Nanoscale Silicon Devices

Shunri Oda, David K. Ferry

Variability in Scaled MOSFETs

Publication details
Toshiro Hiramoto
Published online on: 18 Dec 2015

How to cite:- Toshiro Hiramoto. 18 Dec 2015, Variability in Scaled MOSFETs from: Nanoscale Silicon Devices CRC Press
Accessed on: 18 Jan 2020

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
3 Variability in Scaled MOSFETs

Toshiro Hiramoto

CONTENTS

3.1 Introduction ........................................................................................................54
3.2 Variability in 65 nm Transistors .....................................................................54
  3.2.1 Classification of Transistor Variability ................................................54
  3.2.2 Dependence on the Number of Transistors ........................................56
  3.2.3 Origin of Random Variability ..............................................................58
  3.2.4 Size Dependence of Variability ...........................................................59
  3.2.5 Drain Current Variability .....................................................................60
  3.2.6 Origin of COV Variability ....................................................................62
3.3 Variability of 11G Transistors .......................................................................63
  3.3.1 Measurement of 11G Transistors ........................................................63
  3.3.2 $V_{\text{th}}$ Variability of 11G Transistors ..............................................64
  3.3.3 $V_{\text{thex}}$ and $I_{\text{on}}$ Variability of 11G Transistors ...............................66
  3.3.4 SS Variability of 11G Transistors ......................................................67
3.4 Stability of SRAM Cells ................................................................................68
  3.4.1 Variability of Static Noise Margin ......................................................68
  3.4.2 $V_{\text{dd}}$ Dependence of SNM .............................................................69
  3.4.3 DIBL Dependence of SNM ................................................................71
3.5 Intrinsic Channel FD SOI Transistors ..........................................................72
  3.5.1 $V_{\text{th}}$ and Drain Current Variability ...............................................72
  3.5.2 Stability in FD SOTB SRAM .............................................................74
3.6 Self-Suppression of Variability .......................................................................75
  3.6.1 Mechanism of Self-Improvement .......................................................75
  3.6.2 Measurements of $|V_{\text{th}}|$ Shift by High Voltage Stress ....................77
  3.6.3 Measurements of Self-Improvement of SRAM Stability ....................78
3.7 Conclusions ....................................................................................................79
Acknowledgments ...............................................................................................79
References ............................................................................................................80
3.1 INTRODUCTION

For the past 40 years, the size of metal-oxide-semiconductor field-effect-transistors (MOSFETs) has been scaled down in order to attain higher performance, lower power dissipation, and higher integration in large-scale integrated circuits (VLSI). It is well known that the miniaturized size of MOSFETs has brought about various technical issues including short channel effects, degraded reliability due to a high electric field, and performance degradation due to parasitic resistance and capacitance. Among them, one of the most significant technical problems is the variability of transistor characteristics [1–4]. Owing to the transistor variability, the circuits do not function correctly even though each individual transistor behaves correctly, or the margin in the circuit operation is reduced causing manufacturing yield to drop precipitously. This problem of the variability in characteristics may impose a limit on the transistor scaling, performance improvement, and power dissipation reduction. Therefore, it is an urgent issue to understand the root causes and find solutions. However, the origins that lead to variability in characteristics cover a very wide range from atomic-level impurity distribution to manufacturing equipment. The understanding of the physics behind the variability problems and the development of effective measures to the problems are essential.

In this chapter, the present status of the transistor variability is reviewed. In particular, the random variability and its impact on static random access memory (SRAM) are described. As a solution to the variability issue, a fully-depleted (FD) silicon-on-insulator (SOI) transistor with intrinsic channel is described. A new concept for the suppression of random variability in SRAM is also introduced.

3.2 VARIABILITY IN 65 nm TRANSISTORS

In this section, the status and basic behaviors of transistor variability are illustrated by taking the 65 nm planar bulk technology as an example.

3.2.1 CLASSIFICATION OF TRANSISTOR VARIABILITY

There are so many types of transistor variability in VLSI. Figure 3.1 schematically shows the classification of transistor variability. The transistor characteristics differ from one lot to another. This is lot-to-lot variability (or interlot variability). Within the same lot, the transistor characteristics differ from one wafer to another (wafer-to-wafer variability or interwafer variability). Within the same wafer, transistor characteristics differ from one chip to another (chip-to-chip variability or interchip variability). Even in the same chip, the transistor characteristics differ from one transistor to another (intrachip variability).

Figure 3.2 shows an example of intrawafer variability and intrachip variability of transistors fabricated by the 65 nm bulk planar technology [5]. A large number of transistors were measured using a device-matrix-array (DMA) test-element group (TEG). Each chip has one million (1M) transistors with identical gate length (L) and gate width (W). In order to show the overall and systematic variability within the wafer and chip, each datum point in Figure 3.2 represents the average threshold voltage (Vth) of 1k transistors (each wafer has 1k data points). Apparently, some chips
Variability in Scaled MOSFETs

have higher $V_{th}$, and others have lower $V_{th}$. This is the chip-to-chip variability. There is also a $V_{th}$ variation within a chip. It is found that the magnitude of the intrawafet variability is larger than that of the overall within-chip variability.

Figure 3.3a shows detailed intrachip variability of the 65 nm transistors, where each datum point corresponds to $V_{th}$ of each transistor [5]. Surprisingly, $V_{th}$ differs largely from one transistor to another. The measured $V_{th}$ values are mathematically separated into a random component and systematic component in Figure 3.3b and c [5]. It is found that the random component is much larger than the systematic component. These results indicate that as long as the transistor layout patterns are regular and identical, the random component is dominant and the systematic component is negligible in the planar bulk transistors.
3.2.2 Dependence on the Number of Transistors

The range of the random variability depends on the number of transistors. Figure 3.4a shows I–V characteristics of 100 n-type transistors (nFETs) within a chip. Gate length $L$ is 60 nm and gate width $W$ is 120 nm. Clear $V_{th}$ variability is found, but the range of $V_{th}$ is not so wide. The cumulative plot of $V_{th}$ is also shown. The $V_{th}$ data lie on an almost straight line and $V_{th}$ ranges within $\pm 2.6\sigma$, where $\sigma$ is the standard deviation and approximately 43 mV in this wafer. When the number of transistors increases by 100 times (10,000 transistors), I–V characteristics are varied as shown in Figure 3.4b. The $V_{th}$ range increases to $\pm 3.8\sigma$.

