The Minimization of Short Channel Effects in a 225nm Si MOSFET

Dimitris Antoniadis, Marios Constantinou, Joshua Laney, and Pelayo Ramírez

Abstract - **MOSFET performance parameters are introduced and used in theoretical analysis of short channel effects. Possible optimizations of these effects are then introduced. The theory is then applied in TCAD simulations of a short channel Silicon MOSFET with a gate length of 225nm and performance is compared to that of a device with a 1000nm gate length. Optimizations yield improvements in all performance parameters when compared to the unoptimized short channel device. A DIBL of 6.84 mV/V at a threshold voltage of 0.48V, and a subthreshold swing of 62.77mV/Dec. Improvement is also seen when compared to the long channel in all areas except for output conductance, which remains largely unaffected by the optimizations applied.**

I. INTRODUCTION

Due to the abundance of material, low cost, and relative ease of manufacturing, the silicon MOSFET has found its way into the heart of an overwhelming majority of integrated circuits. As these integrated circuits increase in complexity and capability, the demand for a greater number of transistors to fit in the same area has also increased. This has led to the size of the transistor, and more specifically the gate length of the transistor, to downscale exponentially. As the gate length approaches the size of potential barriers inherent in the MOSFET however, a variety of non-ideal physical effects, called short channel effects (SCE), begin to occur. The minimization of these effects require an increasing number of additions, alterations, and optimizations to the traditional MOSFET structure. Through simulation in Sentaurus TCAD, these SCEs, as well as various minimization techniques, are explored on a silicon MOSFET of gate length, $L = 225\ nm$.

The design follows an iterative approach, starting with the simulation of an ideal, long-channel MOSFET of gate length, $L = 1\mu m$. The gate length is then reduced to $L = 225\ nm$ and the performance analyzed. Next, a variety of SCE minimizing optimizations are simulated, and their impact on device performance examined. Finally, a combined optimization of all these techniques is presented as an optimized, usable design for MOSFETS of $L = 225\ nm$.

The Authors are MSc Students with the Deperment of Electrical and Electronic Engineering, Imperal College London, London, UK, SW7 2AZ (e-mails: dimitris.antoniadis20@imperial.ac.uk, marios.constantinou20@imperial.ac.uk, josh.laney20@imperial.ac.uk pelayo.ramirez20@imperial.ac.uk)

II. PERFORMANCE PARAMETERS

The behavior of a MOSFET is dependent on the proper combination of biasing between its four terminals. This biasing can put the MOSFET in a number of different modes: weak inversion, linear, and saturation.

$$Weak\ Inversion:\ V_{GS} < V_{th} \quad (1)$$

$$Linear:\ V_{DS} < V_{GS} - V_{th} \quad (2)$$

$$Saturation:\ V_{DS} > V_{GS} - V_{th} \quad (3)$$

The characteristics of each of these modes each is dependent on an interconnected combination of physical attributes. The complexity caused by this can make it difficult to compare the performance of one device to another. In order to aid the easy comparison of MOSFETs, a series of performance parameters are developed [1].

A. Transconductance

The transconductance of a MOSFET, g_m, relates a change in the voltage between the gate and source, V_{GS}, to a change in current from the drain to the source, I_{DS} while in the linear region.

$$g_m = \frac{\delta I_{DS}}{\delta V_{GS}}\ (\mu S)\ \ when\ V_{DS} = constant \quad (4)$$

Transconductance can be considered to be the gain of the transistor, and it is desirable to have as large a value as possible. While the transconductance is ideally constant throughout the linear region, no idealities such as the weak inversion region at low V_{GS} and increased contact resistance at high V_{GS} cause variations. As such, for the purposes of performance comparison, the maximum value of g_m^{max} is used.

B. Threshold Voltage

In an ideal MOSFET the threshold voltage, V_{th}, is described as the minimum value of V_{GS} where an inversion channel is formed in the substrate and charge can flow between the source and the drain. Below this value no current flows, and above this value current increases linearly with V_{GS} in the linear region and quadratically in the saturation region, with the value of V_{th} remaining constant throughout the whole range of V_{DS}.

In reality however, the depletion region created at the drain-substrate junction has an effect on the charge present in the inversion channel and thus the threshold voltage. This causes the value of threshold voltage to change as a function of V_{DS} in what is referred to as threshold voltage roll-off. A

more formal equation for threshold voltage can be expressed as

$$V_{th} = 2V_{FB} + 2\phi_F - \frac{Q_{depl}(V_{DS})}{C_{OX}} (V) \quad (5)$$

The threshold voltage in the linear region, V_{th}^{lin}, can be extracted by finding the voltage at which a line tangent to the point of g_m^{max} on the V_{GS} curve crosses 0 I_{DS}. This extrapolation is necessary because sub-threshold currents in the weak inversion region form an exponential relationship with V_{GS}. The tangent is taken from the point of g_m^{max} as this is the most ideal and linear point on the transfer characteristic.

In similar fashion, the saturation threshold voltage, V_{th}^{sat}, is extracted by finding the 0 current point of a line tangent to the point of highest slope on the characteristic of $\sqrt{I_{DS}}$ vs V_{GS}.

