41. B. M. Wilamowski and H. Zhao, "Sensitivity and yield simulations for threshold voltage of short channel MOSFETs" - accepted for publication in <u>Electronic and Telecomunication Letters</u>

Sensitivity and yield simulation for threshold voltage of short channel MOSFETs

B.M.Wilamowski[⊕] Hong-lin Zhao[●]

The worst case analysis and the Monte Carlo method has been used to study the sensitivity of threshold voltage V_T to the process parameter variables, such as: thickness of silicon dioxide t_{ox}, charge density of surface state Q_{SS}, substrate impurity density N_A, channel length L etc. The worst case analysis showed advantage of simplicity in mathematical treatment and programming. With both methods similar results can be obtained. When L enters short channel region, the sensitivity of V_T to L increases rapidly. In Monte Carlo method, the fluctuation of V_T has been simulated and shown to be normal function. The sensitivity was corrected to use standard deviation of the distribution rather than its width. Based on the distribution of V_T, the production yield of it for different channel length L has also been estimated. The results show that with the same processes condition, the shorter channel length results in smaller yield.

1. Introduction

As the MOSFETs are now scaled down to microns or even submicron dimensions, some small geometry effects, especially short channel effects, are emerged. In small geometry region, the threshold voltage V_T of MOS transistor is not only the function of the processes parameters t_{OX}, Q_{SS} and N_A, but also it is related to channel length L, channel width w and junction depth r_j.

The sensitivity of V_T to geometry and processes input parameters has been studied with different ways [1, 2]. Since the input parameters are fluctuating in the device manufacturing practice, it is difficult to simulate the sensitivity of output parameters with a precise analytical method. But worst analysis assumes that the variation of input variables are always at certain value and each change, whether it is positive or negative, always effects output parameters. Since the simplicity and easy to realize in computer

[⊕] Institute of Electronic Technology, Technical University of Gdansk

[●] Electronic Engineering Department, Tianjin University, P.R.China

sensitivity of V_T to L, S(L), is going up and to others down. The results of Fig.1 indicate that the short channel effect can be expressed properly.

