An Optimally Self-Biased Threshold-Voltage Extractor

Siew Kuok Hoon
Texas Instruments, Inc., 12500 TI Boulevard, Dallas, TX 75243, U.S.A.

Üğur Çilingiroğlu
Electrical Engineering Department, Texas A&M University, College Station, TX 77843, U.S.A.

Abstract—A novel threshold-voltage extractor architecture is presented. A differential-difference transconductor (DDT) loop automatically biases the device-under-test in continuous time around the inflection point of the \( \sqrt{I_D} \) vs \( V_{DS} \) characteristics. Another DDT operates as an arithmetic processor to precisely implement multiplication-by-2 and subtraction as needed for extrapolation. The extraction procedure thus complies entirely with all steps of the manual saturation method. With appropriate modifications, the architecture can also serve as an extractor implementing the linear method. The proposed architecture is applicable to both PMOS and NMOS on the same chip, and generates the value of \( V_T \) as a voltage with respect to the appropriate rail. It has been fabricated on silicon, and its accuracy has been experimentally verified by comparing automatically and manually extracted parameter values.

I. INTRODUCTION

Threshold voltage \( (V_T) \) extraction is indispensable to MOSFET device characterization and process monitoring [1]-[3]. Two techniques are widely practiced for this purpose. Namely, the saturation method and the linear method. The former relies on the strong-inversion characteristic

\[
\sqrt{I_D} = \frac{\beta}{2} (V_{GS} - V_T),
\]

of a MOSFET in saturation, and is applied in three steps as illustrated in Fig. 1: i) \( \sqrt{I_D} \) is plotted as a function of \( V_{GS} \); ii) The inflection point of this curve is determined from the maximum of its derivative. iii) The line tangential to the curve at this point is extrapolated to the axis of \( V_{GS} \), where the point of intersection yields \( V_T \) [1],[3].

What makes the second and third steps necessary is the fact that (1) is valid strictly for a limited range of intermediate values of \( V_{GS} \). \( \sqrt{I_D} \) is underestimated by (1) as the device enters weak inversion for smaller values of \( V_{GS} \), and overestimated as the mobility is degraded by larger values of \( V_{GS} \). The inflection point corresponds to the optimum bias condition for the validity of (1).

The linear method is based on the strong-inversion characteristic

\[
I_D = \beta (V_{GS} - V_T - V_{DS}/2) V_{DS},
\]

of a MOSFET in nonsaturation. The device is biased with a small \( V_{DS} \) of about 50 mV, and \( I_D \) is plotted as a function \( V_{GS} \). The second and third steps explained above are again applied to determine \( V_T - V_{DS}/2 \) from the intersection point.

Extractor circuits capable of determining \( V_T \) automatically have received considerable attention because they not only greatly facilitate device characterization and process monitoring, but also enable on-chip temperature sensing and reference voltage generation [4]-[6]. Most of the proposed extractors emulate the saturation method by biasing a device-under-test (DUT) and its quadrupled replica in saturation with identical currents, which results in \( V_{GS} = \sqrt{2I_D/\beta} + V_T \) for the DUT and \( V_{GS} = \sqrt{2I_D/4\beta} + V_T \) for the quad replica. \( V_T \) is then extracted by subtracting the \( V_{GS} \) of the DUT from \( 2 \times V_{GS} \) of the quad replica. This arithmetic operation needs a single-ended amplifier of gain 2, and a differential amplifier of gain 1. The published extractor configurations differ mainly in the way these precision amplifiers are designed [4],[7]-[10]. Some solutions compromise applicability to both PMOS and NMOS on the same chip; some do not provide \( V_T \) with reference to the appropriate rail; some others lack precision or need accurate input bias; and the one that offers precision lacks continuous-time operation. What is common to all, however, is that none incorporates the critical step of identifying the optimum bias point, without which the accuracy of the extracted value of \( V_T \) cannot be guaranteed. In this work, we present design details and experimental evaluation of an extractor circuit which implements all steps of the saturation method without suffering from any of these drawbacks. With minor modifications, it can also implement the linear method. Conceptual precursors and some simulated evaluation of this extractor was published in [11].

![Graphical illustration of the saturation method of \( V_T \)-extraction. The maximum of \( \partial \sqrt{I_D}/\partial V_{GS} \) defines the optimum bias point.](image)

0-7803-7057-0/01/$10.00 ©2001 IEEE.
II. CONCEPT AND ARCHITECTURE

The architecture of the proposed extractor is shown in Fig. 2 for the case of NMOS threshold voltage. It is equally applicable to PMOS with appropriate reversal of polarities. We use three identical diode–connected DUT’s, which receive bias currents $I$, $4I$, and $9I$, and develop the gate-source voltages $V_{GS1}$, $V_{GS2}$, and $V_{GS3}$, respectively. The equality of $\sqrt{9I} = \sqrt{4I} + \sqrt{I}$ makes the difference $V_2 = V_{GS3} - V_{GS2}$ to be equal to the difference $V_1 = V_{GS2} - V_{GS1}$, if all three DUT’s are optimally biased; that is, if they operate on the linear segment of the characteristic. If the bias is higher than the optimum, then $V_2$ exceeds $V_1$. A lower than optimum bias, on the other hand, results in $V_1$ being larger than $V_2$. A capacitively loaded differential–difference transconductance [13] (marked DDT1 in Fig. 2) continuously senses these two differential voltages, and generates an output current $I_O = G_m(V_2 - V_1)$, whose output voltage, which is proportional to the time integral of $I_O$, controls the unit current $I$ of the three bias current sources without affecting the scaling factors 1:4:9. The resulting negative feedback loop forces $I_O$, hence $V_2 - V_1$, to vanish in steady state. Therefore, the DUT’s are automatically biased around the inflection point of the $\sqrt{I}$ vs $V_{GS}$ characteristic.

