PERGAMON

Solid-State Electronics 45 (2001) 507–510

SOLID-STATE ELECTRONICS

Threshold voltage definition and extraction for deep-submicron MOSFETs

X. Zhou *, K.Y. Lim, W. Qian

School of Electrical and Electronic Engineering, Nanyang Technological University, Block S1, Singapore 639798, Singapore

Received 5 February 2000; received in revised form 3 November 2000

Abstract

The subtle difference in MOSFET threshold voltage between the two popular definitions, maximum-g_m and constant current, is investigated in the deep-submicron regime. The result pinpoints to the importance of the lateral-field effect in linear region at very short gate length, and further supports the combined definition known as the "critical current at linear threshold" method, which includes short-channel effects while retaining the simplicity and consistency of the constant-current method. © 2001 Elsevier Science Ltd. All rights reserved.

1. Introduction

Although the threshold voltage (V_t) of a MOSFET is not a figure of merit for device/circuit performance, it is the most important parameter for MOS device modeling and circuit design. The V_t value of a MOSFET is dependent upon its definition [1–3], while the criteria for a "valid" V_t definition should be physical as well as easy to measure. For deep-submicron MOSFETs, threshold voltage and effective channel length (L_{eff}), both being electrical parameters, are the most sensitive model parameters influencing the drain current of a MOSFET model. Many research publications in the literature, especially those on compact models compared to the measured current–voltage (I–V) characteristics, do not mention how V_t and L_{eff} in the model as well as measurement are defined and how they are extracted.

In this paper, the de facto industry standard V_t definition based on the "constant-current" (CC) method is revisited in comparison with the newly proposed "critical current at linear threshold" ("I_{crit} at V_{t0}") method [4]. The subtle difference between the two methods is explored in the context of the 2-D short-channel effects in deep-submicron MOSFETs.

2. Definition and discussion

The threshold voltages presented in this work are extracted from measured I_{ds}–V_{gs} curves with drawn gate lengths (L_{drawn}) from 10 μm down to 0.2 μm (W = 20 μm) on the same die of a 0.25 μm CMOS wafer (with $\Delta V_{gs} = 0.05$ V and $V_{ds} = 0.1$ V), as shown in Fig. 1. For the I_{crit} at V_{t0} definition, linear threshold voltage (V_{t0}) for each device (L_{drawn}) is determined from linear extrapolation of I_{ds}–V_{gs} at peak transconductance (g_m) to zero I_{ds} (commonly known as the "maximum-g_m" method), and the corresponding critical current (I_{crit}) at $V_{gs} = V_{t0}$ is interpolated from the $\log(I_{ds}) - V_{gs}$ curve. To remove ambiguity of the CC method and to compare with the I_{crit} at V_{t0} definition, the value of $I_{crit}(10\ \mu m) = 1.8\ \mu A$ from the I_{crit} at V_{t0} definition at long channel is used as the CC: $I_{d0} = (10/20) \times I_{crit}(10\ \mu m) = 0.9\ \mu A$. At any other L_{drawn}, I_{crit} for the CC definition is scaled according to $I_{crit} = I_{d0}(W/L_{drawn})$ at which V_{gs} is extracted as the value of V_{t0}. For both methods, once the respective I_{crit} is determined, the saturation threshold voltage (V_{tsat}) is obtained from interpolation of the measured saturation I_{ds}–V_{gs} curves ($V_{ds} = 2.5$ V) for V_{gs} at which $I_{ds} = I_{crit}$ for each device, as shown in Fig. 2.

After calibrating the critical currents at long channel, the subtle difference between the two methods at short channel becomes obvious. I_{crit} at $V_{gs} = V_{t0}$ for the maximum-g_m definition are found to occur consistently at maximum dg_m/dV_{gs} (similar to the second-derivative

* Corresponding author. Tel.: +65-790-4532; fax: +65-791-2687.

E-mail address: exzhou@ntu.edu.sg (X. Zhou).

