Impacts of a buffer layer and hydrogen-annealed wafers on the performance of strained-channel nMOSFETs with SiN-capping layer

Tzu-I Tsaia, Horng-Chih Lina,b, Yao-Jen Leeb,*, King-Sheng Chenb, Jeff Wangd, Fu-Kuo Hsuehb, Tien-Sheng Chaoc, Tiao-Yuan Huanga

a Institute of Electronics, National Chiao Tung University, Hsinchu, Taiwan
b National Nano Device Laboratories, Hsinchu, Taiwan
c Department of Electrophysics, National Chiao Tung University, Hsinchu, Taiwan
d Department of Nanotechnology Engineering, University of Waterloo, Waterloo, Canada

A R T I C L E   I N F O

Article history:
Received 15 May 2008
Accepted 7 June 2008
Available online 23 August 2008

Review of this manuscript was arranged by A. Iliadis, C. Richter, and A. Zaslavsky

Keywords:
Strained-Si
CESL
Si3N4 capping
Hydrogen-annealed wafer

A B S T R A C T

In this study, the effects of Si3N4 layer capping and TEOS buffer layer inserted prior to the Si3N4 deposition on the NMOS device characteristics as well as correlated hot-electron degradations were investigated. The devices were built on two kinds of the substrates, namely, Cz and hydrogen-annealed (Hi) wafers. More importantly, we found that hydrogen species is the primary culprit for aggravated reliabilities in strained devices. By exerting the accelerated stress test, we could study the hot-electron degradation thoroughly in terms of threshold voltage shift (ΔVTH), transconductance degradation (ΔGm) and so on. The TEOS buffer layer could effectively block the diffusion of hydrogen species from the Si3N4 capping layer into the channel and the Si/SiO2 interface during the Si3N4 deposition as well as subsequent thermal cycles.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

With CMOS scaling into the nanometer regime, strained channel engineering has been adopted as one of the most effective remedies and thoroughly explored for the much needed performance enhancement such as boosting the drive current of the scaled devices [1–3]. With regard to carrier mobility enhancement, introducing strain in the channel region can enhance the carrier mobility [4]. What the researchers could accomplish is by either applying highly bi-axial tensile strain to the channel region with a SiGe virtual substrate [1], or by uni-axially straining the channel with strain boosters [1–3]. Recent studies have shown that the uni-axial strained channel from a contact etch-stop silicon nitride layer increases the current drivability [5]. Such scheme is attractive and practical because it can be easily implemented using VLSI processing. The local strained channel (LSC) technique is proposed to provide tensile strained channel in nMOSFETs [6].

As long as the knowledge base concerning the mobility enhancement of Si strained channel has been established, it is appropriate now for us to concentrate our attention to its relating issue such as reliability. Among the most critical reliability issues, device degradation induced by hot-electrons is the most representative one in deep sub-micro nMOSFETs [7]. The physical mechanisms and characteristics of hot-electrons degradation have been extensively examined [8]. Furthermore, by exerting the accelerated stress test, we could study the hot-electron degradation thoroughly in terms of threshold voltage shift (ΔVTH), drain current degradation (ΔID), subthreshold swing degradation (ΔS.S), and transconductance degradation (ΔGm) and so on. According to the aforementioned contents, local tensile strain technology with the Si3N4 capping layer has been viewed as one of the most effective methods to boost the drive current in scaled nMOSFET devices, the device reliability associated with the strained device owing to the strain, and excess hydrogen and nitrogen incorporation from the deposited Si3N4 layer is an imminent concern. In line with this, the incorporation of a thin TEOS buffer layer to improve the reliability performance has been proposed [9].

