but without specific knowledge of MOS devices. We have chosen the effects of ion implantation on short channel devices for the purpose of demonstrating the use of two dimensional simulation. Three devices are calculated, whose properties become apparent from the original simulation input decks presented in Figure 1. The following discussion of Figure 1 also

### Fig.1. Some typical input decks

**ONE-MICRON ANALYSIS (DEVICE 1)**

```
BIAS UD = 3. UG = 0.
END
```

**ONE-MICRON ANALYSIS (DEVICE 2)**

```
BIAS UD = 3. UG = 0.
IMPLANT ELEM = B DOSE = 3.5E11 AKEV = 25 TEMP = 925 TIME = 1800
END
```

**ONE-MICRON ANALYSIS (DEVICE 3)**

```
BIAS UD = 3. UG = 0.
IMPLANT ELEM = B DOSE = 3.5E11 AKEV = 25 TEMP = 925 TIME = 1200
IMPLANT ELEM = B DOSE = 1.5E11 AKEV = 100
END
```
Two dimensional MOS transistor modelling

demonstrates the ease of using 'MINIMOS', our simulation program.

The first line is a title line, which is used only to identify the output of the program. The input syntax is totally based on a master key, key and value structure (18). The next input line which is the "DEVICE" statement, characterises the device. Specified is an n-channel device (CHANNEL=N) with an n-doped polysilicon gate (GATE=NPOLY), an oxide thickness of 35nm (TOX=350.E-8), a channel width of 10μm (W=10.E-4) and a channel length of 1μm (L=1.E-4). The 'BIAS' statement specifies the operating point. A drain voltage of 3V (UD=3.) and a gate voltage of zero volts (UG=0.) has been chosen. The substrate voltage is assumed to be zero by 'MINIMOS', if not specified otherwise. The 'PROFILE' statement is used to specify the substrate doping and the source/drain diffusion. In the examples presented here we used the simplest way of defining a doping profile, that is the direct calculation by 'MINIMOS'. Another possibility would be to make use of a technology simulation program like 'SUPREM', the Stanford University PRocess Engineering Models program (1), for the more accurate calculation of vertical profile shapes which are fitted in the lateral direction. For our simulation, a substrate doping of 10¹⁵cm⁻² (NB=1.E15) and a source/drain implantation with phosphorus (ELEM=PH), an implantation dose of 10¹⁵cm⁻² (DOSE=1.E15) and an implantation energy of 40keV (AKEV=40) is specified. The implantation is performed through an isolation oxide of 35nm (TOX=350.E-8) and followed by an annealing step at 1200°C (TEMP=1200) for 1200s (TIME=1200). The second input deck further includes an "IMPLANT" statement which defines a channel implantation with boron (ELEM=B), a dose of 3.5·10¹¹cm⁻² (DOSE=3.5E11), an energy of 25keV (AKEV=25) annealed at 925°C (TEMP=925) for 1800s (TIME=1800). The third input deck has an additional "IMPLANT" statement specifying a second, deeper channel implantation with boron (ELEM=B), a dose of 1.5·10¹¹cm⁻² (DOSE=1.5E11) and an energy of 100keV (AKEV=100). It is assumed that both channel implantation steps are annealed at the same time. It is fairly well known that the first of these three devices is, owing to the short channel effect, "normally-on" and that the shallow implantation of device 2 utilises threshold shift to obtain a "normally-off" device. Furthermore, the deep implantation of device 3 is necessary to avoid punch through. These effects will now be demonstrated by 'birds-eye-view' and contour-plots of the physically relevant quantities in the interior of the three model devices.

The calculated doping density distributions for our devices are shown in Figures 2, 3 and 4. From these, one can read the depth of the pn-junctions (under source and drain) to be approximately 300nm. The surface concentration of the source and drain regions is about 10²⁰cm⁻². The effective channel length is reduced by the lateral...
subdiffusion to about 0.6μm. The shallow channel implantation for threshold tailoring can be seen in Figures 3 and 4. Additionally, Figure 4 shows the deep implantation for punch through suppression. The threshold voltage is only marginally affected by the deep implantation.

