Evaluation of Threshold Voltage Extraction Methods in Deep-submicron Technology

Tae Hyun Kim[1], Hanwool Jeong[2] and Seong-Ook Jung[3]
[123]Department of Electrical and Electronic Engineering, Yonsei University, South Korea
Sinchon-dong, Seodaemun-gu, Seoul, Korea
E-mail: [1]dife2020@yonsei.ac.kr, [2]hanwool87@yonsei.ac.kr, [3]sjung@yonsei.ac.kr

Abstract: Analytic models for threshold voltage are not suitable for short channel devices in these days. As alternatives to these models, various medthods are devised to extract the threshold voltage from drain current versus gate voltage characteristic. This paper compares these methods according to technology scaling in deep-submicrom technology and evaluates which method is suitable.

1. Introduction

The threshold voltage (V_T) is one of the representative characteristic of MOSFET. The MOSFET operation is determined based on V_T as a criterion of depletion mode and inversion mode of the channel.

As technology scales down, short-channel effects such as mobility degradation, velocity saturation or drain induced barrier lowering (DIBL) affect drain current and trans-conductance. V_T is also varied by these effects. Therefore, the traditional analytic equations for V_T, which do not consider these effects, are not suitable for short channel devices. In order to consider these effects for V_T extraction, several threshold voltage extraction methods have been introduced based on I_D-V_G curve ($\sqrt{I_D}-V_G$ curve for V_T in saturation region) [1]-[5],[7],[8].

In this paper, various previously proposed V_T extraction methods are evaluated for 130 nm to 32 nm technology nodes. Section 2 reviews the electronic meaning of the V_T extraction methods. In Section 3, the trend of V_T with technology scaling is presented and suitable V_T extraction method for deep-submicron device is discussed.

2. Consideration of V_T Extraction Methods

In this section, six V_T extraction methods are introduced briefly and the electronic meaning of the methods are discussed.

2. 1 Constant-Current (CC) method

The constant-current (CC) method is the most popular V_T extraction method because it is relatively simple compared with the other methods. V_T is defined as V_G which corresponds to constant drain current $I_D = (W/L)\times 10^{-7}[A]$ [2] . The constant current, $10^{-7}[A]$, is multiplied with ratio of width to length because drain current is affected by this ratio. Figure 1 shows extracted V_T by CC method.

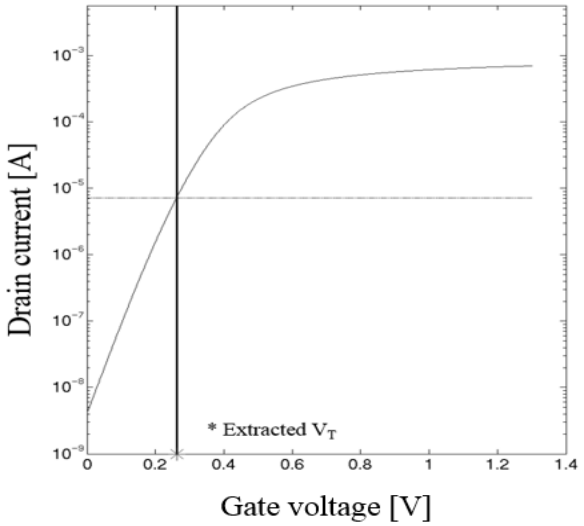


Figure 1. V_T extraction from Drain current vs. Gate voltage transfer characteristic by CC method in linear region.

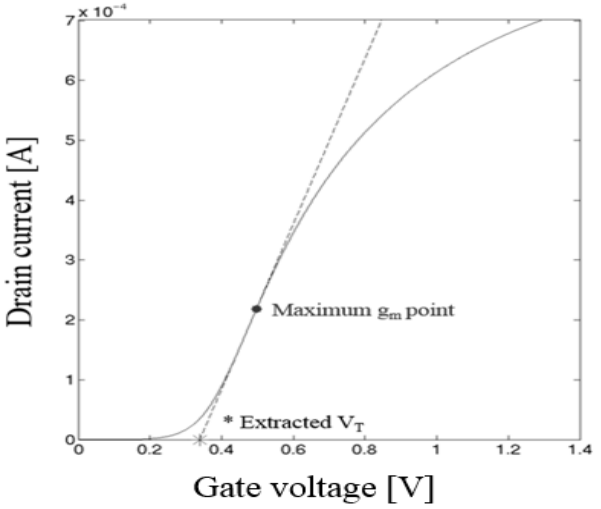


Figure 2. V_T extraction from Drain current vs. Gate voltage transfer characteristic by LE method in linear region.