When the number of transistors increases further by 100 times (1M transistors), I–V characteristics are shown in Figure 3.4c [5], where the $V_{th}$ range further increases to $\pm 5\sigma$. It is also clearly shown that since $V_{th}$ data lie on a straight line in the cumulative plot, $V_{th}$ follows the normal distribution up to $\pm 5\sigma$, and therefore, the variability is random. $V_{th}$ ranges over a wide range from −0.28 to 0.73 V, which may cause severe yield loss and margin degradation in circuit operation.

1M nFETs in other chips in the same wafers and other wafers were also measured. Although the average $V_{th}$'s are slightly different from one chip to another, $V_{th}$ also follows the normal distribution up to $\pm 5\sigma$ in all chips and $\sigma$ is almost the same (not shown). From these results, the transistor variability in lots, wafers,
Variability in Scaled MOSFETs

and chips is schematically shown in Figure 3.5. The intrawafer variability is smaller than the interlot variability. Each wafer has different average $V_{th}$ and each chip also has different average $V_{th}$. The within-chip variability is defined by the average $V_{th}$ and $\sigma$. It is also found that $V_{th}$ follows the normal distribution up to $\pm 5\sigma$ in pFETs and the variability in nFETs is larger than that in pFETs [5,6] (not shown).
3.2.3 ORIGIN OF RANDOM VARIABILITY

In order to examine the origins of the aforementioned huge random variability of measured $V_{th}$, the transistors that exhibit extreme $V_{th}$ values are directly observed by the transmission electron microscope (TEM) [6]. Figure 3.6 shows a top view and the cross-sectional TEM images of $-5\sigma$, median, and $+5\sigma$ nFETs. Observed gate oxide thickness ($t_{ox}$), measured strain, and images of gate polysilicon grains are also shown. It is surprising to find that transistor sizes ($L$, $W$, and $t_{ox}$) of $-5\sigma$ FET and $+5\sigma$ FET are almost identical even though $V_{th}$’s are so different. This result clearly suggests that the large random variability is not caused by the variation of transistor dimensions.

It is now well recognized that the origin of the random variability of transistor characteristics is the random dopant fluctuation (RDF). $V_{th}$ of a transistor is primarily determined by dopant concentration in the depletion layer in the transistor channel. However, the number of dopants in the depletion layer varies from one transistor to another, causing the $V_{th}$ variability. It is well known that when the dopants are randomly distributed, the number of dopants follows the Poisson distribution. When the number of dopants increases, the Poisson distribution is well approximated by the normal distribution. This is why the measured $V_{th}$ follows the normal distribution.

<table>
<thead>
<tr>
<th>$V_{th}$</th>
<th>TEM image (plan view)</th>
<th>TEM image (cross section)</th>
<th>$t_{ox}$</th>
<th>Strain</th>
<th>TEM image (gate poly-Si grain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.335 V ($-5\sigma$)</td>
<td>$L = 66 \text{ nm}$</td>
<td>$W = 128 \text{ nm}$</td>
<td>1.9 nm</td>
<td>130/–130 MPa</td>
<td></td>
</tr>
<tr>
<td>0.529 V (median)</td>
<td>$L = 67 \text{ nm}$</td>
<td>$W = 127 \text{ nm}$</td>
<td>1.9 nm</td>
<td>130/130 MPa</td>
<td></td>
</tr>
<tr>
<td>0.719 V ($+5\sigma$)</td>
<td>$L = 65 \text{ nm}$</td>
<td>$W = 124 \text{ nm}$</td>
<td>1.9 nm</td>
<td>130/130 MPa</td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 3.6 Observations of nFETs that have extremely low $V_{th}$ ($-5\sigma$), median $V_{th}$, and extremely high $V_{th}$ ($+5\sigma$). Top view and cross-sectional TEM images, measured $t_{ox}$, measured strain, and observed poly-Si grains are shown. (Modified from Tsunomura, T. et al., Analyses of $5\sigma$ $V_{th}$ fluctuation in 65 nm-MOSFETs using Takeuchi plot, Symposium on VLSI Technology, Honolulu, HI, 2008, pp. 156–157.)
3.2.4 Size Dependence of Variability

When the random variability is dominant in the transistor variability, the standard deviation $\sigma$ of transistor parameters has transistor size dependence. This phenomenon can be understood by the nature of the Poisson distribution. In a Poisson distribution, $\sigma$ is given by $\sqrt{\mu}$, where $\mu$ is the average number, and the normalized variability ($\sigma/\mu$) is given by $1/\sqrt{\mu}$. In a smaller transistor, the number of average dopant atoms included in the depletion region is smaller, and hence, the normalized variability ($\sigma/\mu$) of the dopant number is larger. When the transistor size increases, the variability of dopant number is averaged out and becomes smaller. Therefore, the variability of a transistor parameter increases as the transistor size is scaled down.

The average number $\mu$ of dopant atoms is proportional to $LW$, where $L$ is the gate length and $W$ is the gate width. Then, the variability is given by a simple function of $1/\sqrt{LW}$ [3]. Figure 3.7a shows $\sigma V_{th}$ as a function of $1/\sqrt{LW}$ [7]. When $t_{ox}$ is fixed and only transistor size is varied, measured data lie on the same straight line. The slope of this line is often called Pelgrom coefficient ($A_{vt}$) and this figure is called Pelgrom plot [3]. Then, $\sigma V_{th}$ is given by $\sigma V_{th} = A_{vt} \sqrt{LW}$. This is a very useful equation. When $A_{vt}$ of a semiconductor process is known and the transistor size is fixed, $\sigma V_{th}$ is easily derived.