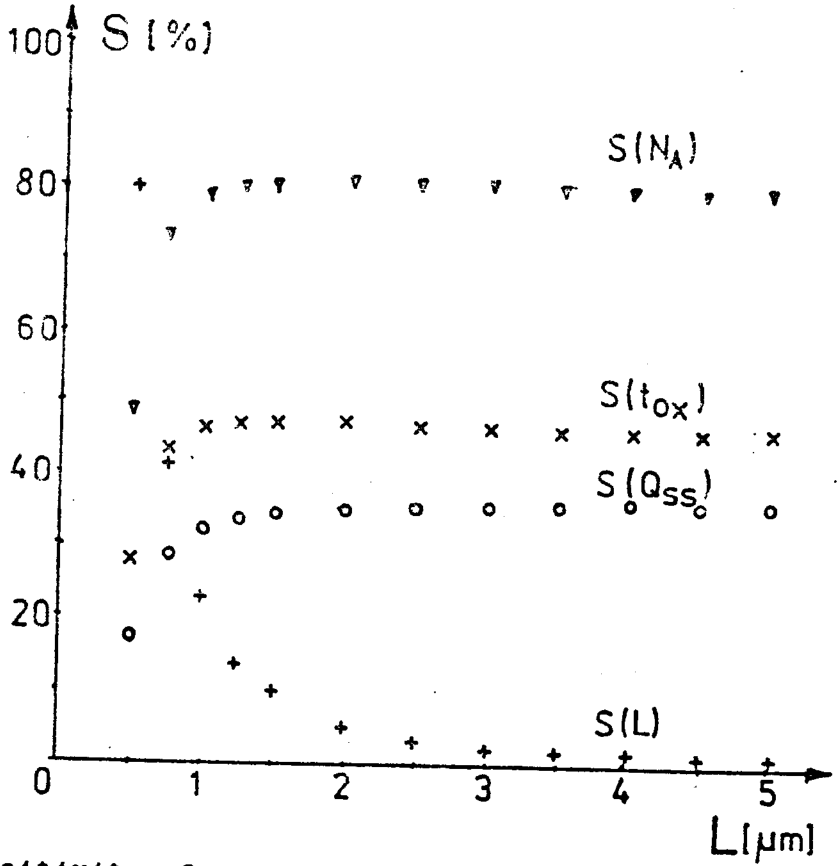

Fig.1. Sensitivity S as a function of channel length for worst case analysis

3. Monte Carlo analysis for sensitivity and yield of V_T

(A) Monte Carlo Procedure

According to Equation (1), the threshold voltage V_T is the function of six variables. In the real production condition, the values of these variables certainly can not be controlled unchangeably. They are always varying randomly and a normal distribution can be observed. As a statistical method, Monte Carlo approach is a powerful tool to simulate this process.

The influence of each variable fluctuation upon the fluctuation of V_T is different or V_T is more sensitive to some of the variables. The sensitivity of V_T to each variable is now defined as

$$V_T = V_{FB} + 2\phi_F + \frac{qH_A W_c}{C_{ox}} \{ 1 - [(1 + \frac{2W_c}{r_j}) - 1]^{1/2} * [\frac{r_j}{L} + \frac{2W_c r_j}{WL}] + \frac{2W_c}{W} \} \quad (1)$$

Where V_{FB} is flatband voltage and it is equal to the difference of Ψ_{MS} and Q_{SS}/C_{ox}. Ψ_{MS} is the work function between metal and silicon. In aluminum-silicon system, for P-type and with $N_A=5x10^{16}$ cm^{-3}, $\Psi_{MS}=-0.95$ V. $C_{ox} = \varepsilon_0 \varepsilon_{ox}/t_{ox}$ is gate capacitance per unit area. Ψ_F is the Fermi potential of substrate. W_c is channel depletion depth. From Equ.(1) it is clear that the output parameter V_T is the function of six variables: t_{ox}, Q_{SS}, N_A, L, W and r_j.

Suppose that these six variables are: $x_1, x_2, \ldots, x_6$ and $V_T(x_1, x_2, \ldots, x_6)$.

If the six variables are controlled precisely as designed values, the output V_T is

$$V_{T0} = V_T(x_1, x_2, \ldots, x_6) \quad (2)$$

In practice, various factors cause each variable to change. Suppose that the variation of x_1 is Δx_1 and

$$V_{T1} = V_T(x_1+\Delta x_1, x_2, \ldots, x_6) \quad (3)$$

$$V_{T2} = V_T(x_1-\Delta x_1, x_2, \ldots, x_6) \quad (4)$$

These two equations indicate that variation in x_1 leads to the variation of V_T:

$$\Delta V_{T1}(x_1) = V_{T1}-V_{T0} \quad (5)$$

$$\Delta V_{T2}(x_1) = V_{T2}-V_{T0} \quad (6)$$

Although for each device, the variation of x_1 has only one possibility, either $+\Delta x_1$ or $-\Delta x_1$. For the worst case analysis, one can define:

$$[\Delta V_T(x_1)]wst = 0.5 * \{[\Delta V_{T1}(x_1)]^2 + [\Delta V_{T2}(x_1)]^2\}/2\}^{1/2} \quad (7)$$

Similarly, for x_2 through x_6, it should be written

$$[\Delta V_T(x_2)]wst = 0.5 * \{[\Delta V_{T1}(x_2)]^2 + [\Delta V_{T2}(x_2)]^2\}/2\}^{1/2} \quad (8)$$

....

$$[\Delta V_T(x_6)]wst = 0.5 * \{[\Delta V_{T1}(x_6)]^2 + [\Delta V_{T2}(x_6)]^2\}/2\}^{1/2} \quad (9)$$

For a practical device, each ΔV_T consists of six Δx. Considering also the worst situation, it can be written

$$[\Delta V_T(x_1, x_2, \ldots, x_6)]wst = \{[\Delta V_T(x_1)]wst + [\Delta V_T(x_2)]wst + \ldots + [\Delta V_T(x_6)]wst\}^{1/2} \quad (10)$$

The sensitivity of V_T to each variable is defined as:

$$S(x_n) = \frac{[\Delta V_T(x_n)]wst}{[\Delta V_T(x_1, \ldots, x_6)]wst} \quad (11)$$

$$n = 1, 2, \ldots, 6$$

The above equations are the basis for sensitivity simulation for V_T using the worst case approach.

