Analysis of Halo Implanted MOSFETs

Colin C. McAndrew and Patrick G. Drennan
Freescale Semiconductor, Tempe, AZ, Colin.McAndrew@freescale.com

ABSTRACT

MOSFETs with heavily doped regions at one or both ends of the channel exhibit qualitative differences in electrical behavior compared to devices with laterally uniform channel doping. These differences include a distinct peakiness in the transconductance near threshold, asymmetries in capacitances, and a surprising decrease in the statistical variation of the peak gain factor as channel length decreases. Historically, accurate modeling of such devices is best done with sectional MOSFET models. Here we present an analytic model of the behavior of the current and transconductance of a (unilaterally or bilaterally) halo implanted MOSFET and show that it predicts the decrease in variation of gain factor with channel length.

Keywords: MOSFET modeling; SPICE modeling; statistical modeling; halo implant.

1. INTRODUCTION

Low voltage MOSFETs historically have uniform lateral doping. Halo (pocket) implants [1] were introduced to improve performance and suppress some short-channel effects, and are now standard. Specific heavily doped regions were introduced at the ends of the channel in GCMOS technologies [2] to boost performance, and these could be done selectively at source and/or drain ends, leading to so-called unilateral and bilateral devices, having one or both ends doped heavily, respectively.

These laterally non-uniformly doped devices exhibit qualitatively different behavior compared to laterally uniformly doped devices. First, as the gate voltage increases from below to above threshold, the lightly doped central region "turns on" at a lower gate voltage than the more heavily doped end region(s), and therefore when the end regions start to turn on the channel length is effectively the length of these heavily doped region(s). As the gate voltage further increases and the level of inversion becomes more uniform in the device, the channel length becomes more the length of both the heavily and lightly doped regions. The (DC electrical) effective channel length L_{eff} is therefore shorter near threshold than for V_g well above threshold, and this gives $g_m(V_g)$ a distinct peaked behavior near threshold (we assume $V_s = V_b = 0$ here).

A very interesting analysis of halo implanted MOSFETs was presented in [3]. This work provided a physical threshold voltage model for bilaterally doped devices, and unequivocally showed that a 3-section MOS model is needed to capture fine details of halo implanted MOS behavior; merely modeling threshold voltage for a 1-section model is insufficient. Although the peaked g_m behavior is not explicitly noted in [3] it is apparent in Figs. 10 and 11 in [3]. But a model explaining this behavior and its statistical variation was not provided.

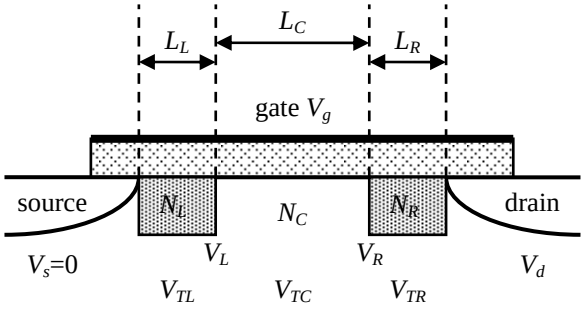

Fig. 1 MOSFET composed of 3 separately doped regions.

For a unilaterally doped device (assuming the heavy doping is on the source side of the device), as the device turns on the drain end inverts before the source end, hence there is a distinct peak in C_{gd} near threshold, and the device is highly asymmetric and $C_{gs} \neq C_{gd}$, even at $V_d = 0$. Again, this behavior cannot be captured by a 1-section MOSFET model, especially with the emphasis of the past decade in improving the symmetry of MOSFET models.

For MOSFETs with constant lateral doping, the statistical variation in the peak transconductance $g_m \propto \mu_0 C_{ox} W/L$ changes with channel length; for short devices it is greater than for long devices. Variation in the oxide thickness T_{ox} and the low field mobility μ_0 contribute equally to the variations in g_m for both long and short devices, but variations in L_{eff} affect a short device more than a long device, so the overall variation in g_m increases as length decreases.

This behavior is not observed in devices with nonuniform lateral doping. In unilateral GCMOS devices the measured variation in g_m is less for short devices than for long devices. We present an analysis that explains this counterintuitive behavior. The basic model underlying the analysis is simple, and is not intended to provide a highly accurate description of all aspects of device behavior. But it is amenable to analysis that elucidates the underlying cause of the unexpected decrease in g_m as channel length decreases.