It is known that $A_{vt}$ depends on gate oxide thickness $t_{ox}$ and $N_A$. In order to develop a more universal relationship between $\sigma V_{th}$ and transistor parameters, it is proposed to use the Takeuchi plot [8]. Figure 3.7b shows a Takeuchi plot, where the data in Figure 3.7a is replotted [7]. The horizontal axis is $\sqrt{T_{inv}(V_{th} + V_0)/LW}$, where $T_{inv}$ is the electrical gate oxide thickness at inversion and $V_0$ is given by $-(V_{FB} + 2\phi_F)$, where $V_{FB}$ is the flat band voltage and $\phi_F$ is the Fermi energy. Even though $t_{ox}$ and

![FIGURE 3.7](image-url)

© 2016 by Taylor & Francis Group, LLC
N_A are different, all data lie on the same straight line. The slope of Takeuchi plot is defined as B_vt. This Takeuchi plot indicates that \( \sigma V_{th} \) increases as \( V_{th} \) becomes higher. The Takeuchi plot is useful when \( \sigma V_{th} \) should be obtained in different \( V_{th} \).

### 3.2.5 Drain Current Variability

Drain current variability, as well as \( V_{th} \) variability, is a major concern in VLSI, because it directly causes huge variations in memory and logic circuit performances. Figure 3.8a shows I–V characteristics of 1M nFETs fabricated by 65 nm bulk technology. Large drain current variability is observed [9]. Figure 3.8b shows the cumulative plot of the on-current (I_on), which shows that I_on also follows the normal distribution up to \( \pm 5\sigma \). Obviously, the origins of drain current variability are \( V_{th} \) variability and \( G_m \) variability. However, it has been found that there is a third origin of drain current variability [9].

Generally, there are two definitions of threshold voltage. One is the threshold voltage defined by subthreshold constant current (\( I_0 = 10^{-7} \times W/L \)). This threshold voltage is called \( V_{thc} \) in this study. The other is the threshold voltage determined by extrapolating drain current (\( V_{gs} \) intercept of tangent line with largest slope in \( I_{ds}–V_{gs} \) characteristics). This is called \( V_{thex} \).

Figure 3.9 shows I–V characteristics of two nFETs, which have identical \( V_{thc} \) and \( G_m \) [9]. I_on differs significantly even though \( V_{thc} \) and \( G_m \) are the same. Please note that the onset point of drain current in the linear scale plot is different, and therefore, \( V_{thex} \) is quite different in these two nFETs. Here, we define “current onset voltage (COV)” as COV = \( V_{thex} \) – \( V_{thc} \) [9,10]. The drain current rises rapidly and I_on is high when COV is small, while the drain current rise is slower when COV is large. Therefore, the COV variability contributes to the I_on variability.

It has been found that \( V_{thc} \), \( G_m \), and COV are almost mutually independent [9]. Then, I_on variability can be separated into three components of \( V_{thc} \), \( G_m \), and COV. Figure 3.10 shows the decomposition of measured I_on variability [9]. I_on variability...
Variability in Scaled MOSFETs

FIGURE 3.9 Measured I–V characteristics of two nFETs with identical $V_{\text{th}c}$ and $G_m$, showing the difference of COV. (Modified from Tsunomura, T. et al., Analysis and prospect of local variability of drain current in scaled MOSFETs by a new decomposition method, VLSI Symposium on Technology, Honolulu, HI, 2010, pp. 97–98.)

FIGURE 3.10 Decomposition of measured $I_{\text{on}}$ variability of nFETs into $V_{\text{th}c}$, $G_m$, and COV components. (a) Saturation region at $V_{\text{ds}} = 1.2$ V. (b) Linear region at $V_{\text{ds}} = 50$ mV. (Modified from Tsunomura, T. et al., Analysis and prospect of local variability of drain current in scaled MOSFETs by a new decomposition method, VLSI Symposium on Technology, Honolulu, HI, 2010, pp. 97–98.)
(\(\sigma_{I_{on}}\)) is normalized to median \(I_{on}\), and the decomposition of \(\sigma_{I_{on}}/I_{on}\) is shown. It is found that \(V_{thc}\) is the major component of \(I_{on}\) variability, and COV is the second biggest component in the saturation region (\(V_{ds} = 1.2\) V) and is larger than the \(G_m\) component. To suppress the \(I_{on}\) variability, the understanding and suppression of COV variability is indispensable.

### 3.2.6 Origin of COV Variability

Why is COV not constant and fluctuates from one transistor to another? This is because the current path in the subthreshold region and strong inversion region is different. In a planar bulk MOS transistor, the channel potential fluctuates due to RDF. The subthreshold current flows in the potential valley, which is often called the percolation path. Therefore, \(V_{thc}\), which reflects the subthreshold current, is determined by how deep the potential valley is. In the strong inversion region, on the other hand, the potential fluctuation by RDF is screened by inversion charges. Therefore, \(V_{thex}\), which reflects the strong inversion current, is determined by the average potential of the channel.

In order to examine the aforementioned model of COV, 3D device simulation is performed. Assuming RDF, I–V characteristics of 200 transistors are simulated. Among them, two transistors, which have the smallest COV and the largest COV, are selected. Figure 3.11 shows simulated potential of the “potential dividing line”

---

**FIGURE 3.11** (See color insert.) Simulated potential distribution in the transistor channel and the potential on the “dividing line” along the channel width direction. (a) The transistor with the smallest COV among 200 pFETs. (b) The transistor with the largest COV among 200 pFETs. (Modified from Kumar, A. et al., Origin of “current-onset voltage” variability in scaled MOSFETs, *IEEE Silicon Nanoelectronics Workshop*, Honolulu, HI, 2010, pp. 7–8.)
Variability in Scaled MOSFETs

along the channel width direction [10]. It is clearly shown that the transistor with the smallest COV has only shallow potential valleys, while the transistor with the largest COV has a very deep potential valley. The potential depth (the difference between average and minimum channel potential) is also simulated. It is found that while \( V_{\text{thr}} \) and \( V_{\text{thex}} \) have poor correlation with the potential depth, COV has a strong correlation with the potential depth [10] (not shown). These results confirm the model that the COV variability is caused by the potential fluctuation due to RDF.