(B) Results

As a processes simulation it is necessary to consider the limitation of equipment condition. Since the material and equipment conditions for various manufacturers are quite different, it is difficult to find typical and universal processes conditions. In this work, the mean values and their deviations, referring to practice, are adopted as follows: $t_{ox}(A^0)$: 500, 30; $N_A(cm^{-3})$: $6x10^{16}$, $1x10^{16}$; $Q_{SS}(cm^{-3})$: $2.5x10^{11}$, $4x10^{10}$; $r_j(\mu m)$: 0.5, 0.05; $W(\mu m)$: 10, 0.3; $L(\mu m)$: 0.5-5, 0.3.

The results are shown in Fig.1 There are four curves in this picture. They are the sensitivity of V_T to t_{ox}, Q_{SS}, N_A and L. The sensitivity of V_T to W and r_j are small, they are not drawn in the picture. Fig.1 shows that when L is getting small, the

$$S(X_P) = \frac{\text{SD of output distribution with one varying input}}{\text{SD of output distribution with all varying inputs}} \qquad (12)$$

$$SD = \sqrt{\frac{\Sigma V_T^2 - \frac{(\Sigma V_T)^2}{N}}{N}} \qquad (18)$$

where SD is standard deviation and N is number of samples.

The computing of sensitivity is as follows. First six sets of uniform random numbers are generated for 1000 samples. Then, according to the mean values and deviations of each variable, the standard normal distribution random number is transformed into the corresponding normal distribution. Using Equ.[1], the V_T is calculated. After obtaining the distribution of one thousand V_Ts, the sensitivity of V_T can readily be calculated from equations (12) and (13).

As the Monte Carlo processes simulation can give out the parameters distribution, it enables to simulate the yield. Generally, the mean value of the output parameters is equal to the designed value, all the device parameter values are distributed on both the sides of the mean value symmetrically. After the mean value of V_T has been obtained, the yield can be calculated by selecting the devices with the value of V_T within the range of $V_T \pm \Delta V_T$.

(B) Results

Fig.2 shows the distribution of the output one thousand V_Ts with all the six input varying parameters and the channel length L equalling 0.5 um. The mean value of it is 0.61961 V and the deviation is 0.20394 V. On the right side of the picture, there are some individual points distributed far away from the mean value. On the whole they do not influence the distribution dispersion very much, but if width method is used to define the sensitivity, they increase the distribution width remarkably.

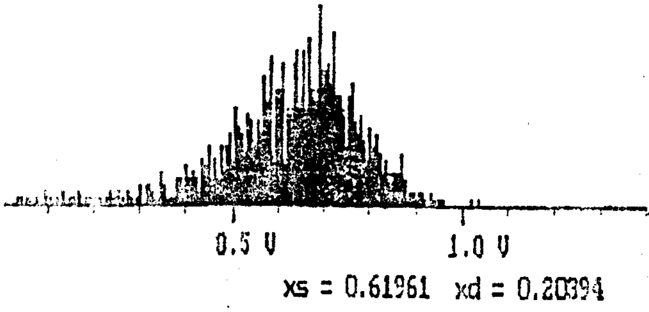

Fig.2 All 6 variables
changing. L = 0.5µm

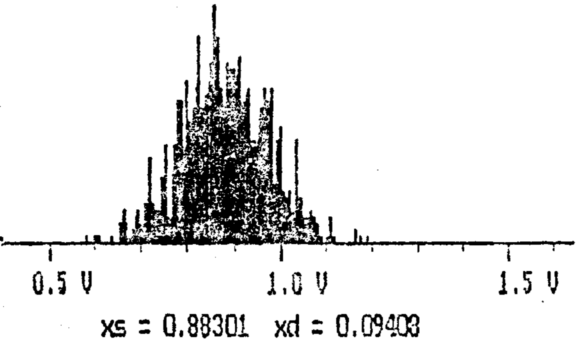

Fig.3 All 6 variables
changing. L = 2.0 µm

Fig.3 is another distribution of V_T with all the six parameters varying but L is changed 2 µm. Its mean value is 0.88301 V and the deviation is 0.09408 V. To compare Fig. 2 with Fig.3, it is clear that merely because of the reduction of L. the deviation of V_T distribution is increasing.