C. Drain Induced Barrier Lowering (DIBL)

DIBL is a quantification of the previously discussed threshold voltage rolloff, and becomes an important parameter as gate lengths shrink in the territory of short channel effects. DIBL is formulated as:

$$DIBL = \frac{|V_{th}^{lin} - V_{th}^{sat}|}{(V_{DS}^{sat} - V_{DS}^{lin})} (\frac{mV}{V}) \quad (6)$$

DIBL should be as low as possible, and an acceptable value of DIBL is considered to be less than $100\ mV/V$

D. Subthreshold Swing (SS)

Subthreshold swing characterizes the behavior of the weak-inversion, subthreshold region. It is measured in saturation mode as

$$SS = (\frac{dLog(I_{DS})}{dV_{GS}})^{-1} = \frac{dV_{GS}}{dLog(I_{DS})} (mV/dec) \quad (7)$$

with the value taken at the point in which the smallest change in V_{GS} causes the largest change in $Log(I_{DS})$. Smaller values of SS are more desirable as they correspond to smaller subthreshold regions, and thus more efficient on-off switching. It is important to note however, that in silicon based n/p MOSFETs the minimum value of SS is $60\ mV/dec$ due to the physical limits of diffusion within the materials.

E. Output Conductance

The output conductance is a measure of how much a change in V_{DS} effects a change in I_{DS} when the device is in saturation.

$$g_o = \frac{\delta I_{DS}}{\delta V_{DS}} (\mu S) \quad (8)$$

It can be taken as the maximum value of this derivative well into the saturation region. The output conductance should be as low as possible to create maximally high impedance nodes.

F. On-Off Current Ratio

The On-Off Current Ratio gives an indication of the leakage power consumption of the device, as well as the sensitivity required to determine if the device is on or off. As such, this ratio should be as large as possible. The point at which the transistor is considered off or on can be calculated from the supply voltage and threshold voltage in saturation mode:

$$V_{GS}^{on} = V_{DS}^{sat} + V_{th}^{sat} - |V_{th}^{sat}| (V) \quad (9)$$

$$V_{GS}^{off} = V_{th}^{sat} - |V_{th}^{sat}| (V) \quad (10)$$

$$I_{ON}/I_{OFF} = \frac{I_{DS}^{sat}(V_{GS}^{on})}{I_{DS}^{sat}(V_{GS}^{off})} \quad (11)$$

III. SHORT CHANNEL EFFECTS

When the length of the MOSFET is of the same order of magnitude as the depletion width of the drain and source, short channel effects (SCEs) come into play degrading its performance. These effects will be presented and explained in ascending order of severity [1].

A. Channel Length Modulation

To understand this short channel effect, the meaning of the pinch-off condition in the channel should be reviewed. This situation is reached when the potential energy difference between the gate and the channel is equal to the threshold voltage, in other words, $V_{DS} = V_{GS} - V_{TH}$. This implies that exactly at the drain end of the channel the condition for inversion has been lost. If the source-drain voltage is increased beyond the pinch-off value, the MOSFET will enter the saturation region, which will mean that the drain and source depletion widths will extend underneath the channel, depleting part of it and, thus, reducing the effective length of the channel from the original value,L, to L_{eff}.

$$I_{DS} = \frac{1}{2}\mu C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \quad (12)$$

It can be seen from equation 12 that I_{DS} is inversely proportional to L, so this reduction in the length will result in increased saturation currents, leading to a higher g_0. Furthermore, it will also result in a threshold voltage modulation, punch through effects and breakdown mechanisms. These three will be presented and discussed below.

B. Charge Sharing

In the long channel MOSFET, the charge in the channel could be assumed to be completely controlled by the gate, as the depletion widths of the drain and source are negligible in comparison to the channel's length. However, as it is scaled down, it can be noticed that the influence of the drain and source on the charge in this region is no longer negligible and its control will be shared. This SCE is called the charge sharing (CS) model and will give a more accurate description of the performance of the MOSFET.

Before moving on to the effects of this model, a recapitulation of the threshold voltage equations will be done in order to compare them with the results obtained from the CS model [1].

$$V_{TH,NO\,CS} = V_{FB} + 2\phi_F - \frac{Q_{DEPL,NO\,CS}}{C_{OX}} \quad (13)$$

$$Q_{DEPL,NO\,CS} = -\sqrt{4eN_A\varepsilon\phi_F} = \frac{-eN_A W_{MAX} Lw}{Lw} \quad (14)$$

Where V_{FB} is the flat band voltage which is obtained from the difference in workfunctions between the materials, the second term is the condition on the Fermi potential to create inversion and the last term contains the charge in the depletion region which is determined ideally by the bulk doping and the gate potential (equation 14). Besides, W_{MAX} stands for the depth of the depletion region and w for the width of the gate.

However, these equations do not yield realistic results for the short channel MOSFET. To improve the theoretical model, CS should be included. Under this condition, the depletion charges at both sides of the channel are no longer controlled by the gate, but they are now shared with the drain and source. As a first approximation, and without introducing much deviations from the real scenario, it can be assumed that the gate controls 50% of this charge and the drain and source control the other 50%. If L is the length of the channel and L_B is the length without the width of the depletion regions, then it is easily seen that the volume of the charge controlled by the gate will decrease as the length decreases from L to $(L + L_B)/2$. Therefore, the charge controlled by the channel is now reduced a factor γ.

$$Q_{DEPL,\,CS} = \gamma \bullet Q_{DEPL,\,NO\,CS} \tag{15}$$

$$\gamma = \frac{L + L_B}{2L} \tag{16}$$

Taking this reduction into consideration, the expression for the threshold voltage in the CS model is straightforwardly obtained from (13) and (15).

$$V_{TH,\,CS} = V_{FB} + 2\phi_F - \gamma \frac{Q_{DEPL,\,NO\,CS}}{C_{OX}} < V_{TH,\,NO\,CS} \tag{17}$$

