Grain-Orientation Induced Work Function Variation in Nanoscale Metal-Gate Transistors—Part II: Implications for Process, Device, and Circuit Design

Hamed F. Dadgour, Student Member, IEEE, Kazuhiko Endo, Member, IEEE, Vivek K. De, Senior Member, IEEE, and Kaustav Banerjee, Senior Member, IEEE

Abstract—This paper investigates the process, device, and circuit design implications of grain-orientation-induced work function variation (WFV) in high-

k/metal-gate devices. WFV is caused by the dependence of the work function of metal grains on their orientations and is analytically modeled in the companion paper (part I). Using this modeling framework, various implications of WFV are investigated in this paper. It is shown that process designers can utilize the proposed models to reduce the impact of WFV by identifying proper materials and fabrication processes. For instance, four types of metal nitride gate materials (TiN and TaN for NMOS devices and WN and MoN for PMOS devices) are studied, and it is shown that TiN and WN result in lower $V_{th}$ fluctuations. Moreover, device engineers can study the impact of WFV on various types of classical and nonclassical metal-gate CMOS transistors using these analytical models. As an example, it is shown that, for a given channel length, single-fin FinFETs are less affected by WFV compared to fully depleted SOI and bulk-Si devices due to their larger gate area. Furthermore, circuit designers can benefit from the proposed modeling framework that allows straightforward evaluation of the key performance and reliability parameters of the circuits under such $V_{th}$ fluctuations. For instance, an SRAM cell is analyzed in the presence of $V_{th}$ fluctuations due to WFV, and it is shown that such variations can result in considerable performance and reliability degradation.

Index Terms—Grain orientation, metal-gate devices, random variations, reliability, subthreshold leakage, threshold voltage, VLSI design, work function variation (WFV).

I. INTRODUCTION

MATERIALS HAS become the primary gate material in advanced CMOS technologies due to the incompatibility of polysilicon with high-$k$ materials [1], [2]. However, using metal as the gate material introduces a new source of random variation due to the dependence of the work function on the orientation of metal grains [3], [4]. Hence, it is critical to model such phenomenon in order to guide the materials selection and fabrication process, as well as to provide a quantitative (yet efficient) way to assess their impact on device/circuit performance and reliability, which can, in turn, be used to optimize nanoscale circuits. In the companion paper (part I), a comprehensive statistical framework is presented for modeling work function variation (WFV) in emerging high-$k$/metal-gate devices. Using such a modeling approach, this paper investigates the process, device, and circuit implications of WFV. More specifically, the proposed model is employed to identify appropriate materials and process conditions that can minimize the impact of WFV. Using the physical insight from the analytical model, several possible methods for reducing the adverse effects of WFV are discussed, and the limitations of each approach are explored. At the device level, impact of WFV are evaluated on bulk, fully depleted (FD)-SOI, and FinFET devices compared to other sources of random variations. It is shown that, for undoped-body devices (FD-SOI and FinFET), WFV is the dominant source of random variations. The proposed modeling framework is also used to investigate the implications of WFV on the key circuit-level performance and reliability parameters. In this paper, it is shown that various SRAM failure mechanisms are adversely affected by WFV, and hence, grain-orientation-induced $V_{th}$ variations must be taken into consideration in future nanoscale circuit designs.

The rest of this paper is organized as follows. Section II provides a brief summary of the proposed model, and Section III investigates the impact of WFV on fabrication process choices. Section IV studies the device-level implications of WFV, and Section V highlights the importance of WFV for the performance and reliability of SRAM cells. Finally, Section VI summarizes this paper.

II. PROPOSED MODEL FOR WFV

In deep nanoscale technologies, the gates of CMOS devices are composed of only a few grains with a random distribution of orientations as shown in Fig. 1. Therefore, since each grain orientation has a different work function value, the work function of the entire metal gate must be modeled as a probabilistic distribution rather than a deterministic value. In the companion paper (part I), it is shown that the mean $(E(\Phi_M))$ and standard deviation $(\text{var}(\Phi_M))$ of the work function...
distribution can be evaluated using (1) and (2), where $N$ is given by (3)

$$E(\Phi_M) = \sum_{i=1}^{r} P_i \Phi_i$$

$$\text{var}(\Phi_M) = \frac{1}{N} \left[ \sum_{i=1}^{r} P_i \Phi_i^2 - \left( \sum_{i=1}^{r} P_i \Phi_i \right)^2 \right]$$

$$N = \left( \frac{L}{G} \right) \times \left( \frac{W}{G} \right).$$

Here, the symbols $\Phi_1, \Phi_2, \ldots, \Phi_n$ and $P_1, P_2, \ldots, P_n$ are used to identify the work function values of grains with different orientations and their corresponding probabilities (percentage share of a particular grain orientation in the total population of grains averaged over a large number of gates). It is also assumed that the grain size ($G$) of each type of metal film is known.