### 3.3 VARIABILITY OF 11G TRANSISTORS

Generally, the ideal normal distribution of \( V_{\text{th}} \) is assumed to predict the yield of large-scale integrated circuits and memories in the circuit simulation. In the previous sections, it is shown that \( V_{\text{th}} \) follows the normal distribution up to \( \pm 5\sigma \) in both nFETs and pFETs [5,6]. However, the state-of-the-art VLSI chips contain more than 1G (one billion) transistors in a chip. To our best knowledge, \( V_{\text{th}} \) distributions of 1G-level transistors have not been reported, mainly because the measurement takes an overwhelmingly long time. Therefore, the 6\( \sigma \) distribution is still unknown.

In this section, a special DMA-TEG for ultrafast \( V_{\text{th}} \) monitoring is designed and fabricated, and 10G-level transistor variability is measured and the distribution is analyzed [11,12].

#### 3.3.1 MEASUREMENT OF 11G TRANSISTORS

Figure 3.12 shows a schematic of the fast \( V_{\text{th}} \) monitoring circuit using the amp mode [11]. The device under test (DUT) is selected by a decoder and the current, \( I_{\text{DUT}} \), is compared with the constant current, \( I_{\text{REF}} \). The gate force voltage, \( V_{\text{GF}} \), is scanned at an interval of 25 mV, and \( V_{\text{GF}} \) at which \( I_{\text{DUT}} \) exceeds \( I_{\text{REF}} \) is defined as measured \( V_{\text{th}} \).


© 2016 by Taylor & Francis Group, LLC
Please note that the measured $V_{\text{th}}$ is $V_{\text{thc}}$, instead of $V_{\text{thex}}$. The designed DMA TEG has 256M transistors in a chip, and the measurement time is approximately 3 h per chip.

DMA TEG chips were fabricated by the 65 nm technology. Gate length $L$ is 60 nm and gate width $W$ is 120 nm. $I_{\text{REF}}$ is set to 100 nA. $V_{\text{th}}$’s of 44 chips were measured, and therefore, the total number of transistors is 11G (11 billion), which corresponds to $6.5\sigma$. After $V_{\text{th}}$’s of all transistors were measured and $V_{\text{th}}$ distribution of 256M transistors was determined in each chip, I–V characteristics were also measured only for transistors with extremely high $V_{\text{th}}$ or low $V_{\text{th}}$ (beyond $\pm 5.0\sigma$) and median $V_{\text{th}}$ in order to investigate the origin of nonnormal distribution of $V_{\text{th}}$.

### 3.3.2 $V_{\text{thc}}$ Variability of 11G Transistors

Figure 3.13 shows measured cumulative plots of $V_{\text{th}}$ of 11G nFETs and pFETs at $|V_{ds}| = 50$ mV [11]. It has been found that nFETs show good normality up to more than $\pm 6.5\sigma$, although slight deviation is observed in the low $V_{\text{th}}$ region below $-4\sigma$. This is the first observation of $V_{\text{th}}$ distribution of 10G-level transistors. On the other hand, pFETs also have an almost normal distribution in the high $V_{\text{th}}$ region up to more than $6.5\sigma$. However, apparent distribution “tail” is observed in the low $V_{\text{th}}$ region below $-4\sigma$. This large “tail” may affect the yield of large-scale logic circuits and SRAM, and determining the cause is urgent.

Figure 3.14 shows measured I–V characteristics at $|V_{ds}|$ of 50 mV and 1.2 V of some of the transistors with extremely high $V_{\text{th}}$ and low $V_{\text{th}}$ (beyond $\pm 5.0\sigma$) [11]. For transistors with extremely low $V_{\text{th}}$, apparently larger drain-induced barrier lowering (DIBL) is observed in both nFETs and pFETs. In particular in pFETs, degraded subthreshold slope (SS) is observed. In order to examine the DIBL degradation in

**FIGURE 3.13** Measured cumulative distributions of $V_{\text{th}}$ of 11G transistors at $|V_{ds}| = 50$ mV. (a) nFETs and (b) pFETs. (Modified from Mizutani, T. et al., Measuring threshold voltage variability of 10G transistors, *International Electron Devices Meeting (IEDM)*, Washington, DC, 2011, pp. 563–566.)
more detail, DIBL distribution of 1000 transistors in the extremely low V\textsubscript{th} region (below −5.0\,\textit{σ}) and extremely high V\textsubscript{th} region (above 5.0\,\textit{σ}) is measured [11] (not shown). Clearly, transistors in low V\textsubscript{th} region have anomalously large DIBL in both nFETs and pFETs. However, a clear difference between nFET and pFET is not found in DIBL.

We found pronounced differences of characteristics in COV between nFET and pFET. Figure 3.15 shows distributions of measured COV in extremely low V\textsubscript{th} region (below −5.0\,\textit{σ}), medium V\textsubscript{th}, and extremely high V\textsubscript{th} region (above 5.0\,\textit{σ}) [11].

**FIGURE 3.14** Examples of measured I–V characteristics at |V\textsubscript{ds}| of 50 mV and 1.2 V with extremely high V\textsubscript{th} and low V\textsubscript{th} (beyond ±5.0\,\textit{σ}). (a) nFETs and (b) pFETs. (Modified from Mizutani, T. et al., Measuring threshold voltage variability of 10G transistors, *International Electron Devices Meeting (IEDM)*, Washington, DC, 2011, pp. 563–566.)

**FIGURE 3.15** Distributions of measured COV in extremely low V\textsubscript{th} region (below −5.0\,\textit{σ}), medium V\textsubscript{th}, and extremely high V\textsubscript{th} region (above 5.0\,\textit{σ}). (a) nFETs and (b) pFETs. (Modified from Mizutani, T. et al., Measuring threshold voltage variability of 10G transistors, *International Electron Devices Meeting (IEDM)*, Washington, DC, 2011, pp. 563–566.)
Again, transistors in low $V_{th}$ region have anomalously large COV in both nFET and pFET. Especially in pFETs, COV of low $V_{th}$ transistors is abnormally large. This result indicates that the distribution tail observed in pFET is closely related with degraded COV and may be caused by local percolation paths that are formed in the regions with extremely small number of dopants due to RDF, which will be discussed again later.

