The Impact of Layout-Dependent STI Stress and Effective Width on Low-Frequency Noise and High-Frequency Performance in Nanoscale nMOSFETs

Kuo-Liang Yeh, *Member, IEEE*, and Jyh-Chyurn Guo, *Senior Member, IEEE*

***Abstract*—The impact of channel width scaling on low-frequency noise (LFN) and high-frequency performance in multifinger MOSFETs is reported in this paper. The compressive stress from shallow trench isolation (STI) cannot explain the lower LFN in extremely narrow devices. STI top corner rounding (TCR)-induced ΔW is identified as an important factor that is responsible for the increase in transconductance G_m and the reduction in LFN with width scaling to nanoscale regime. A semi-empirical model was derived to simulate the effective mobility $(\mu_{\rm eff})$ degradation from STI stress and the increase in effective width $(W_{\rm eff})$ from ΔW due to STI TCR. The proposed model can accurately predict width scaling effect on G_m based on a tradeoff between $\mu_{\rm eff}$ and $W_{\rm eff}$. The enhanced STI stress may lead to an increase in interface traps density $(N_{\rm it})$, but the influence is relatively minor and can be compensated by the $W_{\rm eff}$ effect. Unfortunately, the extremely narrow devices suffer f_T degradation due to an increase in $C_{\rm gg}$. The investigation of impact from width scaling on $\mu_{\rm eff}$, G_m, and LFN, as well as the tradeoff between LFN and high-frequency performance, provides an important layout guideline for analog and RF circuit design.**

***Index Terms*—Effective mobility, effective width, low-frequency noise (LFN), shallow trench isolation (STI) stress.**

I. Introduction

With the aggressive scaling of CMOS technology into the nanoscale regime, the stress introduced from shallow trench isolation (STI) process becomes significant and imposes a dramatic impact on not only CMOS device performance in dc current and ac gate speed [1]–[6] but also high frequency and analog performance. Among previous work on this subject, most of research effort and publications have been focused on the longitudinal stress $(\sigma_{//})$ from STI along the direction of the channel length [1]–[3]. However, relatively fewer studies were done for looking into the influence from STI transverse stress $(\sigma_{\perp})$ along the gate width and transverse to the channel length [4]–[6]. In addition, most of the studies on $\sigma_{\perp}$ are limited to dc characteristics, such as threshold voltage (V_T) and channel current $(I_{\rm DS})$, and restricted to long-channel devices, assuming negligible $\sigma_{//}$ in the channel region [4]. STI edge effect on flicker noise was investigated, and negative impact was reported for NMOS with narrow widths [7], [8]. A minor layout modification, i.e., an edge-extended design, was implemented, trying to reduce the stress and traps introduced by STI and fix the problem of non-1/f characteristics in the measured flicker noise [9]. However, the abnormal deviation from 1/f characteristics as reported for narrow devices cannot be reproduced from our experiment, even with a much more aggressively scaled width. A ring-type device was proposed, trying to eliminate $\sigma_{\perp}$ and identify the influence on flicker noise [10]. However, the study is limited to single-polygate MOSFETs, which are not suitable for analog and RF circuit design, and the impact on high-frequency performance is unknown [10]. Recent studies reported a turn around of $I_{\rm DS}$ with decreasing width of the active area and proposed an increase in effective width, i.e., ΔW, from STI top corner rounding (TCR) as the potential factor trading with mobility degradation from STI compressive stress [4], [5]. Again, most of the works were restricted to single-polygate MOSFETs for logic circuits, but the influence on multiple-polygate-finger (i.e., multifinger) MOSFETs required for RF and analog circuits, with special concern of high-frequency performance and flicker noise, remains an open question.

The aforementioned multifinger MOSFETs have been widely used to reduce the parasitic gate resistance (R_g) in RF circuits for the purpose of improving analog and RF performance like higher maximum oscillation frequency $(f_{\rm MAX})$ and lower RF noise [11]. Note that the multifinger transistor has been recognized as the standard layout for RF circuit design. In this paper, two new layouts derived from the standard one, i.e., narrow-OD and multi-OD, were implemented, trying to enhance the transverse stress $(\sigma_{\perp})$ from STI. The primary objective is to investigate STI transverse stress effect on effective mobility $(\mu_{\rm eff})$, transconductance (G_m), cutoff frequency (f_T), $f_{\rm MAX}$, and, most importantly, the low-frequency noise (LFN) for both RF and analog circuit design.

II. Device Fabrication and Characterization

In this paper, multifinger MOSFETs were fabricated in 90-nm low-leakage CMOS process with nitrided oxide of target physical thickness at 2.2 nm. The gate length drawn on the

Manuscript received May 10, 2010; accepted August 9, 2010. Date of publication October 4, 2010; date of current version November 5, 2010. This work was supported in part by the National Science Council under Grant NSC 98-2221-E009-166-MY3. The review of this paper was arranged by Editor Z. Celik-Butler.

The authors are with the Institute of Electronics Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: jcguo@mail.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2010.2072959

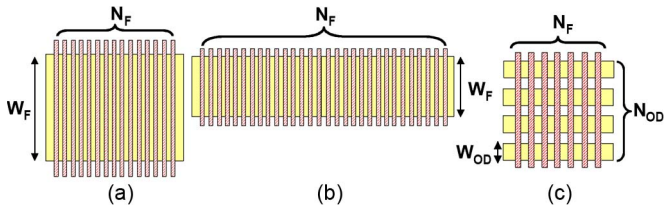

Fig. 1. Schematic of multifinger MOSFETs with three different layouts. (a) Standard multifinger device: $W_F \times N_F = 2\ \mu\text{m} \times 16$ (W2N16). (b) Narrow-OD device: $W_F \times N_F = 1\ \mu\text{m} \times 32$ (W1N32) and $0.5\ \mu\text{m} \times 64$ (W05N64). (c) Multi-OD devices: $W_{\text{OD}} \times N_{\text{OD}} = 2\ \mu\text{m} \times 1$, $0.25\ \mu\text{m} \times 8$, and $0.125\ \mu\text{m} \times 16$.

layout is 90 nm, i.e., $L_{\text{drawn}} = 90$ nm, and the total channel width W_{tot} is fixed at 32 μm. In order to investigate the stress and interface traps generated from the STI process, two kinds of new layouts derived from multifinger MOSFET, i.e., multi-OD and narrow-OD, were designed and implemented. Note that OD means oxide diffusion, which is equivalent to active area (generally denoted as AA). Fig. 1(a)–(c) displays the device layouts for standard, narrow-OD, and multi-OD, respectively. The narrow-OD devices illustrated in Fig. 1(b) were designed with simultaneously varied N_F and W_F under a specified total width, i.e., $W_{\text{tot}} = W_F \times N_F$. For this group of structures, $W_F \times N_F = 2\ \mu\text{m} \times 16$, $1\ \mu\text{m} \times 32$, and $0.5\ \mu\text{m} \times 64$ corresponding to finger width $W_{\text{tot}} = 32\ \mu$m were fabricated in this work. The multi-OD devices shown in Fig. 1(c) represent multiple-OD fingers with simultaneously varied OD finger width (W_{OD}) and OD finger number (N_{OD}) under a specified finger width, which is $W_F = W_{\text{OD}} \times N_{\text{OD}}$. For this category of devices, the polygate finger number N_F is fixed at 16, and $W_{\text{OD}} \times N_{\text{OD}} = 2\ \mu\text{m} \times 1$, $0.25\ \mu\text{m} \times 8$, and $0.125\ \mu\text{m} \times 16$, corresponding to finger width $W_F = 2\ \mu$m. Note that the polygate-edge/OD-edge distance along the direction of channel length is fixed at 0.5 μm for both narrow-OD and multi-OD devices.

