Solid-State Electronics 54 (2010) 1304–1311

Contents lists available at ScienceDirect

Solid-State Electronics

journal homepage: www.elsevier.com/locate/sse

Selection of gate length and gate bias to make nanoscale metal–oxide-semiconductor transistors less sensitive to both statistical gate length variation and temperature variation

Peizhen Yang [a,*], W.S. Lau [a], Seow Wei Lai [b], V.L. Lo [b], S.Y. Siah [b], L. Chan [b]

[a] Nanyang Technological University, School of Electrical and Electronic Engineering, Division of Microelectronics, Block S2.1, Nanyang Avenue, Singapore 639798, Republic of Singapore
[b] GLOBALFOUNDRIES SINGAPORE Pte Ltd, 60 Woodlands Industrial Park D St. 2, Singapore 738406, Republic of Singapore

ARTICLE INFO

Article history:
Received 23 November 2009
Received in revised form 13 April 2010
Accepted 26 June 2010
Available online 24 July 2010
The review of this paper was arranged by Prof. S. Cristoloveanu

Keywords:
Temperature-independent-point
MOSFETs
Halo implant

ABSTRACT

Aggressive scaling of transistors leads to an ever-increasing amount of process variations. In this work, we studied the gate length dependency of on-current (I_{on}), off-current (I_{off}), effective drive current (I_{eff}), saturation threshold voltage ($V_{th,sat}$), and temperature independent point (TIP). Experimental evidence show that the gate length dependency of TIP in nanoscale transistors is related to the $V_{th,sat}$ versus L characteristics rather than velocity saturation. We found that I_{on}, I_{off} and I_{eff} of nanoscale transistors in the transition between reverse short channel effect (RSCE) and short channel effect (SCE) are less sensitive to gate length variation and temperature variation.

1. Introduction

The aggressive scaling of metal–oxide-semiconductor (MOS) transistors inevitably leads to an ever-increasing amount of process variations. Specifically, leakage power consumption, including its variability, is one of the major constraints in chip design [1]. The equation of the subthreshold off-current (I_{off}) can be expressed as [2]:

$$I_{ds} = \mu_{sub} C_{ox} \frac{W}{L}(m-1)\left(\frac{k_B T}{q}\right)^2 \times \exp\left[\frac{q(V_{GS}-V_{th,sat})}{mk_B T}\right][1-\exp(-qV_{DS}/k_B T)] \quad (1a)$$

where μ_{sub} is the low-field mobility in the subthreshold regime [3–5]. C_{ox} is the gate oxide capacitance per unit area. W is the mask gate width. L is the mask gate length. k_B, T and q are the Boltzmann's constant, the absolute temperature and the electron charge, respectively. V_{GS}, V_{DS} and $V_{th,sat}$ are the gate-to-source voltage, the drain-to-source voltage, and the saturation threshold voltage, respectively. The body-effect coefficient (m) can be expressed as [2]:

$$m = 1 + \frac{\sqrt{\varepsilon_0 \varepsilon_{Si} q N_{ch}/(4\psi_B)}}{C_{ox}} \quad (1b)$$

where ε_0 is the permittivity of free space. ε_{Si} is the dielectric constant of silicon. N_{ch} is the channel doping concentration. ψ_B is the difference between the Fermi level in the channel region and the intrinsic Fermi level. In general, the equation for subthreshold swing (S_{ts}) can be expressed as follows [2]:

$$S_{ts} = 2.3\frac{mk_B T}{q} \quad (1c)$$

As shown in Eq. (1c), S_{ts} increases with an increase in temperature, resulting in increase in I_{off} with temperature. For very short MOS transistors, the equation of S_{ts} has to be modified to account for the degradation of S_{ts} when short channel effects are strong [6] but qualitatively S_{ts} still decreases when C_{ox} increases.