Where the flat band voltage and the condition for inversion remain the same under this 50%-50% CS model. Therefore, from equation (17), it is observed that the introduction of the charge sharing leads to a V_{TH} roll-off in short channel MOSFETs. A simple derivation [1] leads to a very useful expression for γ in terms of the length of the channel, L, the depth of the source and drain implants, t_j, and the maximum depth of the inverted region, W_{MAX}.

$$\gamma = 1 - \frac{t_j}{L}\left(\sqrt{1 + \frac{2W_{MAX}}{t_j}} - 1\right) \tag{18}$$

If we want to recover the ideal values, in other words, minimise the effects of charge sharing, $\gamma \rightarrow 1$ is needed. This limit gives two conditions on the device parameters that must be satisfied in order to limit the shift in the threshold voltage.

$$W_{MAX} << t_j \tag{19}$$

$$t_j << L \tag{20}$$

The first condition can be satisfied by increasing the doping, N_A, in the substrate because depletion regions extend less in the highly doped regions. The second condition implies that shallow junctions must be created. Combining together equations 2,6 and 7, a final expression for the threshold voltage shift, when charge sharing is introduced, can be obtained.

$$\Delta V_{TH} = \frac{-eN_A W_{MAX}}{C_{OX}} \bullet \frac{t_j}{L}\left(\sqrt{1 + \frac{2W_{MAX}}{t_j}} - 1\right) \tag{21}$$

This expression offers a way to combat threshold voltage roll-off by modifying certain parameters of the device, for instance, higher oxide capacitance. Another important aspect to remark is that the off current, $I_{off} \propto 10^{-V_{TH}/V_T}$, so it increases due to the shift in V_{TH}, degrading the performance of the device.

C. Drain Induced Barrier Lowering

In a long channel MOSFET it can be assumed that the current is controlled by the gate. However, when the dimensions are scaled down, the drain voltage drop at the source end of the channel is no longer zero. Therefore, the electric field from the drain impacts the source-channel barrier, lowering it and introducing more electrons into the channel by diffusion. Thus, the current is no longer controlled exclusively by the channel, but also the drain influences it. This SCE is measured by the DIBL parameter. High values of DIBL mean poor gate control, which leads to increased currents in saturation region (high g_0).

D. Additional High Electric Field Effects

When the length of the device is reduced, the electric field due to the source-drain voltage increases as E=V/L. This high electric field generated significantly affects the performance of the MOSFET due to the following effects.

i. Mobility Variations: as a consequence of the electric field arisen from the V_{DS} and from the V_{GS} an effective electric field appears attracting electrons into the oxide, damaging it and increasing surface roughness scattering and, thus, reducing the mobility in the channel.

ii. Velocity Saturation: as the electric field increases, electrons gain more energy from it and increase their velocity. However, this process is limited by the saturation velocity. The more energy an electron has, the higher the probability of it losing it through a phonon emission because of the electron-lattice interaction. Therefore, there is a maximum energy an electron can have without losing it (almost instantly) through phonon emission, and this maximum energy sets the maximum velocity, v_{sat}. In long channel MOSFETs, this velocity saturation normally occurs after pinch-off condition, however, in short channel MOSFETs this effect can happen in the linear region and it will lead to a reduced I_{DS} in the saturation region.

iii. Avalanche Breakdown: near the drain, the high energetic electrons previously mentioned can also lose their energy by impact ionisation. During this process, they collide with an atom and transfer part of their energy to an electron in the atom, if this energy transfer is sufficiently enough, the electron will be released from the atom. Now these two electrons will collide with another atom and this mechanism will increase exponentially.

iv. Hot Electron Effect: electrons with very high kinetic energy are injected into the oxide, damaging it and creating trap states in it. These states allow other electrons to travel inside it resulting in a leakage current through the gate.

v. Punch Through: if the V_{DS} is high enough or the length is small enough, the depletion region extending from the drain and the source will reach each other and they will combine into one. This combination is called space charge limiting currents and leads to a zero potential barrier between the source and the channel, giving rise to electrons flowing through it.

IV. MINIMISING SHORT CHANNEL EFFECTS

When downscaling the MOSFET [1], two possible paths can be followed: constant field scaling or constant voltage scaling. In order to be able to compare the results with the original long channel MOSFET, the latter was chosen, keeping V_{DD} constant. It will be shown that, even though this increases the electric field by a factor $\alpha = 4.44$, and, thus, the high electric field effects, the results obtained are of high quality.

A. Oxide Thickness and High-K dielectrics

The shorter the channel, the more the drain control over the channel increases. Therefore, it is of the utmost importance to increase the gate control over the channel in order to avoid DIBL as much as possible. This gate control is mainly determined by the gate capacitance per unit area, C_{OX}.

$$C_{OX} = \frac{\varepsilon_O \varepsilon}{t_{OX}} \tag{22}$$

Where ε is the relative permittivity of the oxide and t_{OX} its thickness. As a first solution, it would seem adequate to reduce the oxide thickness to increase gate control. However, this will lead to tunnelling leakage currents through the gate degrading the performance of the MOSFET. This effect sets a minimum usable $t_{OX} = 1.2\,nm$. As a consequence, a more suitable solution is to increase the dielectric constant of the material. This approach leads to the use of high-K dielectrics instead with a metal on top. For this project, HfO_2 with TiN were used. TiN is a midgap metal for Silicon, in other words, this metal places the Fermi level at the same level that the intrinsic level of Silicon. This helps reduce the gate to drain/source extension tunnelling [2] .Furthermore, the use of high-K dielectrics significantly reduces gate leakage currents. Nevertheless, high-K dielectrics have some downsides that have to be faced. These materials have high interface trap density that degrades the mobility in the channel and the on current, I_{ON}. In order to avoid trap states in the interface, a very thin layer of SiO_2 is usually grown between the channel and the dielectric.