S-parameters were measured by an Agilent network analyzer E8364B for high-frequency characterization and ac parameter extraction. Open de-embedding was performed to remove the parasitic capacitances from the pads and interconnection lines, and short de-embedding was done to eliminate the parasitic resistances and inductances originated from the metal interconnection. Fig. 2 demonstrates the LFN measurement system, which consists of an Agilent dynamic signal analyzer (DSA 35670) and low-noise amplifier (LNA SR570). The aforementioned system was employed to measure the power spectral density of drain current noise, i.e., S_{ID}, which is the so-called LFN. In this paper, the LFN was measured from multifinger MOSFETs under various gate voltages (V_{GS}) and fixed drain voltage at $V_{\text{DS}} = 50$ mV. Note that V_{GS} is converted as gate overdrive, i.e., $V_{\text{GT}} = V_{\text{GS}} - V_T$ to offset V_T variation among different devices. The LFN measurement generally covers a wide frequency range of 4 Hz–10 kHz. The aforementioned measurement for dc and LFN characterization has been carried out on 10–15 dies for a statistical analysis.

The charge-pumping (CP) current system was built up, as shown in Fig. 3(a), for interface trap density (N_{it}) measurement. The CP method proposed by Elliot [12] with sweeping base and fixed amplitude was applied to extract N_{it} from the

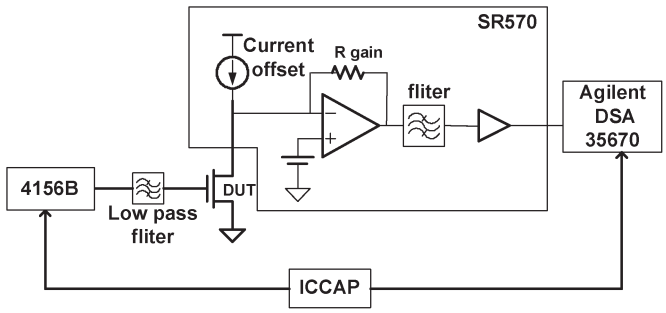

Fig. 2. LFN measurement system setup consisting of an Agilent dynamic signal analyzer (DSA 35670), low noise amplifier (LNA SR570), and Agilent 4156B for dc power supply. The measurement is automatically controlled by Agilent ICCAP.

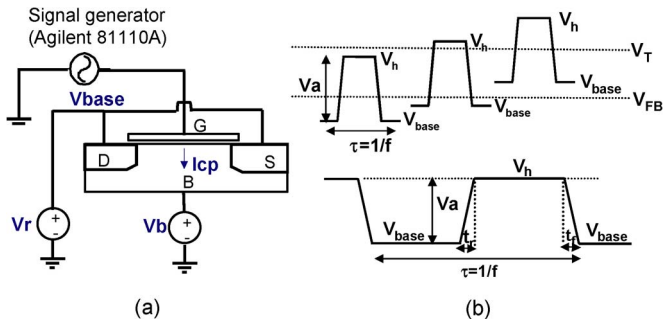

Fig. 3. (a) CP measurement system consisting of dc source Agilent 4156B, pulse generator 81110A, and switching matrix Keithley 707A. (b) CP pulse waveform, where t_r and t_f are the rising and falling times, and V_{base} and V_h are swept from accumulation $(V_{\text{base}} < V_{\text{FB}})$ to inversion $(V_h > V_T)$, under fixed pulse amplitude $V_a = V_h - V_{\text{base}}$.

maximum CP current $(I_{\text{CP,max}})$. Fig. 3(b) illustrates the pulse waveform employed for the Elliot method, in which the base level (V_{base}) is swept from accumulation to inversion while keeping the pulse amplitude (V_a) constant. Note that $I_{\text{CP,max}}$ can occur under the condition of $V_{\text{base}} < V_{\text{FB}}$ and $V_{\text{base}} + V_a = V_h > V_T$, and the accuracy is justified with a clear plateau for $I_{\text{CP,max}}$ and a precise linear dependence on amplitude (V_a) and frequency (not shown).

III. Local Strain Engineering Effect on NMOS Performance

The STI stress introduced in MOSFETs with three different layouts as previously mentioned (standard, narrow-OD, and multi-OD) are illustrated in Fig. 4 to assist an analysis and understanding of layout effect on STI stress and, then, the electrical characteristics. Note that STI stress is classified as longitudinal stress, which is denoted as $\sigma_{//}$ and in parallel with the channel length, and transverse stress, i.e., $\sigma_{\perp}$, which is transverse to the channel length. In this paper, the longitudinal stress $\sigma_{//}$ is considered to be similar for all of the devices with various layouts, due to fixed gate length and polygate-edge/OD-edge distance [13]. Regarding the stress favorable for mobility enhancement, it has a critical dependence on the device types and orientations. The experimental results in previous work indicate that tensile stress in the longitudinal direction $(\sigma_{//})$ can enhance electron mobility in NMOS, whereas it will degrade hole mobility in PMOS. As for the transverse stress $\sigma_{\perp}$ of our

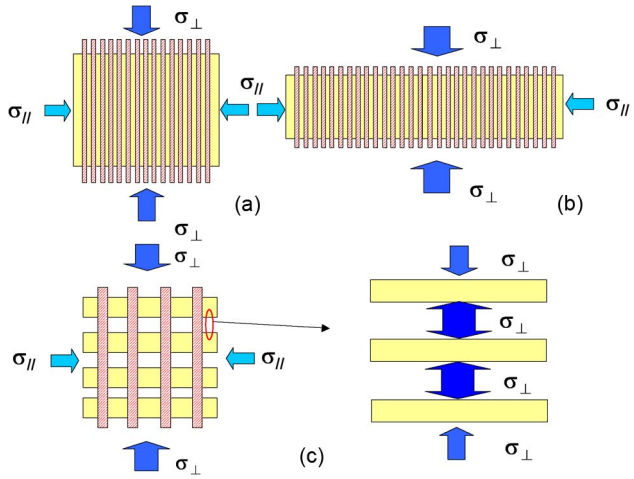

Fig. 4. Schematic of STI stresses along the longitudinal and transverse directions, which are defined as $\sigma_{//}$ and $\sigma_{\perp}$ in MOSFETs with three different layouts. (a) Standard multifinger device. (b) Narrow-OD device. (c) Multi-OD devices.

major interest in this paper, the compressive stress from STI is expected to degrade the effective mobility $\mu_{\rm eff}$ for both NMOS and PMOS [6].

