Strained Silicon Layer in CMOS Technology

Tatjana Pešić-Brđanin, Branko L. Dokić

Abstract—Semiconductor industry is currently facing with the fact that conventional submicron CMOS technology is approaching the end of their capabilities, at least when it comes to scaling the dimensions of the components. Therefore, much attention is paid to device technology that use new technological structures and new channel materials. Modern technological processes, which mainly include ultra high vacuum chemical vapor deposition, molecular beam epitaxy and metal-organic molecular vapor deposition, enable the obtaining of ultrathin, crystallographically almost perfect, strained layers of high purity. In this review paper we analyze the role that such layers have in modern CMOS technologies. It’s given an overview of the characteristics of both strain techniques, global and local, with special emphasis on performance of NMOS biaxial strain and PMOS uniaxial strain. Due to the improved transport properties of strained layers, especially high mobility of charge carriers, the emphasis is on mechanisms to increase the charge mobility of strained silicon and germanium, in light of recent developments in CMOS technology.

Index Terms—Strained silicon layers, uniaxial strain, biaxial strain, CMOS technology.

Strained Silicon Layer in CMOS Technology

Tatjana Pešić-Brđanin, Branko L. Dokić

However, the other demands of designing integrated circuits with ultra large scale of integration, such as, for example, the increase in density on the chip and reducing the size of the chip, have led to the scaling of other transistor dimensions. Fig. 1 shows the trend of reducing the gate length that has led to the emergence of new technologies [1]. Reducing the gate length caused a simultaneous scaling of other technological parameters, in order to meet the required performances of integrated circuits regarding high speed and low power consumption, or the desired degree of integration. However, with recent CMOS technologies, in which the gate length is less than 90 nm, this performance improvement becomes more difficult due to physical limitations in miniaturization of MOS transistors.

Fig. 1. Technology and gate length evolution [1].

With the scaling dimensions, at the same time, new process technologies were developed, especially ultra high vacuum chemical vapor deposition (UHV-CVD), and molecular beam epitaxy (MBE), which enabled the obtaining of ultrathin, crystallographically almost perfect, strained layers of high purity. Due to the low temperature processes (<600 °C), diffusion of impurities is very weak. This allowed obtaining of super abrupt pn junctions with much greater difference in concentration of ingredients on both sides of the joint than in the case of other epitaxial growth techniques. Using the metal-organic CVD (MO-CVD) growth techniques, the

I. INTRODUCTION

In the last 30 years, the primary characteristic of improving performance of CMOS technology has been scaling the dimensions of the components. Continuous reduction in size of devices and simultaneously increase of the size of chips has led to the production of highly complex integrated circuits with ultra large scale of integration. The realization of such complex high-speed integrated circuits is enabled by using CMOS technology with critical dimensions of components of a few nanometers. Numerous factors influence the increase of speed of MOS transistor operation; the most important is the time of the carriers transport through the channel under the gate, which can be reduced by shortening the gate length.
homogeneous and heterogeneous layers are created with controlled concentration and thickness up to 50 Å, while using the MBE technology obtained layers were even less than 10 Å of thickness [2]. With these new technological processes, and great progress in lithographic techniques, new structures of MOS transistors with ultrathin layers of germanium or SiGe alloys on silicon substrates were reached. Compared to conventional, these new structures are characterized, above all, with a significantly higher speed operation, as well as improved current-voltage performances. The operations of these transistors, however, are characterized by a number of new physical phenomena (quantum effects, transport across the heterojunction, ballistic transport, tunneling etc.), which must be taken into account in optimization of such structures and specifications of their electrical characteristics [3,4].

One of the new technological structures of MOS transistors, the structures with strained silicon in channel region, thanks to advances in the local strain technique, are used in CMOS technology (e.g., logic circuits in 90 nm CMOS technology) [5]. The increase of the carrier mobility, which is implemented by appropriate silicon strain, provides higher speed of the carriers, respectively higher current drive, under the same conditions of polarization and a fixed oxide thickness. On the other hand, it means that with the same current conditions in the channel, thicker oxides and/or lower voltages supply could be used, which leads to relaxation of compromise between the current, consumption and the effects of short channels. Therefore, the strain technique and a corresponding increase in carrier mobility are considered as factors required for future CMOS technologies [6,7].