Another capacitively loaded DDT (marked DDT2 in Fig. 2) operating under self–feedback and sensing $V_{GS1}$ and $V_{GS2}$, sets its output voltage to satisfy $V_O = V_{GS1} - V_{GS2}$ in steady state. Since $V_{GS1} = \sqrt{4I} + V_T$, and $V_{GS2} = \sqrt{8I} + V_T$, the output of DDT2 directly yields $V_T$.

III. CIRCUIT DESIGN

A schematic of the extractor is given in Fig. 3. The three diode–connected transistors marked M represent the DUT’s. M1 supplies the unit current $I$. Four and nine units of this device are combined to construct the scaled current sources M2 and M3, respectively. All are cascoded to desensitize the current scaling factors against bias conditions. The most severe compliance condition occurs on the $9I$ branch. Assuming that M3 and its cascode are geometrically identical, this condition imposes the following constraint

$$\sqrt{\beta_3/\beta} > 2 \frac{V_{GS3\text{\,(max)}} - V_T}{V_{DD} - V_{SS} - V_{GS3\text{\,(max)}}},$$

which is used for sizing M3 based on the estimates of $\beta$, $V_T$ and $V_{GS3\text{\,(max)}}$ of the DUT. M1, M2 and their cascodes are then scaled down according to the 1:4:9 rule.

The bias optimizing DDT1 comprises a differential–difference input stage, whose drivers are marked M4, and a gain stage, whose driver is M5. The active load of the input stage is self-cascode for minimizing the compliance voltage because its output voltage has to remain close to $V_{GS}$ due to the essentially small $V_{GS1}$. The design of DDT1 is guided mainly by the following three constraints:

- The bias voltage at the gate of M5 must be small enough to prevent the leftmost M4 from exiting saturation even for the smallest expected value of $V_{GS1}$.
- Differential input range of the input stage must be wide enough to handle the largest expected value $\Delta V_{GS\text{\,(max)}}$ of $V_{GS3} - V_{GS2}$ and $V_{GS2} - V_{GS1}$. Hence,

$$\sqrt{2I_{B1}}/\beta_b \leq V_{GS1\text{\,(min)}},$$

for comparable PMOS and NMOS threshold voltages.

- The bias controlling loop must be stable. Including the DUT’s and their current sources, the loop comprises three gain stages, and two high–impedance nodes marked N1 and N2 in Fig. 3. We apply frequency compensation by placing a Miller capacitor $C_s$ across M5, whose capacitance is approximately equal to the capacitance of node N2. The dc magnitude of the open–loop gain is given by

$$A(0) = g_m \frac{r_{on\text{\,(in)}}}{r_{on\text{\,(out)}}},$$

where $g_m$ is the transconductance of M5, $r_{on\text{\,(in)}}$ and $r_{on\text{\,(out)}}$ are the output resistance of the DDT1 input and output stages, respectively, and $g \equiv i_{o1}/v_o$ (see Fig. 3) is the transconductance of the DUT/current–source stage and the input stage of DDT1 combined. It is described by

$$g = \sqrt{\beta_1 \beta_3 I_{B1}} \frac{3}{2} \left( \frac{3}{a_3} + \frac{1}{a_1} - \frac{4}{a_2} \right),$$

where $a_1$, $a_2$, and $a_3$ are the transconductance gain parameters of the DDT1.

![Fig. 2. The architecture of the proposed extractor.](image)

![Fig. 3. The circuit schematic of the proposed extractor.](image)
where, $a_1$, $a_2$ and $a_3$ denote the slope of the $\sqrt{I_D}$ vs $V_{GS}$ characteristic of the DUT for the three optimized bias voltages $V_{GS1}$, $V_{GS2}$ and $V_{GS3}$, respectively. Note from Fig.1 that $a_1$ and $a_2$ are slightly smaller than $a_3$ around the inflection point. Therefore, $g$ is positive but small. The dominant pole is located at

$$\omega_{p1} = \frac{1}{r_{o(out)}} \frac{g_{m5}}{r_{o(out)} C_c},$$

whereas the second pole and the right-half plane zero coincide at

$$\omega_{p2} = \omega_{p0} = \frac{g_{m5}}{C_c}. $$

Therefore, stability with a minimum phase margin $\phi$, necessitates

$$g \leq g_{m5} \tan \left(\frac{90 - \phi}{2}\right) = \sqrt{2I_{DS} \beta_5} \tan \left(\frac{90 - \phi}{2}\right).$$

The extrapolating DDT2 also comprises a differential–difference input stage and a gain stage. It has a simpler configuration which can handle an output as low as $V_{GS}$. Since the output is terminated at the measurement pin with a very large parasitic capacitance, no additional measure is necessary for the stability of DDT2.