A. STI Transverse Stress Effect on DC Performance of Multi-OD and Narrow-OD NMOS

As previously mentioned, narrow-OD and multi-OD MOSFETs were designed to study OD width scaling effect on transverse stress $\sigma_{\perp}$ from STI in different layouts. Narrow-OD devices are shown in Fig. 4(b), wherein the OD width $(W_{\rm OD} = W_F)$ is reduced to be extremely narrow and the polygate finger number (N_F) is simultaneously increased to keep the total width $(W_{\rm tot} = W_F \times N_F)$ unchanged. Multi-OD devices illustrated in Fig. 4(c) features a matrix of multiple ploy fingers, as well as multiple OD fingers $(N_{\rm OD})$ with narrow OD width $(W_{\rm OD})$. Note that $W_{\rm OD}$ and $N_{\rm OD}$ are simultaneously varied to keep $W_{\rm OD} \times N_{\rm OD}$ the same as the finger width of standard device, i.e., $W_{\rm OD} \times N_{\rm OD} = W_F = 2\ \mu$m.

It has been known that the STI process generally leads to V_T lowering with channel width scaling, and it is the so-called inverse narrow-width effect (INWE) [14]. As shown in Fig. 5, the V_T versus $W_{\rm OD}$ for narrow-OD and multi-OD NMOS presents an obvious INWE, following a universal curve for different layouts. Note that INWE is determined by collective effects from STI TCR (ΔW), STI stress, corner field crowding, doping profile, etc. Considering V_T variations from the aforementioned effects, $V_{\rm GT} = V_{\rm GS} - V_T$ is used for electrical characterization and analysis.

Fig. 6(a) presents the transconductance (G_m) under varying $V_{\rm GT}$'s, which are measured from narrow-OD NMOS with two splits of $W_F \times N_F$, such as 1 μm × 32 (W1N32) and 0.5 μm × 64 (W05N64) and the standard one (W2N16) for comparison. The result indicates that the smaller $W_{\rm OD}\ (= W_F)$ leads to lower G_m and the maximum G_m $(G_{m,\max})$ in W05N64 $(W_{\rm OD} = W_F = 0.5\ \mu$m) is degraded by about 8%, compared with the standard one, i.e., W2N16 $(W_{\rm OD} = W_F = 2\ \mu$m), as shown in Fig. 6(b). The monotonic degradation of G_m with

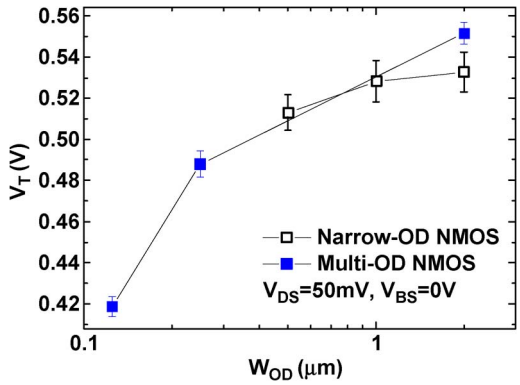

Fig. 5. Linear V_T versus $W_{\rm OD}$ for narrow-OD and multi-OD NMOSs under the biases of $V_{\rm DS} = 50$ mV and $V_{\rm BS} = 0$ V.

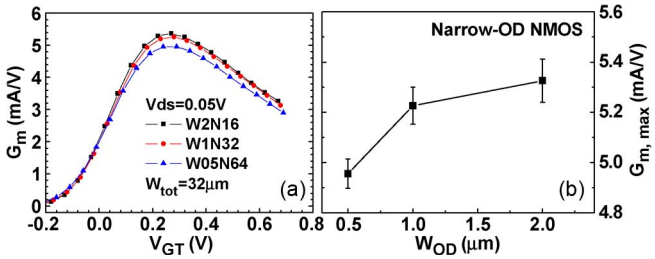

Fig. 6. (a) Transconductance G_m versus $V_{\rm GT}$. (b) Maximum transconductance $G_{m\_\max}$ measured from narrow-OD NMOS W1N32 $(W_F = 1\ \mu$m, $N_F = 32)$ and W05N64 $(W_F = 0.5\ \mu$m, $N_F = 64)$, and standard multifinger device W2N16 $(W_F = 2\ \mu$m, $N_F = 16)$. All of the devices have the same total finger width $W_F \times N_F = 32\ \mu$m.

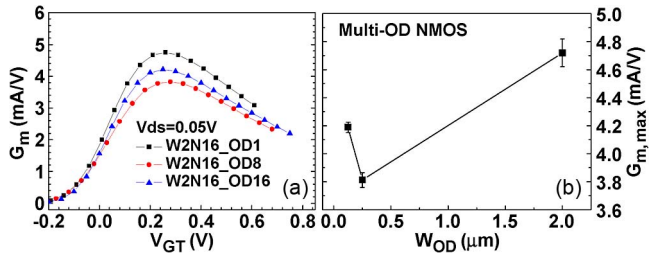

Fig. 7. (a) Transconductance G_m versus $V_{\rm GT}$. (b) Maximum transconductance $G_{m\_\max}$ measured from multi-OD NMOS OD8 $(N_{\rm OD} = 8, W_{\rm OD} = 0.25\ \mu$m) and OD16 $(N_{\rm OD} = 16, W_{\rm OD} = 0.125\ \mu$m), and standard multifinger device OD1 $(N_{\rm OD} = 1, W_{\rm OD} = 2\ \mu$m). All of the devices have the same polygate finger number $(N_F = 16)$ and total OD width along each gate finger $((N_{\rm OD} \times W_{\rm OD} = 2\ \mu$m).

$W_{\rm OD}$ scaling in narrow-OD devices suggests that the increase in STI compressive $\sigma_{\perp}$ is the dominant factor responsible for $\mu_{\rm eff}$ degradation and the resulting G_m degradation. As for multi-OD NMOS shown in Fig. 7, the $G_{m,\max}$ of OD8 $(N_{\rm OD} = 8, W_{\rm OD} = 0.25\ \mu$m) is degraded by about 20%, compared with the standard one (OD1), but the continuous scaling of $W_{\rm OD}$ to 0.125 μm in OD16 $(N_{\rm OD} = 16, W_{\rm OD} = 0.125\ \mu$m) leads to an increase in G_m, compared with OD8, and the $G_{m,\max}$ degradation, compared with the standard one, is shrunk to 11%. The result looks very interesting and cannot be explained by STI compressive stress alone. Note that the statistical variations of V_T and G_m remain less than 2% from measurement on 10–15 dies, and it ensures the experimental results a high confidence level.

To explain the unusual characteristics measured from multi-OD devices with extremely narrow $W_{\rm OD}$, the increase in $W_{\rm eff}$ due to STI TCR is proposed as the primary mechanism trading with the STI compressive stress effect to determine the channel current and transconductance G_m. Note that G_m is the key parameter responsible for RF and analog circuit performance. At first, the STI stress effect on mobility is calculated by the model as follows:

$$\frac{\Delta\mu}{\mu_0} = -(k_\perp \sigma_\perp + k_{//}\sigma_{//}) \tag{1}$$

where

- μ_0 mobility free from STI-stress-induced degradation;
- $\Delta\mu$ mobility variation due to STI stress;
- $\sigma_\perp$ transverse stress perpendicular to the direction of the channel length;
- $\sigma_{//}$ longitudinal stress along with the direction of the channel length;
- $k_\perp$ first order of coefficient expressing mobility variation proportional to $\sigma_\perp$;
- $k_{//}$ first order of coefficient expressing mobility variation proportional to $\sigma_{//}$.

