Effect of the Cell Size Reduction on the Threshold Voltage of UMOSFETs

Yoshiro Baba and Ichiro Omura
Effect of cell size reduction on the threshold voltage of UMOSFETs

Yoshiro Baba,* Ichiro Omura

Kyushu Institute of Technology, Kitakyushu, Japan

Abstract

UMOSFET on-resistances have been dramatically improved in recent decades with the miniaturization of cell size by innovations in the fabrication process. However, with miniaturization, failure in the gate oxide, large deviations in the threshold voltage and reductions in avalanche capability have emerged as design problems for mass production. In particular, threshold voltage rises have appeared with the introduction of a trench source contact. The source contact trench and MOS gate trench are fabricated next to each other with a narrow silicon mesa region, and the voltage rises appear when the silicon mesa width becomes narrower than 80 nm. So far, it appears that the P+ layer dopant in the contact sidewall diffuses toward the gate oxide and the channel doping increases, which causes the rise in the threshold voltage Vth. We analyzed the distribution of Vth at the wafer level/shot level and found for the first time that the rise in Vth is caused by the punch-through effect from the channel depletion layer to the contact P+ layer, and thus, sidewall dopant diffusion will not affect the rise in Vth. We established an analytical model for the rise in Vth. Our model showed that for the field plate type UMOSFET with a shorter gate contact length, not only is there a rise in Vth, but also it is difficult to control Vth using the conventional channel implantation method.

1. Introduction

The breakdown voltage and RonA tradeoff relationship in UMOSFETs that appeared in the 1990s has been dramatically improved with the miniaturization of cell size by innovations in the fabrication process[1]. Figure 1 shows the trend in the cell size and RonA for a 60V-UMOSFET. To reduce the cell size, it is necessary to reduce the width of the contacts and the gate trench. In the 2000s, UMOSFETs changed from a planar contact structure to a trench contact structure. Typical cell size changed from 2.6 um to 1.5 um, and the gate contact length (Si mesa width) changed from 0.6 um to 0.15 um [2]. In the 2010s, UMOSFETs changed to a field plate structure with greatly improved capacitance characteristics to withstand the breakdown voltage and RonA tradeoff relationship [3]. However, with the field plate structure, it is necessary to form the source electrode and gate electrode in the same trench structure. So conversely, the gate trench width becomes larger. Scaling down the cells by reducing the Si mesa width has been considered. The Si mesa width is the insulation distance between the poly gate electrode and the source metal; however, it is susceptible to causal defects and is also a critical length that governs gate failure.

The following three issues must considered when shortening the Si mesa width. The first issue is the rise in and the difficulty in controlling the threshold voltage Vth. The second issue is gate failures due to the increased sensitivity to random manufacturing defects [4]. The last issue is the avalanche capacity during inductive switching due to random manufacturing defects that increase the base resistance by shielding P+ ion implantation in the base contact region [5][6].

In this paper, we report on the first issue using different Vth distributions in conventional UMOSFETs with a tapered contact structure and normal trench contact structure as examples. The rise in Vth was thought to be caused by the increase in the channel region concentration related to P+ diffusion from the trench contact sidewall. However, when there was
punch through from the gate depletion layer to the contact diffusion layer, which has no diffusion effect on the channel region, a rise in $V_{th}$ was observed with the tapered contact structure.

Fig. 1 Relationship between cell size and RonA for the 60V-UMOSFET

2. Measurement and Analysis

2.1. $V_{th}$ of UMOSFET with trench contact structure

Figure 2 shows the measured UMOSFET structures for two types of conventional UMOSFET with the same Si mesa width of 150 nm. One has a normal trench contact structure. Another has a 30° tapered contact structure. The difference between the two structures lies in the P+ diffusion layer length and the base concentration. The difference in the mean $V_{th}$ values is due to the base concentration.