This paper provides an overview of the principles and application techniques of strained layers, with special emphasis on the physical mechanisms to increase mobility due to strain, in modern and future CMOS technologies.

II. PHYSICAL MECHANISMS FOR MOBILITY ENHANCEMENT

One way to improve the performance of MOS transistor is the possibility to change the properties of materials. The property of silicon to exhibit modified characteristics of transport parameters while strained [8,9] is used for this purpose. Measurements which were performed on Hall structures in strained silicon layer at room temperature showed a large electron mobility, which in the case of a very low temperatures (0.4 K) reaches extremely high values, up to 500 000 cm²/Vs [10]. The increase in electron mobility in strained silicon layers is supported by theoretical studies [11,12].

A. Electron Mobility Enhancement

The silicon layer strain is technologically commonly performed by the growth of silicon on the SiGe substrate. Fig. 2 schematically shows the crystal structure of the silicon and the relaxed layer of Si₁ₓGe₁₋ₓ. Since the lattice constant of Si₁ₓGe₁₋ₓ composition is higher than the silicon lattice constant in the balance state, the pseudomorph layer of silicon which has grown on the relaxed Si₁ₓGe₁₋ₓ layer (or virtual substrate) is under biaxial strain. The distance of silicon atoms assimilate to the higher atomic distance in the SiGe layer, thereby increasing the lattice constant of silicon a₆ in the growing level. The lattice constant of the SiGe composition, and hence the strained silicon, is approximately a linear function of the germanium content (or the value of the x in the Si₁ₓGe₁₋ₓ composition). As the germanium lattice constant is about 4% higher than the lattice constant of silicon, it will, for example, Si₀.₇₅Ge₀.₂₅ virtual substrate cause an increase in lattice constants of strained silicon which has grown on the composite layer for approximately 1% [9].

The strain leads to the shape degeneration of six equipotential valleys in conduction zone by reducing the value of the energy minimum in the two transverses (Δ₂), compared to the values of the energy minimum in four longitudinal directions (Δ₄) [13] (Fig. 2). The result is an effective anisotropy of the electron effective mass; transversal mass (m₁ = 0.19m₀) appears, and it is smaller than the longitudinal mass (m₇ = 0.916m₀). m₀ is the electron mass. Table 1 shows the values of the electron effective mass depending on the crystallographic planes and directions for silicon and germanium [14].

The difference of the effective mass causes different physical properties in Δ₂ and Δ₄ valleys. The conductive electron mass in parallel MOS interface is smaller in Δ₂ the plane than in the Δ₄ plane, and, therefore, the electron mobility is greater in the Δ₂ plane, than in the Δ₄ plane. Also, as the inverted layer thickness and the subzone energy are determined by effective mass in the direction normal to MOS interface, an inversion layer is thinner and the subbands energy is lower in Δ₂ plane, and so the effective mass is higher.
The effect of strain on the electron mobility can be reflected by changing the subbands energy, i.e. the energy levels in the conduction zone. The increase of the energy of the sublevels results in increased mobility through two mechanisms: increasing the average value of mobility in $\Delta_2$ plane due to a large number of electrons with higher mobility and modification of the bottom of the conduction band leads to a reduction of dissemination of holders on the phonons, which, in total, increases the electron mobility even when the electric field has low values [13].

Unlike the conduction band, the structure of the valence band is similar for silicon and germanium. The energy minimum and maximum are located in the same crystallographic directions. Band structure of germanium is less anisotropic when compared to silicon, so its effective mass is lower. Germanium structure reacts to strain in a similar way as silicon structure, with improved response to the strain inversion.

There are several empirical phrases to assess the width modification of the band gap zone of strained silicon. For strained silicon, a narrowing of the band gap zone depending on the percentage of germanium in the relaxed Si$_{1-x}$Ge$_x$ layer can be calculated:

$$E_{g\text{st}} = E_g - 0.4 \cdot x \text{ (eV)}$$  \hspace{1cm} (1)

where $E_g$ is the silicon bang gap energy, and $E_{g\text{st}}$ is the strained silicon bang gap energy.

Of particular importance is the electron mobility model which has to include the functional mobility dependence for strain, temperature, ingredients concentration in the channel and polarization. Fig. 3.a shows the experimental results for mobility enhancement factor, which is defined as the mobility ratio of strained and unstrained silicon. The findings for the two models of mobility, derived mainly from the theory of phonon diffusion, are also shown [3,15,16].