**IV. Test Chip and Experimental Results**

We have designed and fabricated a test chip in 1.2μm CMOS to verify the proposed concept of automatic $V_T$ extraction. The chip contains both NMOS and PMOS extractor circuits, and each circuit is capable of processing two sets of DUT's of different size. Each set of DUT's is digitally selectable by means of a control bit. We also included in close proximity of these DUT's stand-alone replicas for manual $V_T$ extraction purposes. Photomicrographs of the two extractor circuits are presented in Fig.4.

In Table 1, we compare the experimental results generated automatically by the extractor with those obtained by applying the saturation method manually on the replicas. Also compared in the same table are the bias values set automatically by the extractor and determined manually. The results indicate that the difference between the two sets of threshold voltages is typically 3%, increasing to less than 5% in the case of wider NMOS device. For a further evaluation of extractor performance, we have repeated these tests at different temperatures within the range $-50 \leq T(°C) \leq 100$.

**Table 1: Experimental Values of $V_{GS2}$ and $V_T$**

<table>
<thead>
<tr>
<th>Device (µm)</th>
<th>Manual $V_{GS2}$ (mV)</th>
<th>Auto $V_{GS2}$ (mV)</th>
<th>Manual $V_T$ (mV)</th>
<th>Auto $V_T$ (mV)</th>
<th>$\Delta V_T/V_T$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMOS 9.6/1.8</td>
<td>900.0</td>
<td>869.0</td>
<td>568.1</td>
<td>560.4</td>
<td>1.4</td>
</tr>
<tr>
<td>NMOS 19.2/1.8</td>
<td>903.0</td>
<td>901.0</td>
<td>582.9</td>
<td>584.8</td>
<td>4.9</td>
</tr>
<tr>
<td>PMOS 9.6/1.8</td>
<td>1230.0</td>
<td>1474.0</td>
<td>-850.8</td>
<td>-874.6</td>
<td>2.8</td>
</tr>
<tr>
<td>PMOS 19.2/1.8</td>
<td>1250.0</td>
<td>1737.0</td>
<td>-855.8</td>
<td>-877.8</td>
<td>2.4</td>
</tr>
</tbody>
</table>

The result plotted in Fig.5, indicates an automatically extracted temperature coefficient $-1.6$ mV/°C and $-1.7$ mV/°C for NMOS and PMOS, respectively. These are in close agreement with the manually extracted sensitivities $-1.5$ mV/°C and $-1.6$ mV/°C, respectively. Finally, the sensitivity of the extracted value of $V_T$ to power supply voltage has been experimentally evaluated. The results are presented in Fig.6. Both NMOS devices and the narrow PMOS device exhibit a relative sensitivity around 0.7 percent per volt, whereas the wider PMOS has a smaller sensitivity around 0.1 percent per volt.

![Fig.4. PMOS and NMOS extractor photomicrographs.](image)

**Fig.4.** PMOS and NMOS extractor photomicrographs.

**Fig.5.** Variation of $V_T$ with temperature. Upper plot: NMOS. Lower plot: PMOS.
V. APPLICATION WITH CONSTANT DRAIN–SOURCE BIAS AND EXTENSION TO THE LINEAR METHOD

Using diode-connected DUT’s, as done in almost all extractors, forces the DUT’s to operate under different values of drain–source bias, which in turn, disturbs the identity of \( V_D \) vs \( V_GS \) characteristics due to channel–length modulation effect. This can be overcome by keeping the \( V_D \) of all DUT’s at the same value by introducing the OTA–based virtual biasing scheme shown in Fig.7. Even a low-gain simple single-stage differential OTA would do. The very same scheme enables the proposed architecture to be extended to the implementation of the linear method of extraction. The only modification necessary is a rescaling of the three bias–current sources from 1:4:9 to 1:2:3. While the OTA’s keep all three DUT’s biased at the same small \( V_DS \) and DDT1 optimizes the bias conditions, DDT2 will generate \( V_T + V_{DD}/2 \) from \( V_GS1 = 1/3V_DS + V_T + V_DS/2 \) and \( V_GS2 = 2I/V_DS + V_T + V_DS/2 \).

VI. CONCLUSIONS

A continuous-time threshold–voltage extractor architecture has been developed to incorporate optimum self-biasing of the device-under-test. The extraction procedure is thus made to comply with all steps of the manual saturation method. With appropriate modifications, the architecture can also implement the linear method. Another novelty is the introduction of a simple differential–difference transconductor as a precise arithmetic processor to implement multiplication by 2 and subtraction as needed for the final step of extraction. The proposed architecture is applicable to both PMOS and NMOS on the same chip, and generates the value of \( V_T \) as a voltage with respect to the appropriate rail. It has been implemented on silicon, and its accuracy has been experimentally verified by comparing automatically and manually extracted parameter values.

REFERENCES