Then

$$\frac{\Delta\mu}{\mu_0} \cong -k_\perp \sigma_\perp \tag{2}$$

where $\sigma_\perp$ is a function of gate width expressed by

$$\sigma_\perp = -k \cdot \log(W_{\rm OD}) + k_0 \tag{3}$$

and can be written as another expression given by

$$\sigma_\perp = k \cdot \log\left(\frac{W_{\rm ref}}{W_{\rm OD}}\right) \tag{4}$$

where

$$W_{\rm OD} \times N_{\rm OD} = W_F. \tag{5}$$

In the preceding expressions, both longitudinal stress $\sigma_{//}$ and transverse stress $\sigma_\perp$ are considered in the original model. For multifinger MOSFETs, $\sigma_{//}$ limits its effect to the polygate at the two ends. Therefore, $\sigma_{//} \ll \sigma_\perp$ for all of the polygates other than those at the two ends and it leads to the approximation of (1) to (2).

According to (3) or (4), it is proposed that G_m should decrease with the reduction in $W_{\rm OD}$ since the mobility decreases as the transverse compressive stress $\sigma_\perp$ increases with $W_{\rm OD}$ scaling.

However, in this work, we found that OD width scaling from $W_{\rm OD} = 0.25$ μm in OD8 to $W_{\rm OD} = 0.125$ μm in OD16 led to an increase in $G_{m,\max}$, as shown in Fig. 7, rather than degradation predicted by the stress model in (1)–(5). This observation suggests that the transverse compressive stress $\sigma_\perp$ from STI, which is maximized in OD16 due to the minimum $W_{\rm OD}$, cannot fully explain the largest $G_{m,\max}$ degradation in OD8, instead of OD16. The experimental results suggest that the variation of G_m with $W_{\rm OD}$ scaling is determined by not only STI stress effect on mobility ($\mu_{\rm eff}$) but also STI TCR effect on effective width ($W_{\rm eff}$), i.e., ΔW effect [5], [15]. For both OD8 and OD16, the large compressive stress $\sigma_\perp$ from STI spreads into the active region and degrades the $\mu_{\rm eff}$. However, for OD16 devices, the ΔW effect dominates and causes the increase in G_m. To simulate the proposed STI stress and TCR effects on G_m in miniaturized MOSFET, the semi-empirical formulas are derived as follows [4]:

$$I_{\rm DS} = W_{\rm eff} C_{\rm ox}(V_{\rm GS} - V_T - \alpha V_{\rm DS})\mu_{\rm eff}\frac{V_{\rm DS}}{L_{\rm eff}} \tag{6}$$

$$G_m = \frac{\partial I_{\rm DS}}{\partial V_{\rm GS}} = W_{\rm eff} C_{\rm ox}\mu_{\rm eff}\frac{V_{\rm DS}}{L_{\rm eff}}. \tag{7}$$

Let

$$\beta = C_{\rm ox}\frac{V_{\rm DS}}{L_{\rm eff}} \tag{8}$$

$$\mu_{\rm eff}(W_{\rm OD}) = \mu_0 + \Delta\mu(W_{\rm OD}) = \mu_0\left[1 - k \cdot \log\left(\frac{W_{\rm ref}}{W_{\rm OD}}\right)\right] \tag{9}$$

$$W_{\rm eff} = (W_{\rm OD} + \Delta W) \times N_{\rm OD} \times N_F. \tag{10}$$

Then

$$G_m = \beta\mu_0\left[1 - k \cdot \log\left(\frac{W_{\rm ref}}{W_{\rm OD}}\right)\right](W_{\rm OD} + \Delta W)N_{\rm OD}N_F \tag{11}$$

where ΔW is the increase in OD finger width due to STI TCR, and N_F, $N_{\rm OD}$, and $W_{\rm OD}$ are defined in Fig. 1.

In this paper, the standard device with $W_F = W_{\rm OD} = 2$ μm is adopted as the reference, i.e., $W_{\rm ref} = 2$ μm to minimize the transverse stress $\sigma_\perp$ and its impact on mobility. The derived model given by (9) and (11) can accurately predict $\mu_{\rm eff}$ and G_m measured from multifinger devices with various OD layouts, such as standard, narrow-OD, and multi-OD.

Fig. 8(a) and (b) shows that simultaneous best fitting to $\mu_{\rm eff}$ and G_m for multi-OD NMOS can be realized under the condition of $\Delta W = 43$ nm and $k = 0.2888$. Note that ΔW is dependent on the OD layout and varies between the standard and multi-OD NMOSs. For a standard device with more gradient trench profile, $\Delta W = 24$ nm appears to be smaller than that of multi-OD devices. For narrow-OD devices, even with minimum OD width in W05N64, its OD width ($W_{\rm OD} = W_F = 0.5$ μm) is four times larger than that of OD16 ($W_{\rm OD} = 0.125$ μm), and ΔW effect is not strong enough to overcome $\mu_{\rm eff}$ degradation from STI stress. It can be understood that the ΔW ratio calculated by

$$\Delta W\_ratio = \frac{\Delta W \times N_{\rm OD} \times N_F}{W_{\rm eff}} = \frac{\Delta W}{W_{\rm OD} + \Delta W} \tag{12}$$

$$W_{\rm OD} = \frac{W_F}{N_{\rm OD}} \tag{13}$$

is about 10.6% for narrow-OD device W05N64 with $W_{\rm OD} = W_F = 0.5$ μm, whereas that of multi-OD device OD16 with $W_{\rm OD} = 0.125$ μm is significantly increased to 32.1%, which is about three times larger than that of narrow-OD device W05N64.

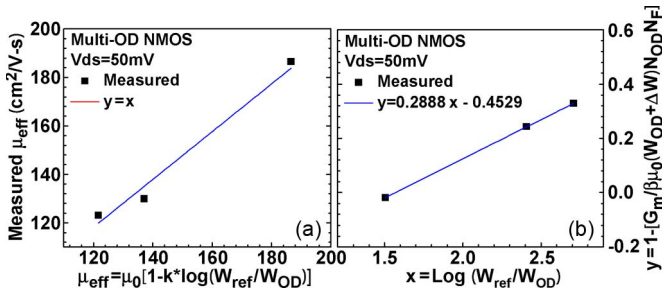

Fig. 8. (a) Measured $\mu_{\rm eff}$ and the fitting by model (9). (b) Measured G_m and the fitting by model (11). The simultaneous best fitting for multi-OD NMOS was realized under the condition of $\Delta W = 43$ nm and $k = 0.2888$.

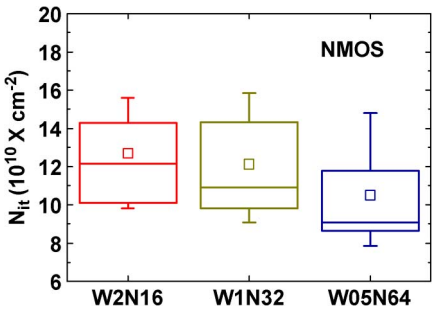

Fig. 9. Interface trap density $N_{\rm it}$ of W2N16, W1N32, and W05N64 devices extracted by using $W_{\rm eff}$, which is measured at $V_D = V_S = V_B = 0$ V and $V_G = V_{\rm amp} = 1.2$ V.