We define $cr$ as the effective mesa width excluding the P+ diffusion layer length. The effective Si mesa widths $cr$ were 120 nm and 92 nm respectively. Figure 3 shows the measured $V_{th}$ distributions for both structures. Comparing both structures, the $V_{th}$ for the tapered contact structure had a periodic pattern dependent on the shot map of the lithography instrument. The total deviation in $V_{th}$ is the sum of the geometric mean of the deviation in the shot area and the deviation in the wafer. The total $V_{th}$ deviation for the tapered contact structure was larger than for the normal contact structure. The periodic deviation in $V_{th}$ for the tapered contact structure is thought to depend on the deviation in $cr$.

Fig. 2 Measured boron concentration distribution near the trench contact for the two types of UMOSFET

Fig. 3 Measured $V_{th}$ distributions for the two types of UMOSFET: (a), (b) $V_{th}$ wafer map and $V_{th}$ histogram for the normal contact structure; (c), (d) $V_{th}$ wafer map and $V_{th}$ histogram for the tapered contact structure

2.2 Calculating $V_{th}$

From the $V_{th}$ measurements for the tapered contact structure, the rise in $V_{th}$ is dependent on $cr$. When $cr$ is small, there is punch through from the gate depletion layer to the P+ diffusion layer, where $w$ is the gate depletion width at the gate bias $V_g$. When $cr$ is larger than $w$, the gate electric field is triangular and the gate surface potential does not depend on $cr$. However, when $cr$ is smaller than $w$, the gate electric field is trapezoidal. To form the surface inversion layer, a higher gate bias is needed.

When there is punch through from the gate depletion layer to the contact side wall of the P+ diffusion layer over the voltage $V_{pt}$, the depletion layer width $w$ can be calculated from $cr$ with the following equations:

$$V_{pt} = 2\varepsilon Si/\varepsilon_{ox} \cdot t_{ox} \cdot E_{pt} \cdot \Phi n + E_{pt} \cdot cr/2 \cdot \Phi p,$$

$$E_{pt} = cr \cdot q \cdot N_A/e/\varepsilon Si,$$

$$w = E_{si} \cdot e \cdot \varepsilon Si/q/N_A,$$

(1)

(2)
where \( na \) is the base concentration. The potential distribution in both cases is shown in Fig. 4.

![Potential distribution](image)

**Fig. 4 Potential distribution from gate to source contact**

When \( w \) is smaller than \( cr \), the gate surface electron concentration \( N_{s1} \) and the gate bias voltage \( V_{g1} \) can be calculated from the surface electric field \( E_{si} \) as follows:

\[
\phi_{s1} = w \cdot E_{si}/2 - \phi_p.
\]

\[
N_{s1} = n_i \exp(q \phi_{s1}/kT).
\]

\[
V_{g1} = t_{ox} \cdot \varepsilon_{ox} - \Phi_p - \phi_{s1}. \quad (3)
\]

When \( w \) is larger than \( cr \), the gate surface electron concentration \( N_{s2} \) and gate bias voltage \( V_{g2} \) can be calculated from the surface electric field \( E_{si} \) as follows:

\[
\phi_{s2} = cr \cdot (2E_{si} - q \cdot N_A \cdot cr/\varepsilon_{si}/\varepsilon)/2 - \phi_p.
\]

\[
N_{s2} = n_i \exp(q \phi_{s2}/kT).
\]

\[
V_{g2} = t_{ox} \cdot \varepsilon_{ox} - \Phi_p - \Phi_s2. \quad (4)
\]

Calculated \( N_{s1}(V_{g1}) \) and \( N_{s2}(V_{g2}) \) curves are shown in Fig. 5 for different values of the parameter \( cr \). When the base concentration \( na \) is \( 1.0 \times 10^{17} \) cm\(^{-3} \) and the gate oxide thickness \( t_{ox} \) is 50 (nm), \( V_{th} \) is defined as the gate bias voltage \( V_g \) at which \( N_s \) is equal to \( N_A \). In the conventional model \((w < cr)\), \( V_{th} \) is 2.0 (V) and is a constant independent of \( cr \). However, the model in which \( V_{th} \) is dependent on \( cr \) \((w > cr)\) shows a rise in \( V_{th} \) when \( cr \) is smaller than 80 (nm).