### 3.3.3 $V_{\text{thex}}$ and $I_{\text{on}}$ Variability of 11G Transistors

It is impossible to measure $V_{\text{thex}}$ of all 11G transistors, because I–V curve measurements are necessary to derive $V_{\text{thex}}$, which takes an extremely long time. By measuring I–V characteristics of low $V_{\text{thc}}$ region and high $V_{\text{thc}}$ region with additional measurements of I–V characteristics of another 4k transistor TEG (which corresponds to the distribution center), the distribution of $V_{\text{thex}}$ is derived [12]. Figure 3.16 shows cumulative plots of measured $V_{\text{thex}}$ of 11G nFETs and pFETs at $|V_{ds}| = 50 \text{ mV}$ [12]. It is newly found that $V_{\text{thex}}$ does not have a long distribution tail, contrary to $V_{\text{thc}}$ even in pFETs. Although $V_{\text{thc}}$ has an impact on the standby power, circuit operations and performance are mainly affected by $V_{\text{thex}}$ rather than $V_{\text{thc}}$. Therefore, one expects that the impact of the long $V_{\text{thc}}$ distribution tail on circuit design is minimal.

Using a similar method, the $V_{\text{thc}}$ and $I_{\text{on}}$ distributions at $V_{ds} = 1.2 \text{ V}$ are determined [12]. As expected, $V_{\text{thc}}$ has a long tail in low $V_{\text{thc}}$ region at $|V_{ds}| = 1.2 \text{ V}$ even in nFETs (not shown), because DIBL is very large in low $V_{\text{thc}}$ region. Figure 3.17 shows cumulative plots of measured $I_{\text{on}}$ of 11G nFETs and pFETs at $|V_{ds}| = 1.2 \text{ V}$ [12]. nFETs have much larger $I_{\text{on}}$ variability than pFETs, because $V_{\text{th}}$ variability in nFETs is larger than that in pFETs. Moreover, $I_{\text{on}}$ deviates from the normal distribution in the low $I_{\text{on}}$ region in both nFETs and pFETs. Some transistors have

**FIGURE 3.16** Cumulative distributions of measured $V_{\text{thex}}$ of 11G transistors at $|V_{ds}| = 50 \text{ mV}$. (a) nFETs and (b) pFETs. (Modified from Mizutani, T. et al., Analysis of transistor characteristics in distribution tails beyond $\pm 5.4\sigma$ of 11 billion transistors, *International Electron Devices Meeting (IEDM)*, Washington, DC, 2013, pp. 826–829.)
Variability in Scaled MOSFETs

abnormally low $I_{\text{on}}$, which may have a greater impact on circuit performance and yield. It is found that $|V_{\text{thc}}|$ of the FETs with lowest $I_{\text{on}}$ is not necessarily the highest in both nFET and pFET [12] (not shown), indicating that the origin of abnormally low $I_{\text{on}}$ is not high $V_{\text{thc}}$ but some other mechanisms such as extraordinarily high contact resistance.

3.3.4 SS Variability of 11G Transistors

Subthreshold swing (SS) is one of the most important parameters, which determine the off-current of transistors. In spite of the importance, few experimental data [13] have been reported on the variability of SS. Looking at I–V curve in Figure 3.14 carefully, SS in high subthreshold current ($SS_7'$, at $3 \times 10^{-7} \times (W/L) A$) is apparently degraded and fluctuates significantly in low $V_{\text{thc}}$ region in pFET, although SS in deep subthreshold region ($SS_8$, at $1 \times 10^{-8} \times (W/L)A$) is not degraded. It is also found that $SS_7'$ has very poor correlation with DIBL [12] (not shown), indicating that the degradation of $SS_7'$ is not caused by the short channel effect. Moreover, degraded $SS_7'$ has very good correlation with degraded COV [12] (not shown).

In order to examine the reason that $SS_7'$ is degraded but $SS_8$ is not degraded in low $V_{\text{thc}}$ region, 3D device simulation was performed assuming two types of deep potential valleys in the transistor channel as shown in Figure 3.18a [12]. It is found that when the valley is narrow enough, punch-through is prevented. Although SS in high subthreshold current region is degraded, better SS in deep subthreshold region is obtained, as shown in Figure 3.18b [12]. These results indicate that the narrow potential valleys caused by RDF are responsible for degraded SS and degraded COV in high subthreshold current region.
3.4 STABILITY OF SRAM CELLS

The instability in SRAM cells due to the variability of individual transistors in the cells is known as a crucial problem that will prevent further device integration and V_{dd} lowering [14]. The yield and the minimum operating voltage (V_{min}) are mainly determined by SRAM cells in recent VLSI. Therefore, the analysis of cell imbalances at the transistor level is essential for better understanding of SRAM stability at low V_{dd}. In this section, static noise margin (SNM) of SRAM cells and V_{th} of six individual transistors in the cells are directly measured and their variability is intensively analyzed using a DMA TEG of 16 kbit SRAM cells.

3.4.1 VARIABILITY OF STATIC NOISE MARGIN

Figure 3.19 shows a schematic of the 16 kbit SRAM DMA-TEG [15]. The TEG is based on the transistor DMA TEG that we have developed [5,6]. DUT is connected to six switch transistors, so that DUT is electrically isolated from other devices. DUTs are arrayed in a matrix manner and each DUT can be accessed by using decoder.
circuit. In the present SRAM DMA TEG, DUT is replaced by an SRAM mini array of 6 × 8 cells. Terminals of V_{dd}, word line (WL), left bit line (BLL), right bit line (BLR) as well as two internal storage nodes (VL, VR) at the center cell of the SRAM miniarray are connected to the six switched transistors and can be accessed, and the rest of SRAM cells are dummy cells. Since the internal storage nodes are accessible, noise margins as well as characteristics of the 6 individual transistors can be directly measured. The 16 kbit SRAM DMA TEG was fabricated with 65 nm technology.

I–V characteristics of 1k access NMOS transistors (T_a), drive NMOS transistors (T_n), and PMOS transistors (T_p) were measured [15] (not shown). The transistor characteristics vary significantly. The normal distributions of V_{th}’s of T_a, T_n, and T_p are confirmed. It is also confirmed that T_p has a smaller V_{th} variability than T_a and T_n.