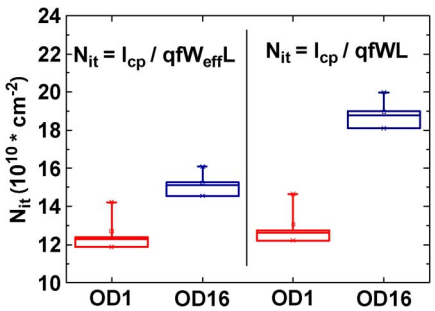

Fig. 10. Interface trap density $N_{\rm it}$ of OD1 and OD16 devices extracted by using $W_{\rm eff}$ and W, measured at $V_D = V_S = V_B = 0$ V and $V_G = V_{\rm amp} = 1.2$ V. $W_{\rm eff} = (W_{\rm OD} + \Delta W)N_{\rm OD} * N_F {\rm W} = W_{\rm OD}N_{\rm OD} * N_F$.

B. Interface Traps of Narrow-OD and Multi-OD NMOS

Strain effect on interface traps and the resulting impact on carrier mobility is one more critical concern for the deployment of strain engineering in the state-of-the-art process. Figs. 9 and 10 present the interface trap density $N_{\rm it}$ extracted by the CP current method, according to

$$I_{\rm cp\_max} = qfW_{\rm eff}LN_{\rm it} \tag{14}$$

for narrow-OD and multi-OD devices, respectively, where $I_{\rm cp,max}$ is the maximum of the measured CP current, f is the frequency of the applied pulse waveform, and L is the gate length of the device under test. Note that the effective width $W_{\rm eff}$ is used to reflect the ΔW effect.

As shown in Fig. 9, the $N_{\rm it}$ extracted from narrow-OD NMOS indicates nearly the same distribution for W2N16 and W1N32, and a minor decrease in W05N64 with minimal

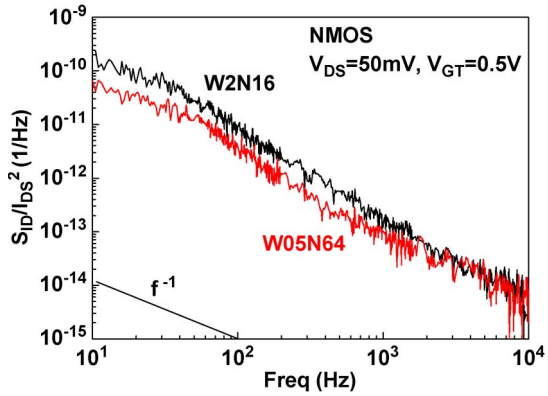

Fig. 11. LFN $S_{\rm ID}/I_{\rm DS}^2$ measured from the standard NMOS W2N16 and narrow-OD NMOS W05N64, under the biases of $V_{\rm DS} = 50$ mV and $V_{\rm GT} = 0.5$ V.

$W_F(= W_{\rm OD})$ in this category. The results just disapprove the conventional consideration that the narrower OD width may lead to higher $N_{\rm it}$ due to an increase in STI stress $\sigma_\perp$. As for multi-OD NMOS in Fig. 10, OD16 with minimal OD width ($W_{\rm OD} = 0.125$ μm) reveals apparently larger $N_{\rm it}$, compared with the standard one (OD1). The result suggests that extremely narrow OD combined with minimal OD-to-OD space in multi-OD devices may introduce a significant increase in $N_{\rm it}$ due to a substantially steeper trench profile, compared with narrow-OD devices. The extra trap density in the sidewall also contributes to the higher $N_{\rm it}$ of OD16. Note that ΔW should be considered to accurately determine the effective active area and $N_{\rm it}$ in devices with extremely narrow $W_{\rm OD}$. The increase in $N_{\rm it}$ of OD16, which is normalized to that of OD1 and denoted as $\Delta N_{\rm it(OD16,OD1)}/N_{\rm it(OD1)}$, is as large as 55% when ΔW is neglected and reduced to near 10% when ΔW is considered and $W_{\rm eff}$ is taken.

C. LFN $S_{\rm ID}/I_{\rm DS}^2$ in Narrow-OD and Multi-OD NMOS

Fig. 11 presents LFN in terms of $S_{\rm ID}/I_{\rm DS}^2$ measured from narrow-OD NMOS with minimum OD width, i.e., W05N64 ($W_{\rm OD} = 0.5$ μm) and the comparison with the standard one, i.e., W2N16 ($W_{\rm OD} = 2$ μm). Similarly, Fig. 12 makes a comparison of $S_{\rm ID}/I_{\rm DS}^2$ between multi-OD NMOS with minimal OD width, i.e., OD16 ($W_{\rm OD} = 0.125$ μm) and the standard one (OD1, $W_{\rm OD} = 2$ μm). Note that the noise spectra for all of the devices (standard, narrow-OD, and multi-OD) follow 1/f characteristics over a wide frequency domain from 10 Hz to 10 kHz. The measured LFN manifests itself as a typical flicker noise.

As shown in Fig. 11, the narrow-OD NMOS W05N64 presents lower $S_{\rm ID}/I_{\rm DS}^2$ than the standard device W2N16. The result looks consistent with the lower $N_{\rm it}$ for narrower OD width, as shown in Fig. 9. More rigorously, the larger $W_{\rm eff}$ for narrower OD due to the ΔW effect have to be considered and will be discussed later. As for multi-OD devices shown in Fig. 12, it is interesting to note that OD16 with extremely narrow $W_{\rm OD}$ to 0.125 μm presents lower $S_{\rm ID}/I_{\rm DS}^2$ than the standard one (OD1, $W_{\rm OD} = 2$ μm)), even though the $N_{\rm it}$ of OD16 is somewhat higher than that of OD1 (Fig. 10). The reduction in LFN with OD width scaling in both

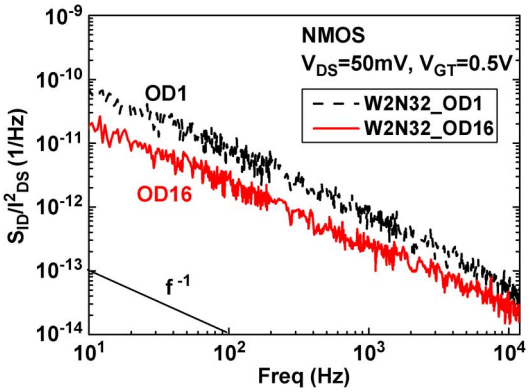

Fig. 12. LFN $S_{\rm ID}/I_{\rm DS}^2$ measured from the standard NMOS OD1 ($N_{\rm OD}=1, W_{\rm OD}=2\ \mu$m) and multi-OD NMOS OD16 ($N_{\rm OD}=16, W_{\rm OD}=0.125\ \mu$m) under the biases of $V_{\rm DS}=50$ mV and $V_{\rm GT}=0.5$ V.

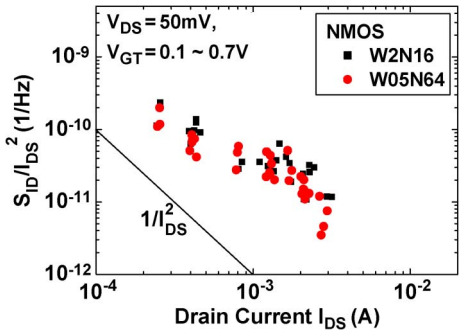

Fig. 13. Measured $S_{\rm ID}/I_{\rm DS}^2$ versus $I_{\rm DS}$ under $V_{\rm GT}=50$ mV and various $V_{\rm GT}$ (0.1–0.7 V) for standard NMOS W2N16 and narrow-OD NMOS W05N64.

narrow-OD and multi-OD NMOS is in contradiction with the general expectation and suggests that ΔW from STI and its effect on $W_{\rm eff}$ appears to be an important factor in explaining the extraordinary results.