B. Reducing the junction depth

The junction depth, t_j of the high doping density (HDD) regions can be reduced by modifying the parameters of the implantation processes. This reduction implies a decreased amount of charge is shared between the drain/source and the gate, increasing gate control. However, the resistance of the ohmic contact region is inversely proportional to t_j.

$$R_{Si} = \frac{\rho \bullet L}{w \bullet t_j} \tag{23}$$

Where ρ is the resistivity of the material. Therefore, the DIBL and V_{TH} roll-off are being reduced at the expense of reducing the transconductance. A feasible solution to reduce DIBL while improving the g_m is called 'Elevated Drain and Source'. The idea is to reduce the junction depth without reducing the cross sectional area through which the current flows and, thus, not increasing R. To implement this solution, an implantation of doped Silicon will be done over the actual drain and source of the device.

C. Substrate Doping, Pockets and Halos

A major reason for SCEs is the depletion region of the source and the drain, extending in the channel region. In order to deal with this problem, the concentration of the substrate can be increased. By doing this, threshold voltage increases and DIBL decreases. As a result, undesired high VDS influence on device operation is alleviated.

Additionally, the increase of the substrate concentration tackles the problem of breakdown and punch through effects as they have been described in section III. However, this improvement comes at a cost. By increasing substrate doping, parasitic capacitances will be created between the drain/source and the body. Furthermore, the higher the doping, the more scattering will occur and the greater the electric field on pn junctions. As a result, mobility decreases and ON current decreases accordingly. This problem can be resolved by introducing a graded substrate. In this way, the doping is relatively high below the gate and it gradually decreases. Since the doping concentration below the drain/source region is reduced, the parasitic capacitances also decline.

Another problem is the high electric field and the extension of the depletion regions in the channel. This problem can be solved by introducing Low Doping Density (LDD) Pockets between the drain/source below the gate. The depletion region extends towards the area with the lower doping concentration. Therefore, the depletion area will extend in the LDD region, reducing the electric field and increasing the DIBL. More sophisticated methods can be introduced such as placing high doping substrate pockets below the LDD drain/source pockets. These structures are called halos.

Finally, a highly doped substrate region can be placed below the channel. In this way, the extension of the depletion region is further controlled. Additionally, this structure reduces the infulce of the bulk to the channel of the device.

V. RESULTS

A. Long Channel and Unoptimized Short Channel MOSFETs

To serve as a baseline for comparison, a long channel MOSFET of gate length $L = 1\mu m$ is simulated. The gate length is then reduced to $L = 225nm$ without altering any other aspects of the device. The impact of SCEs is clear in Table I when comparing the performance of the two devices. Charge sharing in the channel causes the threshold voltage in the linear region to be reduced by a factor of almost 2. DIBL then further shifts the threshold voltage in saturation, turning the MOSFET into a depletion mode device. The increase in the subthreshold swing and output conductance can also be attributed to the impact of the loss of gate control from the DIBL. Additionally the shift in threshold voltage creates a larger off current, reducing the Ion/Ioff ratio. The increase in transconductance comes as an added benefit from the downscaling. As shown in equation 23, resistance is directly related to channel length, and thus transconductance is inversely proportional.

TABLE I
COMPARISON BETWEEN LONG CHANNEL AND UNOPTIMIZED SHORT CHANNEL MOSFETS.

Gate Length	1000nm	225nm
gm (uS)	0.955	3.8019
Vth Linear (V)	0.66	0.3571
Vth Sat (V)	0.6072	-0.054
DIBL (mV/V)	35.4734	275.9454
SS (mV/dec)	92.4844	149.1816
go (uS)	2.3124	27.974
Ion/Ioff	$8.1383*10^8$	733.9102

TABLE II
CONSTANT VOLTAGE SCALING MOSFET PERFORMANCE PARAMETERS

Gate Length	225nm
gm (uV)	5.38
Vth Linear (V)	0.495
Vth Sat (V)	0.404
DIBL (mV/V)	60.96
SS (mV/dec)	84.40
go (uS)	1.22
Ion/Ioff	$2.92*10^7$

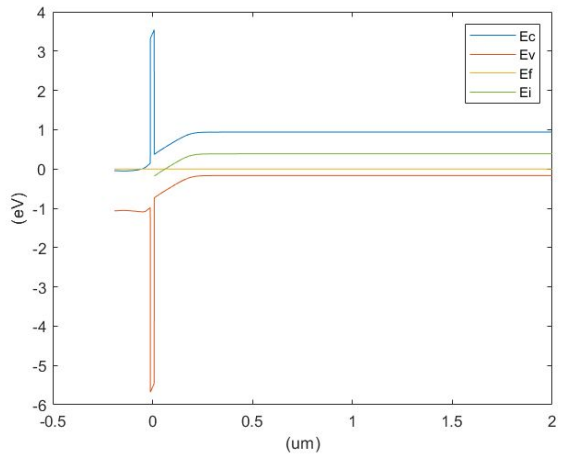

Fig. 1 Band Diagram for 225nm device, VDS = 0. The surface potential Vs is greater than $2\Phi_F$ and therefore a negative voltage Vth must be applied in order for the Vs to be equal to $2\Phi_F$

Fig. 1 shows the band Diagram of the scaled device. According to [1] the threshold voltage is equal to the externally applied voltage on the gate which leads to a surface potential equal to $2*\Phi_F$. In this case, even when $V_{DS} = 0V$, the drop off on the surface potential Vs is greater than $2*\Phi_F$ and this means that the threshold voltage is negative. More specifically, from the diagram Vs=0.5686 V and $\Phi_F = 0.1663$, therefore Vth=-0.236V.