Figure 3.20a shows measured butterfly curves of 1 kbit SRAM cells at V_{dd} of 1.2 V [15]. The WL is set to V_{dd}. There are huge variations of the butterfly curves. Here, one-side SNM is defined: SNM(L) is the square of the left eye of the butterfly curve, and SNM(R) is the square of the right eye. Please note that the SNM is defined as the smaller square of two eyes of the butterfly curve. Figure 3.20b shows cumulative distribution of measured 16 kbit SNM at several values of V_{dd} [15]. It is found that, while the one-sided SNM follows the normal distribution up to ±4σ even when V_{dd} is lowered from 1.2 to 0.4 V [15] (not shown), SNM does not follow the normal distribution. It is also shown that SNM is degraded when V_{dd} decreases and some cells fail at V_{dd} of 0.4 V.

3.4.2 V_{dd} Dependence of SNM

Figure 3.21 shows measured SNM as a function of V_{dd} [15]. Among 16 kbit SRAM cells, cells that have SNM values between 0.20 and 0.21 V at V_{dd} of 1.2 V are selected, and their V_{dd} dependences are shown. It is very interesting to note that,
even though SNM at 1.2 V is very similar, the $V_{dd}$ dependence is very different depending on the cell: some cells show improvement of SNM when $V_{dd}$ is lowered down to 0.8 V, and some cells show very severe degradation of SNM when $V_{dd}$ decreases. At $V_{dd}$ of 0.4 V, SNM of some cells remains at a high value above 0.1 V, but some cells fail (SNM is 0 V).

However, it has been found that this peculiar $V_{dd}$ dependence is not simply explained by $V_{th}$ variability alone [15]. Actually, when taking the variability of cell
transistors into account in the circuit simulation, only $V_{th}$ variability is generally considered. The circuit simulation results of SNM or $V_{min}$ are not necessarily consistent with the measured data [15]. Therefore, the effects of transistor parameters other than $V_{th}$ should be examined.

### 3.4.3 DIBL Dependence of SNM

In order to investigate the impact of device parameters on SNM, a unique method using a half-cell is employed [16]. The 16 kbit SRAM has 32 kbit half-cells. Among 32 kbit half-cells, half-cells in which $V_{th}$’s of three transistors ($T_n$, $T_p$, and $T_a$) are within the median value $\pm 10$ mV are selected. As a result, 183 half-cells are selected. Then, the selected 183 half-cells have almost identical $V_{th}$’s for the three transistors and the effects of $V_{th}$ variability are eliminated. It is found that there is no clear correlation between measured SNM of the selected half-cells and measured $G_m$ [16] (not shown), indicating that $G_m$ variability has no clear effect on SNM. It is also found that the body factor has no clear effect on SNM.

Figure 3.22 shows measured SNM of the selected half-cells as a function of measured DIBL of $T_n$ and $T_p$ [16]. Here, DIBL is defined by $V_{th}(V_{ds} = 50$ mV) $- V_{th}(V_{ds} = 0.6$ V). Apparently, SNM has a negative correlation with DIBL of $T_n$ and $T_p$, and SNM is more degraded when DIBLs of $T_n$ and $T_p$ are larger. There is clear experimental evidence that DIBL variability degrades SRAM stability. This negative correlation was not found in $T_a$ [16] (not shown). These results show that the DIBL variability should be taken into account to explain SNM variability and its $V_{dd}$ dependence.

![Figure 3.22](image_url)
3.5 INTRINSIC CHANNEL FD SOI TRANSISTORS

The major origin of random $V_{th}$ variability is RDF. It is reported that DIBL and COV variabilities are also caused by channel potential fluctuations due to RDF [9,10]. These results suggest that, if the dopant atoms are removed from the channel, not only $V_{th}$ variability but DIBL variability and COV variability will be suppressed. In this section, intrinsic channel FD SOI MOSFETs were fabricated and their variability was compared with that of conventional bulk MOSFETs using DMA TEG.

3.5.1 $V_{th}$ AND DRAIN CURRENT VARIABILITY

Figure 3.23 shows a schematic of intrinsic channel FD SOI nFET and pFET [17,18]. The channels are not intentionally doped. The SOI is very thin ($t_{SOI} = 12$ nm in this study) to suppress the short channel effect. The buried oxide (BOX) is also very thin ($t_{BOX} = 10$ nm in this study), which enables us to control $V_{th}$ by back bias. This device is also called a silicon-on-thin-BOX (SOTB) transistor [17–19]. FD SOTB transistors were fabricated with 65 nm technology. For comparison, conventional bulk transistors, where the channels are doped ($2 \times 10^{18}$ cm$^{-3}$), were also fabricated for reference. A poly-Si gate was used in both FD SOTB and bulk MOSFETs, while high-k/SiON gate dielectric was used to adjust $V_{th}$ of FD SOTB transistors. $T_{inv}$ is almost the same (approximately 2.6 nm). The gate length is 60 nm and gate width is 120 nm. The characteristics of both bulk and intrinsic channel FD SOTB transistors were measured using DMA TEG.

Figure 3.24 compares $I_d$–$V_{gs}$ characteristics of 1k transistors of bulk nFETs and intrinsic channel FD SOTB nFETs [20]. Apparently, FD SOTB transistors have a smaller variability. Figure 3.25 shows cumulative plots of $V_{thc}$ in bulk and SOTB nFETs [20]. $V_{thc}$ shows a normal distribution in both bulk and SOTB, and $V_{thc}$ variability is suppressed in SOTB nFETs ($\sigma = 17.8$ mV in linear region) compared with bulk ($\sigma = 37.5$ mV). This is because the intrinsic channel FD SOTB transistors have a very small number of dopants in the channel.