Note that LDMOS transistors also have laterally non-uniform doping, but this is because their body region is formed by out-diffusion and not because of a separate heavy doping region at the source or drain. The analysis here is not directly applicable to LDMOS devices, but does capture some of the qualitative behavior they exhibit, because a unilaterally doped GCMOS device is qualitatively similar to an LDMOS device.

2. NUMERICAL BILATERAL HALO MODEL

Consider a MOSFET segmented into 3 regions, see Fig. 1. Adjacent to the source is a region of length L_L and threshold voltage V_{TL}, in the middle is a region of length L_C and

NSTI-Nanotech 2007, www.nsti.org, ISBN 1420061844 Vol. 3, 2007

threshold voltage V_{TC}, and adjacent to the drain is a region of length L_R and threshold voltage V_{TR}. The source and bulk are assumed to be at zero Volts and voltages V_g and V_d are applied to the gate and drain.

For operation in strong inversion non-saturation, at small V_d, the currents (normalized to $\mu_0 C_{ox} W$) in each of the 3 sections, which must be equal, are approximately

$$(1) \quad i_L = \frac{V_g - V_{TL} - 0.5V_L}{L_L} V_L$$

$$(2) \quad i_C = \frac{V_g - V_L - (V_{TC} + B_C V_L) - 0.5(V_R - V_L)}{L_C} (V_R - V_L)$$

$$(3) \quad i_R = \frac{V_g - V_R - (V_{TR} + B_R V_R) - 0.5(V_d - V_R)}{L_R} (V_d - V_R)$$

where V_L and V_R are the voltages between the left (source) and center sections, and between the center and right (drain) sections, respectively, and B_C and B_R are the body effect factors for the central and right region, respectively. The terms involving the last two parameters are required because the effective backgate bias for the center and right sections (taking the source as the reference) is not zero.

V_L and V_R in (1) through (3) vary with gate bias. Iteratively solving for these, as a function of gate bias and overall gate length $L = L_L + L_C + L_R$, then computing the current and g_m (and normalizing with respect to the g_m at V_g just above threshold), gives the results shown in Fig. 2.

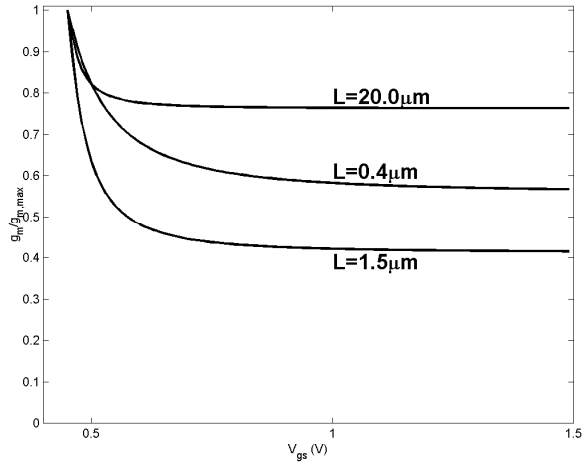

Fig. 2 Normalized g_m vs. V_g for various lengths.

For the simulations used to generate Fig. 2, $L_L = L_R = 0.08\mu m$ (the simulations here were based on the technology of [2], which is a 0.4μm CMOS technology with a minimum L_{eff} of 0.25μm from the GCMOS structure), $V_{TL} = V_{TR} = 0.4$ V and $V_{TC} = 0.1$ V. Note that for the simulations series resistance is ignored and mobility reduction due to both vertical and lateral fields is not included, so surface roughness scattering and velocity saturation are ignored; this is done to specifically elucidate the effects of the non-uniform lateral doping and to not have them confounded with other physical phenomena.