In addition, hydrogen-annealed wafers (Hi-wafer) have been reported having reduced oxygen defects in Czochralski (CZ) wafers [10–11], with improved micro roughness and less defect on the surface after high temperature hydrogen anneal. The following sections describe in detail the behaviors of nMOSFETs device with local strained channel technique using SiN-capping layer on Hi- or Cz-wafers as well as the related reliability.
2. Device concept

2.1. Strained-Si technology

Strained-Si technology improves the drive current of CMOS through altering the energy band structures of the surface channel [12–14]. There is a branch of strained-Si technology: bi-axial strain and uni-axial strain. The bi-axial tensile strained channel achieved by growing a Si channel layer on a relaxed SiGe substrate could improve the drive currents of both nMOSFETs and pMOSFETs when the incorporated Ge content is more than 20% of the relaxed SiGe substrate [15–16]. It is worthy to note that the thickness of the top strained-Si layer has to be thinner than a critical thickness that depends on the Ge content of the underlying relaxed SiGe substrate to avoid the generation of abundant amount of dislocations due to lattice mismatch [17]. A plenty dislocations of the virtual SiGe substrate may become an obstacle to practical applications.

In contrast, the uni-axial strained channel technology has been proposed to avoid the shortcomings of bi-axial strained channel. Uni-axial strain can be set up by modifying contact-etch-stop-layer (CESL) deposition [18–19], silicidation [20], source/drain material [21], shallow trench isolation (STI) [22], packing process [23], and so on. Moreover, uni-axial strain can be arbitrarily exerted in any direction correlative with the carrier transportation; hence the performance of both NMOS and PMOS devices can be enhanced by respectively applying the tensile and compressive strains. Mobility, \( \mu \), for carriers in semiconductors is formulated as

\[
\mu = \frac{q\tau}{m^*}
\]

where \( 1/\tau \) is the overall scattering rate and \( m^* \) is the conductivity effective mass. From the above mobility formula, we could set about mobility enhancement of strained-Si devices by reducing either the conductivity effective mass or the scattering rate. The conduction band of unstrained bulk Si is consisted of six degenerate valleys (\( D_6 \)) with the same energy [24] (as shown in Fig. 1). Under the effect of bi-axial tensile strain, the sixfold degenerate conduction band is split up into a fourfold (\( D_4 \)) in-plane and a twofold (\( D_2 \)) out-of-plane degenerate valleys in the energy band diagram [25]. The energy difference (\( \Delta E \)) between \( D_2 \) and \( D_4 \) sub-bands determines the total population of electrons in each sub-band. As a result, the larger \( \Delta E \), the more the percentage of total electron population would occupy the \( D_2 \) valleys. Since the \( D_2 \) valleys have a smaller effective mass as compared with that of the \( D_4 \) valleys, the electron mobility enhancement could be accomplished as more electrons occupy the \( D_2 \) valleys. In addition, suppression of inter-valley phonon scattering can effectively reduce the electron scattering rate (\( 1/\tau \)) [25–26], which in turn may also enhance the mobility. From the effort of Leitz et al. [27], we know that uni-axially and compressively strained pMOSFETs may have lighter in-plane effective mass through full-band Monte Carlo simulation [24], thus hole mobility is enhanced. However, for the case of bi-axial tensile strain, the reduction of the inter-valley scattering is the only plausible approach to the hole mobility enhancement by about 25–30% of germanium concentration necessary for introducing more than 1 G Pa stress [27].

2.2. Hi-wafer technology

A raw Czochralski-grown (CZ) silicon wafer includes supersaturated oxygen atoms and nuclei for oxygen precipitation, which are introduced during crystal growth. For ultra large scale integrated (ULSI) devices, CZ silicon wafers ought to be free of defects in the device active layer and adequate oxygen precipitates in the bulk region to enable intrinsic gettering for metallic contamination. However, inadequate oxygen contents would induce many micro-defects during later heat treatments for ULSI fabrication. The micro-defects, which are induced near surface region, lead to various harmful defects such as OISF (Oxidation induced Stacking Fault), pattern edge dislocations, gate oxide breakdown failures and so on. For the purpose of improving the surface quality and preventing the harmful defect generation, both oxygen and nuclei ought to be completely removed from the surface region. There is an approach to produce oxygen-less Si wafer by virtue of a high temperature anneal in hydrogen ambient so as to efficiently eject the oxygen atoms from the surface region [10–11]. The hydrogen anneal was carried out at 1200 °C for 1 h by using a hot wall type vertical furnace. Wafers which received such treatment are dubbed “Hi-wafer”. Fig. 2 shows the oxygen out diffusion profile after annealing in hydrogen or oxygen ambient. The oxygen contents

![Fig. 1. Simple schematic of conduction and valence band bending with strain [24].](image)
in the surface region of Hi-wafer are obviously smaller than that of a wafer annealed in oxygen ambient. Thus the generation of oxygen-induced defects near the surface region could be significantly reduced in Hi-wafers, making such wafers a good choice for improving the wafer yield.