Hence, for a transistor with a gate length of $L$ and width of $W$, the total number of grains ($N$) within the metal-gate area can be calculated as $\left( \frac{L}{G} \right) \times \left( \frac{W}{G} \right)$.

III. IMPLICATIONS OF METAL-GATE WFV FOR PROCESS ENGINEERING

The standard deviation of WFV predicted by (2) depends on the various material properties of metal gates and process conditions under which devices are fabricated. As it will be shown in this section, it is possible to minimize the impact of WFV by choosing proper materials and fabrication conditions.

A. Implications of Different Gate Metals on WFV

The metal elements used for metal-gate devices must satisfy a number of criteria including thermal stability and a suitable work function [5]. For instance, in order to obtain desirable threshold voltage levels, the work function of the metal must be within appropriate range from the conduction and the valence bands of silicon. In other words, the metal work functions for NMOS and PMOS devices must be in ranges $4.0–4.5$ eV and $4.8–5.3$ eV, respectively. Few selected metals are plotted in Fig. 2 against their respective work function values, where the top/bottom highlighted horizontal stripe indicates the metal elements that fall into appropriate work function ranges for the PMOS/NMOS device. Moreover, the metal elements are not usually employed in their pure form due to their low thermal stability. As an alternative, metal nitrides have been used in the fabrication of metal gates. In this paper, four metal materials are chosen for further investigations: tantalum nitride (TaN) and titanium nitride (TiN) for NMOS devices and molybdenum nitride (MoN) and tungsten nitride (WN) for PMOS transistors, as they are among the most common gate electrode materials.

It should be noted that nitrogen tends to change the physical properties of metals such as the grain size and the work function; however, a lower percentage of nitrogen in the metal nitride material (for example, 1:10) has a negligible impact on these properties.

The important physical properties of metal nitrides, which have been used in this work, are summarized in Table I. It should be noted that theoretical studies show that the gate work function could vary significantly with the metal’s microstructure, the metal/dielectric interface chemistry, and even the gate dielectric properties such as its thickness [7], [8]. The reported values in Table I are from different sources, where each experiment had its own unique settings, and hence, there might be variations from those values if a different approach or setting is chosen. These particular values for material properties (such as work function and grain size) are borrowed from the literature and are used only as examples. Different process/temperature conditions will certainly alter the aforementioned values but do not affect the accuracy of our model in any way. In other words, the proposed model can be used to evaluate the impact of WFV regardless of the process/temperature conditions. Finally, it should be noted that, in multiple metal-layer gates, the WFV of the layer in contact with the oxide is the most important factor affecting the random $V_{th}$ fluctuations of transistors.

Using the data from Table I and (2), the work function fluctuations due to random orientations of the metal grains are evaluated for different technology nodes in Fig. 3. In this figure, the WFV of NMOS devices with a metal gate material of TiN and TaN is plotted, along with the WFV for the PMOS devices that employ MoN- and WN-based gates. In these simulations, the sizing for all transistors is assumed to be $3L_{\text{min}} \times 2L_{\text{min}}$, where $L_{\text{min}}$ is the minimum channel length allowed in the technology node. The reason for choosing $L > L_{\text{min}}$ is to reduce leakage and to improve the reliability of devices as it is often the case in realistic industrial designs [20]. It can be observed that, as technology scales, the relative importance of WFV increases. This is due to the fact that larger devices have significantly more grains (the gate area increases, whereas the grain size is constant), and hence, the work functions of these devices are more likely to exhibit a smaller range of variations due to the averaging nature of the work function (1).

An important insight from Table I is the fact that $V_{th}$ fluctuation is a strong function of the choice of metal material. MoN has the highest work function fluctuation because its two grain...
Table 1

<table>
<thead>
<tr>
<th>Material</th>
<th>Orientation</th>
<th>Probability</th>
<th>Work function (eV)</th>
<th>Grain size (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;111&gt;</td>
<td>40%</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>TaN</td>
<td>&lt;100&gt;</td>
<td>50% [12]</td>
<td>4.0 [13]</td>
<td>7 [14]</td>
</tr>
<tr>
<td></td>
<td>&lt;200&gt;</td>
<td>30%</td>
<td>4.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;220&gt;</td>
<td>20%</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;200&gt;</td>
<td>15%</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;220&gt;</td>
<td>15%</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;311&gt;</td>
<td>5%</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>MoN</td>
<td>&lt;110&gt;</td>
<td>60% [18]</td>
<td>5.0 [19]</td>
<td>17 [16]</td>
</tr>
<tr>
<td></td>
<td>&lt;112&gt;</td>
<td>40%</td>
<td>4.4</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Standard deviation values of WFV in NMOS (TaN and TiN) and PMOS (MoN and WN) devices.

orientations have relatively close probabilities (60% and 40% as shown in Table I) and the difference between their work functions is substantial (5.0 and 4.4 eV). Additionally, the WFV corresponding to NMOS devices with TiN gate metal seems to become flat for technology nodes smaller than 32 nm (Fig. 3). This is due to the fact that, below the 32-nm node, the device is so small that one grain (TiN grain size is 22 nm) can almost cover the entire gate surface area, and hence, WFV becomes independent of the technology node.