To further explore the mechanism responsible for LFN, the $S_{\rm ID}/I_{\rm DS}^2$ measured at 50 Hz and under various $V_{\rm GT}$'s are plotted versus $I_{\rm DS}$ for narrow-OD and multi-OD NMOS, as shown in Figs. 13 and 14, respectively. Note that LFN data have been collected from ten devices for each layout to perform a statistical analysis. Herein, the measured $S_{\rm ID}/I_{\rm DS}^2$ follows a function proportional to $1/I_{\rm DS}^2$ over the whole range of bias conditions ($V_{\rm GT}=0.1-0.7$ V), which indicates that the number fluctuation model given by

$$\frac{S_{\rm ID}}{I_{\rm DS}^2}=\frac{q^2k_BT\lambda N_{\rm it}}{W_{\rm eff}LC_{\rm ox}^2}\frac{1}{f}\frac{G_m^2}{I_{\rm DS}^2} \tag{15}$$

is the dominant mechanism governing NMOS devices' LFN, where $N_{\rm it}$ is the density of interface traps at quasi-Fermi level [16].

The proposed number fluctuation model suggests that the $S_{\rm ID}/I_{\rm DS}^2$ of NMOS is proportional to $N_{\rm it}/W_{\rm eff}$ and predicts the decrease in LFN with increasing effective width $W_{\rm eff}$. It was expected that the increase in compressive transverse stress $\sigma_\perp$ as well as interface traps $N_{\rm it}$ suffered by OD16 device might aggravate interface scattering effect and increase the flicker noise [17]. However, the larger $W_{\rm eff}$ may compensate or even overcome these effects. The aforementioned mechanism

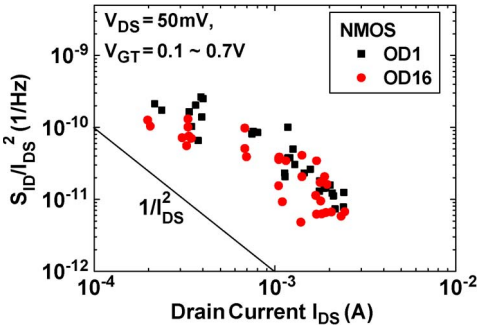

Fig. 14. Measured $S_{\rm ID}/I_{\rm DS}^2$ versus $I_{\rm DS}$ under $V_{\rm GT}=50$ mV and various $V_{\rm GT}$ (0.1–0.7 V) for standard NMOS OD1 ($N_{\rm OD}=1, W_{\rm OD}=2\ \mu$m) and multi-OD NMOS OD16 ($N_{\rm OD}=16, W_{\rm OD}=0.125\ \mu$m).

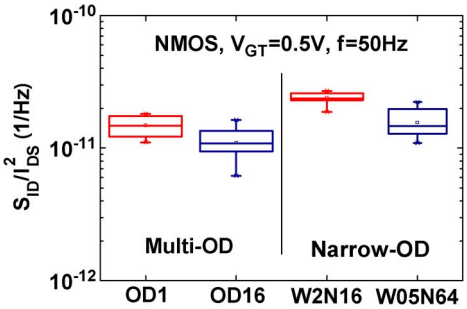

Fig. 15. Statistical distribution of LFN $S_{\rm ID}/I_{\rm DS}^2$ measured from ten dies for multi-OD and narrow-OD NMOSs.

can explain why the multi-OD devices with extremely narrow $W_{\rm OD}$, such as OD16, can have lower LFN, compared with OD1. Fig. 15 indicates the statistical distribution of $S_{\rm ID}/I_{\rm DS}^2$ measured from ten dies for narrow-OD and multi-OD NMOS devices to further investigate OD width scaling effect on LFN. The ratio of OD16:OD1 is in the range of 40%–87%. According to (15), the ΔW effect can lead to 32% increase in $W_{\rm eff}$ for OD16 and then 24% lower LFN, i.e., OD16 : OD1 = 76% from ΔW effect alone. Considering the statistical distribution of $N_{\rm it}$ shown in Fig. 10, the increase in $W_{\rm eff}$ of OD16 leads to a decrease in $N_{\rm it}$, and the ratio of OD16:OD1 is about 1.2 for the mean value and 1.13 for the maximum in the distribution. The combined effect from ΔW and $N_{\rm it}$ indeed leads to 10%–15% lower LFN in OD16 than OD1. The statistical data suggest that the ΔW effect may not be the only factor but is the primary factor contributing to lower LFN.

For narrow-OD devices, the $N_{\rm it}$ is similar between W2N16 and W05N64 (Fig. 8), and the larger $W_{\rm eff}$ in W05N64 can significantly overwhelm $N_{\rm it}$ and the stress effects. As a result, the smaller OD width in W05N64 leads to lower LFN ($S_{\rm ID}/I_{\rm DS}^2$), compared with W2N16. The investigation of layout effects on LFN in multifinger MOSFETs discloses an interesting and important phenomenon that STI stress collaborating with STI TCR-induced ΔW constitutes the primary mechanism responsible for the OD width scaling effect on G_m and LFN.

D. High-Frequency Performance in Narrow-OD and Multi-OD NMOSs

The potential impact from the aforementioned STI stress and TCR-induced ΔW on high-frequency performance is of one

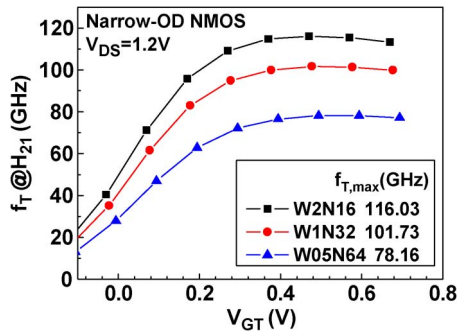

Fig. 16. Cutoff frequency f_T versus $V_{\rm GT}$ measured from narrow-OD NMOS under $V_{\rm DS} = 1.2$ V and varying $V_{\rm GT}$. f_T is extracted from the unit current gain $|H_{21}| = 1$.