Figure 5 shows the distribution of the electric potential for the first device. The drain contact is on the right. In the depletion region of the reverse biased drain bulk diode, the potential decreases monotonously and it is more or less constant in the highly doped source and drain regions. The barrier at the source channel diode is relatively small. Figure 6 shows the potential distribution in the second device. The 'birdseye-view' plot looks very similar to the plot in Figure 5. The contour-plot, however, shows quite a pronounced potential basin directly below the interface. Of even greater importance than the basin itself is the saddlepoint below the basin. At this saddlepoint, the electric field vanishes and current

**Fig.5. Electric potential for device 1**

**Fig.6. Electric potential for device 2**

**Fig.7. Electric potential for device 3**

**Fig.11. Lateral current density for device 1**

**Fig.12. Lateral current density for device 2**

**Fig.13. Lateral current density for device 3**
can only flow by carrier diffusion. This sort of saddlepoint is, following the proposition of many authors (2), (9), a typical indication of the punch-through effect. The electric field which is induced by the gate is unable to separate the depletion regions of source and drain. These depletion regions are in contact below the region of control by the gate. As it will become visible later on, the saddlepoint is a reliable indication of the punch-through effect, but it does not need to exist. Figure 7 shows the potential distribution in the third device. The 'birds-eye-view' plot differs just marginally from the plot in Figure 6. But from the contour plot one can see a well pronounced barrier between source and channel which guarantees the "normally-off" behaviour.

Figure 8 shows the electron density distribution in the first transistor. For a better global imagination two 'birds-eye-view' plots and a contour plot are linked in all figures showing carrier density distributions. The 'birds-eye-view' plot on the left is given in the same mode as for the potentials. The right plot shows the distribution as a mirror image at the surface, thus giving a better insight in the channel region. Because of the slightly negative threshold voltage, the surface concentration in the channel is relatively high; the operating point is in the strong inversion region. One can clearly see the carrier minimum near the drain representing the pinch-off region. Figure 9 shows the concentration of electrons in the second device. As expected the surface concentration has decreased in the channel area because of the channel implantation. In spite of this, something like a carrier channel occurs at a depth of about 300nm. That this is caused by punch through, will become clearly apparent from the current density distribution. Figure 10 shows the electron distribution in the third device. The second implantation results in a monotonous decrease of the carrier density from the transistor surface into the bulk, which indicates the suppression of punch through. The device operates deep in the region of depletion, which is distinctly manifested by the quantitative value of the carrier density in the channel.

Figure 11 shows the lateral current density distribution in the first device. For better visibility, the plot on the right again shows the mirror image in the same fashion as for the electron distributions. In the channel near the source side, the current is forced to flow at the surface by the transversal component of the field. But already in the middle of the channel, a typical short channel effect, one can watch current spreading caused by the drain influence. It also should be noted that the current channel is fairly wide. The reason for this phenomenon is to be found in a superposition of an inversion channel and a punch through channel. The maximum of the lateral current density surprisingly lies below the contacts. This fact becomes clear when we consider current continuity. Current can only pass through the contact in a transversal direction. Current flow in the semiconductor, however, takes place globally in the lateral direction from source to drain.

As current flow is continuous, the lateral current component below the contacts has to be large because the flux in the channel, which, as mentioned, is relatively wide, is also large. The lateral current distribution for the second device is shown in Figure 12. As one can see, this device is operating in the punch through mode. The current flow takes place in a wide channel in the bulk (4). Effectively, surface current does not exist. Furthermore, the maximum of the current density has decreased by more than an order of magnitude compared to the first device. Figure 13 shows the lateral current density distribution for the third device. The second channel implantation results in a total suppression of punch through in this operating point. The entire current flows at the semiconductor surface, but the peak value of the current density is about a factor of 200 lower than in the second device. Current density distributions of this shape are typical for regularly operating transistors in subthreshold and can be used as criteria for evaluation.

Figure 14 shows the mobility distribution in the first device. In the highly doped source and drain regions the mobility is eminently reduced by impurity scattering. Below the source region the mobility shows a fast increase up to the bulk value. This does not hold for the region below the drain because the electric field in the reverse biased drain/bulk diode reduces the mobility due to velocity saturation. As an interesting detail, note the local mobility maximum below the drain. In this small region almost no impurity scattering takes place and the electric field is still too low to reduce the mobility effectively.
In the channel of the surface, the mobility decreases monotonously from source to drain. Just in the first part of the channel near the source, the mobility stays fairly constant since there is only a negligible electric field component parallel to current flow. In transversal direction towards the surface the mobility is reduced by surface scattering. A detailed discussion of the different scattering mechanisms which lead to mobility reduction and the corresponding mathematical formulae can be found in (18). The mobility distribution in the second device is given in Figure 15. The qualitative shape of this figure is, as expected, rather similar to Figure 14. The additional mobility reduction at the surface caused by scattering at the impurities of the shallow implant becomes apparent. A further noticeable difference is the fact that the mobility stays fairly constant at the surface from source to drain in a longer region because of the reduced short channel effect. Figure 16 shows the mobility distribution in the third device. Almost no changes are visible in the channel area compared to the second device. Only in the depth of the p-n junction does an additional mobility reduction become apparent. This is caused by the impurities of the second implant. However, this detail is not of relevance as far as the device behaviour is concerned.