**Table 1**

<table>
<thead>
<tr>
<th>Physical constants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constants</strong></td>
</tr>
<tr>
<td>( q ) Elementary charge ( 1.6 \times 10^{-19} ) C</td>
</tr>
<tr>
<td>( \varepsilon_{si} ) Si dielectric ( 11.7 \times 8.854 \times 10^{-14} ) F/cm</td>
</tr>
<tr>
<td>( \varepsilon_{ox} ) SiO(_2) dielectric ( 3.8 \times 8.854 \times 10^{-14} ) F/cm</td>
</tr>
<tr>
<td>( k ) Boltzmann constant ( 1.38 \times 10^{-23} ) J/K</td>
</tr>
<tr>
<td>( n_i ) Intrinsic concentration ( 1.18 \times 10^{10} ) cm(^{-3} )</td>
</tr>
</tbody>
</table>

**Table 2**

Parameters used in model verification

<table>
<thead>
<tr>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P base concentration ( N_A ) ( 1.0 \times 10^{17} ) cm(^{-3} )</td>
</tr>
<tr>
<td>Gate poly-Si concentration ( 1.0 \times 10^{19} ) cm(^{-3} )</td>
</tr>
<tr>
<td>Gate oxide thickness ( t_{ox} ) 50 nm</td>
</tr>
</tbody>
</table>

**Fig. 5 Inversion layer electron concentration calculated with the proposed model as functions of the gate voltage \( V_g \) for different \( cr \) lengths**

2.3 Calculating \( V_{th} \) with manufacturing variations

In the 2000s, the typical chip size of a UMSFET was 4 mm\(^2\) and its gate width was about 4 m. The chips consist of many cell transistors and there are dimensional variations in the active area of the chip.

The \( V_{th} \) of a chip is approximately the average \( v_{th} \) of the cell transistors in the chip’s active area. The \( V_{th} \) of the chip is calculated from the convolution of the \( v_{th}(cr) \) function and the Gaussian distribution \( g(x - cr) \) as shown in Eqs. (5) and (6). One of the root causes of the variation is lens aberrations in the lithography equipment. There are two main types of dimensional variation in the cell transistors due to the manufacturing process. It is important to calculate the rise in \( V_{th} \) from \( cr \) and its variations. Figure 7 shows the definition of parameter \( a \) as the spread from the designed width of the contact and gate size. The distribution of
for the cell transistor approximates a Gaussian distribution. We used a 13.3 nm standard deviation for the chip, obtained from actual measurement. Figure 7 shows the original gate contact distance with the P+ diffusion and with cr0 on the x-axis, where cr is the mean effective Si mesa width with a P+ diffusion length of 58 nm for the tapered contact structure and (cr – a) is the mean effective Si mesa width with the spread parameter a. The chip’s Vth (cr, a) for the parameter a is calculated from Eqs. (7) and (8). The measured maximum value of a was 24 nm. Figure 8 shows the calculated distribution of vth for the cell transistor with a = 0 nm (no spread) and a = 24 nm (max spread). Vth for the chip transistor is the average of the blue and orange distributions calculated by a Monte Carlo simulation using 5000 cell transistors. The results show that Vth for the chip transistor rises from a vth0 of 1.48 (V) to 1.64 (V) as the spread parameter a is increased from 0 (nm) to 24 (nm).

\[
d(cr) = \int_{0}^{\infty} vth(x) \cdot g(x - cr) \, dx \\
Vth(cr) = d (cr)/d (x_0) \cdot vth_0 \\
d(cr, a) = \int_{0}^{\infty} vth(x) \cdot g(x - cr + a) \, dx \\
Vth(cr, a) = d (cr, a)/d (x_0) \cdot vth_0
\]  

The parameter b is the alignment error between the gate pattern and the contact pattern. In general, the contact pattern alignment is used with the gate pattern as the datum. Figure 9 shows the effect of the alignment error b on the Si mesa width.