3. Experimental results

3.1. Device fabrication

All device fabrications were carried out at National Nano Device Laboratories. nMOSFETs were fabricated on 6-in Hi- and Cz-wafers with resistivity of 15–25 $\Omega$ cm. The hydrogen anneal was carried out at 1200 °C for 1 h. The p-type well was formed first by BF$_2^+$ implantation at 70 keV and $1.2 \times 10^{13}$ cm$^{-2}$. Next, a standard local oxidation of silicon (LOCOS) process with channel stop implant (by BF$_2^+$ implantation at 120 keV and $4 \times 10^{13}$ cm$^{-2}$) was used for device isolation. Threshold voltage adjustment and anti-punch through implantation steps were done by implanting 40 keV BF$_2^+$ and 35 keV B$^+$, respectively. Gate dielectric thickness about 2.5 nm was grown in an N$_2$O ambient. Then, 200 nm undoped poly-Si was deposited. Shallow S/D extensions were formed by implanting As (10 keV, $1 \times 10^{15}$ cm$^{-2}$). After a 200 nm TEOS spacer formation, S/D regions were formed by As$^+$ implantation at 30 keV and $5 \times 10^{15}$ cm$^{-2}$. Then the substrate electrode patterning was performed through lithography and etching processes, followed by the formation of the substrate contact regions by BF$_2^+$ implantation at 40 keV and $5 \times 10^{15}$ cm$^{-2}$. Rapid thermal anneal (RTA) was subsequently carried out in a nitrogen ambient at 1000 °C for 10 s to activate dopants in the gate, S/D, and substrate regions. A 300 nm LPCVD Si$_3$N$_4$ was next deposited (denoted as SiN/Cz and SiN/Hi), while some wafers were deliberately skipped of the Si$_3$N$_4$ deposition step to serve as the controls (denoted as Cz-control and Hi). In addition, the effect of intrinsic stress inside the Si$_3$N$_4$ layer with 300 nm thickness was remarkable [31]. For some Si$_3$N$_4$-capped nMOSFETs on either Hi- or Cz-wafers, a thin LPCVD-TEOS buffer layer (about 7 nm, denoted as SiN/buffer/Cz and SiN/buffer/Hi) was capped prior to the Si$_3$N$_4$ deposition, as shown in Fig. 3. Finally, a four-level metallization was carried out in PVD system for contact.

3.2. Electrical characterization and related reliability

All measurements in the following section were evaluated by an HP4156A precision semiconductor parameter analyzer and an HP4284 LCR meter, respectively. Temperature-regulated hot chucks were used to set the measurement temperature at 25 °C. Fig. 4 shows the capacitance–voltage (C–V) characteristics of devices. In Fig. 4, we could distinctly observe that the poly-depletion effect becomes obvious in the splits with Si$_3$N$_4$ capping layer depositions, irrespective of the use of Cz or Hi-wafers. This phenomenon is attributed to the additional thermal budget associated with the nitride deposition step by 780 °C LPCVD. It is known that the solid solubility of dopants is temperature-dependent and the thermal conditions mentioned above tend to lower the activated carrier concentration in the poly-Si gates [28]. In addition, the poly-silicon depletion can be avoided when PECVD is used for the nitride deposition owing to the short deposition time of PECVD for the activated carrier concentration in the poly-Si gates being less influenced. Fig. 5 shows the drain current ($I_D$) and transconductance (Gm) versus $V_G$ for nMOSFETs with $W/L = 10 \mu$m/0.5 $\mu$m from different splits. Higher subthreshold slope (S.S.) and leakage current, about 2 orders larger, are found for the splits with Si$_3$N$_4$ capping layer due to contact etched damages. A significant increase of Gm was found for the splits with Si$_3$N$_4$ capping layer owing to the tensile stress in the channel, which increases the electron mobility in the channel.