![FIGURE 3.23](image_url) A schematic diagram of an intrinsic channel FD SOI FET. The device is also called an STOB transistor. A bulk FET cointegrated with an SOTB FET is also shown in this figure. (From Sugii, N. et al., Ultralow-voltage operation SOTB technology toward energy efficient electronics, *International Conference on Solid State Devices and Materials (SSDM)*, Fukuoka, Japan, pp. 736–737, Copyright 2013 The Japan Society of Applied Physics.)
Variability in Scaled MOSFETs

Figure 3.26 shows cumulative plots of on-current ($I_{on}$) in bulk and SOTB nFETs [20]. The $I_{on}$ current is normalized to median $I_{on}$ and $\sigma_{I_{on}}/I_{on}$ is shown in the figure. $I_{on}$ variability is also reduced in the FD SOTB transistors. It is found that not only $V_{thc}$ variability suppression but also COV variability suppression contributes to the $I_{on}$ variability reduction in FD SOTB transistors (not shown). Figure 3.27 compares simulated potential fluctuations of transistor channels in bulk and FD SOTB MOSFETs [21]. Thanks to the intrinsic channel, the potential of an FD SOTB transistor channel is very smooth, leading to the reduction of $V_{th}$ and $I_{on}$ variabilities as well as DIBL and COV variabilities.
Small variability in FD SOTB transistors allows a drastically reduction in the operating voltage of FD SOTB SRAMs. Figure 3.28 compares the measured butterfly curves of 1 kbit bulk and FD SOTB SRAMs at $V_{dd}$ of 0.4 V [22]. Some cells fail at 0.4 V in bulk SRAM, while all 1 kbit cells operate even at 0.4 V in FD SOTB SRAM. The minimum operation voltage ($V_{min}$) of 48 kbit SRAM is reduced from 0.542 V in

$$L = 56 \text{ nm}$$

$$W = 110 \text{ nm}$$

(a) σ$_{I_{on}/I_{on}}$ (%)

- 2.5%
- 2.8%
- 6.6%
- 5.3%

(b) Normal quantile

Bulk NMOS SOTB NMOS

$V_{ds} = 50 \text{ mV}$

$V_{ds} = 1.2 \text{ V}$

FIGURE 3.26 Cumulative distributions of measured σ$_{I_{on}/I_{on}}$ at $V_{ds} = 0.05 \text{ V}$ and $V_{ds} = 1.2 \text{ V}$ in (a) bulk nFETs and (b) SOTB nFETs. (Modified from Mizutani, T. et al., Reduced drain current variability in fully depleted silicon-on-thin-BOX (SOTB) MOSFETs, IEEE Silicon Nanoelectronics Workshop, Honolulu, HI, 2012, pp. 71–72.)

FIGURE 3.27 (See color insert.) Simulated potential distributions of transistor channels in (a) bulk nFETs and (b) SOTB nFETs. (Modified from Hiramoto, T. et al., Suppression of DIBL and current-onset voltage variability in intrinsic channel fully depleted SOI MOSFETs, IEEE International SOI Conference, San Diego, CA, 2010, pp. 170–171.)

3.5.2 STABILITY IN FD SOTB SRAM

Small variability in FD SOTB transistors allows a drastically reduction in the operating voltage of FD SOTB SRAMs. Figure 3.28 compares the measured butterfly curves of 1 kbit bulk and FD SOTB SRAMs at $V_{dd}$ of 0.4 V [22]. Some cells fail at 0.4 V in bulk SRAM, while all 1 kbit cells operate even at 0.4 V in FD SOTB SRAM. The minimum operation voltage ($V_{min}$) of 48 kbit SRAM is reduced from 0.542 V in
Variability in Scaled MOSFETs

bulk to 0.290 V in FD SOTB (not shown) [22]. The intrinsic channel is essential to achieve ultralow voltage operation below 0.4 V in large-scale SRAM cell array.

3.6 SELF-SUPPRESSION OF VARIABILITY

Even in the intrinsic channel FD SOI MOSFETs, variability remains, and the variability seems to be enhanced in the next generations as the transistor size shrinks. Therefore, a new concept to cope with the variability problem is strongly required. In this section, a postfabrication scheme for self-suppression of SRAM variability (or self-improvement of SRAM stability) is proposed [23] and experimentally demonstrated [24]. This concept requires techniques of nonvolatile $V_{th}$ shift of transistors and $V_{th}$ shifts with high voltage stress were utilized in the experiments.

3.6.1 MECHANISM OF SELF-IMPROVEMENT

Figure 3.29 shows a schematic of a six-transistor SRAM cell [25]. The two storage nodes in the cell are named as VL and VR. In the self-improvement technique, the stress voltage is applied to the $V_{dd}$ terminal of SRAM cell array. $V_{dd}$ is raised from 0 V to the stress voltage (3.2 V in this study), keeping the word line (WL) at 0 V. The scan time (stress time) is only several seconds. Since the $V_{th}$ shift of nFETs was small enough while the $|V_{th}|$ shift of pFETs is much larger at a stress voltage of 3.2 V, the self-improvement technique in the retention operation is explained based on the pFET $|V_{th}|$ shift in the following.

The SRAM cell is bistable in the retention condition when $V_{dd}$ is high enough and can store one bit per cell. The storage node VR can be either “high” or “low” in the bistable operation. However, when $V_{dd}$ is very low (e.g., 0.1 V), each SRAM cell is not bistable because of the unbalance of the four transistors that compose two inverters.
This is caused by the $V_{th}$ variability. Then, VR is fixed to only “high” in almost half of the cells at very low $V_{dd}$ and VR is fixed to only “low” in the other cells. In this technique, $V_{dd}$ is just scanned from 0 to 3.2 V.

At the beginning of $V_{dd}$ scan (0 V), VR is fixed to “high” in almost half-cells. As an example, let us assume that a cell whose VR is fixed to “high” at the beginning of $V_{dd}$ scan, as shown in Figure 3.29. This means that the strength to pull up VR is stronger than the strength to pull up VL in this cell. Therefore, $T_{pR}$ or $T_{nL}$ may be stronger (lower $|V_{th}|$) than $T_{pL}$ or $T_{nR}$. Here, we call a pFET connected to the high node as “p-ON” and a pFET connected to the low node as “p-OFF”, because the former is at the on state and the latter is at the off state. Similarly, “n-ON” and “n-OFF” are defined. It should be noted that, if p-ON is weakened, the cell stability is certainly improved, because the strength to pull up VR is weakened. Therefore, it is shown in the following that p-ON is the stronger pFET (that should be weakened for cell stability improvement) and p-OFF is the weaker pFET (that should be strengthened).