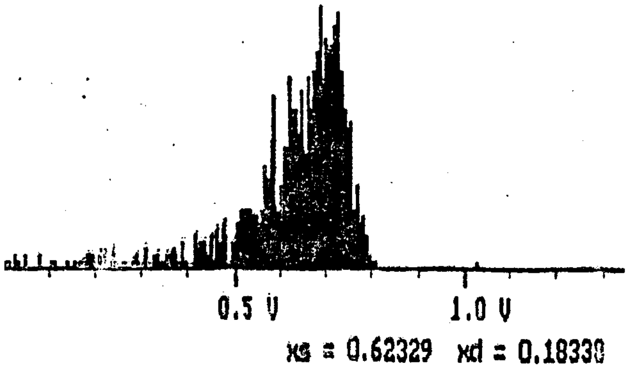

Fig.4 only L changing
L = 0.5 µm

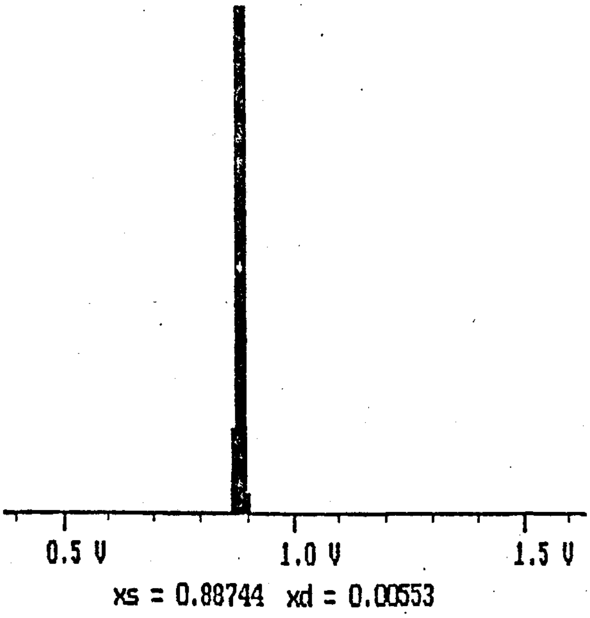

Fig.5 only L changing
L = 2.0 µm

Fig.4 and Fig.5 are other distributions of the output V_T with only L varying and their mean values and the deviations are shown in the pictures. From Fig.4 to Fig.5 all the input parameters remain unchanged except the channel length L from 0.5 µm to 2.0 µm. But one can see the dispersion of Fig.4 is much higher than that of Fig.5. These two pictures are not symmetric enough, especially the Fig.4. For a statistical method one thousand sample are sometimes not large enough to obtain accurate results.

