Computational Microelectronics
S. Selberherr *(Editor-in-Chief)*

Viktor Sverdlov

Strain-Induced Effects in Advanced MOSFETs

SpringerWienNewYork

SpringerWienNewYork

Computational Microelectronics

Edited by

Siegfried Selberherr
Technical University Vienna
Vienna, Austria

For further volumes:
www.springer.com/series/1263

Viktor Sverdlov

Strain-Induced Effects in Advanced MOSFETs

SpringerWienNewYork

Viktor Sverdlov
Technical University Vienna
Institute for Microelectronics
Gusshausstrasse 27-29
1040 Vienna
Austria
sverdlov@iue.tuwien.ac.at

Product Liability: The publisher can give no guarantee for all the information contained in this book. The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Printed in Germany

SpringerWienNewYork is a part of Springer Science+Business Media
springer.at

Cover: WMXDesign GmbH, Heidelberg, Germany

Typesetting: SPi, Chennai, India

Printed on acid-free and chlorine-free bleached paper
SPIN: 80018573

With 101 Figures

Library of Congress Control Number: 2010938373

ISBN 978-3-7091-0381-4 e-ISBN 978-3-7091-0382-1
DOI 10.1007/978-3-7091-0382-1
SpringerWienNewYork

To Alexandra, Karin, Ludmila, & Nikolai

Preface

Strain is the main tool to boost current and enhance performance of advanced silicon-based metal-oxide-semiconductor field-effect transistors (MOSFETs). Modeling and understanding of strain effects on band structure and mobility has become the important task of modern simulation tools used to design ultra-scaled MOSFETs. This book focuses on modern modeling approaches and methods describing strain in silicon. Contrary to the valence band, strain-induced conduction band modifications have received substantially less attention. Peculiarities of subband structures in thin semiconductor films under stress are investigated in detail using numerical pseudopotential calculations as well as a **k·p** theory, which includes the two lowest conduction bands. Implementation of strain in transport modeling for modern microelectronics design tools is overviewed. Application ranges from device modeling to applied mathematics and software development.

The book is based on my research and partly on my course of lectures given for the Master's and PhD students in electrical engineering, microelectronics, physics, and applied mathematics at the Institute for Microelectronics, Technische Universität Wien. This book would not have been written without the support of the Institute for Microelectronics and its Director Univ.Prof. Dipl.-Ing. Dr.techn. E. Langer. I would like to thank Univ.Prof. Dipl.-Ing. Dr.techn. T. Grasser, Univ.Prof. Dipl.-Ing. Dr.techn. H. Kosina, and Univ.Prof. Dipl.-Ing. Dr.techn. Dr.h.c. S. Selberherr for their overwhelming encouragement, support, and help in writing the book.

I would like to acknowledge the contributions to the book made by my colleagues and co-authors: O. Baumgartner, J. Cervenka, S. Dhar, T. Grasser, A. Gehring, G. Karlowatz, M. Karner, H. Kosina, K. Likharev, M. Nedjalkov, M. Pourfath, F. Schanovsky, S. Selberherr, Z. Stanojevic, M. Vasicek, E. Ungersboeck, and T. Windbacher. I also would like to thank H. Ceric, R. Entner, O. Ertl, W. Goes, P. Hehenberger, R. Heinzl, S. Holzer, A. Makarov, G. Milovanovich, N. Neophytou, R. Orio, V. Palankovski, K. Rupp, P. Schwaha, I. Starkov, F. Stimpfl, O. Triebl, S. Tyaginov, S. Vitanov, P. Wagner, S. Wagner, J. Weinbub, and W. Wessner for many fruitful and stimulating discussions and C. Haslinger, E. Haslinger, M. Katterbauer, and R. Winkler for technical support in preparing the manuscript. Special thanks go to O. Baumgartner, M. Nedjalkov, and K. Sitzwohl, who kindly agreed to take the heavy duty of proof-reading and improving the manuscript.

The new scientific results described in the last sections of the book would have been impossible to obtain without financial support from the Austrian Science Fund FWF through the projects P-17285-N02 and P-19997-N14, from the European Research Council through the grant 247056 MOSILSPIN, from the European Commission, project SINANO IST-50684, and from the European Science Foundation EUROCORES Program FoNE funded by the Austrian Science Fund FWF (project I79-N16), CNR, EPSRC and the EC Sixth Framework Program.