PII: S0038-1101(01)00035-1

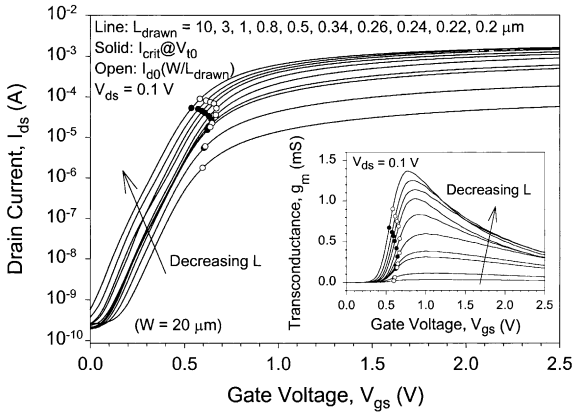

Fig. 1. Measured linear I_{ds}–V_{gs} curves (—) for each device of drawn length L_{drawn} as indicated. Critical currents based on the I_{crit} at V_{t0} definition (•) and the CC definition (○) are shown for each device. The inset shows the corresponding linear transconductance.

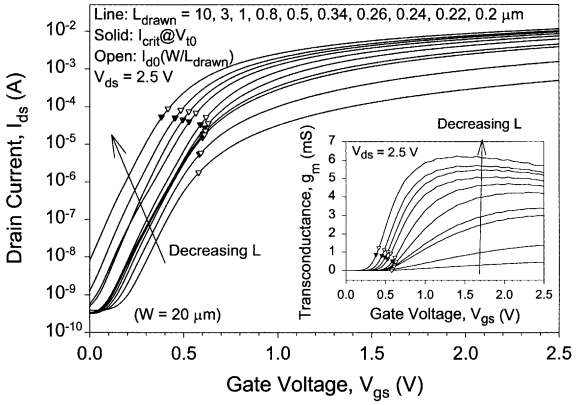

Fig. 2. Measured saturation I_{ds}–V_{gs} curves (—) for the same set of devices. V_{tsat} are interpolated at V_{gs} at which $I_{ds} = I_{crit}$ (from Fig. 1 for both methods) for each device. The inset shows the corresponding saturation transconductance.

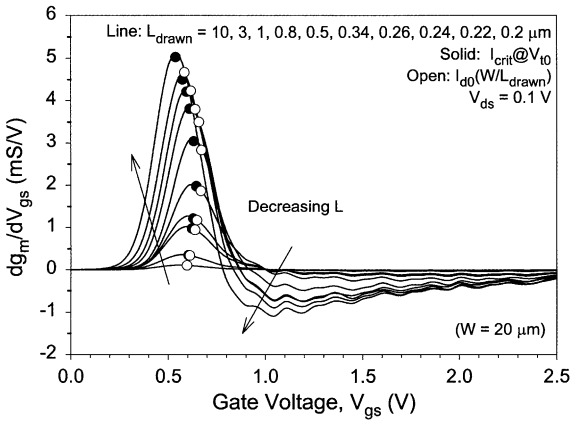

Fig. 3. Second derivative of the measured linear I_{ds}–V_{gs} data (—) for each device, with the corresponding values indicated at the extracted V_{t0} for the I_{crit} at V_{t0} definition (•) and the CC definition (○).

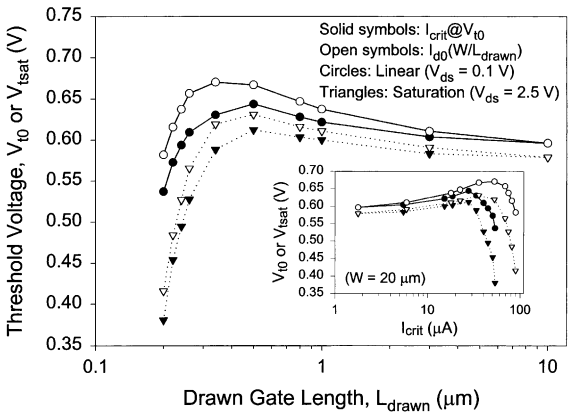

Fig. 4. Extracted V_{t0} (•, ○) and V_{tsat} (▼, ▽) versus L_{drawn} for the I_{crit} at V_{t0} definition (•, ▼) and the CC definition (○, ▽). The inset shows the same data against I_{crit}, as shown by the symbols in Figs. 1 and 2.

method [5]), while those from the CC definition are off the peak, as shown in Fig. 3. The extracted V_t–L_{drawn} curves also show different V_t roll-up (due to reverse short-channel effect) and roll-off behaviors, as demonstrated in Fig. 4, which are reflected in the V_t–I_{crit} curves (inset of Fig. 4) from the measured I_{ds}–V_{gs} data due to different definitions. This difference prompts the importance of the definition-dependent nature of V_t, since the modeling of other quantities, such as mobility and series resistance, depends a lot on the V_t model.