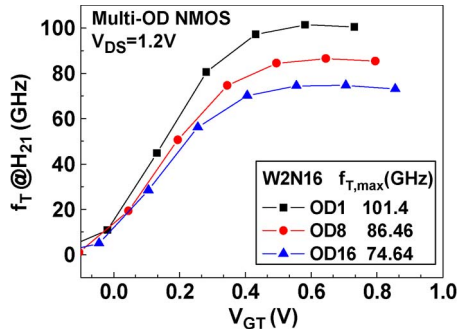

Fig. 17. Cutoff frequency f_T versus $V_{\rm GT}$ measured from multi-OD NMOS under $V_{\rm DS} = 1.2$ V and varying $V_{\rm GT}$. f_T is extracted from unit current gain $|H_{21}| = 1$.

more special concern for multifinger MOSFETs in RF circuit design. Fig. 16 illustrates the cutoff frequency f_T measured from standard (W2N16) and narrow-OD (W1N32, W05N64) NMOS under varying $V_{\rm GT}$. Note that H-parameters were converted from S-parameters after two-step de-embedding (open and short), and then, f_T was extracted from the extrapolation of $|H_{21}|$ to unity current gain. The experimental results reveal that narrow-OD devices suffer significant degradation of f_T and the peak f_T drops from 116 GHz for W2N16 to 78 GHz for W05N64. Referring to

$$f_T = \frac{G_m}{2\pi\sqrt{C_{\rm gg}^2 - C_{\rm gd}^2}} \tag{16}$$

which is an analytical model for calculating f_T [18], it is predicted that f_T degradation may be originated from degradation of G_m or increase in $C_{\rm gg}$. For narrow-OD devices, the smallest G_m appearing in W05N64 (Fig. 6) suggests one of the factors responsible for the worst f_T. Furthermore, $C_{\rm gg}$ measured from narrow-OD NMOS indicates the largest one for W05N64 (not shown). The collective effect from G_m and $C_{\rm gg}$ can explain f_T degradation in narrow-OD devices. Regarding OD width scaling effect on f_T in multi-OD NMOS, Fig. 17 demonstrates that OD16 with the smallest OD width ($W_{\rm OD} = 0.125$ μm) reveals the worst f_T. The peak f_T drops from 101 GHz for OD1 to 75 GHz for OD16. $C_{\rm gg}$ measured from multi-OD NMOS indicates that OD16 suffers about 20% larger $C_{\rm gg}$ than that of OD1 (not shown). The collective effect from smaller G_m (Fig. 7) and larger $C_{\rm gg}$ is responsible for f_T degradation.

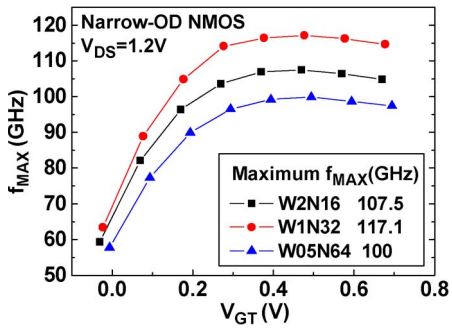

Fig. 18. $f_{\rm MAX}$ versus $V_{\rm GT}$ measured from narrow-OD NMOS under $V_{\rm DS} = 1.2$ V and varying $V_{\rm GT}$. $f_{\rm MAX}$ is determined by the unilateral gain method.

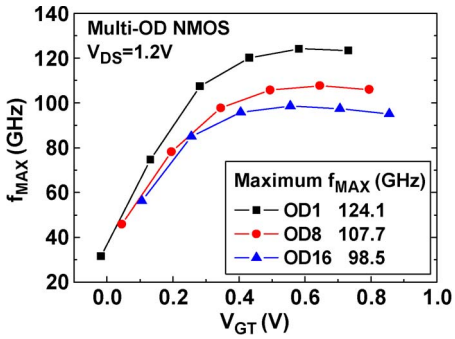

Fig. 19. $f_{\rm MAX}$ versus $V_{\rm GT}$ measured from multi-OD NMOS under $V_{\rm DS} = 1.2$ V and varying $V_{\rm GT}$. $f_{\rm MAX}$ is determined by the unilateral gain method.

The maximum oscillation frequency $f_{\rm MAX}$ is another important performance parameter for RF circuit design, particularly for power amplifiers. The impact of layout variation on $f_{\rm MAX}$ in multifinger MOSFET is investigated as follows: In this paper, $f_{\rm MAX}$ was determined by the conventionally used unilateral gain method. Fig. 18 presents $f_{\rm MAX}$ extracted from standard and narrow-OD NMOS. It reveals a very interesting result that W1N32 attains about 10 GHz higher $f_{\rm MAX}$ than the standard (W2N16), even though W1N32 suffers the lower f_T than W2N16 (Fig. 16). However, W05N64 with the lowest f_T remains the worst in terms of $f_{\rm MAX}$. Through an equivalent circuit analysis on unilateral gain (U), $f_{\rm MAX}$ can be approximated by

$$f_{\rm MAX} = \frac{f_T}{2\sqrt{R_g(g_{\rm DS} + 2\pi f_T C_{\rm gd}) + g_{\rm DS}({\rm R}_i + {\rm R}_s)}} \tag{17}$$

in which R_g and f_T appear as two key parameters controlling $f_{\rm MAX}$ [11]. This expression predicts that the higher f_T and lower R_g can increase $f_{\rm MAX}$. For narrow-OD devices, the smaller $W_F (= W_{\rm OD})$ and larger N_F play a multiplied effect to reduce R_g (not shown). The higher $f_{\rm MAX}$ achieved by W1N32, compared with W2N16, suggests that the significant reduction in R_g can compensate or even overcome the impact from f_T degradation. However, the advantage of R_g suppression cannot be obtained by multi-OD devices due to the feature of fixed N_F [Fig. 1(c)]. The $f_{\rm MAX}$ measured from multi-OD NMOS, as shown in Fig. 19, proves the expectation that the trend of $f_{\rm MAX}$ simply follows that of f_T (Fig. 17), i.e., the smaller $W_{\rm OD}$ leads to the lower $f_{\rm MAX}$ and OD16 suffers worse degradation in $f_{\rm MAX}$.

IV. Conclusion

The potential impact from layout-dependent STI stress on LFN and high-frequency performance has been investigated on multifinger MOSFETs with various layouts, such as narrow-OD and multi-OD. The monotonic decrease in G_m with finger width (W_F) scaling in narrow-OD NMOS proves $\mu_{\rm eff}$ degradation from the compressive STI stress along the transverse direction $(\sigma_\perp)$. However, the multi-OD NMOS reveals an abnormal G_m increase for extremely narrow OD width to $W_{\rm OD} = 0.125\ \mu$m. The observed results suggest that STI stress is not the only mechanism governing the electrical property in miniaturized devices. STI TCR-induced ΔW is identified as another key factor, which may overcome STI stress effect in determining channel current and G_m. Semi-empirical formulas have been derived to successfully predict $W_{\rm OD}$ scaling effect on $\mu_{\rm eff}$ and G_m. Taking this method, ΔW can be precisely extracted based on a simultaneous best fitting to $\mu_{\rm eff}$ and G_m, and the resulting increase in effective width $(W_{\rm eff})$ is dramatically large to about 34% for OD16 with $W_{\rm OD} = 0.125\ \mu$m. The larger $W_{\rm eff}$ becomes, the more it contributes to reducing LFN and overcoming $N_{\rm it}$ effect in narrow-OD and multi-OD devices with sufficiently small $W_{\rm OD}$. The reduction in LFN with OD width scaling is the other evidence reflecting STI TCR-induced ΔW effect.

Unfortunately, the OD width scaling leads to a negative impact on high-frequency performance like f_T and $f_{\rm MAX}$, due to G_m degradation and undesired increase in $C_{\rm gg}$. An improved open de-embedding method can reduce the parasitic capacitances from intermetal coupling but cannot eliminate gate-related fringing capacitances. The multifinger MOSFETs with miniaturized OD width cannot prevent from f_T degradation. The tradeoff between LFN and high-frequency performance identified from this work provides an important layout guideline for analog and RF circuit design.

Acknowledgment

The authors would like to thank the National Device Laboratory (NDL) for their support during noise measurement and the Chip Implementation Center (CiC) for test key tape-out and device fabrication.