As observed in experimental data, the g_m exhibits a "peaked" characteristic (again, this is not from series resistance or vertical field mobility degradation, which increase the peakiness, the latter uniformly over length and the former more for short than for long devices; it is purely from the intrinsic nature of the laterally non-uniform doping – for a laterally uniformly doped device $g_m(V_g)$ would be independent of bias in Fig. 2). Note that the "peakiness" is non-monotonic in length. Asymptotically as $L_C \to 0$ the device appears, for $N_L = N_R$, to be laterally uniformly doped, and as $L_C \to \infty$ the relative effect of the end regions becomes vanishingly small, therefore the $g_m / g_{m,\max}$ curve should be flat for both of these cases. Fig. 3 shows the peakiness (the ratio of the peak g_m to its value for large V_g) as a function of (reciprocal) channel length.

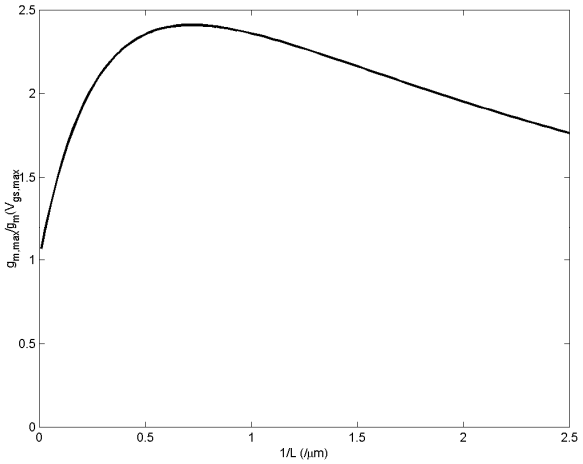

Fig. 3 g_m peakiness vs. reciprocal channel length.

Starting from short devices (rightmost abscissa values in Fig. 3), as the channel length increases (moving to the left), the peakiness increases, as is observed in practice. For very large lengths the peakiness again decreases, as is expected from the qualitative analysis above.

3. ANALYTICAL BILATERAL HALO MODEL

Instead of numerically solving (1) through (3), analytically differentiating with respect to V_g, and dropping negligible terms (V_d, V_L, and V_R are small), gives

$$(4) \quad g_{m,L} = \frac{V_L + (V_g - V_{TL}) \dfrac{\partial V_L}{\partial V_g}}{L_L}$$

$$(5) \quad g_{m,C} = \frac{V_R - V_L + (V_g - V_{TC}) \dfrac{\partial V_R}{\partial V_g} - (V_g - V_{TC}) \dfrac{\partial V_L}{\partial V_g}}{L_C}$$

$$(6)\quad g_{m,R} = \frac{V_d - V_R - (V_g - V_{TR})\dfrac{\partial V_R}{\partial V_g}}{L_R}$$

(detailed calculations that keep the omitted terms for now and eliminate them later are extremely tedious and lead to the same results given below).

Because the current through each section must be the same, the change in each current with V_g must the same, so equating (4) and (6) gives

$$(7)\quad \frac{\partial V_R}{\partial V_g} = \frac{L_L(V_d - V_R) - L_R V_L - L_R(V_g - V_{TL})\dfrac{\partial V_L}{\partial V_g}}{L_L(V_g - V_{TR})}.$$

Equating (5) and (6) and using (7) gives

$$(8)\quad \frac{\partial V_L}{\partial V_g} = \frac{\begin{array}{c} L_L\left((V_g - V_{TR})(V_R - V_L) + (V_g - V_{TC})(V_d - V_R)\right) \\ - L_C(V_g - V_{TR})V_L - L_R(V_g - V_{TC})V_L \end{array}}{\begin{array}{l} L_L(V_g - V_{TC})(V_g - V_{TR}) + L_C(V_g - V_{TL})(V_g - V_{TR}) \\ + L_R(V_g - V_{TL})(V_g - V_{TC}) \end{array}}$$

Substituting (8) into (4) and rearranging, gives (again normalized to $\mu_0 C_{ox} W$)

$$(9)\quad \frac{g_m}{V_d} = \frac{\begin{array}{l} \dfrac{V_L}{V_d}(V_g - V_{TC})(V_g - V_{TR}) \\ +\left(\dfrac{V_R}{V_d} - \dfrac{V_L}{V_d}\right)(V_g - V_{TL})(V_g - V_{TR}) \\ +\left(1 - \dfrac{V_R}{V_d}\right)(V_g - V_{TL})(V_g - V_{TC}) \end{array}}{\begin{array}{l} L_L(V_g - V_{TC})(V_g - V_{TR}) \\ + L_C(V_g - V_{TL})(V_g - V_{TR}) \\ + L_R(V_g - V_{TL})(V_g - V_{TC}) \end{array}}.$$

Note the inherent symmetry and consistency of (9); setting any of region length to zero (noting that if $L_L = 0$ then $V_L = 0$, if $L_R = 0$ then $V_R = V_d$, and if $L_C = 0$ then $V_L = V_R$) reduces the expression to an obviously similar form; as does setting the parameters of one region to those of an adjacent region.