2. 2 Linear Extrapolation (LE) method

Linear extrapolation (LE) method determines V_T as V_G axis intercept of linear extrapolation at maximum trans-conductance (g_m) point in I_D vs. V_G curve [3]. LE method assumes drain current holds 0A under V_T and increases linearly above V_T. The linear extrapolation at the maximum trans-conductance is assumed as this linear line and V_G axis intercept becomes V_T. Figure 2 shows drain current versus gate voltage characteristic and V_T extracted by LE method in linear region. Trans-conductance decreases as the gate voltage increases when the gate voltage is larger than certain value because of the mobility degradation.

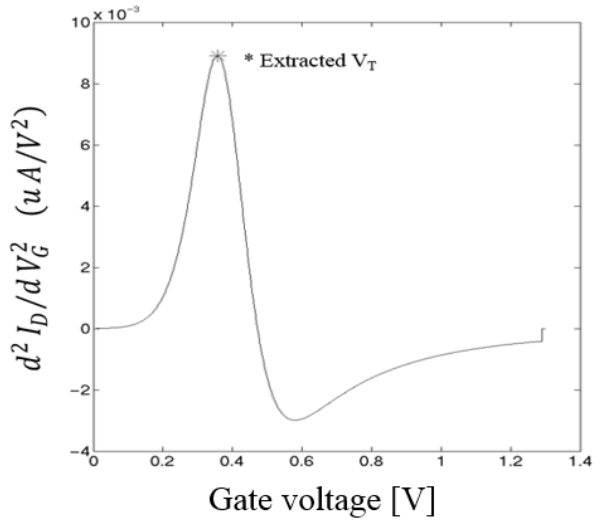


Figure 3. V_T extraction from second derivative of I_D vs. V_G in linear region.

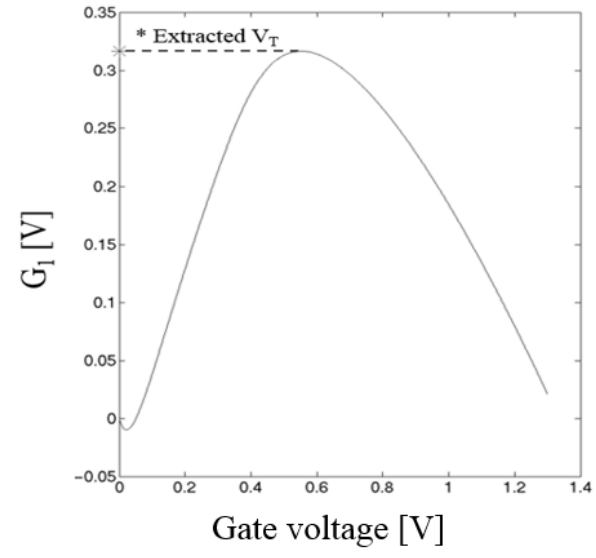


Figure 4. V_T extraction from G_1 vs. V_G by transition method in linear region.

Therefore, the drain current increases slowly when the gate voltage is larger than the point at which g_m is maximized.

2. 3 Second Derivative (SD) method

The drain current is divided into two regions, exponential region and linear region according to gate voltage. In a real device, the conversion from exponential region to linear region is continuous. Thus, the gradient of I_D vs V_G transfer curve, g_m, must rapidly increases when V_G is smaller than this conversion point and becomes slow when V_G is larger than the conversion point. As the extension of this concept, SD method determines V_T as gate voltage when second derivative of I_D to V_G becomes maximum [4]. Figure 3 shows the maximum first derivative of g_m and extracted V_T.

2. 4 Transition method

The transition method [5] is based on the below equation,

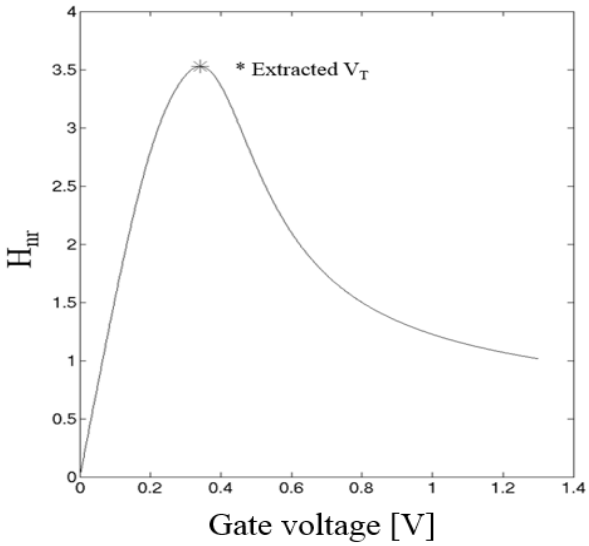


Figure 5. V_T extraction by NRH method in linear region.