It should be noted that the result of all previous simulations is a strong function of the composition of the metal gate (percentage of grains with different orientations), which itself depends on the growth condition. In order to study the impact of the varying composition of grains with different orientations, the $V_{th}$ variation of TiN- [Fig. 4(a)] and MoN-based [Fig. 4(b)] metal gates is considered. In these simulations, the ratio of the preferred grain orientation (for instance, ⟨200⟩ in the case of TiN) is varied from 50% to 100% (on the Y-axis) for different technology nodes (X-axis), and each contour corresponds to a specific level of $V_{th}$ variation. For example, it can be observed that, for the TiN device, to limit the $V_{th}$ variation to 10 mV in 32-nm technology, the ratio of the preferred grain must be at least 90%, whereas for the MoN device, this ratio has to be no less than 95%. Hence, one of the effective ways to limit the $V_{th}$ fluctuations is to improve the quality of the metal growth process, i.e., to increase the ratio of the preferred orientation.

B. Reducing the Impact of WFV

According to the proposed model (2), there are several possible strategies to lower the impact of WFV [in other words, var(ΦM)].

1) One can achieve this goal by manipulating the process conditions, particularly the temperature, to increase $N$. It is a known fact that the deposition of metal at lower temperatures reduces the grain size and, hence, increases $N$. However, the benefit of such an approach is limited by the thermal budget dictated by fabrication steps, which follow the deposition of metal [21].

2) Another approach is to choose metals that have more uniform grain orientation distributions. Such metal gates are primarily composed of grains with an identical orientation. For example, if $P_1$ is the probability of having a grain with the preferred orientation, to lower var(ΦM) in (2), one must have the following: $P_1 \gg P_2, P_1 \gg P_3, \ldots, P_1 \gg P_i$. While it is believed that such grain orientation distributions can be realized by manipulating the process conditions, currently, there is no known systematic method to accomplish this.

3) It is also possible to employ metal elements whose different grain orientations offer more or less comparable work function values. This means that, in (2), we should have $\Phi_1 \approx \Phi_2 \approx \cdots \approx \Phi_i$. In this fashion, even if the metal gate is composed of a mixture of various grain orientations, the overall work function (ΦM) will not have large deviations.
The threshold voltage of a MOS device is a linear function of the gate work function [24]. Therefore, the randomness in their WF values do not fluctuate.

Considering approaches 1) and 2) mentioned previously, one can explain the reason for the insignificant impact of WFV in transistors with poly-Si gates. First, taking into account the average size of silicon grains (5–15 nm) [23] and large gate area (∼100 nm), the number of grains is sufficient enough (∼N > 50) to make WFV irrelevant for such devices according to (2). Furthermore, it is experimentally shown that, in deposited polysilicon films, the percentage of the preferred grain orientation is much higher than that of other orientations, and hence, WFV does not considerably affect poly-Si-gate devices [23]. Additionally, the Fermi levels in the grains remain pinned due to degenerate doping of poly silicon and as a result, their WF values do not fluctuate.

4) The last solution is to grow gates with amorphous metal materials, which exhibit disordered atomic-scale structures. In contrast to most metals, which are crystalline and, therefore, are composed of grains with various orientations, amorphous alloys are noncrystalline and do not have distinguishable grains [22]. As a result, the WFV is almost zero for such gates because the number of grains [N in (2)] is virtually infinite. However, one should note that amorphous metal gates come at the cost of higher sheet resistances, and hence, their applicability for real circuits is limited.

In order to investigate their relative importance, three major sources of the random variations (RDF, LER, and WFV) are shown in Fig. 5(a) for different technology nodes. In this figure, for RDF and LER random variations, the models presented in [25] and [26] have been used, and $V_{th}$ fluctuation due to WFV is evaluated using (2)–(4) and the data from Table I. Note that all devices are considered to be sized as $3L_{min} \times 2L_{min}$ with TiN metal gate, and other device properties are adopted from [27]. According to this figure, the effect of WFV can be as high as that due to RDF for TiN gates. However, it is important to note that the composition of the metal gate for this simulation is considered to be 60% (200) and 40% (111), which might not always be the case. To reduce the WFV, the fabrication process can be improved to yield more uniform (higher percentage of the preferred orientation) metal compositions. It is also interesting to analyze the impact of all sources of the random $V_{th}$ variation on three different types of metal-gate-based CMOS transistors: bulk, FD-SOI, and FinFET devices. For bulk devices, it is assumed that all three sources of the random $V_{th}$ variations are present and are mutually independent. Therefore, the total random $V_{th}$ variation can be evaluated as