Assuming a symmetric halo implant, $L_H = L_L = L_R$ and $V_{TH} = V_{TL} = V_{TR}$, then for small V_d, equating (1) and (2), and noting that at low V_d for a symmetric device $V_R \approx V_d - V_L$,

$$(10)\quad V_L \approx \frac{L_H/(V_g - V_{TH})}{L_C/(V_g - V_{TC}) + 2L_H/(V_g - V_{TH})} V_d$$

(which is clearly the expected result for $V_{TH} = V_{TC}$). The current in the transistor is then, normalized to $\mu_0 C_{ox} W$,

$$(11)\quad I_d \approx \frac{1}{L_C/(V_g - V_{TC}) + 2L_H/(V_g - V_{TH})} V_d .$$

This indicates that modeling the device as the series connection of separate region channel resistances, as in [3], is reasonable. However, our explicit solution for the internal node voltages verifies this approach.

Calculation of the transconductance gives

$$(12)\quad \frac{g_m}{V_d} \approx \frac{L_C/(V_g - V_{TC})^2 + 2L_H/(V_g - V_{TH})^2}{\left(L_C/(V_g - V_{TC}) + 2L_H/(V_g - V_{TH})\right)^2}.$$

This is simpler than the form (9) yet almost as accurate. However, it is still not intuitive how this varies with geometry and statistical fluctuations.

Again for left and right end regions with the same physical characteristics (length, doping, threshold voltage), for a long device from (9)

$$(13)\quad \frac{V_L}{V_d} \to 0, \frac{V_R}{V_d} \to 1, \frac{g_m}{V_d} \to \frac{1}{L_C + 2L_H \dfrac{V_g - V_{TC}}{V_g - V_{TH}}}.$$

Because $V_{TH} > V_{TC}$, this approximation predicts that g_m increases as V_g increases, which is qualitatively different from the behavior in Fig. 2. This is because it does not take into account the better approximation (10). A tighter upper bound is

$$(14)\quad \frac{g_m}{V_d} \to \frac{1}{L_C + 2L_H}.$$

For a short device

$$(15)\quad \frac{V_L}{V_d} \to 0.5, \frac{V_R}{V_d} \to 0.5, \frac{g_m}{V_d} \to \frac{1}{2L_H + L_C \dfrac{V_g - V_{TH}}{V_g - V_{TC}}}.$$

(As $L_C \to 0$, because the central region has a lower doping and hence lower threshold than the end regions, it is "turned on" more than the ends, and has a lower resistance; V_L thus approaches V_R).

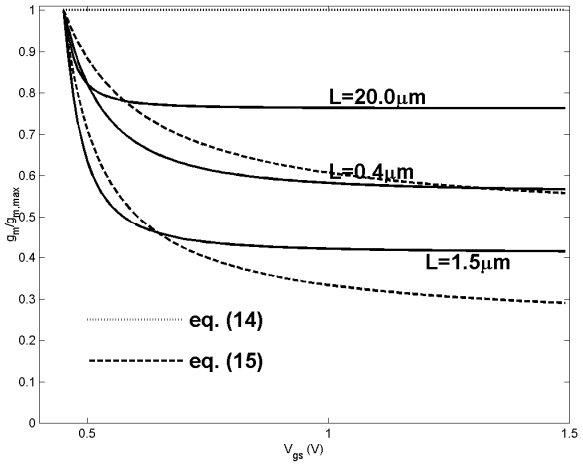

Fig. 4 Normalized g_m from analytic approximations.