Figure 3.b shows the dependence of effective mobility on the effective electric field at a room temperature, with and without strain. It shows that the mobility enhancement factor is almost unchanged in a wide range of values for the external electric field.

**TABLE I**

<table>
<thead>
<tr>
<th>Plane</th>
<th>Channel direction</th>
<th>Effective mass ($m_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$m_x$</td>
</tr>
<tr>
<td>(100)</td>
<td>$\langle 001 \rangle$</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>$\langle 011 \rangle$</td>
<td></td>
</tr>
<tr>
<td>Silicon</td>
<td>$\langle 001 \rangle$</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>$\langle 110 \rangle$</td>
<td></td>
</tr>
<tr>
<td>(110)</td>
<td>$\langle 001 \rangle$</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>$\langle 011 \rangle$</td>
<td>0.085</td>
</tr>
<tr>
<td>Germanium</td>
<td>$\langle 001 \rangle$</td>
<td>0.601</td>
</tr>
<tr>
<td></td>
<td>$\langle 110 \rangle$</td>
<td>0.082</td>
</tr>
</tbody>
</table>

**B. Hole Mobility Enhancement**

Unlike the electron mobility enhancement, the physical mechanisms of change of the hole mobility in PMOS transistors due to strain have not been fully explained. Also, it has been shown that the effects of the uniaxial and biaxial strain on the hole mobility are completely different. With the biaxial strain, the hole mobility enhancement due to strain is a consequence of the decrease of the hole effective mass in the occupied zones of the valence zone, suppression of inter-subband scattering and increase of the availability of subbands with a smaller effective mass, i.e. higher mobility.

Figure 4.a shows the experimental results for hole mobility enhancement factors for biaxial strained PMOS transistor, and the results obtained by theoretical models. It can be noted that a large percentage of germanium significantly increase the hole mobility. The dependence of hole mobility enhancement factors on the effective electric field indicates to the fact that the maximum hole mobility enhancement factor is at a lower
values $E_{\text{eff}}$. Figure 4.b. As mobility at large values $E_{\text{eff}}$ is significant for most practical applications, it is necessary to use a biaxial strain silicon layers with a high content of germanium. Recent studies have shown that, when uniaxial compression along $\langle 110 \rangle$ PMOS channel direction is applied, the hole mobility enhancement factors is higher, even for less strain, and it does not decrease significantly with increasing values $E_{\text{eff}}$ [17-20].

Fig. 4. Mobility enhancement factor as a function of Ge content in SiGe substrates for PMOS transistor (a) and effective field dependence of hole mobility in biaxial strained Si NMOS (b).

By measuring the piezoresistivity at $\langle 100 \rangle$ surface of PMOS channel during uniaxial compression parallel to the channel direction and strain by stretching parallel to the channel width, the hole mobility enhancement has been shown. Fig. 5 shows these complex dependencies of hole mobility on strain for PMOS transistor. It can be seen that, for less strain, compression at $\langle 110 \rangle$ surface increases the hole mobility. Also, for higher strain, the hole mobility is increased by biaxial strain, as by compression, and stretching. It can be concluded that the uniaxial compression at $\langle 110 \rangle$ surface and biaxial stretching are the most effective in terms of increasing the hole mobility for PMOS transistors.

Fig. 5. Hole mobility enhancement factor for PMOS transistor with uniaxial compressive and biaxial tensile strain [23]. The value of $E_{\text{eff}}$ is taken to be $1 \text{MVcm}^{-1}$.

III. MOS STRUCTURES WITH STRAINED LAYERS

The strain is widely accepted as a promising technique which improves CMOS performances through a significant increase of mobility, even from 90 nm CMOS technology. In recent years, substrate and/or process-induced strained channels are successfully integrated in MOS transistors, for the purpose of increasing the channel mobility. Technologically, so far, the biaxially strained layer showed the best results in the case of long channels transistor (over 100% improvement in mobility). However, the contact resistance of the source and drain, saturation velocity, self-heating of relaxation for strained layer complicate the improvement of performances in nano devices.

The advantage of the uniaxial strain access is that it is installed during production of CMOS circuits. However, the scalability and the geometrical dependence are the main problems. The lateral dimensions of the transistors affect the efficiency and strength of strain in the strained layer. It is shown that the process-induced strain layers increase the mobility of both types of carriers and the power drain in NMOS and PMOS transistors. The strain in the channel can be obtained through the individual process steps and the use of strained silicon substrate.