Fig. 6 shows the distribution of the threshold voltage ($V_{TH}$), $V_{TH}$(L)–$V_{TH}$(10 $\mu$m), versus the gate length. Among the splits, the control splits depict a pronounced reverse-short-channel-effect (RSCE). This is probably due to boron segregation at the implant-damaged regions located near the edge of the channel. The phenomenon, however, is not observed in the Si$_3$N$_4$ capping splits. It

![Fig. 3. TEM pictures for TEOS buffer layer. The thickness is about 7 nm.](image-url)
is noted that RSCE would be suppressed by boron redistribution as the thermal budget increases and bandgap narrowing effect caused by channel stress [29].

Fig. 7 shows the percentage increase of transconductance (Gm) for all splits as a function of channel length. The Gm is linearly related to channel mobility. When the channel length is scaled to less than 1 μm, Gm increases sharply owing to the unique feature mentioned above. Specifically, Gm increases by 20% for nMOSFETs with W/L = 10 μm/0.5 μm. In addition, the channel strain was scarcely influenced as the thickness of the TEOS buffer layer was lower than 20 nm [9]. Fig. 8 shows the percentage increase of the on-current with respect to the Cz-control (V_D = 1.8 V, V_G = 1.8 V) for all splits as a function of channel length. The trend of on-current enhancement is the same as the Gm enhancement.

Fig. 9 shows the distribution of the S.S. versus the gate length. For the splits of Cz-controls and Hi, S.S. is independent of gate length. Hi split depicts lower S.S. than Cz-controls because of the better surface quality after high temperature hydrogen anneal. However, the splits with Si3N4 capping layer depict higher S.S., especially for long channel devices. The root cause for such phenomenon is not clear yet at this stage. Increase in EOT due to poly-depletion effect may play a role. It is noted that the TEOS buffer layer could effectively block hydrogen molecules diffusion from Si3N4 capping layer deposition. As a consequence, the S.S. increases due to less effective interface passivation.

Fig. 10 shows charge pumping current (I_{cp}) versus base voltage. The device size is W/L = 10 μm/0.5 μm for (a), W/L = 10 μm/10 μm for (b), the pulse amplitude is 1.5 V, and the square waveform is used for I_{cp} extraction. The splits of SiN/Cz and SiN/Hi depict slightly higher I_{cp} compared with the splits of Hi and Cz-control.
However, $I_{cp}$ increases drastically for the splits of SiN/buffer/Cz. These results indicate that the TEOS buffer layer can effectively block hydrogen diffusion into the channel region during the Si$_3$N$_4$ deposition process. As a result, the hydrogen incorporation in the gate oxide as well as at the Si/SiO$_2$ interface can be suppressed remarkably with the incorporation of TEOS buffer. Therefore, the interface states from the Si$_3$N$_4$ capping layer are not repaired due to the mechanism of TEOS blocking hydrogen diffusion. Therefore, the percentage increase of $G_m$ for the SiN/buffer/Cz split is lower than that for the SiN/Cz split owing to the large amount of un-repaired interface traps, compared with the Cz-controls. In addition, it is noted that $I_{cp}$ of SiN/buffer/Hi split only

**Fig. 9.** Subthreshold slope versus gate length.

**Fig. 10.** Charge pumping current ($I_{cp}$) versus base voltage. The device size is W/L = 10 μm/0.5 μm for (a), W/L = 10 μm/10 μm for (b), the pulse amplitude is 1.5 V, and the square waveform is used for $I_{cp}$ extraction.