NSTI-Nanotech 2007, www.nsti.org, ISBN 1420061844 Vol. 3, 2007

Fig. 4 compares the simple approximations (14) and (15) with values calculated from numerical simulations based on iterative solution of (1) through (3) (results from (12) are not included; they are almost identical to those from the numerical simulations; results from (15) are also not included for the 20μm long device as this approximation is only applicable to short devices). The long channel simple approximation (14) is clearly somewhat inaccurate. However, the agreement of the very simple analytic model (15) with the more detailed numerical model from the iterative solution of (1) through (3) is quite reasonable.

4. ANALYSIS

Because the halo region is more heavily doped than the center of the channel $V_{TH} > V_{TC}$, and the ratio of the halo to central region gate overdrives in the denominator of (15) is less than 1. The effect of this ratio on the overall device transconductance decreases as V_g decreases, hence the smaller the V_g the shorter the effective channel length in the denominator of (15), and the higher the g_m. This is the root cause of the peakiness of the $g_m(V_g)$ characteristic, and this simple and intuitive theoretical analysis agrees both with the qualitative description in the Introduction and with experimental observation that $g_m(V_g)$ is more "peaked" for short than for long devices.

More interestingly, consider what happens when L_H, V_{TH}, L_C, or V_{TC} in (15) vary. Increasing any of these parameters would normally, as a first response, be expected to lead to a reduction in current, and hence g_m. This is seen to be the case for the parameters L_H, L_C, and V_{TC}, but is unexpectedly not true if V_{TH} increases. Increasing V_{TH} causes $V_g - V_{TH}$ to decrease (at a fixed gate voltage), which reduces the overall denominator in (15) and therefore increases g_m. An alternative viewpoint is that if V_{TH} increases compared to V_{TC}, the central region is turned on "harder" when the halo regions invert, this makes the effective channel length appear shorter, and hence enhances g_m. Numerical simulations confirm this counter-intuitive behavior.

Statistical variations in the gain factors come from statistical variations in T_{ox} and μ_0, which (ignoring the uncorrelated, or mismatch, component) are the same for short and long devices, and from statistical variations in L_H, V_{TH}, L_C, and V_{TC}. For illustrative and comparison purposes we ignore the effect of the variations in T_{ox} and μ_0 and consider only the variations in the length and threshold parameters. Propagation of variance (PoV) analysis [4] of (15) therefore gives

$$\frac{\sigma^2_{\frac{\delta g_m}{g_m}}}{1/L_e^2} = 4\sigma^2_{\delta L_H} + \left(\frac{L_C(V_g - V_{TH})}{(V_g - V_{TC})^2}\right)^2 \sigma^2_{\delta V_{TC}} + \left(\frac{V_g - V_{TH}}{V_g - V_{TC}}\right)^2 \sigma^2_{\delta L_C} + \left(\frac{L_C}{(V_g - V_{TC})^2}\right)^2 \sigma^2_{\delta V_{TH}} \quad (16)$$

where $L_e = 2L_H + L_C(V_g - V_{TH})/(V_g - V_{TC})$.

Applying this PoV analysis to the 1.5μm and 0.4μm channel length devices, for which Fig. 4 shows that the approximation (15) is reasonable, gives the results shown in Fig. 5 and Fig. 6. The contribution of the variations in each of the parameters, at the 1-σ level, is shown in these figures. The counter-intuitive conclusion from this analysis agrees with experimental observations: the statistical variation in the peak g_m for halo devices is greater for longer devices than for shorter devices.

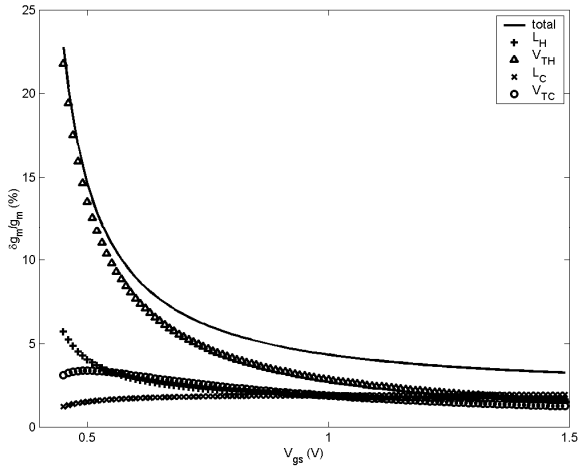

Fig. 5 Statistical variations in g_m for the L=1.5μm device.