$$G_1(V_G, I_D) \equiv \frac{D(V_G, I_D)}{I_D} = (V_G - V_{Gi}) - 2\frac{\int_{V_{Gi}}^{V_G} I_D dV_G}{I_D} \quad (1)$$

where $D(V,I) = \int_0^I V dI - \int_0^V I dV = VI - 2\int_0^V I dV$ [1]. (2)

If I_D is constant as leakage current below V_T and linearly increases as V_G increases above V_T, G_1=-V_G for $V_G < V_T$ and G_1=V_T for $V_G > V_T$ where V_{Gi} is 0 [1]. However in a real device, G_1 increases linearly when $V_G < V_T$ if we set I_D as below,

$$I_D = I_{s0} \exp[\beta(V_{GS} - V_T)]/n] \text{ [6]} \quad (3)$$

Furthermore, G_1 decreases when $V_G > V_T$ because integration of I_D for sub-threshold region is not negligible. Mobility degradation also restrains the maximum G_1 value. Therefore, V_T is determined as maximum G_1 value like in Figure 4, which means V_T approximately.

2. 5 Normalized Mutual Integral Difference (NMID) & Normalized Reciprocal H function (NRH) method

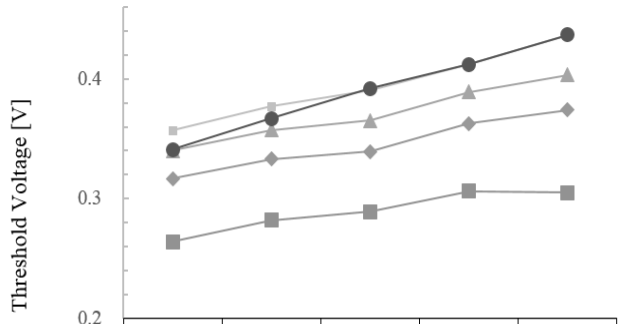
NMID method [7] is based on equation (2) as below.

$$D_{normal}(V_G, I_D) \equiv \frac{D(V_G, I_D)}{I_D V_G} = 1 - 2\frac{\int_{V_{Gi}}^{V_G} I_D dV_G}{I_D V_G} \text{ [1]} \quad (4)$$

NRH method [8] is based on the following equation, (5), which is derived from (4) by removing 1 and taking reciprocal.

$$H_{nr}(V_G) = \frac{V_G[I_D - I_D(V_G = 0)]}{2\int_0^{V_G} I_D(V_G) dV_G} \text{ [1]} \quad (5)$$

Numerator and the number 2 of denominator of (5) represent area of triangle which has V_G as lower base and I_D



	130nm	90nm	65nm	45nm	32nm
CC	0.264	0.282	0.289	0.306	0.305
LE	0.340	0.357	0.365	0.389	0.404
SD	0.357	0.377	0.391	0.412	0.437
TM	0.317	0.333	0.339	0.363	0.374
NRH & NMID	0.341	0.367	0.392	0.412	0.437

Figure 6. V_T in linear region for various methods with technolgy scaling

Table 1. Specification of deep-submicron technologies from 130nm to 32nm

	Channel length [nm]	Channel width [nm]	W/L
130nm	70	175	2.50
90nm	50	129	2.58
65nm	35	99	2.83
45nm	35	79	2.26
32nm	30	44	1.47

as height. Denominator except 2 represents area below I_D-V_G curve. Therefore H_{nr} is area ratio of these two. H_{nr} will increase until the maximum changing point of g_m and the point is the conversion point from exponential region to linear region for I_D. Thus NRH method determines V_T as V_G when H_{nr} is the maximum as shown in Figure 5. The second term of (4) is proportional to reciprocal of (5), so NMID determines V_T as V_G when D_{normal} is maximized. V_T values extracted from NMID and NRH are same because they are based on the same mathematical principle.

3. Extraction Methods in Deep-submicron Technolgy with Technology Scaling

In this section, NMOS in 130 nm to 32 nm technology nodes are used to compare extraction methods. Device specification is based on Intel's device papers [9]-[13] and minimum width is used to observe effect of scaling clearly. Table 1 shows the specification of devices.