The difference becomes apparent when I_{crit} for the two definitions are plotted against L_{drawn} in Fig. 5, especially on a log–log scale (inset of Fig. 5). Deviation from "linearity" (on a log–log scale) of I_{crit} at V_{t0} is a result of increased contribution of S/D series resistance (R_{sd}) at short channel (due to increased current), which has been commonly considered as a major drawback of the maximum-g_m method. However, the critical currents at such defined $V_{gs} = V_{t0}$ correspond consistently to the condition for peak transconductance and channel-mobility change for every device, and they represent the actual current that flows under the *physical* polygate (L_g, not L_{drawn}) as well as the S/D junctions. On the other hand, the CC definition "unphysically" scales the critical current with a W/L_{drawn} dependency. This has been the basis on which the "I_{crit} at V_{t0}" method [6] for simultaneously extracting L_{eff} and R_{sd} is based.

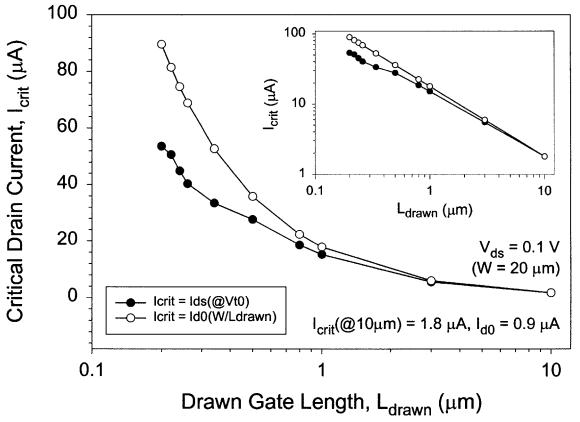

Fig. 5. I_{crit} versus L_{drawn} for the I_{crit} at V_{t0} definition (•) and the CC definition (○) extracted from the linear I_{ds}–V_{gs} curves (Fig. 1). The inset shows $\log(I_{crit})$ versus $\log(L_{drawn})$.

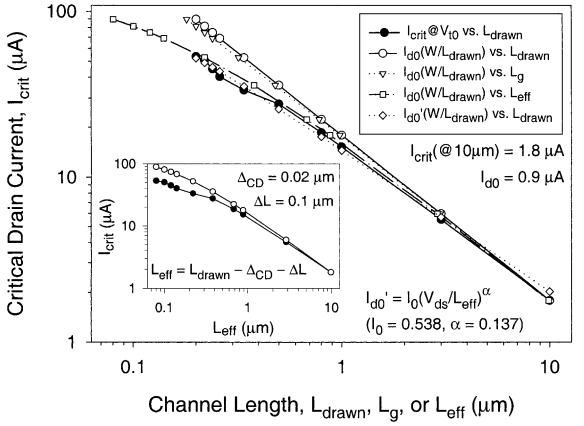

Fig. 6. The same data from Fig. 5 against L_{drawn} (•,○). The I_{crit} data of the CC definition when plotted against L_g (▽) or L_{eff} (□) assuming $\Delta_{CD} = 0.02$ μm and $\Delta L = 0.1$ μm. The empirical model $I'_{d0} = I_0(V_{ds}/L_{eff})^{\alpha}$ fitted to the I_{crit} at V_{t0} data (◇). The inset plots I_{crit} against L_{eff} for both methods.

From MOS device physics, the drain current in linear mode is inversely proportional to the *effective* channel length, which should be close to the metallurgical channel length (L_{met}). Without a priori knowledge of L_{eff} (which is also definition dependent), L_{drawn} has been used over the years in the CC definition, which gives an exact -1 slope on the $\log(I_{crit})$–$\log(L_{drawn})$ curve. To examine the difference in the two I_{crit} definitions, a simple *critical-dimension correction* (Δ_{CD}) (due to uncertainties in mask/polysilicon lithography and polyetching) is assumed to model the physical polygate length $L_g = L_{drawn} - \Delta_{CD}$, and a constant $\Delta L (= 2\sigma x_j)$ to model LDD lateral diffusion such that $L_{eff} = L_{met} = L_g - \Delta L = L_{drawn} - \Delta_{CD} - \Delta L$ [6,7]. When the CC-defined $I_{crit} = \log[I_{d0}(W/L_{drawn})]$ is plotted against $\log(L_g)$ and $\log(L_{eff})$ with the estimated $\Delta_{CD} = 0.02$ μm and $\Delta L = 0.1$ μm, it is found that I_{crit} increases *sublinearly* (on a log–log scale) at shorter channel length, as shown in Fig. 6 by the open triangles and open squares, respectively. However, I_{crit} such interpreted is still larger than that of the I_{crit} at V_{t0} definition (see inset of Fig. 6) because the long-channel I_{d0} has been kept constant.