Process sensitivity

As already mentioned, VLSI is based on the miniaturisation of the single transistor device. To shrink just the physical device dimensions will usually cause serious problems as far as the device behaviour is concerned. As is commonly known, all device parameters have to be "scaled" (6), (11) together with the device geometry according to certain rules. In general, lower voltages, heavier doping, shallower junctions, and thinner oxides help to maintain applicable device characteristics as channel length is reduced. Down to about two microns channel length, the device behaviour can be suitably controlled by the relevant technological steps (implantation, diffusion, oxidation). However, as often observed through experimental investigations, this controllability is no longer ensured for devices with a further reduced channel length. Reproducibility tends to become worse at decreasing feature size. The deviation of device parameters of transistors on one and the same chip, which should behave identically, increases.

In order to achieve more transparency of this fact, we performed a process sensitivity analysis of certain device parameters with MINIMOS, our two-dimensional MOS simulation program, with measurements as a verifying feedback. In this chapter the sensitivity of the threshold voltage, which is usually the most interesting device parameter for the designer, will be outlined for a well established short channel MOS process to establish a practical limit of miniaturisation for a given technology. However, the analysis of threshold sensitivity is just an example for a strategy, which is applicable to examining the sensitivity of any device parameter.
Two dimensional MOS transistor modelling

The technological properties of the analysed devices

An n-channel silicon gate process with arsenic source/drain junctions and a double channel implantation has been chosen. Figure 18 shows the doping distribution logarithmically drawn in a quasi-three-dimensional plot for a one micron transistor. The channel implantation is performed with a dose of $3 \times 10^{11} \text{cm}^{-2}$ and an energy of 35keV for the shallow layer and a dose of $10^{11} \text{cm}^{-2}$ and an energy of 160keV for the deeper layer.

Figure 19 shows details of the doping profile. A junction depth of 320nm and a subdiffusion of about 200nm is obtained by this process. The extremely steep gradient of the junctions is typical for arsenic. The oxide thickness is about 50nm for these devices. The whole process was designed for two micron lateral dimension and our main question was: How do these devices behave if the channel length is reduced?

The definition of the threshold voltage

We define threshold voltage as the applied gate voltage, at which the device sinks 0.1μA times the channel width per channel length. The channel length is defined as the distance between the metallurgical junctions. With this definition it is ensured that no threshold voltage shift versus channel length for long devices occurs and we can thus directly obtain a quantitative measure for the influence of the short channel effects.

Figure 20 shows the threshold voltage versus channel length for our devices. An operating point of 3V drain bias and -2V bulk bias has been chosen as a fair trade-off for the comparison of different channel lengths. To avoid confusion all the following figures will also refer to this operating point. Figure 20 reflects the well known decrease of the threshold voltages with shrinking device length, which becomes dramatic at a length of below one micron.

Sensitivities

Usually in papers on short channel
MOS transistors, a comparison between theoretical curves and selected experimental results is given. Some of them report on statistical measurements (5), but only one paper (26), as shown to the authors so far, deals explicitly with the sensitivity of an electrical parameter, namely the threshold voltage. However, with respect to the inherent dependence of most parameters on the dispersion of geometry and technology, it seems to be a real necessity to analyse and present these dependencies directly.

Numerical investigations in conjunction with experiments have been carried out to extract the most interesting sensitivities. A two-dimensional simulation program like MINIMOS is excellently suited for this purpose. One just has to adapt the physical model parameters of the program (e.g. mobility parameters) to obtain good quantitative agreement of measured and simulated results. This can be done even with a few non-critical, experimentally well analysed devices. In order to obtain a sensitivity, one only has to vary the interesting parameter in the vicinity of its nominal value and then to differentiate with the results. This parameter variation within a small range cannot, in general, be performed experimentally. However, with a fast modelling program, the partial derivative of any electrical parameter with regard to any technological or geometrical parameter can be calculated easily with the outlined strategy.

Figure 21 shows the partial derivative of the threshold voltage with regard to the channel length versus channel length for our devices; that is the sensitivity of the threshold voltage on tolerances of the channel length. Assume a transistor with an effective channel length of one micron accurate to ten percent. With this figure one can read off an uncertainty of the threshold voltage of ±60mV, which will cause quite a few circuits to fail in operation.

Figure 22 shows the sensitivity of the threshold voltage on tolerances of the oxide thickness. As one probably has not expected at first glance, this sensitivity decreases for devices with short channels. This is because of the decreasing influence of the bulk charge with shrinking channel lengths. Note that this figure is qualitatively very similar to the figure showing the threshold voltage versus channel length (Figure 20), which can even be understood by analytical reasoning (18). With an uncertainty of about ±40mV for a 5 micron device and not even half this value for a 1 micron device. However, considering the absolute value of the threshold voltage the oxide thickness is a very important technological parameter.