A. Global Strain Technology

The structure of MOS transistor with biaxially strained silicon layer, which has grown on a relaxed SiGe substrate, has been intensively explored in recent years. Also, there have been research on the structures with strained silicon layer on insulator (Strain-Si On Insulator - SSOI) [24], where strained/relaxed layers are formed on the buried oxide, and the structures where the strained silicon layer is directly linked to the buried oxide (Strain-Si Directly On Insulator - SSDOI) [25]. Common structures and substrates using biaxial strained layers are shown in Fig. 6. The technology for production of
MOS transistor on substrates on which strained layers are formed is called a "Global strain technology".

Fig. 6. Cross-sections of typical MOS structures using biaxial strain [3].

As for MOS transistors using silicon substrates from the global strain technology, the research and development of optimized components are implemented on 45 nm and 32 nm CMOS technology. Many research groups have achieved an improvement of current activation in the range of 10-25% with a global strained silicon layer, and gate lengths smaller than 100 nm [25,26]. Also, there has been success in the design of CMOS circuits with gate lengths of 25 nm and the integration of strained silicon layers with gate oxides with high dielectric constant and a metal gate [27].

Main advantages of global strain technology are in obtaining uniformly strained layers with more strength and possibility of implementing standard CMOS process steps with minimal modifications. On the other hand, there are certain difficulties relating to the limited improvement of PMOS transistor performances with small and medium silicon strain, the occurrence of defects and dislocations at the boundary surfaces, increased production costs and an increase of current leakage at compounds. In transistors with ultra short gate lengths, special attention must be paid to the reduction of the parasitic resistance of the source and drain [3,25].

B. Local Strain Technology

In order to eliminate the limitations of the global strain technology, special attention is paid to the local introduction of structures and materials that will cause strain in the channel of MOS transistors. This process can be conducted in two ways:

- The source and drain of the transistor are formed by epitaxy growth of SiGe composition. These buried SiGe layers, which introduce uniaxial compression in the channel area, are applied in the case of PMOS transistors. Recently it has been shown that NMOS transistor with a layer that is strained by stretching, in which the source and drain formed by selective growth of SiC instead of SiGe, can be implemented (Fig. 7) [28].

- A thin layer of SiN can be applied on the MOS structure, which introduces strain in the channel. In many cases, the SiN layers with stretching are applied for NMOS transistor, while the SiN layers with compression are applied for PMOS transistor (Fig. 7) [29].

Fig. 7. Illustration of a device structure using different local strain techniques. STI means Shallow Trench Isolation. Black and white arrows indicate compressive and tensile strain, respectively [3].

Beside the above mentioned methods, the strain can be also used for the other parts of the structure, such as, for example, the field insulation, the gate electrode, ... Since the mentioned methods are used in the production of logical LSI circuits for technologies less than 90 nm, the local strain technology has a growing practical importance in CMOS technology [5,30].

The basic advantages of the local strain technology are the possibility of using standard CMOS process with slight modifications and low cost, and acquiring the uniaxial strained PMOS transistors with excellent performance with relatively low strain. However, as in any strain technology, what presents a problem is the process of relaxation of strained layers, which occurs due to the advanced processes in a technological sequence or the component geometry, and a great attention is
It is particularly interesting to consider the electrical characteristics due to the combined strain, the uniaxial (process-induced) and biaxial (substrate-induced). The combined strain effects for SS NMOS with variable gate lengths (minimum 45 nm) are given in [31], and the performance improvement for more than 62% compared to the conventional NMOS has been shown. Figs 9 and 10 show the combined strain effects on the electrical characteristics of SS MOS.

V. CONCLUSION

Today, modern CMOS technologies cannot be imagined without the strained layers. These layers will, in the future, for the new CMOS technology nodes, have a much more important role, given the limitations of improvements in the standard CMOS technology and the requirements to maintain high current drive. This role imposes the need to develop methods that strain layers can make more compatible with other innovations that are introduced into the standard CMOS process, regardless of whether it is a new material (e.g. for gate of dielectrics), or new structures (i.e. multi-gate structures). Therefore, future research can move in two directions: the optimal design of strained structures and their implementation, on the one hand, and the robustness of the processes and structures to changing performance, with a high level of reliability, on the other hand.

REFERENCES