$$\sigma_{V_{th\_WFV}} = \sigma_{V_{th\_WFV}}$$

where $\sigma$ represents the standard deviation. The threshold voltage of MOS transistors also exhibits random variations mainly due to two other sources: random dopant fluctuation (RDF) and line-edge roughness (LER). RDF represents the random $V_{th}$ variations due to arbitrary and uncontrollable number and placement of dopant atoms in the channel area during the fabrication process [25]. LER refers to the $V_{th}$ fluctuation that is caused by irregularity of the edge of the poly gate bordering the drain/source regions [26].

In the case of FD-SOI transistors, because of the presence of an undoped body, there is no $V_{th}$ fluctuation due to RDF. Therefore, only two terms are present in the formula of the total random $V_{th}$ variation for the FD-SOI devices

$$\sigma_{V_{th\_FD-SOI}} = \sqrt{\sigma_{V_{th\_WFV}}^2 + \sigma_{V_{th\_LER}}^2}.$$
Similarly, FinFET devices do not get affected by RDF. Furthermore, the gate area for each FinFET is twice that of a bulk MOSFET or an FD-SOI transistor (for a given channel length and FinFET height equal to the width of planar devices) because FinFET devices are double-gate structures. Therefore, the $V_{th}$ fluctuation for a FinFET device (with single fin) due to the randomness of grain orientations, as well as LER, is approximately half of the variation for the same-sized bulk transistor (the standard deviation of WFV is inversely proportional to the number of grains, i.e., the area of the gate). Hence, the total random $V_{th}$ variation for a FinFET device would be

$$
\sigma_{V_{thFinFET}} = \sqrt{0.5 \times \sigma_{V_{thLER}}^2 + 0.5 \times \sigma_{V_{thWFV}}^2}.
$$

Fig. 5(b) shows the total random $V_{th}$ variations for the bulk, FD-SOI, and FinFET NMOS devices evaluated by (5)–(7), respectively. In these simulations, all devices are assumed to be TiN metal gates in the 65-nm technology node and sized to be $3L_{min} \times 2L_{min}$. Moreover, it is assumed that the channel lengths are identical for all transistors and the heights of the FinFET devices are equal to the width of the planar transistors. As it can be observed, the random variation for the bulk devices is the highest due to the presence of RDF. The FD-SOI transistors also exhibit a higher random $V_{th}$ fluctuation compared to FinFETs because of the lower impact of WFV in FinFETs.

**B. Impact of WFV on Subthreshold Slope**

A subthreshold slope ($S$) is defined as the inverse of the slope of the $\log(I_D)$–$V_{GS}$ curve in the subthreshold region, where $I_D$ is the drain current of the device and $V_{GS}$ denotes the gate–source voltage difference. In other words, the subthreshold slope is the voltage (in millivolts) required to increase $I_D$ by one decade in the subthreshold region of operation. The subthreshold slope in CMOS devices can be calculated using the formula shown in [28]

$$
S = \left( \frac{\partial \log(I_D)}{\partial V_{GS}} \right)^{-1} = \ln 10 \times \frac{kT}{q} \left( 1 + \frac{C_d}{C_{ox}} \right)
$$

where $k$, $T$, and $q$ are the Boltzmann constant, the temperature, and the electron charge, respectively. $C_d$ and $C_{ox}$ denote the depletion layer and the dielectric capacitance, respectively. Since all the parameters in (8) except $C_d$ are independent of the metal work function, the impact of WFV on $S$ can be evaluated by calculating the variations of $C_d$. In order to achieve this goal, one should understand the physical origin of the depletion capacitance ($C_d$). In Fig. 6(a), the band diagram of a metal-gate NMOS device is shown, where $E_C$, $E_i$, $E_{FS}$, $E_{FM}$, and $E_V$ denote the bottom of the conduction band, the intrinsic energy level in silicon, the Fermi level in silicon, the Fermi level in metal, and the top of the valence band, respectively. The work function values of the metal and silicon are represented by $\Phi_M$ and $\Phi_S$, respectively, and the work function difference between metal and silicon is indicated as $\Phi_{MS} (= \Phi_M - \Phi_S)$. Furthermore, $V_{ox}$ refers to the voltage drop across the dielectric material.