B. Constant Scaling

Following the constant voltage scaling laws, a scaling factor of α=4.44 is applied to the previously presented unoptimized short channel device. This is achieved by increasing the substrate doping concentration N_A by a factor of α, and decreasing the oxide thickness t_{ox} by a factor of 1/α. Table II shows the performance parameters of the device with the constant scaling voltage method. Comparing the results with Table I for the 225nm unoptimized MOSFET, it is important to highlight that the transimpedance g_m and the linear threshold voltages V_{th}^{lin} increase by 40%, due to an increase in drain current I_{ds} and decrease of charge sharing respectively. The saturation threshold voltage V_{th}^{sat} increases around 848.83%, resulting in a decrease of the DIBL parameter by 77.91%. Additionally, due to the drain current rise, the sub-threshold parameter is reduced by 43.42%. Lastly, the output conductance g_o of the device is decreased by 56.43%, again due to drain current variations and the major improvement regarding the Ion/Ioff current ratio. The leakage current is improved from the unoptimized short channel MOSFET but it is still lower than the Long Channel, indicating one of the main effects of decreasing the gate length. Thus, the device is improved because both the DIBL and SS parameters are minimized, the control over the leakage current is improved and finally, the g_m and g_o are improved.

C. High-K Dielectric

The next optimization to further improve the S.C FET device is replacing the polySilicon gate with a metal one using TiN and replacing the SiO2 with HfO2 which has a higher dielectric constant. This technique is currently used in the industry, and one of the main benefits is that it provides better oxide capacitance, therefore reducing the leakage current in comparison to the conventional MOSFET [1]. Furthermore, the HfO2 is deposited on Silicon in comparison with the SiO2 which is required to be grown by thermal oxidation of Silicon. As a result, this improves the operation speed of the MOSFET and improves the cooling conditions due to lower leakages.

As seen in Table III, a parametric sweep of the tox was done in order to understand the impact of the oxide thickness on the performance parameters. Here, the on/off current ratio is three orders of magnitude larger. The transimpedance g_m is improved by 6 times, there is a 5% increase of the linear threshold voltage as well as a 25% rise in the saturation threshold voltage. Moreover, DIBL decreases by 81.95% and the SS parameter decreases by 26.26%. Finally, the output impedance is decreased by 55 times because the g_o parameter increases. Therefore, although some performance parameters showed a significant improvement, the output impedance is dramatically reduced. Even though the t_{ox} of 4nm provides a

TABLE III
PERFORMANCE OF VARYING HFO2 THICKNESS WITH TiN GATE

HfO2 t_{OX}	2.5nm	3nm	3.5nm	4nm
gm (uS)	32.66	31.09	29.44	27.74
Vth Linear (V)	0.5216	0.5268	0.5314	0.5354
Vth Sat (V)	0.5052	0.5090	0.5131	0.5177
DIBL (mV/V)	11.008	11.9709	12.2398	11.8986
SS (mV/dec)	62.2303	62.4407	62.6543	62.8708
go (uS)	68.168	66.696	65.160	63.447
Ion/Ioff	$5.06*10^{10}$	$5.56*10^{10}$	$6.09*10^{10}$	$6.6*10^{10}$

better on/off current ratio, 2.5nm is more desirable due to the higher g_m values and lower DIBL, which implies a better control over the gate in comparison to the constant voltage scale device. Finally, using the equation (22), one can see that by increasing the relative permittivity ε_r of the insulator material, the C_{ox} is increased with t_{ox} remaining constant, therefore the Ion is reduced, and SS is improved, however there is an impact on Vth.

D. Shallow Drain and Source

In short channel MOSFETs, the length of the channel is a similar order of magnitude compared to the depletion region widths of the drain/source depth. In this case, the depletion regions extend in the channel and affect the device by altering its threshold voltage[1].

More specifically, the depletion regions at the source and the drain are not explicitly controlled by these junctions but they are affected by the gate control voltage (V_{GS}). This problem is called charge sharing. By inspection, it can be found that the larger the junction depth of drain/source, the larger the charge sharing area[1].

According to equations 19, 20 and 21, a potential solution to this problem is to decrease the junction depth, since ΔV_{th} is directly proportional to t_j. Additionally, based on the equation 6, DIBL is directly proportional to ΔV_{th}, therefore, it is also proportional to t_j. By decreasing t_j, the charge sharing area is reduced. Therefore, the threshold voltage roll-off is mitigated.The reducedution in depth is achieved by lowering the energy parameter of implantation [3].

Table IV verifies the previous discussion. By simulating devices with various depths, it can be shown that the depth reduction leads to ΔV_{th} decrease and therefore to DIBL improvement. However, the device of 87nm should have at least 2 times better DIBL compared to the 205nm device. According to equation 21, a 2 times reduction of t_j, should

TABLE IV
PERFORMANCE OF VARYING DRAIN/SOURCE DEPTH

t_j	87nm	130nm	205nm
gm (uS)	28.72	29.81	32.6602
Vth Linear (V)	0.518	0.518	0.5211
Vth Sat (V)	0.505	0.504	0.5052
DIBL (mV/V)	8.65	9.245	10.621
SS (mV/dec)	62.32	62.307	62.23
go (uS)	160	114	88.1
Ion/Ioff	$6.05*10^{10}$	$5.29*10^{10}$	$5.07*10^{10}$