Figure 6 shows V_T change in linear region with technology scaling. V_{DS} is set as 0.05V in linear region. V_T increases for all the methods except for CC method as technology scales down. In the case of CC method, V_T increases until 45nm and decreases for 32nm because current constraint varries proportional to (W/L) ratio. Although (W/L) decreases when the technology scales down from 65nm to 45nm, V_T increases because leakage current is much smaller in 45nm while subthreshold swing is almost same [11],[12].

Especially V_T of CC method is smaller than that of the other methods because the other methods are strongly related with conversion from exponential region to linear region for I_D and mobility degradation affects this conversion.

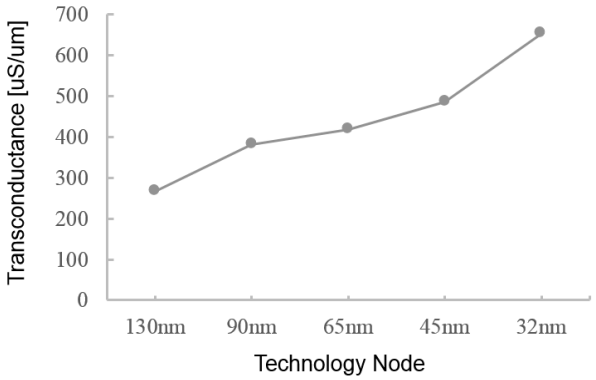


Figure 7. g_m in linear region with technolgy scaling

For LE method, g_m is the most important factor because LE method is based on the maximum g_m. As shown in Figure 7, the maximum g_m increases as technology scales down in deep-submicron technology. Although g_m is strongly affected by the mobility degradation in the linear region and then the maximum g_m can be observed in smaller voltage in a smaller technology node, V_T increases in the linear region because the values of the maximum g_m increases as technology scales down.

For NMID method, V_T is determined as V_G which makes (4) be maximized, or the derivative of (4) be 0. By substituting (3) to (4), following equation (6) is obtained,

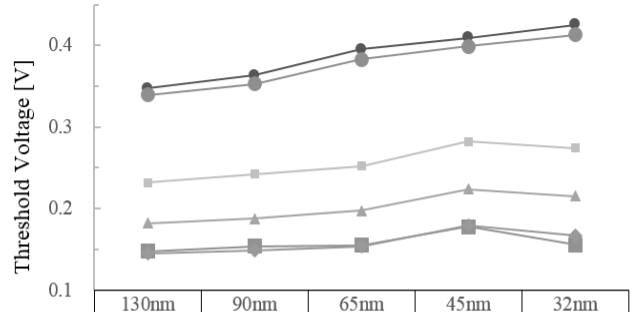
$$D_{normal}(V_G) = 1 - 2\frac{\frac{nK}{\beta}(e^{V_G\frac{\beta}{n}} - 1)}{(Ke^{V_G\frac{\beta}{n}} - I_{D,leakage})V_G} \qquad (6)$$

where $$K = I_{s0}e^{-\frac{\beta}{n}V_T} \qquad (7)$$

The derivative of (6) becomes 0 only where V_G is 0 if the integral of leakage region is negilgible (I_D=0 when V_G=0). It means the point at which (6) is maximized is not obtained in the exponential region. Instead, (6) is maximized at the point at which I_D linearly increases with V_G. In other words, (4) becomes maximized at the maximum changing point of g_m similar to H_{nr}, which is mentioned in Section 2.5. Therefore the maximum of (6) appears at the conversion point from exponential to linear region. However, if the leakage is not negligible, the maximum point can be at the exponential region, which is lower than that conversion point.

In the case of NRH method, as (5) is basically same with (4), V_T is extacted at the conversion point when leakage is negligible as in the cas of NMID. Similarly, V_T is extracted at the point smaller than the conversion point when leakage is not negligible.

V_T extracted by SD method is very close to the conversion point as considering concept of SD method. Therefore, V_T of NMID or NRH is much smaller than that of SD in 130 nm. On the other hand, V_T of these methods becomes closer to that of SD as technology scales down. This is because the maximum point of NMID or NRH



	130nm	90nm	65nm	45nm	32nm
CC	0.148	0.154	0.155	0.178	0.156
LE	0.182	0.188	0.198	0.224	0.215
SD	0.232	0.242	0.252	0.282	0.274
TM	0.146	0.149	0.154	0.179	0.167
NRH	0.347	0.363	0.395	0.409	0.425
NMID	0.339	0.353	0.383	0.399	0.413

Figure 8. V_T in saturation region for various methods with technolgy scaling

becomes closer to the conversion point as leakage current decreases with scaling.