In Fig. 6(a), the surface potential (the potential at the interface of the dielectric and silicon) is $\psi_S$, whereas $\psi_B$ denotes the difference between the $E_{FS}$ and $E_i$ energy levels. In order to determine $C_d$, the value of $\psi_s$ must be calculated since $C_d$ is a function of $\psi_s$. At the boundary of strong and weak inversion conditions, assuming $\psi_s = 2\psi_B$, the gate voltage bias ($V_{GS} = V_{th0}$) that is required to create such a condition can be shown as [Fig. 6(c)]

$$
V_{GS} = V_{FB} + V_{ox} + \psi_s
$$

where

$$
V_{FB} = \phi_{MS} - \frac{Q_f}{C_{ox}}.
$$
In (10), \( Q_f \) represents the fixed charges at the interface of the dielectric and silicon. As shown in Fig. 6(b), \( V_{ox} \) is the voltage drop across the dielectric capacitance \( C_{ox} = \varepsilon_0 \varepsilon_{ox}/d \) due to charges in the depletion region \( (Q_d) \), which can be estimated as \( qN_Aw \), where \( q \) is the charge of an electron, \( N_A \) is the doping concentration of the substrate, and \( w \) denotes the depth of the depletion layer [28]. Note that in the subthreshold region, the inversion charge, \( Q_i \), is negligible and the effective gate capacitance is equal to the series combination of \( C_{ox} \) and \( C_d \).

The value of \( w \) can be calculated using (11), where \( \varepsilon_0 \) and \( \varepsilon_S \) are the permittivity of vacuum and the relative permittivity of silicon, respectively

\[
w = \sqrt{\frac{2\varepsilon_0\varepsilon_S\psi_S}{qN_A}}. \tag{11}
\]

Using (11), the formulas for \( V_{ox} \) and \( V_{GS} \) can be derived as (12) and (13), respectively

\[
V_{ox} = \frac{Q_d}{C_{ox}} = \frac{qN_Aw}{C_{ox}} = \sqrt{2\varepsilon_0\varepsilon_SqN_A\psi_S} \tag{12}
\]

\[
V_{GS} = V_{FB} + \frac{\sqrt{2\varepsilon_0\varepsilon_SqN_A\psi_S}}{C_{ox}} + \psi_S. \tag{13}
\]

Equation (13) is a quadratic equation in terms of the square root of \( \psi_S \) and has only one plausible solution as shown in

\[
\psi_S = \left[ \frac{1}{2} \times \left( -\frac{\sqrt{2\varepsilon_0\varepsilon_SqN_A}}{C_{ox}} \right) \right]^2 + \left[ \frac{\sqrt{2\varepsilon_0\varepsilon_SqN_A}}{C_{ox}} - 4 \left( \frac{\phi_{MS} - Q_f}{C_{ox}} - V_{GS} \right) \right]^2. \tag{14}
\]

In the presence of WFV, for a given gate bias \( (V_{GS}) \), all parameters in (14) are deterministic values except for \( \phi_{MS} (= \Phi_M - \Phi_S) \) which must be modeled as a random variable due to \( \Phi_M \) fluctuations. The variations of \( \Phi_M \) result in \( \psi_S \) fluctuations [according to (14)], which, in turn, can cause the deviations of \( w \) [shown by (11)]. Finally, the variations of \( w \) lead to \( C_d \) fluctuations (since \( C_d = \varepsilon_0 \varepsilon_S/w \)), which, according to (8), can be used to determine the deviations of the subthreshold slope \( (S) \) of devices.

A Monte Carlo simulation with 10000 samples is used to demonstrate the impact of WFV on the probability distribution of the subthreshold slope (Fig. 7). The list of parameters that have been considered in this simulation is shown in Fig. 7(a). For these simulations, the mean value of the metal work function \( \phi_{MS} \) and its standard deviation (or the level of WFV) are assumed to be 4.2 eV and 50 meV, respectively. The distribution of the subthreshold slope is shown in Fig. 7(b). As it can be observed, the standard deviation of the distribution is 0.1% \((100 \times 0.07/84.2 \approx 0.1)\) of its mean value. These simulations indicate that the impact of WFV on the subthreshold slope is negligible, and hence, the work function engineering of metal gates cannot be effective in reducing the subthreshold slope of CMOS devices.

V. PERFORMANCE AND RELIABILITY ANALYSIS OF SRAM CELLS CONSIDERING WFV

Using the proposed model, it is possible to evaluate the impact of the increased level of random variation due to the contribution of the metal work function fluctuation on the key characteristics of an SRAM cell (shown in Fig. 8). The SRAM cell is one of the most sensitive circuits to the random \( V_{th} \)
variations since it is typically composed of small-sized devices. The sizing of devices for the SRAM cell is chosen to yield $\beta \approx 1.33$ (see Fig. 8), which results in a mean noise margin of $\approx 200 \text{ mV}$ at $V_{dd} = 1.1 \text{ V}$ and $C_L = 50 \text{ fF}$ [29]. Note that, to consider a realistic SRAM cell design, $L$ is selected to be larger than $L_{\text{min}}$. In this analysis, the gate material is TiN.

In this paper, the reliability and standby leakage power consumption of SRAM cells are investigated under the influence of WFV using a sensitivity-based analysis.