Vienna *Viktor Sverdlov*
July 2010

Contents

List of Symbols

Notation

x	Scalar
$\mathbf{x}$	Vector
$\hat{x}$	Tensor
$\mathbf{A}$	Matrix
A_{ij}	Elements of the matrix **A**
$\mathbf{x} \cdot \mathbf{y}$	Scalar product
$[hk\ell]$	Miller indices to specify a crystal direction
$\langle hk\ell \rangle$	Miller indices to specify equivalent crystal directions
$(hk\ell)$	Miller indices to specify a crystal plane
$\{hk\ell\}$	Miller indices to specify equivalent crystal planes

Physical Quantities

Symbol	Unit	Description
$\mathscr{C}[f]$	s^{-1}	Collision operator
e_{ij}, γ_{ij}	1	Engineering strain component (i, j)
ε_{ij}	1	Component (*ij*) of the strain tensor
σ_{ij}	GPa	Component (*ij*) of the stress tensor
$\sigma_x, \sigma_y, \sigma_z$		Pauli matrices
C_{ijkl}	GPa	Component (*ijkl*) of the elastic stiffness tensor
c_{ij}	GPa	Component (*ij*) of the contracted stiffness tensor
S_{ijkl}	GPa^{-1}	Component (*ijkl*) of the elastic compliance tensor
s_{ij}	GPa^{-1}	Component (*ij*) of the contracted compliance tensor
D_n	$\mathrm{m}^2\mathrm{s}^{-1}$	Electron diffusion coefficient
D	eV	Shear deformation potential
E	eV	Energy
$E_n(\mathbf{k})$	eV	Energy dispersion
E_F	eV	Fermi energy
E_g	eV	Band gap energy
$\mathbf{E}$	Vm^{-1}	Electric field
$\mathbf{F}$	N	Force
$f(\mathbf{r}, \mathbf{k}, t)$	1	Distribution function
$f_W(\mathbf{r}, \mathbf{k}, t)$	1	Wigner distribution function

ϕ	V	Electrostatic potential
g	$\mathrm{m}^{-3}\mathrm{eV}^{-1}$	Density of states
k	m^{-1}	Wave number
$\mathbf{k}$	m^{-1}	Wave number vector
κ	$\mathrm{AsV}^{-1}\mathrm{m}^{-1}$	Dielectric permittivity
μ_n	$\mathrm{m}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$	Electron mobility
μ_p	$\mathrm{m}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$	Hole mobility
m	kg	Mass
n	m^{-3}	Electron concentration
N_D	m^{-3}	Concentration of donors
N_A	m^{-3}	Concentration of acceptors
Ψ	$\mathrm{m}^{-1/2}$	Wave function
$\mathbf{r}$	m	Space vector
a_0	m	Lattice constant
t	m	Film thickness
α	ev^{-1}	Non-parabolicity parameter
T	K	Temperature
$\mathbf{v}$	ms^{-1}	Velocity vector

Constants

h	Planck's constant	$6.6260755 \times 10^{-34}$ Js
$\hbar$	Reduced Planck's constant	$h/(2\pi)$
k_B	Boltzmann's constant	$1.380662 \times 10^{-23}\ \mathrm{JK}^{-1}$
e	Elementary charge	$1.6021892 \times 10^{-19}$ C
m_0	Electron rest mass	$9.1093897 \times 10^{-31}$ kg
κ_0	Dielectric constant of vacuum	$8.8541878 \times 10^{-12}\ \mathrm{AsV}^{-1}\mathrm{m}^{-1}$
i	$\sqrt{-1}$	

Chapter 1
Introduction

Introduced in mass production at the beginning of the 1970s, the Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET) is the key element of modern integrated circuits. Although the transistor feature size has shrunk dramatically over the past three decades, its overall design stayed nearly the same until recently. Even the 90 nm technology node MOSFETs introduced in 2004–2005 and still found in nowadays computers are based on the same principle and consist of the same basic elements as three decades ago. The inversion channel, which connects the source and drain electrodes, is formed at the silicon interface by applying a certain voltage to the gate electrode. The gate electrode made of heavily doped poly-silicon is electrically separated from the inversion channel by an oxide layer. A high quality silicon dioxide is resilient against an electrical break-through even at high electric fields and possesses little defects at the Si/SiO$_2$ interface. The good quality of this interface guarantees high mobility of the carriers in the inversion channel. Due to their perfect compatibility, the pair Si/SiO2 has quickly become the main stream microelectronic element of Si-based MOSFETs. Low defect density, high yield, and a relatively simple and inexpensive fabrication process have put MOSFETs into the heart of all modern high density integrated circuits.

Although the basic design of the transistor did not change, the operation speed and performance have increased dramatically. This became possible thanks to the scalability of the MOSFETs. Gordon Moore, one of the founders of Intel, has postulated the rule known as the Moore's law, according to which the MOSFET size reduces exponentially. A new generation of transistors with improved performance is introduced every two to three years which allows to double the number of transistors on integrated circuits every two years, decrease costs per transistor and increase performance for the same costs. With the 32 nm technology node presented at the International Electron Devices Meeting in December 2008 by Intel, the Moore's law did not loose its actuality and MOSFET scaling is successfully continuing.