References

[1] J. Xue, Z. Ye, Y. Deng, H. Wang, L. Yang, and Z. Yu, "Layout-dependent STI stress analysis and stress-aware RF/analog circuit design optimization," in *Proc. ACM/IEEE Int. Conf. Comput.-Aided Des.*, 2009, pp. 521–528 .

[2] K.-W. Su, Y.-M. Sheu, C.-K. Lin, S.-J. Yang, W.-J. Liang, X. Xi, C.-S. Chiang, J.-K. Her, Y.-T. Chia, C. H. Dim, and C. Hu, "A scalable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2003, pp. 245–248.

[3] A. B. Kahng, P. Sharma, and R. O. Topaloglu, "Chip optimization through STI-stress-aware placement perturbations and fill insertion," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, no. 7, pp. 1241–1252, Jul. 2008.

[4] R. Li, L. Yu, H. Xin, Y. Dong, K. Tao, and C. Wang, "A comprehensive study of reducing the STI mechanical stress effect on channel-width-dependent $I_{\rm dsat}$," *Semicond. Sci. Technol.*, vol. 22, no. 12, pp. 1292–1297, Nov. 2007.

[5] P. B. Y. Tan, A. V. Kordesch, and O. Sidek, "Analysis of deep submicron CMOS transistor Vtlin and Idsat versus channel width," in *Proc. Asia-Pacific Microw. Conf.*, 2005, pp. 1569–1572.

[6] Y. Luo and D. K. Nayak, "Enhancement of CMOS performance by process-induced stress," *IEEE Trans. Semicond. Manuf.*, vol. 18, no. 1, pp. 63–68, Feb. 2005.

[7] R.-V. Wang, Y.-H. Lee, Y.-L. R. Lu, W. McMahon, S. Hu, and A. Ghetti, "Shallow trench isolation edge effect on random telegraph signal noise and implications for flash memory," *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 2107–2113, Sep. 2009.

[8] H. Lee, J.-H. Lee, H. Shin, Y.-J. Park, and H. S. Min, "An anomalous device degradation of SOI narrow width devices caused by STI edge influence," *IEEE Trans. Electron Devices*, vol. 49, no. 4, pp. 605–612, Apr. 2002.

[9] C.-Y. Chan, Y.-S. Lin, Y.-C. Huang, S. S. H. Hsu, and Y.-Z. Juang, "Edge-extended design for improved flicker noise characteristics in 0.13-μm RF NMOS," in *Proc. IEEE MTT-S Int. Microw. Symp.*, 2007, pp. 441–444.

[10] Y.-L.R. Lu, Y.-C. Liao, W. McMahon, Y.-H. Lee, H. Kung, R. Fastow, and S. Ma, "The role of shallow trench isolation on channel width noise scaling for narrow width CMOS and flash cells," in *Proc. Symp. VLSI Technol. Syst. Appl.*, 2008, pp. 85–86.

[11] H. Shimomura, A. Matsuzawa, H. Kimura, G. Hayashi, T. Hirai, and A. Kanda, "A mesh-arrayed MOSFET (MA-MOS) for high-frequency analog applications," in *VLSI Symp. Tech. Dig.*, 1997, pp. 73–74.

[12] A. B. M Elliot, "The use of charge pumping currents to measure surface state densities in MOS transistors," *Solid State Electron.*, vol. 19, no. 3, pp. 214–247, Mar. 1976.

[13] M. Miyamoto, H. Ohta, Y. Kumagai, Y. Sonobe, K. Ishibashi, and Y. Tainaka, "Impact of reducing STI-induced stress on layout dependence of MOSFET characteristics," *IEEE Trans. Electron Devices*, vol. 51, no. 3, pp. 440–443, Mar. 2004.

[14] A. Ono, R. Ueno, and I. Sakai, "TED control technology for suppression of reverse narrow channel effect in 0.1 μm MOS devices," in *IEDM Tech. Dig.*, 1997, pp. 227–230.

[15] G. Scott, J. Lutze, M. Rubin, F. Nouri, and M. Manley, "NMOS drive current reduction caused by transistor layout and trench isolation induced stress," in *IEDM Tech. Dig.*, 1999, pp. 827–830.

[16] G. Ghibaudo, O. Roux, C. H. Nguyen-Duc, F. Balestra, and J. Brini, "Improved analysis of low frequency noise in field-effect MOS transistors," *Phys. Stat. Sol. (A)*, vol. 124, no. 2, pp. 571–581, Apr. 1991.

[17] S. Maeda, Y. S. Jin, J. A. Choi, S. Y. Oh, H. W. Lee, J. Y. Woo, M. C. Sun, J. H. Ku, K. Lee, S. G. Bae, S. G. kang, J. H. Yang, Y. W. Kim, and K. P. Suh, "Impact of mechanical stress engineering on flicker noise characteristics," in *VLSI Symp. Tech. Dig.*, 2004, pp. 102–103.

[18] T. Manku, "Microwave CMOS—Device physics and design," *IEEE J. Solid-State Circuits*, vol. 34, no. 1, pp. 277–285, Mar. 1999.

Kuo-Liang Yeh (M'09) received the B.S.E.E. degree from National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 1995 and the M.S.E.E. degree from National Taiwan University, Taipei, Taiwan, in 1997. He is currently working toward the Ph.D. degree in electronics engineering in the Institute of Electronics Engineering, NCTU.

In 1999, he joined Taiwan Semiconductor Manufactory Company Inc., Hsinchu, where he has worked on process integration and yield improvement. From 2004 to 2007, he has been a Senior Engineer with MediaTek Inc., Hsinchu. He is currently a Senior Manager with the Silicon Motion Technology Corporation, Hsinchu. He has authored more than ten technical publications in international journals and conference proceedings. His research interests include the characterization and parameter extraction of CMOS devices for modeling and circuit simulation, as well as the protection of intellectual property rights.

Jyh-Chyurn Guo (M'06–SM'07) received the B.S.E.E. and M.S.E.E. degrees from National Tsing-Hua University (NTHU), Hsinchu, Taiwan, in 1982 and 1984, respectively, and the Ph.D. degree in electronics engineering from National Chiao-Tung University (NCTU), Hsinchu, in 1994.

For more than 19 years, she was with the semiconductor industry, where her major focus was on device design and VLSI technology development. In 1984, she joined the ERSO/ITRI, where she had been engaged in semiconductor integrated circuit technologies with a broad scope that covers high-voltage, high-power, submicrometer project, and high-speed SRAM technologies. From 1994 to 1998, she was with Macronix International Corporation and engaged in high-density and low-power Flash memory technology development. In 1998, she joined Vanguard International Semiconductor Corporation, where she assumed the responsibility of Device Department Manager for advanced DRAM device technology development. In 2000, she joined Taiwan Semiconductor Manufacturing Company (TSMC), where she served as Program Manager in charge of 100-nm logic CMOS FEOL, high-performance analog, and RF CMOS technology development. In 2003, she joined the Institute of Electronics Engineering, NCTU, as an Associate Professor, and she has been a Full Professor since 2008. She has authored or coauthored more than 60 technical papers. She is the holder of 19 U.S. patents in her professional field. Her current research interests include RF/MS CMOS device design and modeling for low-power and low-noise nanoscale CMOS noise modeling and strain engineering effects, broadband and scalable inductors modeling, novel nonvolatile memory technologies, and device integration technologies for system-on-chip.