A. Implication of WFV for SRAM Cell Reliability

To investigate the impact of WFV on the reliability of SRAM cells, four failure mechanisms (write, read, hold, and access failures) are considered. Write failure occurs when the time needed to write the desired data to the SRAM exceeds the design specifications. Read failure refers to a case where, during the read operation, the content of the SRAM cell is flipped (the read operation of the SRAM cell must be nondestructive). Such a situation happens when the voltage at node $Q_R$ (Fig. 8) exceeds the tripping voltage (input voltage for which the output of an inverter switches) of the left-hand-side inverter (for $BL = BLB = 1$). Additionally, hold failure occurs during the idle mode when the supply voltage is reduced to decrease leakage power consumption. Lower $V_{dd}$, the static noise margin of the cell decreases, and hence, the content of the SRAM can be flipped by noise, resulting in a hold failure. Finally, access failure refers to a scenario where the time required to read the content of the SRAM cell exceeds the maximum value allowed by the design requirements.

In this method, the assumption is that the $V_{th}$ fluctuations of all transistors can be regarded as mutually independent. This is a safe assumption in small-sized devices since the impact of random variation is dominant compared to that of systematic variations. It is also assumed that, due to a relatively small level of $V_{th}$ fluctuations, the variation of SRAM cell characteristics [such as write delay (WD)] can be estimated by a first-order Taylor expansion of that function around its nominal value

$$WD = WD_0 + \frac{\partial WD}{\partial V_{th,AR}} \Delta V_{th,AR} + \frac{\partial WD}{\partial V_{th,AL}} \Delta V_{th,AL} + \cdots + \frac{\partial WD}{\partial V_{th,PL}} \Delta V_{th,PL}$$

(15) where $WD_0$ represents the nominal value of the write delay, $\Delta V_{th,AR}$ represents the random variation of $V_{th,AR}$ for the specific device $X$ (Fig. 8), and $\partial WD/\partial V_{th,X}$ represents the sensitivity of the write delay to the $V_{th}$ fluctuation of that device and can be obtained from circuit simulations. Since the range of the threshold voltage variation is small (< 50 mV), this function can be estimated by only the first-order Taylor expansion, ignoring the higher order terms. Using (15) and assuming that the variations of all threshold voltages are normally distributed, one can further assume that the distribution of the write delay value (or any other metric) will also be a Gaussian distribution

$$WD \sim N(\mu_{WD}, \sigma_{WD})$$

(16)

where the mean and standard deviation values of the distribution can be obtained using (17) and (18), respectively, shown at the bottom of the page. In (18), similar to (15), $\partial WD/\partial V_{th,X}$ represents the sensitivity of the write delay to the variation in $V_{th,AR}$, and $\sigma V_{th,AR}$ shows the standard deviation of the $V_{th}$ of the device.

In order to evaluate the probability of write failure, first, the distribution of write delay (WD) for the SRAM cell must be estimated under random $V_{th}$ variation. To obtain this distribution, the sensitivities of the WD with respect to the threshold voltage of all transistors can be determined using circuit simulations [Fig. 9(a)]. Once the sensitivities are extracted, the WD distribution can be plotted using (15)–(18) [Fig. 9(b)]. To validate the

$$\mu_{WD} = WD_0$$

$$\sigma_{WD} = \sqrt{\left(\frac{\partial WD}{\partial V_{th,AR}} \sigma V_{th,AR}\right)^2 + \left(\frac{\partial WD}{\partial V_{th,AL}} \sigma V_{th,AL}\right)^2 + \cdots + \left(\frac{\partial WD}{\partial V_{th,PL}} \sigma V_{th,PL}\right)^2}$$

(18)
Fig. 10. (a) Read failure occurs when the voltage on the $Q_R$ node ($V_{QR}$) exceeds the trip voltage of the inverter ($V_{Trip}$) as indicated by the hashed area. (b) Hold failure occurs when the hold margin is less than zero (hashed area).

Fig. 11. (a) Access failure occurs when the read delay is higher than the maximum allowed delay (hashed area). (b) Probabilities of different failure mechanisms (hold, access, read, and write failures) in SRAM cells in two cases where the impact of WFV is considered or neglected.

accuracy of such a distribution, a realistic WD distribution is obtained from Monte Carlo simulations of 500 cells and plotted along with the estimated distribution [Fig. 9(b)].

It can be observed that the estimated distribution closely follows the one obtained from the Monte Carlo simulations. Once the WD distribution is estimated, it is possible to determine the probability of write failure according to the maximum allowed WD, which is set by design specifications.