Nevertheless, although the scaling is keeping pace with Moore's law, new technological solutions for MOSFET design had to be introduced beginning from the 90 nm technology node. These crucial changes are addressing growing heat generation caused by rapidly increasing leakage currents in scaled devices.

For a high-speed operation it is indispensable to have high drive current in the open, or on state of the transistor. In scaled devices the reduced gate length however

V. Sverdlov, *Strain-Induced Effects in Advanced MOSFETs*, Computational Microelectronics, DOI 10.1007/978-3-7091-0382-1_1, © Springer-Verlag/Wien 2011

results in a gradual channel control worsening which leads to high source-to-drain current in the passive, or off-state, for similar gate voltages. One option to keep the ratio between the on- and off-currents sufficiently high for operation is by decreasing the off-current, which can be done by increasing the gate voltage swing between the on- and off-state of the transistor. However, this again leads to high power production and thus is unacceptable. In order to continue scaling under the constrain of reduced heat generation the transport properties of the channel in the on-state must be improved. Since scattering with defects and surface roughness are already optimized, future progress requires a profound modification of the electron band structure leading to the increase of the carrier velocity.

Application of strain allows to increase the on-current significantly without changing the transistor design and meeting the projected performance increase. Although it has been long known that the electrical properties of silicon strongly depend on applied stress, strain as a mobility booster was first introduced in the MOSFET fabrication process at the 90 nm technology node. Since then strain engineering has become an integral part of the MOSFET fabrication process.

FinFET and ultra-thin body MOSFET multigate non-conventional structures possess superior channel control and reduced leakage as compared to bulk planar MOSFETs and are therefore suitable candidates for providing successful scaling to the end of the ITRS roadmap. Stress can be easily incorporated in non-conventional MOSFETs and is thus completely compatible with the upcoming non-classical MOSFET structures. Therefore, strain engineering is expected to keep its pace and remain one of the key elements of Complimentary MOS (CMOS) technology at the 22 nm technology node and beyond.

Strain is not the only new element introduced recently into CMOS production process. In order to guarantee a proper control over the channel in the 65 nm node transistor the silicon dioxide layer has become so thin that the gate leakage current and related heat generation could no longer be ignored. This prevents future silicon dioxide size reduction, and a new paradigm of scaling under the constraint of heat generation must appear. The solution is to replace the native silicon dioxide by another oxide with higher dielectric permittivity. This replacement allows to further reduce the equivalent electrostatic dielectric thickness thus improving electrostatic channel control while keeping the physical oxide dimension thick enough to prevent tunneling. At the same time, to reduce the depletion layer in the gate and to partly recover the channel mobility the polysilicon gate is replaced by a metal gate.

Although this step looks natural and simple, the introduction of a new dielectric and metal gates represents the most revolutionary change in the history of semiconductor industry and MOSFET production process since the replacement of germanium by silicon. Intel first introduced the new hafnium-based dielectrics with metal gates for its 45 nm technology node, and high-k materials with improved properties are now used in the 32 nm transistor. Together with new dielectric and metal gates, an improved technique to induce more strain into the channel for obtaining enhanced performance are employed for the 45 nm and the next 32 nm technology node.

Manufacturing complexity and production yield increase development cycle time and costs. Statistical parameter fluctuations are becoming more pronounced with shrinking transistor dimensions causing broader variations in device and circuit performance. It is customary to have a tool which allows predicting transistor properties thus making design easier. Technology modeling and simulations help reducing R&D costs and shorten the design cycle. Therefore, Technology Computer Aided Design (TCAD) tools are indispensable for development and optimization of upcoming generations of devices and integrated circuits.

In order to be predictive, TCAD tools must be based on accurate physical models. Although piezo-resistive coefficients describe modifications of electrical properties of bulk silicon on stress for small strain values, it is not enough to model transport in inversion channels, where the corresponding coefficients depend on carrier concentration, doping, channel length, etc. More detailed transport models are therefore required to describe current enhancement in inversion layers as well as in FinFETs and ultra-thin body FETs. The transport model must include carrier quantization in the confined direction. It has to include all appropriate carrier scattering mechanisms. For strain engineering the models must include stress induced modification of the band structure. These modifications have a profound impact on subband quantization energies, effective masses, non-parabolicity parameters, wave functions, and thus on scattering matrix elements. Although strain engineering is a mature technology to increase CMOS performance, the maximum performance enhancement has not been yet analyzed. A careful analysis to determine optimal conditions that lead to enhanced transport properties and the current boost is therefore needed.