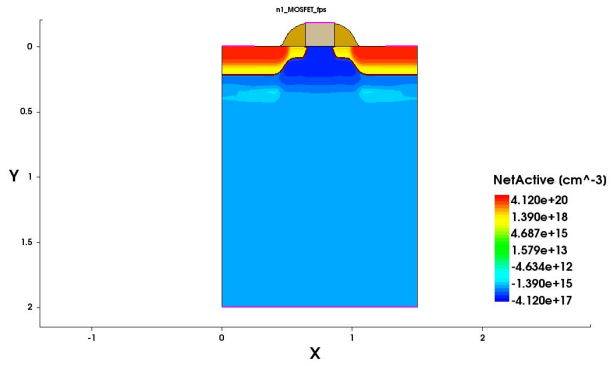

Fig 2. Device with Graded Substrate and LDDs. Under the yellow area are the LDDs pockets. Under the channel, high substrate doping exists and it gradually decreases.

lead to a 2 times reduction or 50% at ΔV_{th} and DIBL respectively. The results show a reduction of approximately 20%. This is because this improvement comes at a cost. By reducing t_j, current is forced to go through a smaller area. As a result, g_o increases and g_m decreases.

E. Substrate Doping Gradient and Halos

The implementation of LDDs, halos and substrate doping gradients improve the performance of the device, as they offer greater control on the channel and minimize undesired effects such as high electric field at pn junctions, and avalanche breakdown. LDDs are created by implantation using 45 degrees tilt angle, 90 degrees rotation angle and mult.rot=2 in order for the implantation device to point directly below the gate [3].

Comparing this device to the 205nm device of the previous subsection, there are no meaningful improvements (Table V). DIBL slightly improves, which means that ΔVth has slightly improved. The most important improvement is g_o, which means that the device gate has better control on the channel.

TABLE V
PERFORMANCE OF DEVICE WITH SUBSTRATE DOPING GRADIENT AND HALOS

gm (uS)	29.758
Vth Linear (V)	0.523
Vth Sat (V)	0.509
DIBL (mV/V)	9.284
SS (mV/dec)	62.63
go (uS)	71.14
Ion/Ioff	$4.94*10^{10}$

TABLE VI
PERFORMANCE PARAMETERS OF THE ELEVATED DRAIN AND SOURCE DEVICE

gm (uS)	32.582
Vth Linear (V)	0.521
Vth Sat (V)	0.505
DIBL (mV/V)	10.623
SS (mV/dec)	62.232
go (uS)	88.918
Ion/Ioff	$5.098*10^{10}$

F. Elevated Source and Drain

The main idea of the elevated source and drain is, as it was described previously, to elevate the ohmic contacts so that the cross sectional area the current sees increases while the depth of the junctions remains small. As a consequence, charge sharing can be limited by having small effective t_j, while the resistance reduces by increasing the denominator in equation (12). To elevate the drain and source of the device, n-doped Silicon is deposited using a negative mask [3] as the one used to build the gate (Fig 3). If these results are compared with those from Table III, for $t_{OX} = 2.5\ nm$, it can be seen that little variation has been introduced by the elevated drain and source. It would be expected to achieve a higher g_m with this process as it is directly related to the resistance of the material by the following equation:

$$g_m^{EXT} = \frac{g_m^{INT}}{g_m^{INT} \bullet R_{SI} + 1} \quad (24)$$

Where g_m^{EXT} is the value measured and g_m^{INT} stands for the internal value of the device, the value that it could be achieved. However, the value obtained is almost the same as the one obtained from Table III. This may be due to the geometry of the elevated drain and source. While its cross sectional area has been increased, some rough corners have been introduced that may be limiting the positive effects of

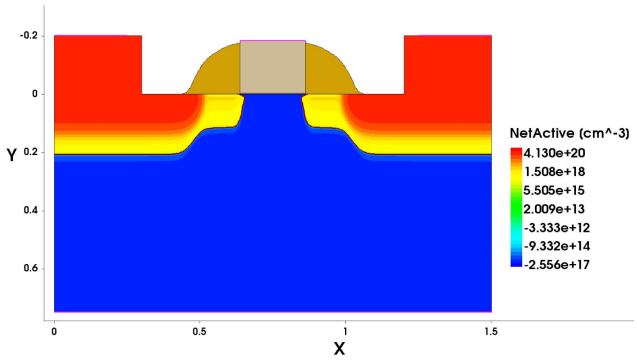

Fig 3. Elevated drain and source on a TiN/HfO_2 (2.5 nm thickness) MOSFET. Implemented by deposition of highly doped n-type Silicon over the original drain and source.

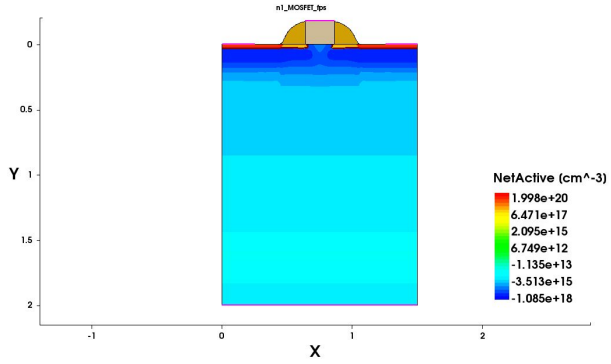

Fig 4. Final optimized FET. This device uses a TiN gate with HfO2 oxide, has self aligned LDD and HDD drain and source regions of reduced depth, and a doping gradient with high doped halos in the substrate.

this process. In real industry processes, this is implemented via epitaxy technology in vacuum which allows more accurate depositions on top of the device.