As it was mentioned earlier, read failure occurs when the voltage at node $Q_R$ ($Q_L$) exceeds the tripping voltage ($V_{Trip}$) of the left (right) inverter. As a result, calculating the read failure probability involves estimating two probability distributions (for example, $V_{QR}$ and $V_{Trip}$ of the left inverter) and finding the area for which $V_{QR} > V_{Trip}$ [Fig. 10(a)]. It should be noted that this hashed area in this figure is equal to the area under the $V_{QR} - V_{Trip}$ distribution between zero and minus infinity (not shown in the figure). Therefore, to determine the likelihood of read failure, first, the probability distributions of $V_{QR}$ and $V_{Trip}$ should be estimated using the sensitivity analysis method. Then, the read failure probability can be calculated by estimating the area under the $V_{QR} - V_{Trip}$ distribution between zero and minus infinity.

Hold failure occurs when the SRAM cell fails to hold the accurate logic value due to a diminished noise margin. Therefore, the probability of such a failure is determined by estimating the distribution of the hold margin and calculating the likelihood of the hold margin being below zero [Fig. 10(b)]. Similarly, access failures can be obtained using the sensitivity-based analysis by estimating the probability distribution of the read delay and calculating the probability that the delay exceeds the maximum allowed time for reading the content of the cell [Fig. 11(a)]. For the SRAM cell shown in Fig. 8, the probabilities of the four failure mechanisms are calculated for two cases: one considering only RDF and LER and the other with WFV included along them [Fig. 11(b)]. As expected, it can be observed that including WFV considerably increases the probability of all failure mechanisms.

B. Implication of WFV for SRAM Cell Subthreshold Leakage

Another important concern in SRAM design is the subthreshold leakage of the cell. Using the proposed model, one can accurately evaluate the impact of the grain orientation dependence of the work function of devices on the total subthreshold leakage of the SRAM cells. In order to achieve this goal, it is easier to start by studying the subthreshold leakage distribution of a single transistor under WFV. Since the distribution of $V_{th}$ due to WFV is assumed to be Gaussian [as shown in the companion paper (part I)], it is expected that the probability density function of the subthreshold leakage will be lognormal (due to the exponential dependence of the subthreshold leakage on $V_{th}$).

Using the proposed model, the subthreshold leakage distribution of individual TiN-based NMOS transistors can be evaluated as shown in Fig. 12(a), where the values on the $x$-axis (subthreshold leakage) are normalized to the leakage current of a device with the nominal $V_{th}$ value. In this figure, one curve represents the distribution of the leakage when WFV is neglected, and the other is when it is taken into account.
Fig. 12. (a) Probability density and (b) cumulative distribution functions of the subthreshold leakage for TiN-gate NMOS devices in the 65-nm node.

Fig. 13. (a) Probability density and (b) cumulative distribution functions of the total subthreshold leakage current of an SRAM cell (with TiN gates) obtained with Monte Carlo simulations for 10,000 cells.

Clearly, a higher level of variation (due to the inclusion of WFV) considerably shifts the distribution toward the right, which implies that the devices are more likely to exhibit high leakage current. Fig. 12(b) shows the cumulative distribution function of the subthreshold leakage. The existence of WFV shifts the cumulative distribution function toward the right indicating an increase in the number of NMOS transistors with higher leakage.

The total subthreshold leakage of the SRAM cell can be obtained by aggregating six independent distributions of transistors in the cell. Although each of these six distributions is lognormal [Fig. 12(a)], the distribution of the total leakage will be closer to a Gaussian distribution than a lognormal one (as per the central limit theorem). The probability density function of such a distribution is shown in Fig. 13(a), along with its cumulative density function in Fig. 13(b). These figures indicate that considering the impact of WFV results in higher probability for SRAM cells with high subthreshold leakages. These distributions are obtained from a Monte Carlo simulation of 10,000 SRAM cells, where the subthreshold leakage of the individual devices is estimated using the proposed model. Each of these figures demonstrates two cases: with and without considering WFV. As expected, WFV increases the SRAM cell’s probability of having a higher subthreshold leakage.

VI. CONCLUSION

The implications of a recently identified [3] source of random $V_{th}$ fluctuation (WFV) in emerging high-$k$/metal-gate devices have been investigated. In order to achieve this goal, the modeling approach proposed in the companion paper (part I) is employed. The proposed statistical framework is used for three purposes. First, it is employed to identify suitable materials and fabrication processes that can minimize the effect of WFV. For instance, it is shown that, using TiN/WN as the metal gate materials, one can achieve lower levels of WFV for NMOS/PMOS devices, respectively. Second, the impact of WFV on various types of classical and nonclassical transistors is studied. For example, it is shown that, compared to bulk and FD-SOI devices, FinFETs (with single-fin), are less influenced by WFV due to their larger gate area, for a given channel length. Third, the circuit-level implications of WFV are investigated for SRAM cell design. For instance, it is shown that considering WFV reduces the percentage of cells with $1.5 \times$ the nominal leakage from 98% to $\approx 96\%$ while increasing the read failure probability from $\approx 10^{-6}$ to $10^{-4}$. According to the comprehensive analyses presented in this work, WFV is going to be one of the dominant sources of random $V_{th}$ variation in emerging nanoscaled CMOS transistors. Moreover, WFV is the most dominant source of $V_{th}$ variation for transistors with undoped bodies such as FD-SOI and FinFET devices.