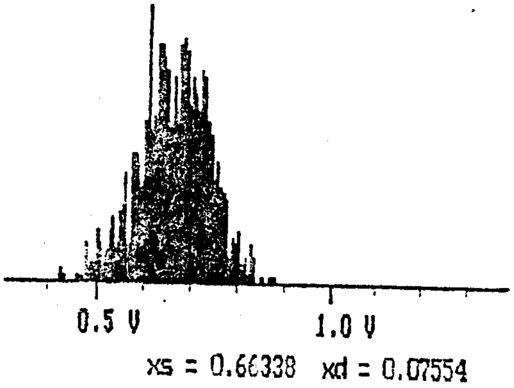

Fig.6 only H_A changing
L = 0.5 µm

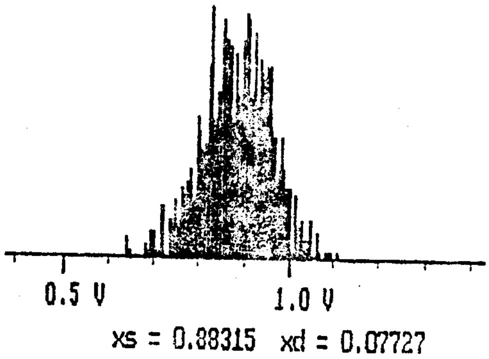

Fig.7 only N_A changing
L = 2.0 µm

Fig.6 and Fig.7 are distributions of output V_T with only N_A varying. Their mean values and deviations have also been shown in the pictures. From Fig.6 to Fig.7 all the input parameters remain unchanged besides the mean value of L varying from 0.5 µm to 2.0 µm. On the contrary to Fig.4 and Fig.5, the dispersion of Fig.7 is now smaller than that of Fig.6. It indicates that the sensitivity of V_T to L increases, whereas to N_A decreases. Similar results are also with respect to t_{ox}, Q_{SS} etc.

From the dispersion of V_T distributions, the sensitivity of V_T to the input parameters can also be calculated on the base of equations (14) and (15). The results are similar to Fig.(1) calculated by worst case analysis. Table.1 shows part of their data.

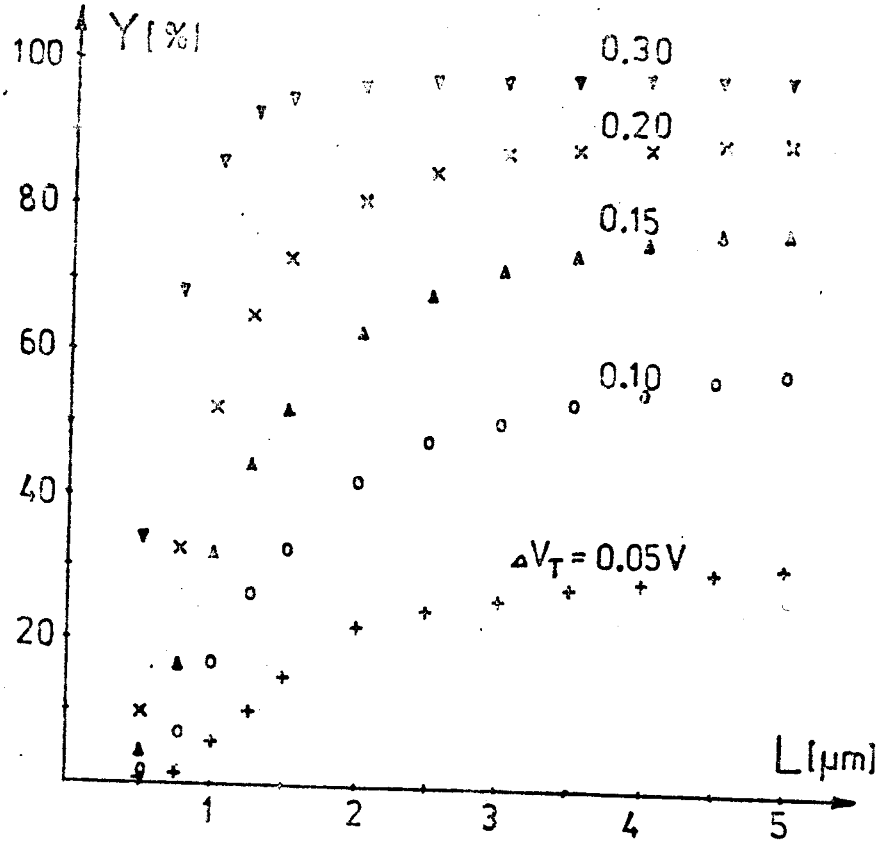

Fig.8 Yield Y as a function of channel length L

Table 1. Sensitivity of V_T

L (μm)		0.5	1.0	2.0	3.0
$S(t_{ox})$	WST	0.1728	0.3201	0.3474	0.3539
	MC	0.1291	0.3194	0.3484	0.3551
$S(Q_{SS})$	WST	0.2892	0.4608	0.4673	0.4674
	MC	0.2162	0.4599	0.4687	0.4675
$S(N_A^-)$	WST	0.4909	0.7939	0.8109	0.8106
	MC	0.3704	0.7997	0.8213	0.8212
S(L)	WST	0.8033	0.2337	0.0565	0.0249
	MC	0.8991	0.2503	0.0587	0.0266
S(W) and $S(r_1)$	WST	≈0	≈0	≈0	≈0
	MC	≈0	≈0	≈0	≈0

In table 1, the differences between the results from both the worst case analysis and the Monte Carlo method become appreciable when channel length becomes very small.