The application of halo implant will cause a maximum point in the $V_{th,sat}$ versus L characteristics owing to the opposing effects of the reverse short channel effect (RSCE) [7] and the short channel effect (SCE). Hence, transistors located at the transition regime between SCE regime and RSCE regime is expected to be the least sensitive to the statistical variation in the effective channel length (L_{eff}) [8]. From Eq. (1a), transistors in the transition regime is expected to have the smallest variation in I_{off}. In addition, simulations have shown that halo implants will result in a minimum point in

* Corresponding author.
E-mail address: yang0121@e.ntu.edu.sg (P. Yang).

0038-1101/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sse.2010.06.026

the I_{off} versus L characteristics [9]. Despite the lack of experimental data, Thompson et al. pointed out that halo implants are used in the sub-100 nm MOS technologies to engineer the shape of the I_{off} versus L characteristics [10]. Our previous work showed experimentally that the minimum I_{off} occurs at gate length that lies between the RSCE and SCE regimes [11,12]. Hence, we expect the nanoscale transistors that lie between RSCE and SCE regimes to have the smallest variation in both $V_{th,sat}$ and I_{off}.

Besides I_{off} variation, the reversal in temperature dependency of the drain current (I_{ds}) across the temperature independent point (TIP) also has detrimental effects on the open loop gain of operational amplifier as each of its transistors responds differently to a change in temperature [13]. When the transistor is biased with V_{GS} that is below the TIP, I_{ds} will increase with increasing temperature owing to thermionic emission between the source and the channel (i.e. decrease in threshold voltage [14]). When the transistor is biased with V_{GS} that is above TIP, I_{ds} will decrease with increasing temperature owing to the degradation of phonon-limited mobility [15]. Here, we would like to point out that the drain current of nanoscale transistor in the linear operation is limited by surface roughness scattering rather phonon scattering. Hence, it is important to mention that an increase in temperature will bring about a decrease in mobility regardless if I_{ds} is limited by Coulomb scattering [16], phonon scattering [15] or surface roughness scattering [17]. For long-channel MOS transistor, the saturation drain current (I_{ds}) is related to the low-field mobility in the above threshold voltage regime (μ_{eff}) as follows [18]:

$$I_{ds} = \frac{\mu_{eff} C_{ox,inv} W}{2L}(V_{GS} - V_{th,sat})^2 \quad (2)$$

where $C_{ox,inv}$ is the gate oxide capacitance at inversion per unit area. Unlike long-channel transistors, the drain current transport of the nanoscale transistors is not clearly understood. Concepts involving velocity saturation [19,20] and ballistic transport [21] regard mobility irrelevant in the nanoscale transistors. However, stress engineering techniques [22–25] are able to increase the on-current (I_{on}) of nanoscale transistors through the strain-induced mobility enhancement [26,27]. By unifying the merits of quasi-ballistic theory [28,29] and ballistic theory [21], we have proposed a simplified equation for the saturation drain current of the nanoscale MOS transistor that can account for the strain-induced I_{on} enhancement but comprise of electrical parameters that are easily obtained from standard electrical measurement [30]:

$$I_{ds} = v_{sat\_eff} W C_{ox,inv}(V_{GS} - V_{th,sat\_IV}) \quad (3a)$$

where V_{th,sat_IV} , which is a fitting parameter, can be extracted from the saturation I_{ds} versus V_{GS} characteristics [30]. v_{sat_eff} is the average value of the carrier velocity (v_{eff}) when V_{GS} is close to V_{DD}. For n-channel MOS (NMOS) transistor with the conventional ⟨1 1 0⟩ channel orientation on (1 0 0) p–Si substrate, the application of tensile stress along the channel direction will increase both electron μ_{eff} and v_{sat_eff} [30]. For p-channel MOS (PMOS) transistor with ⟨1 1 0⟩ channel orientation, the application of compressive stress along the channel direction will increase both hole μ_{eff} and v_{sat_eff}. In fact, Tatsumura et al. has experimentally established a correlation between the low-field mobility and the high-field velocity [31]. Hence, Eq. (3a) can account for the strain-induced I_{on} improvement in nanoscale transistors by various strain engineering techniques [22–25]. The saturation transconductance method can be used to extract v_{eff} [32]. The analytical expression of v_{eff} can be expressed as follows [30]:

$$v_{eff}(\mu_{eff}, V_{GS}, T) = \left(\frac{1}{v_{inj}(V_{GS}, T)} + \frac{1}{\mu_{eff}(V_{GS}, T)\varepsilon(0^+)}\right)^{-1} \quad (3b)$$

where the injection velocity (v_{inj}), which is a concept introduced by Natori's 1994 ballistic theory, increases with increasing V_{GS} and increasing temperature [21]. $\mu_{eff}\,\varepsilon(0)^+$, which is introduced by Lundstrom's 1997 quasi-ballistic theory [28], decreases with increasing temperature. The analytical expression for the lateral electric field near the top of the potential barrier is [30]: $\varepsilon(0^+) = v_{sat\_eff}/\mu_{eff}$. The experimental $\varepsilon(0^+)$ of a NMOS transistor (L = 60 nm) is around 8.588×10^4 V/cm at V_{GS} of 1.2 V [30]. According to Lee et al. [33], the simulated $\varepsilon(0^+)$ of a PMOS transistor (L = 50 nm) is between 8×10^4 V/cm and 3×10^5 V/cm for various gate overdrives. In our previous work, we have shown that the second term of Eq. (3b) dominates over the first term because v_{eff} decreases with increasing temperature. However, v_{inj} can be used to account for the I_{on} enhancement at very low temperature such as liquid helium temperature [34–36] because v_{inj} increases with increasing V_{GS} [21].

From Eqs. (2) and (3a), TIP is expected to occur in both long-channel MOS transistor and nanoscale MOS transistors because both μ_{eff} and v_{sat_eff} decrease when temperature increases. In addition, Yan et al. reported that TIP has gate length dependency owing to the onset of velocity saturation [37]. However, Monte Carlo simulation indicates that velocity overshoot will occur in sub-100 nm MOS transistor [38]. Furthermore, velocity overshoot has been observed experimentally in bulk NMOS transistor (L = 32 nm) at room temperature [39]. This shows that the gate length dependency of TIP is unlikely to be caused by velocity saturation. Hence, the suggestion put forward by Goel et al. seems to be more appropriate [40]: the roll-off of TIP at smaller gate length is caused by the threshold voltage roll-off at shorter gate length. However, their simulation does not consider the effects of halo implants [40]. As opposed to Goel and Tan [40], the effects of halo implants are considered in this study. In this paper, we will provide experimental evidence to show that it is possible to fabricate nanoscale MOS transistors that are less sensitive to both statistical variation in gate length and temperature variation. We will also discuss the relationship between the TIP versus L characteristics and the $V_{th,sat}$ versus L characteristics. This allows us to tune TIP of the transistor so as to reduce the temperature sensitivity of I_{on}.

2. Experimental description

The relationship between the TIP versus L characteristics and the $V_{th,sat}$ versus L characteristics were studied using MOS transistors fabricated using 65 nm CMOS technology. The channel orientation was ⟨1 1 0⟩. The power supply voltage (V_{DD}) is 1.2 V. Shallow trench isolation (STI) was first formed on the p-type (1 0 0) Si substrate, followed by well implants, threshold adjustment implant, the formation of silicon oxynitride gate dielectric, and the patterning of the polycrystalline silicon gates. Halo implants were used to control SCE. Next, source/drain (S/D) extensions, and deep S/D regions were formed. Subsequently, S/D spike anneal was done. After nickel salicidation, 0.7 GPa tensile stressed contact etch stop layer (CESL) was used to boost the I_{on} performance of the NMOS transistors and then the conventional backend process followed.

It is well-known that the stress transfer of CESL has a gate length dependency: shorter channel transistor has a bigger tensile stress than longer channel transistor [41]. In order to decouple the effects of halo implants on I_{off} from the effects of CESL-induced tensile stress, we performed a CESL split for two adjacent wafers in the same lot: (i) neutral CESL, and (ii) tensile stress. The transistors were fabricated using 45 nm CMOS technology and had ⟨1 0 0⟩ channel orientation on (1 0 0) p-type silicon substrate. The power supply voltage (V_{DD}) is 1.1 V. Unlike the tensile stressed CESL, neutral CESL will only consider the gate length dependency of halo implants.