G. Final Optimized Device

In the previous subsections a thorough investigation is presented regarding the SCEs and the potential solutions. Based on the above discussion, and combining the knowledge and the results gained by this investigation, a final optimized device is created. This device is presented in Fig. 4.

On the top of the device, the gate is formed by a layer of 180nm TiN and a high-k dielectric HfO2 of 2.5nm. The depth of drain/source n-doping regions is optimized to achieve better performance. More specifically, the depth t_j is equal to 31.5nm. Between drain/source and the channel, LDD Pockets have been introduced. Additionally, highly substrate doping pockets have been placed below LDD n-pockets and a super-steep retrograde body doping. In order to further improve the performance, graded doping substrate technique is used. Fig. 5 depicts the $I_{DS} - V_{GS}$ plot for the linear region and Fig. 6 depicts $I_{DS} - V_{GS}$ for the saturation region respectively.

For the construction of the device, initially multiple implantations are performed to create the graded substrate, then the gate is created. For the LDDs, an implantation with tilt equal to 45 degrees, rotation equal to 90 degrees and

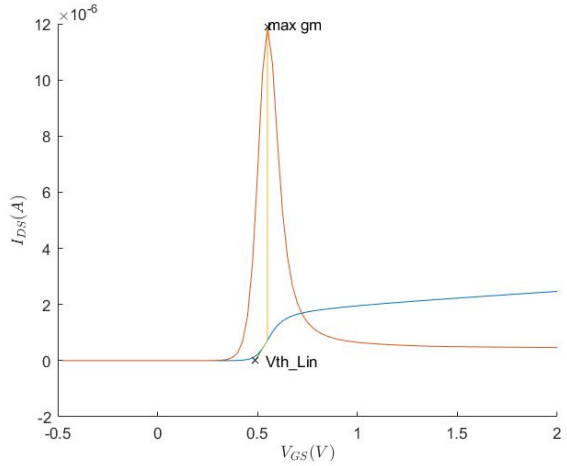

Fig. 5. I-VGS(blue line) in Linear Region for VDS=10mV and its Transconductance gm(orange line). The maximum transconductance point is where I(current) trace is the steepest. By extrapolating this point on the current trace, threshold voltage on the Linear region is obtained[1].

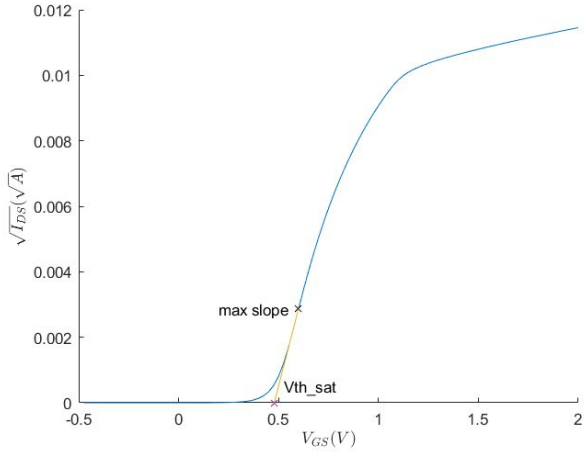

Fig 6. Square root of I-VGS(blue line) in Saturation Region(VDS=1.5V). By finding the maximum value of the slope and extrapolating it, the threshold voltage in the Saturation region can be obtained[1].

mult.rot equal to 2 is performed. In this way n-type light doping regions are created. Then, additional p-type implantations are performed. The only different parameter is the tilt angle, which was equal to 30 degrees aiming below n-type LDDs. In this way, p-pockets and retrograde p-type doping are added. Finally, High Doping Density(HDD) n-type is added to source/drain[3].

Table VII compares the extracted parameters of the device to the long channel and unoptimized short channel iterations. For the extraction of the parameters, V_{DS} in the linear region is set equal to 10mV, while for the saturation region it is set equal to 1.5V. The optimized device achieves around four times greater maximum gm compared to the scaled device. The gm parameter increases with scaling the device because of the smaller the channel length, and the reduced resistance of the channel. However, the gm parameter decreases by reducing the drain/source depth. Therefore, as an optimal depth, tj was chosen equal to 31.5nm. Additionally, by introducing a high-k dielectric the optimized device achieves

TABLE VII
PERFORMANCE PARAMETERS OF FINAL OPTIMIZED DEVICE COMPARED TO INITIAL DEVICES

Parameters	Long Channel	Unoptimized Short Channel	Optimized Device
gm (uS)	0.955	3.8019	11.89
Vth Linear (V)	0.66	0.3571	0.488
Vth Sat (V)	0.6072	-0.054	0.478
DIBL (mV/V)	35.4734	275.9454	6.840
SS (mv/dec)	92.4844	149.1816	62.77
go (uS)	2.3124	27.974	25.52
Ion/Ioff	$8.1383*10^{8}$	733.9102	$1.19*10^{10}$

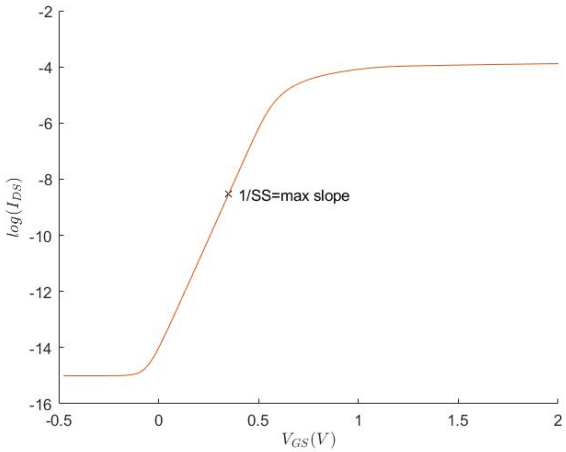

Fig. 7. Log(I)-VGS in Saturation Region. The x point is where the slope is maximum. The reciprocal of that point is equal to the Subthreshold Swing (SS) which is almost equal to 62mv/dec. The best possible SS a transistor can have is 60mv/dec[1]. The steeper the slope the better the device, since it can go faster from on state to off state.

almost perfect Subthreshold Swing equal to 62.77 mv/dec. The introduction of TiN/HfO2 provides superior control over the gate. The leakage current from the gate to the substrate decreases significantly and improves the ratio of Ion/Ioff. Another important optimization for improving the ratio Ion/Ioff is the graded substrate and the highly doped pocket areas introduced below the gate.