The alignment error b gives the two distributions for the Si mesa width shown in Fig. 9. The variation in the chip’s Vth (cr, b) with alignment error b is calculated from Eqs. (9), (10) and (11).

\[
d2(cr, b) = \int_{0}^{\infty} vth(x) \cdot g(x - cr - b) \, dx \\
Vth(cr, b) = (d1 (cr) + d2 (cr))/2 \cdot d (x_0) \cdot vth_0
\]  

Fig. 9 Distributions of vth for the cell transistor separated by alignment error b

We measured alignment error b from the cross-sectional view of the taped contact UMOSFET cells in Fig. 2. The measured maximum value of b was 12 nm. In Fig. 10, the blue distribution is for the larger Si mesa width (cr + b) and the orange distribution is for the shorter Si mesa width (cr – b). The green distribution is the normalized total of both distributions. The chip’s Vth is the mean of the green distribution. The Vth shifts from 1.48 (V) to 1.52 (V).

2.4 Vth sensitivity to parameters a and b

Figure 11 shows the sensitivity of the chip’s Vth to parameters a and b. The sensitivity of a is higher
than that of b. Unexpectedly, the alignment error b is not as sensitive to the chip’s Vth. However, it is thought that b is an important parameter for the RonA characteristics and avalanche capability.

2.5 Root cause of variations in Vth on the shot map

Figure 12 shows the distribution of Vth for the whole wafer for a UMOSFET with a tapered contact structure. The periodical distribution in Vth is assumed to be due to lens aberrations in the lithography equipment. The shot area of the lithography equipment is 23 mm × 23 mm, and there are 23 × 10 chips in the X and Y directions.

We measured the spread a in the shot area and made the contour spread map shown in Fig. 12.

Figure 13 shows the distribution of Vth calculated with the measured spread map and the actual measured distribution of Vth with alignment error b = 4 nm. The calculated distribution pattern was similar to the measured pattern. The mean and standard deviation of Vth calculated for the chips within the shot area were 1.54 V and 46 mV. And the measured mean and standard deviation of Vth for the 230 chips in the shot area were 1.61 V and 51 mV. The chips within the yellow area had a higher Vth than the chips within the blue area (w > cr) due to the spread a.

The periodical distribution of Vth for the UMOSFET with a tapered contact structure can be explained by a shift (spread) in the contact and gate pattern caused by lens aberrations.

3. Discussion

The conventional Vth control parameters were tox and NA, but in the gate depletion layer punch-through model, cr has been added to the control parameters. The effect of reducing cr will become severe in field plate MOSFETs [7]. Figure 14 shows the Vth calculation curves for parameters cr and NA with no manufacturing variations. The gate oxide thickness was fixed at 50 nm. The base concentration was changed to 0.75 × 10¹⁷, 1.0 × 10¹⁷ and 1.25 × 10¹⁷ (cm⁻³). When cr is larger, Vth increases. The spread cr shows a smaller spread in the Vth curve. The small cr parameter is effective for reducing the spread in Vth.

Using the assumed manufacturing variations for the field plate UMOSFET, we calculated Vth for the chip. Figure 15 shows the Vth curves for different values of the spread parameter a. The base concentration and the gate oxide were 1.0 × 10¹⁷ cm⁻³ and 50 nm. The standard deviation of the Si mesa width was 13.3 nm, the same as for the conventional UMOSFET. The contact size was assumed to be about half that of the conventional UMOSFET. So the range of spread parameter a changed to 0.5 and 10 nm. When cr was smaller than 50 nm, the chip’s Vth rose drastically and became more sensitive to cr than in the conventional
4. Conclusion

We investigated the relationship between the rise in $V_{th}$ in UMOSFETs and the reduction in cell size. The rise in $V_{th}$ was thought to be due to the increase in the base concentration as a result of P+ diffusion from the contact sidewall. However, we found that even when there were no effects from the diffusion layer, when there was punch through to the P+ diffusion layer from the gate depletion layer, $V_{th}$ increased similarly. Our gate depletion layer punch-through model indicated that, for the latest field plate MOSFET with a shorter gate contact length, it will be difficult not only to suppress the rise in $V_{th}$ but also to control $V_{th}$ using the conventional channel implantation method.

Acknowledgements

The authors would like to thank ROHM Shiga Co. Ltd for providing evaluation devices, Y. Aki and T. Izutsu for their useful advice, K. Watanabe, H. Yanagimori and H. Sugie for their technical support and J. Fukushima for his helpful discussions.

References