**Fig. 11.** Comparisons of effective mobility for devices with W/L = 50 μm/50 μm. (a) Cz group; and (b) Hi group.

**Fig. 12.** On-current ($V_G = 1.8$ V, $V_D = 1.8$ V) versus off-current ($V_G = 0$ V, $V_D = 1.8$ V) with W/L = 10 μm/0.4 μm.
slightly increases over the SiN/Hi split. Therefore, the effect of hydrogen diffusion or blocking is not a serious concern for the SiN/buffer/Hi split. Besides, the larger $I_{cp}$ of long length with Si$_3$N$_4$ capping layer than that of short length was a direct evidence to explain the S.S. increases for long channel devices due to less hydrogen species diffusion at long channel devices.

Fig. 11 compares the effective mobility for devices with $W/L = 50 \mu m/50 \mu m$. For Cz group, as shown in Fig. 7a, degraded behaviors are observed for the splits with Si$_3$N$_4$ capping. However, for Hi group, as shown in Fig. 7b, it is very interesting to note that the buffer layer prevents the mobility degradation from the Si$_3$N$_4$ capping.

Fig. 12 shows the distribution of the on-current ($V_G = 1.8 \, V$, $V_D = 1.8 \, V$) versus off-current ($V_G = 0 \, V$, $V_D = 1.8 \, V$) with $W/L = 10 \, \mu m/0.4 \, \mu m$. Lower off-current is found for the splits on the Hi-wafers, owing to its better surface quality.

Figs. 13 and 15 show the threshold voltage shift ($\Delta V_{TH}$) and the interface state density increase as a function of stress time for all splits after hot-electron (HC) stressing at $V_D = 3.75 \, V$ and $V_G$ at the peak substrate current, and the device size is $W/L = 10 \, \mu m/0.5 \, \mu m$. The SiN-capping splits for both Cz and Hi-wafers exhibit the worst HC degeneration, while the improvement of the HC degradation by using the TEOS buffer layers is obviously seen. However, the threshold voltage shift of the SiN/Hi split is worse than that of the SiN/Cz split, similar to the case of the buffer-layer splits,
owing to the fact that the fresh maximum Gm values of both the Si/H split and the SiN/buffer/Hi split are larger than those of the SiN/Cz and SiN/buffer/Cz splits.

It is important to note that, although the Gm of the Si/H split is high for the SiN/buffer samples, significant improvement in HC resistance than that without the buffer is observed. Since the HC tends to break the Si–H bonds during the stressing, and much severe degradation will occur with higher Si–H density. The use of TEOS buffer layer can effectively block the diffusion of hydrogen species into the channel region, therefore less HC degradation in terms of Vth shift and interface-state generation is achieved.

Fig. 16 shows the transconductance degradation as a function of stress time for all splits after receiving hot-electron (HC) stressing at VDS = 3.75 V and VGS at the peak substrate current. The device size is W/L = 10 μm/0.5 μm. The SiN/Hi split shows the worst Gm degradation among all splits owing to the strain enhancement effects and extra hydrogen incorporation mentioned above.

Fig. 17 shows the increase in charge pumping current after 5000 s of hot-carrier stress for devices with W/L = 10 μm/0.5 μm. The effect of TEOS buffer layer is observed in terms of lower ∆Icp. After stressing, the SiN/buffer/Hi split possesses higher ∆Icp than the SiN/buffer/Cz split, although in principle the former split should have a better interface quality. Higher Isub and higher population of generated HC are postulated as the plausible origin for this observation.

4. Conclusions

We have performed a detailed investigation of the LSC channel technique induced by SiN-capping layer with or without buffer layer. Device characteristics including Gm, S.S., Icp, and reliability were studied on devices fabricated with Hi and Cz starting wafers. The Gm and current drivability can indeed be enhanced by capping the Si3N4 layer. We found that the splits with Hi-wafers depict lower S.S. than the splits with Cz-wafers. However, with the incorporation of TEOS buffer layer, Icp increases and Gm is degraded for SiN/buffer/Cz splits, contrary to the results of the SiN/buffer/Hi splits. The hot-electron degradation is adversely affected when the Si3N4 capping layer is deposited over the gate, regardless of the types of wafers. When a TEOS buffer layer is inserted prior to the Si3N4 deposition, although still worse than the control ones, significant improvement in hot-carrier resistance over those without the buffer layer is achieved.

References