Figure 8 shows V_T change in saturation region. In this case, V_{DS} is set to the nominal voltage of the corresponding technology node. It is observed that V_T of saturation region is smaller than that of linear region due to the DIBL.

V_T of CC, LE and SD methods in saturation region maintains similar tendency with that in the linear region except for 32nm. V_T of LE and SD decreases in saturation region for 32nm because DIBL is stronger than other technology nodes. V_T difference between CC and LE or CC and SD is maintained because DIBL affect similarly to these methods.

V_T of transition method becomes much closer to CC in saturation region because transition methold is much sensitive to DIBL compared with other methods. This is because transition method is not be normalized with V_G compared with NMID and NRH and V_G remains itself while I_D is replaced with $\sqrt{I_D}$ in saturation region.

On the other hand, V_T of NMID and NRH in saturation region is almost the same as that in linear region. This is because the change from I_D to $\sqrt{I_D}$ for considering the saturation region doesn't affect (4) or (5) significantly because it is cancelled out at numerator and denominator when I_D is expressed as exponential like (3) and I_D=0 (V_G=0). Therefore, NMID and NRH method do not consider effect of DIBL clearly.

4. Conclusion

The CC method is confirmed as independent to mobility degradation in deep-submicron device. The other methods need to consider mobility degradation because they are based on trans-conductance from drain current versus gate voltage transfer characteristic. Even though these methods still show clear tendency with technology scaling in linear region, DIBL effect is not applied correctly in saturation region for integral based methods (transition, NMID and NRH).

Therefore, CC is most powerful method in deep-submicron technology, and LE or SD is also reliable to show trend of V_T in linear and saturation regions. On the other hand, NMID, NRH and transition method is unsuitable because they do not consider effect of DIBL clearly.

References

[1] Ortiz-Conde, Adelmo, et al. "Revisiting MOSFET threshold voltage extraction methods." *Microelectronics Reliability* 53.1 (2013): 90-104.

[2] Tsuno, Morikazu, et al. "Physically-based threshold voltage determination for MOSFET's of all gate lengths." Electron Devices, IEEE Transactions on 46.7 (1999): 1429-1434.

[3] Liou, Juin Jei, Adelmo Ortiz-Conde, and Francisco Garcia-Sanchez. Analysis and design of MOSFETs: modeling, simulation, and parameter extraction. Springer Science & Business Media, 2012.

[4] Wong HS, White MH, Krutsick TJ, Booth RV. Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFET's. Solid-State Electron 1987;30:953.

[5] Garcia-Sanchez FJ, Ortiz-Conde A, Mercato GD, Salcedo JA, Liou JJ, Yue Y. New simple procedure to determine the threshold voltage of MOSFETs. Solid-State Electron 2000;44:673–5.

[6] Sánchez, FJ Garcıa, et al. "New simple procedure to determine the threshold voltage of MOSFETs." Solid-State Electronics 44.4 (2000): 673-675

[7] He J, Xi X, Chan M, Cao K, Hu C, Li Y, et al. Normalized mutual integral difference method to extract threshold voltage of MOSFETs. IEEE Electron Dev Lett 2002;23:428–30

[8] Ortiz-Conde A, Cerdeira A, Estrada M, Garcia Sanchez FJ, Quintero R. A simple procedure to extract the threshold voltage of amorphous thin film MOSFETs in the saturation region. Solid-State Electron 2001; 45:663–7.

[9] Tyagi, Sunit, et al. "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects." Electron Devices Meeting, 2000. IEDM'00. Technical Digest. International. IEEE, 2000.

[10] Ghani, Tahir, et al. "A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors." Electron Devices Meeting, 2003. IEDM'03 Technical Digest. IEEE International. IEEE, 2003.

[11] Tyagi, S., et al. "An advanced low power, high performance, strained channel 65nm technology." IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.. 2005.

[12] Auth, Chris, et al. "45nm high-k+ metal gate strain-enhanced transistors." VLSI Technology, 2008 Symposium on. IEEE, 2008.

[13] Packan, P., et al. "High Performance 32nm Logic Technology Featuring 2 nd Generation High-k+ Metal Gate Transistors." Electron Devices Meeting (IEDM), 2009 IEEE International. IEEE, 2009.