In principle, for every channel-length device, I_{d0} should be proportional to V'_{ds}, the voltage drop across its *intrinsic* L_{eff}, and mobility μ_{eff}, both of which decrease at shorter channel due to increased voltage drop across R_{sd} and increased lateral channel field (V_{ds}/L_{eff}), respectively, since $V_{ds} = 0.1$ V is fixed. This implies that I_{d0} should be L_{eff} dependent, and this dependency is actually contained in the I_{crit} at V_{t0} data since the critical current that flows through the MOSFET under the maximum-g_m condition includes the effects of R_{sd} and lateral field. We propose that this effect be empirically modeled by a new $I'_{d0} = I_0(V_{ds}/L_{eff})^{\alpha}$ with two fitting parameters, I_0 and α. By fitting $I_{crit} = I'_{d0}(W/L_{drawn})$ to the I_{crit} at V_{t0} versus L_{drawn} data, as shown in Fig. 6 (open diamonds), the extracted values are found to be $I_0 = 0.538$ and $\alpha = 0.137$. As the behavior of I_{crit}–L_{drawn} from the maximum-g_m definition is unknown, this simple model and extraction approach provides away to empirically model the I_{crit}–L_{drawn} behavior.

This simple empirical model also confirms the idea of lateral-field (V_{ds}) dependence of the linear channel resistance [6] as well as the nonscaling characteristics of total resistance in the deep-submicron regime [8]. Complete modeling and extraction of L_{eff}, Δ_{CD}, and ΔL based on the I_{crit} at V_{t0} method has been developed and reported elsewhere [9].

3. Conclusion

In conclusion, the arbitrary choice in the industry-standard constant-current V_t definition can be removed by calibrating I_{d0} to that from the maximum-g_m definition at long channel, which avoids the ambiguity while retaining the simplicity. However, the effect of unphysical scaling in the constant-current definition becomes pronounced for deep-submicron MOSFETs. If such defined V_t is used in I–V modeling, it may require additional efforts in mobility and resistance modeling, or even lead to incorrect information (e.g., V_t roll up) in the application of inverse modeling [10]. The I_{crit} at V_{t0} definition is based on consistent operation at long and short channel as well as different regions of operation, and contains information on actual device and short-channel effects (L_g, R_{sd}). The proposed empirical approach to modeling the I_{crit}–L_{drawn} behavior is simple and

can be applied to the modified constant-current method for V_t extraction.

Acknowledgements

Chartered Semiconductor Manufacturing Ltd. for providing the experimental data for this work is gratefully acknowledged.

References

[1] Yan ZX, Deen MJ. Proc Inst Elect Engng G 1991;138:351–7.
[2] Liou JJ, Ortiz-Conde A, Sanchez FG. Proc IEEE HKEDM'97, Hong Kong, 1997. p. 31–8.
[3] Tsuno M, Suga M, Tanaka M, Shibahara K, Miura-Mattausch M, Hirose M. IEEE Trans Electron Dev 1999;46:1429–34.
[4] Zhou X, Lim KY, Lim D. IEEE Trans Electron Dev 1999;46:807–9.
[5] Wong HS, White MH, Krutsick TJ, Booth RV. Solid-State Electron 1987;30:953–68.
[6] Zhou X, Lim KY, Lim D. IEEE Trans Electron Dev 1999;46:1492–4.
[7] Zhou X, Lim KY. IEEE Trans Electron Dev 2001;48.
[8] Esseni D, Iwai H, Saito M, Ricco B. IEEE Electron Dev Lett 1998;19:131–3.
[9] Zhou X, Lim KY. MSM2001, to appear in Proc MSM2001, Hilton Head Island, SC, 2001.
[10] Lee ZK, McIlrath MB, Antoniadis DA. IEEE Trans Electron Dev 1999;46:1640–8.