The LDDs pockets improve the output conductance by allowing the extension of the depletion region in the pockets. The graded substrate combined with the pockets provide a good trade-off on the output conductance, since the output conductance of the optimized device is almost equal to the scaled one. Additionally, by introducing pockets and a graded substrate, DIBL is almost eliminated. The optimized device achieves a DIBL of 6 mV/V. When scaling down a device, the

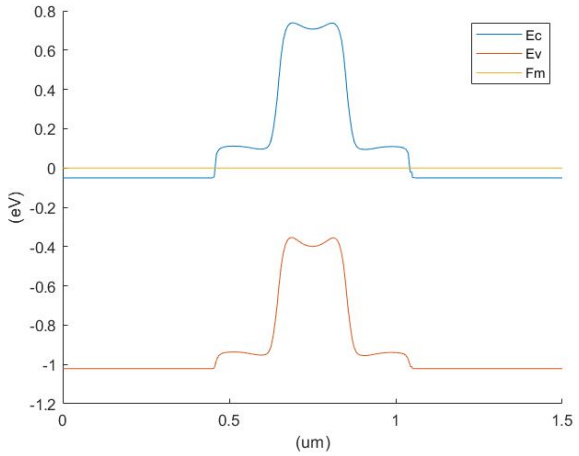

Fig. 8. Band Diagram of Optimized Device. Bending means the existence of depletion regions. Small depletion regions exist on the sides because of the Halos and Pockets introduced. In this way the depletion regions do not extend below the channel but outside of the channel. Fermi Level is above the conduction band because of the extreme doping concentration of n type.

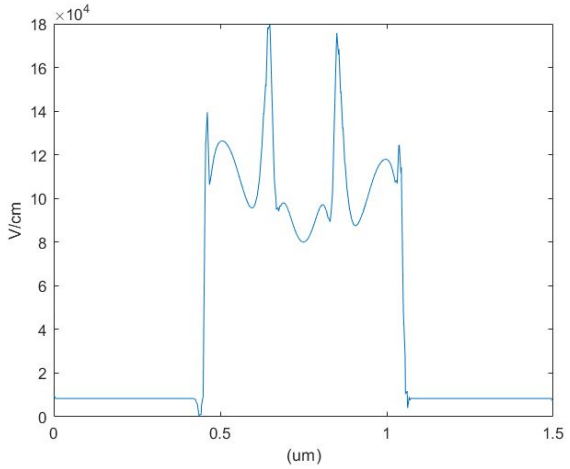

Fig. 9. Electric Field Diagram of Optimized Device. The electric field is not maximum on the sides because of the introduction of LDDs. The electric field is proportional to the doping. In this way, high electric fields do not exist near the drain/source but they exist mainly below the gate, protecting the device from avalanche breakdown or other undesired effects.

device suffers a roll-off in the threshold Voltage. However, the optimized device has higher threshold voltage compared to the scaled device and lower roll-off.

The band diagram plot depicts the fermi level above the conduction band on the sides. This is because of the extremely high doping concentration on drain/source. Additionally, its shape verifies that the depletion regions do not extend in the channel but in the LDDs. The electric field diagram verifies

the same statement, since the electric field is low and increases towards the center below the gate.

VI. CONCLUSION

The impact of short channel effects is effectively managed through the introduction of a high-k dielectric, reduced drain and source depth, and doping gradients within the substrate. These optimizations yield significantly more desirable performance parameters when compared to a simple scaling of the gate length. In most metrics the optimized device also improves over the performance of the long-channel baseline.

While the optimized device boasts some major improvements, the output conductance remains higher than desired. This could introduce non-ideal behavior in circuits dependent on the drain providing a high impedance node. Future work should focus on reducing the output conductance to values closer to that of the long channel MOSFET. This reduction will mostly likely come as a result in further optimization of short channel methods, especially through improved methods of substrate gradients, halos, and retrograde doping. Additionally, further SCE reduction techniques can be explored including improving the implementation of elevated source drain contacts, the introduction of strain on the substrate, and silicon on insulator structures.

Overall, the reduction in threshold voltage, DIBL, SS, and the considerable increase in on/off current are indicative of a more power efficient device. Lower power rails are required for its operation, and less power will be lost due to currents in the off state. This, combined with the reduced gate length, prove particularly beneficial for use in smaller and increasingly complex integrated circuits.

REFERENCES

[1] K. Fobelets,Advanced Semiconductor Devices: Summary course notes.unpublished, 2017.
[2] Hou, Yong-Tian & Li, Ming-Fu & Low, Tony & Kwong, Dim-Lee, *"Metal gate work function engineering on gate leakage of MOSFETs."*, IEEE Transactions on Electron Devices, Dec. 2004.
[3] Synopsis,Sentaurus Process, version y-2006.06 ed., 6 2006. Manual.