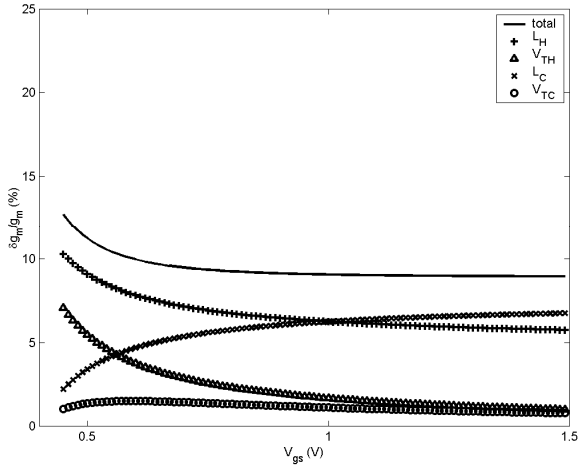

Fig. 6 Statistical variations in g_m for the L=0.4μm device

Most strikingly, it is the decrease of the influence of variation in V_{TH} as length decreases on the gain factor that stands out. Analytically, the 4th term in the right-hand side of (16) should decrease as L_C decreases, and this is observed in numerical simulations and data. However, overall contributions in practice are more difficult to compartmentalize.

The empirical conclusion from Fig. 5 and Fig. 6 is that statistical variations in peak g_m decrease as channel length decreases, for lengths in a certain range. This conclusion is supported by experimental data, which show that the peak g_m variation for GCMOS type devices decreases as L_C decreases.

5. UNILATERAL HALO ANALYSIS

For a device with an implant only at the source end of the channel, which in some ways is analogous to an LDMOS device,

$$(17)\quad \frac{g_m}{V_d} = \frac{V_g - V_{TL} + (V_L/V_d)(V_{TL} - V_{TC})}{L_L(V_g - V_{TC}) + L_C(V_g - V_{TL})}$$

and for long devices $V_L/V_d \to 0$ and for short devices $V_L/V_d \to 1$, hence (13) and (15) still apply, with L_H substituted for $2L_H$; qualitatively the behavior is identical.

The analyses here have been targeted to bilateral halo implanted devices, but the original analysis we developed was based on unilateral devices, to try to understand physically why the observed statistical variation in the peak g_m was smaller for shorter devices than for moderate length devices.

6. CONCLUSIONS

We have provided an analysis of the drain current of halo implanted MOSFETs, at low V_{ds}, based on equality of the currents flowing in different regions of the device. This enabled expressions for the gain factor g_m/V_{ds} to be derived.

The analysis elucidates the reason for the qualitative change in the shape of $g_m(V_{gs})$ plots, and also for the observed decrease in statistical variation as channel length decreases.

Our analyses support the conclusion of [3], that there are qualitative features of the characteristics of halo implanted devices that cannot be modeled using a single MOSFET model; a multi-section model (two or three sections, depending on the structure of the device) is needed. However, rather than analyzing a device from the channel resistance perspective, as noted in [3], we found it advantageous to consider equality of currents in different regions of the device. This led to the physical understanding of why the peak g_m variation for short devices is unexpectedly less than for moderate length devices, and what physical parameter variations underlie this behavior.

REFERENCES

[1] Y. Okumura *et al.*, "A Novel Source-to-Drain Nonuniformly Doped Channel (NUDC) MOSFET for High Current Drivability and Threshold Voltage Controllability," *IEEE IEDM Tech. Dig.*, pp. 391-394, Dec. 1990.

[2] F. K. Chai *et al.*, "A Cost-Effective 0.25μm Leff BiCMOS Technology Featuring Graded-Channel CMOS (GCMOS) and a Quasi-Self-Aligned (QSA) NPN for RF Wireless Applications," *Proc. IEEE BCTM*, pp. 110-113, Sep. 2000.

[3] R.Rios *et al.*, "A Three-Transistor Threshold Voltage Model for Halo Processes," *IEEE IEDM Tech. Dig.*, pp. 113-116, Dec. 2002.

[4] C. C. McAndrew and P. G. Drennan, "Unified Statistical Modeling for Circuit Simulation," *Proc. MSM-WCM*, pp. 715-718, 2002.