Next, let us consider the situation where $V_{dd}$ is raised to 3.2 V. The negative gate bias is automatically applied to only p-ON that is stronger, because this transistor is at the on state. This bias condition is just the same as that of negative bias temperature instability (NBTI) stress, as shown in Figure 3.30a. Positively charged interface traps are generated. Then, $|V_{th}|$ of p-ON is selectively raised and this transistor is weakened. On the other hand, the bias condition of p-OFF, which is weaker, is shown in Figure 3.30b. It is known that $|V_{th}|$ is raised by the off state due to negative charge generation in oxide near the drain, and this transistor is strengthened. As a result, the cell stability is improved [25].

**FIGURE 3.29** A schematic diagram of a six-transistor SRAM cell. It is assumed that VR is fixed to “high” at very low $V_{dd}$. (From Hiramoto, T. et al., *IEICE Trans. Electron.*, E96-C, 759, Copyright 2013 IEICE. With permission.)
3.6.2 Measurements of $|V_{\text{th}}|$ Shift by High Voltage Stress

SRAM DMA TEG was fabricated with 40 nm bulk technology and the self-improvement technique was applied to 4 kbit SRAM cells. We pay attention to the $|V_{\text{th}}|$ shift of p-ON and p-OFF. By checking “high” or “low” of VL and VR, it is easy to determine which pFET is p-ON or p-OFF. Figure 3.31a shows measured $|V_{\text{th}}|$ shift of p-ON, which was originally stronger and should be weakened for the self-improvement [25]. It is found that a majority of p-ON transistors show positive $|V_{\text{th}}|$ shift, indicating that p-ON is weakened. This positive $|V_{\text{th}}|$ shift is caused by the NBTI stress.

Figure 3.31b shows measured $|V_{\text{th}}|$ shift of p-OFF, which is weaker [25]. Almost all p-OFF transistors show negative $|V_{\text{th}}|$ shift, indicating that p-OFF which was originally weak is strengthened by the high voltage stress. The shift is even larger than that of p-ON. The self-improvement mechanism works.

Figure 3.31 shows that not all cells exhibit the favorable $|V_{\text{th}}|$ shift, that is, some cells exhibit negative $|V_{\text{th}}|$ shift of p-ON and positive $|V_{\text{th}}|$ shift of p-OFF.

FIGURE 3.31 Measured $|V_{\text{th}}|$ shifts of pFETs in 4 kbit SRAM cells. (From Hiramoto, T. et al., IEICE Trans. Electron., E96-C, 759, Copyright 2013 IEICE. With permission.)
which is the opposite direction to the self-improvement. It is found that cells that show the opposite $|V_{\text{th}}|$ shift are originally stable cells [25] (not shown). It is thought that these cells are so stable that “high” or “low” of the storage nodes is not determined at the beginning of $V_{\text{dd}}$ scan. As a result, p-ON and p-OFF are interchanged resulting in the opposite $|V_{\text{th}}|$ shift. However, since these cells are still stable enough after the opposite $|V_{\text{th}}|$ shift, this phenomenon does not result in yield loss of SRAM.

3.6.3 MEASUREMENTS OF SELF-IMPROVEMENT OF SRAM STABILITY

Figure 3.32 shows examples of measured butterfly curves in the retention condition (WL is 0 V) before and after applying high voltage stress [25]. Here, RetNM(L) is defined as the square of the left eye of the butterfly curve and RetNM(R) is the square of the right eye. RetNM is the smaller square of two eyes of the butterfly curve. The change of butterfly curves by the stress is explained in the following.

As mentioned in Section 3.6.1, VR is fixed to “low” or “high” at the beginning of the $V_{\text{dd}}$ scan. In a cell in Figure 3.32a, “VR = low” is stable at low $V_{\text{dd}}$, and hence RetNM(R) is larger and RetNM(L) is smaller. The right pFET ($T_{pR}$) is the p-OFF. When high voltage stress is applied, $|V_{\text{th}}|$ of p-OFF is lowered, and the inverter curve by $T_{pR}$ and $T_{nR}$ moves in the right direction, as shown in Figure 3.32a. Similarly, raised $|V_{\text{th}}|$ of p-ON moves the inverter curve by $T_{pL}$ and $T_{nL}$ down. In this way, both p-ON and p-OFF contribute to enlarge RetNM(L), resulting in the cell stability improvement.

In a cell in Figure 3.32b, on the other hand, “VR = high” is stable, and hence, RetNM(R) is smaller. In this specific cell, only negative $\Delta |V_{\text{th}}|$ of p-OFF

![FIGURE 3.32 Measured butterfly curves in the retention condition before and after applying high voltage: (a) a cell where “VR = low” is stable at low $V_{\text{dd}}$ and (b) a cell where “VR = high” is stable at low $V_{\text{dd}}$. (From Hiramoto, T. et al., IEICE Trans. Electron., E96-C, 759, Copyright 2013 IEICE. With permission.)](image-url)
Variability in Scaled MOSFETs

contributes, but no $\Delta |V_{th}|$ of p-ON is observed. The cell stability of this cell is also improved, because RetNM(R) is enlarged.

Figure 3.33 shows measured RetNM distributions before and after the high voltage stress to $V_{dd}$ in 4k SRAM cells [25]. Clear improvement of RetNM is observed, particularly in the worst cell that have originally the smallest RetNM. Since $V_{min}$ is determined by the worst cell, this self-improvement technique largely contributes to the yield enhancement of SRAM cells.

3.7 CONCLUSIONS

The present status of the variability in scaled transistors is reviewed. The variability of a large number of transistors is extensively measured, and it is shown that the main origin of random variability in bulk transistors is RDF. The relationship between cell transistor variability and cell stability in SRAM was analyzed. As an approach to variability suppression, two methods are described: (1) utilization of intrinsic channel FD MOSFET to avoid RDF and (2) novel self-improvement technique of SRAM stability. These results will largely contribute to further device scaling and further minimization of energy consumption in future VLSI.

ACKNOWLEDGMENTS

A part of this work was performed under the MIRAI project, LEAP project, and ELP project supported by NEDO.
REFERENCES