In the worst case analysis, the change of each input including ΔL is fixed. In Monte Carlo method, ΔL is merely the standard deviation. A part of the ΔL is larger than that of being used in the worst cast analysis. The influence of this part on the dispersion of V_T becomes serious with L getting small.

Fig.8 presents the yields of V_T against V_T limitations. It is clear from this picture that, in addition to the selected limitation, even the equipment and the processes conditions are maintained totally the same, the yields will be reduced merely due to shortening of the channel length L. In the short channel region, this effect is very serious.

4. Results of analysis and comparison

(A) The results of the output parameter sensitivity, either from worst case analysis or Monte Carlo method, show the same objective law. If the channel length of MOSFETs is not short, for instance more than 3 μm, the output V_T is actually independent of L. But, if L becomes as small as 2 μm, V_T depends on L obviously. The short channel effect is so strong that, when it occurs, the sensitivity of V_T to other input parameters drops sharply to compensate the increasing sensitivity of V_T to L.

In this work, the deviation of the channel length and the channel width are fixed while the mean value of them is varying. It is certain that when L and W become smaller, the ratio of $\Delta L/L$ and $\Delta W/W$ get larger. This is also one of the reasons to cause the S(L) and S(W) to rise. As a processes simulation, L and W are practically fixed. From the mathematical view point, if we use the deviation of L and W so as to maintain the ratio of $\Delta L/L$ and $\Delta W/W$ constant, the short channel effect is still reflected in the sensitivity but is less serious than that of Fig.1 and Fig.7.

(B) For the output parameter sensitivity simulation, this work provides two methods, the worst case analysis and the Monte Carlo method. Both methods gave the same results. Either from the mathematical procedure and program preparation, the worst case method is much faster and simpler than that of the Monte Carlo method. For the pure sensitivity simulation, the worst case analysis is good enough.

(C) The superiority of the Monte Carlo processes simulation is that it can simulate the statistical process and supply the distribution of all device parameters. With this merit of Monte Carlo method, the simulation of the product yield becomes realistic.

Fig.8 shows that the yield of V_T basically depends on ΔV_T when the channel length is larger than 3 μm. As the channel length L becomes small, especially less than 2 μm, with the same limitation of ΔV_T the yield reduces very fast. It is noted that when channel length L is in the submicron region, to acquire the high yield of MOSFETs is difficult.

Conclusion.

The sensitivity of the threshold voltage of short channel MOSFETs has been studied by the worst case analysis and the Monte Carlo method. Both the methods provided the same results. when the channel length falls down to 3μm, the short channel effect comes into being.

A set of V_T generated by the Monte Carlo method shows a normal distribution and the short channel effect can also be observed. The yields of the output parameter V_T has also been studied. The result of it has shown that in the short channel region, the yield reduced fast if L is shortened. For submicron channel length MOSFETs, it is hard to obtain a high yield.

References.

[1] Yokoyama, A.Yaki and Horiguchi, IEEE, Trans. ED-27, p.1509-14 (1980).

[2] A.R.Alvarez and L.A.Akers, Electronics Letters, Vol.18, No.1, p.42-43 (1982)

[3] C.H.Stepper and P.B.Hwang, Semiconductor silicon, pp.469-473 (1977)

[4] D.G.Ong, "Modern MOS Technology - process, device and design", Chap.6 (1984)

Моделирование чувствительности и прироста пороговых напряжений транзисторов МОСФЕТ с коротким каналом

В статье проведен анализ чувствительности порогового напряжения транзисторов МОС с использованием меода Монте-Карло и метода наиболее неблагоприятного случая. В качестве параметров, зависящих от технологического процеса, приняты: толщина окиси вентиля, поверхностная плотность примесей основания, а также длина и ширина вентия. Результаты подтверждают, что при одинаковых условиях процесса, уменьшение длины канала приводит к уменьшеню прироста продукции.