ACKNOWLEDGMENT

The authors (H. F. D. and K. B.) would like to thank S. H. Rasouli and C. Xu for useful technical discussions and careful review of the manuscript.

REFERENCES


R. Fujii, Y. Gotoh, M. Y. Liao, H. Tsuji, and J. Ishikawa, “Measurement and implementation of energy-efficient circuits and systems using emerging nanoscale transistors. He has published several papers in leading international conferences and journals.

Mr. Doggur’s paper introducing a new source of random variability in high-k/metal-gate transistors was a finalist for the IEEE/ACM William J. McCalla ICCAD Best Paper Award in 2008. He was an recipient of the Award of Distinction from UCSB in 2009 and a Peter J. Frenkel Foundation Fellowship from the Institute for Energy Efficiency at UCSB in 2010.

Kazuhiko Endo received the Ph.D. degree in electrical engineering from Waseda University, Tokyo, Japan, in 1999.

Since 1993 to 2003, he was with the Silicon Systems Research Laboratories, NEC Corporation, where he worked on the research and development of multilevel interconnects and high-k gate-stack technologies for ULSI. From August 1999 to August 2000, he was a Visiting Scholar with the Center for Integrated Systems, Stanford University, Stanford, CA. He is currently a Senior Researcher with the Silicon Nanoscale Devices Research Group, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. His research interests include nanometer-scale manufacturing for aggressively scaled multigate devices in advanced VLSI technologies.

Dr. Endo is a member of the IEEE Electron Devices Society and the Japan Society of Applied Physics. He was a recipient of a Best Paper Award at the 2003 Advanced Metalization Conference and at the 1998 Meeting of the Japan Society of Applied Physics.

Vivek K. De received the B.S. degree in electrical engineering from the Indian Institute of Technology Madras, Chennai, India, in 1985, the M.S. degree in electrical engineering from Duke University, Durham, NC, in 1986, and the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1992.

He is currently a Fellow and the Director of Circuit Technology Group, Corporate Research Laboratory, Intel Corporation, Hillsboro, OR. In his current role, he provides strategic direction for future circuit technologies and is responsible for aligning Intel’s circuit research with technology scaling challenges. He has published more than 150 technical papers in refereed conferences and journals and six book chapters on low-power circuits. He is the holder of 136 patents, with 57 more patents filed (pending).

Dr. De was the recipient of an Intel Achievement Award for his contributions to a novel integrated voltage-regulator technology.
Kaustav Banerjee (S’92–M’99–SM’03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

From 1993 to 1997, he held summer/visiting positions with Texas Instruments Incorporated, Dallas, TX, From 1999 to 2001, he was a Research Associate with the Center for Integrated Systems, Stanford University, Stanford, CA. In 2001, he held summer/visiting position with the Swiss Federal Institute of Technology, Lausanne, Switzerland. From February to August 2002, he was a Visiting Faculty with the Circuits Research Laboratory, Intel Corporation, Hillsboro, OR. In July 2002, he joined the faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he has been a Full Professor since 2007 and is also an affiliated Faculty with the California NanoSystems Institute and the Institute for Energy Efficiency. He is the author of over 200 journal and refereed international conference papers and several book chapters. He is also a Coeditor of *Emerging Nanoelectronics: Life With and After CMOS* (Springer, 2004). His current research interests include nanometer-scale issues in VLSI, as well as circuits and systems issues in emerging nanoelectronics. He is also involved in exploring the physics, technology, and applications of various carbon nanostructures for ultra energy-efficient electronics and energy harvesting/storage applications.

Prof. Banerjee was the recipient of numerous awards in recognition of his work, including a Best Paper Award at the Design Automation Conference in 2001, the Association of Computing Machinery Special Interest Group on Design Automation Outstanding New Faculty Award in 2004, an IEEE Micro Top Picks Award in 2006, and an IBM Faculty Award in 2008. He has served on the Technical and Organizational Committees of several leading IEEE and ACM conferences, including the International Electron Devices Meeting, the Design Automation Conference, the International Conference on Computer-Aided Design, the International Reliability Physics Symposium, the International Symposium on Quality Electronic Design, the EOS/ESD Symposium, and the International Conference on Simulation of Semiconductor Processes and Devices. From 2005 to 2008, he served as a member of the Nanotechnology Committee of the IEEE Electron Devices Society (EDS). He currently serves on the IEEE/EDS GOLD Committee and the IEEE/EDS VLSI Technology and Circuits Committee. He has been a Distinguished Lecturer of IEEE EDS since 2008.