Figure 23 shows the sensitivity of junction depth tolerances versus channel length. A one micron device with an uncertainty of 10% in the junction depth has an uncertainty of about ±40mV of the threshold voltage. The underlying physical effect of this sensitivity is the reduction of the channel charge by the depletion regions of source and drain (cf. (23)).

Figure 24 shows the sensitivity on drain bias variation. A 300mV change, that is 10% of the applied bias, results in about 30mV change of the threshold voltage for this operating point. Again the modulation of the depletion region of the drain is the relevant physical effect.

Figure 25 shows the sensitivity on bulk bias variation. A 200mV change, that is again 10%, results in a threshold shift of about 11mV, which is usually not dramatic. For the practical problem, however, one has to deal with the sum of all these uncertainties. Therefore this influence may also become important. An interesting detail of this figure is the fact that the sensitivity decreases first with shrinking channel length and at a certain length increases again and even more rapidly. This behaviour is caused by a superposition of the short channel effect, which decreases this particular sensitivity, and the punch-through effect, which increases the sensitivity.

Fig.17. Subthreshold characteristics
Figure 26 shows the influence of an implantation energy fluctuation. Qualitatively, the superposition of the short channel effect and the punch through effect is again apparent. The absolute value of this particular sensitivity is low due to the fact that the depletion region below the channel covers the whole implanted region at this operating point.

Figure 27 shows the sensitivity on uncertainties of the implantation dose. The figure is rather similar to the last one as expected.

Figure 28 shows the temperature coefficient of the threshold voltage for our devices. We have qualitatively a similar behaviour as already discussed, namely the superposition of short channel effect and punch through. The absolute value is around \(-1 \text{mV/K}\). The qualitative behaviour as well as the absolute value of this sensi-
Fig. 20. Threshold voltage versus channel length

Fig. 21. Sensitivity on channel length tolerances

Fig. 22. Sensitivity on oxide thickness tolerances

Fig. 23. Sensitivity on junction depth tolerances

Global sensitivities

The partial derivatives denote isolated sensitivities on a set of certain parameters. These values show which parameters are the most critical ones. However, in addition, a global sensitivity number seems to be useful. This indicates the cumulative effect of the isolated sensitivities. The global sensitivity is related to a certain technology and its expected application. It should indicate the limit of channel length reduction. To obtain such a global number typical ranges of deviation of design parameters have to be specified. The table in Figure 29 is an example for such a specification.

Figure 30 shows the global threshold voltage sensitivity based on the specifications of Figure 29. \( \sigma_D \) denotes the uncertainty of the threshold voltage for identical devices on the same chip. \( D \) stands for device. This sensitivity is given by just the length influence, as the other parameters are commonly very homogeneous across one wafer. \( \sigma_W \). \( W \) stands for wafer, denotes the uncertainty for identical devices on wafers, which have been fabricated with different charges. Here one has to use a Euclidian norm over all deviations. Note that this value is highly constant down to a certain channel length, but then increases dramatically. The channel length at which the excellently pronounced knee is located, at 1.4 microns for our devices, can thus be interpreted as the practical limit of channel length reduction due to threshold uncertainty.

Conclusions

In this feature, results obtained with the modelling program 'MINIMOS' are discussed. The presented examples are not only of didactic relevance but also of large practical value. The power of two dimensional simulations has been demonstrated by calculating the distributions of the relevant physical quantities in the interior of one micron gate length MOS-transistors differing in the ion implantation steps. Many features become apparent, which are not all accessible to measurement such as the carrier density and the current density distributions.

Figure 24. Sensitivity on drain bias variation

Fig. 25. Sensitivity on bulk variation

Fig. 26. Sensitivity on implantation energy tolerances

Fig. 27. Sensitivity on implantation dose tolerances

Fig. 28. Sensitivity on temperature variation
Two dimensional MOS transistor modelling

within a device.

Furthermore, the threshold voltage sensitivity to various parameters in short channel MOS transistors of a specific technology has been evaluated as an example of practical importance for a generally applicable method. The most critical parameters have been obtained, and in addition, a global sensitivity number has been derived indicating the practical limit of miniaturisation for the analysed process technology.

Acknowledgment

This work has been supported by the "Fond zur Forderung der wissenschaftlichen Forschung" (Projekt Nr. S22/11). Essential help of Siemens AG, Munich, in providing MOS devices is gratefully acknowledged. Critically reading our manuscript by Dr. J. Machek is very appreciated. Last but not least the authors wish to thank Dr. D. Schornbock and the whole staff of the computer centre for the excellent computer access.

References