Chapter 2
Scaling, Power Consumption, and Mobility Enhancement Techniques

2.1 Power Scaling

The power dissipation of a CMOS circuit consists of the dynamic (due to switching) and the static contribution in the off-state and can be written as [68]

$$P = \sum_i \alpha_i C_i V_{DD}^2 f + I_{OFF} V_{DD}, \tag{2.1}$$

where $0 < \alpha_i < 1$ is the "switching activity factor" of the ith circuit block, C_i is the total effective capacitance including that of all the interconnects and input capacitance of transistors, f is the clock frequency, and I_{OFF} is the total current in the off-state of all the transistors biased by the power supply voltage V_{DD}. In contrast to I_{OFF}, the on-current $I_{ON} = \sum_i (I_{ON})_i$ participates in (2.1) indirectly, via the speed requirement

$$f = p/\tau, \tag{2.2}$$

where

$$\tau = C_i V_{DD}/(I_{ON})_i, \tag{2.3}$$

and $p \ll 1$ is the fraction of the fraction of the clock period $1/f$ taken by the capacitance recharging constant τ.

The model of the power consumption described by (2.1)–(2.3) is approximate, however it captures the basic balance between the static and dynamic components of power generation.

At the beginning of the CMOS era the power consumption was reduced by scaling the transistor dimensions and thus the supply voltage V_{DD} down. However, with approaching 100 nm channel size, the V_{DD} scaling has slowed down. One of the reasons was a gradual increase of the currents in the *off*-state. This increase was mostly due to parasitic leakages, the most important is due to carrier tunneling through a thinner oxide. Indeed, in order to maintain a proper electrostatic control over the channel the thickness of the gate dielectric separating the gate from the channel must be reduced together with scaling of the gate length, which leads to a sharp increase of tunneling through a thin dielectric. With an increase of the *off*-current one option to preserve the high ratio I_{ON}/I_{OFF} is to increase the supply voltage

V. Sverdlov, *Strain-Induced Effects in Advanced MOSFETs*, Computational Microelectronics, DOI 10.1007/978-3-7091-0382-1_2, © Springer-Verlag/Wien 2011

V_{DD}. This option is, however, unacceptable, since, according to (2.1), it leads to an increase of the power consumption.

The industry has faced the problem of increase of heat generation already at the 90 nm technology node. The engineering solution to continue scaling, increase performance, and keep the heat generation under control was the introduction of strain into the channel [20]. Strain modifies the transport properties of the transistor in the open state, while keeping them practically unchanged in the off-state. If the I_{ON} current is increased by applying stress, it leads, according to (2.2), to higher speed and performance. Therefore, if a higher I_{ON} is achieved for the same I_{OFF} and V_{DD}, the performance gain is accomplished at nearly no increase of the power generation. Alternatively, the performance similar to an unstrained device is achieved at lower V_{DD} and thus reduced power consumption.

Although a new technology of high-k dielectric/metal gate, which allows reducing I_{OFF} (and thus power consumption) while preserving the good control over the channel, was introduced at 45 nm technology node [43], stress technique remains one of the main boosters of performance enhancement with scaling. In the 32 nm technology node introduced by Intel at the end of 2008, the fourth generation of advanced channel stressors is employed [44] allowing to get tensions of 1.2–1.5 GPa in the channel. In 2009 nearly 2 GPa stress in the channel was achieved [50].

The on-current boost by stress is due to the strain-induced mobility enhancement in the channel. Depending on the stress conditions, up to the fourfold mobility enhancement for holes and nearly twofold for electrons was reported [71] and up to 50% increase in transistor drive current [22, 43, 62, 76, 80] was documented. The mobility enhancement is predicted up to the stress level of at least 3 GPa [71], which is higher than the level currently delivered into the channel. It makes stress a viable, competitive, and important technology which will certainly be used to boost the performance of future technology generations beyond the 32 nm technology node currently in production.

In this chapter we will briefly review the history of stress in silicon and the main techniques to introduce stress currently utilized in laboratories and industry. Stress is not the only option to enhance mobility in the channel. As shown in Fig. 2.1, substrates different from the commonly used (001) wafers may also be used to obtain higher mobility. This hybrid orientation technology becomes important with the introduction of Fin-FET devices with the [110] channel direction, where the two fin interfaces are $(1\bar{1}0)$ oriented. Devices with channel directions different from [110] can also be considered. Finally, alternative channel materials with mobilities higher than silicon mobility, e.g., germanium or III–V semiconductors can be used.

2.2 Strain Engineering

Strain engineering technologies are based on enhancing the transport properties by mechanically stressing the silicon channel of a MOSFET. The advantage of these techniques is that they allow to get higher performance without changing the

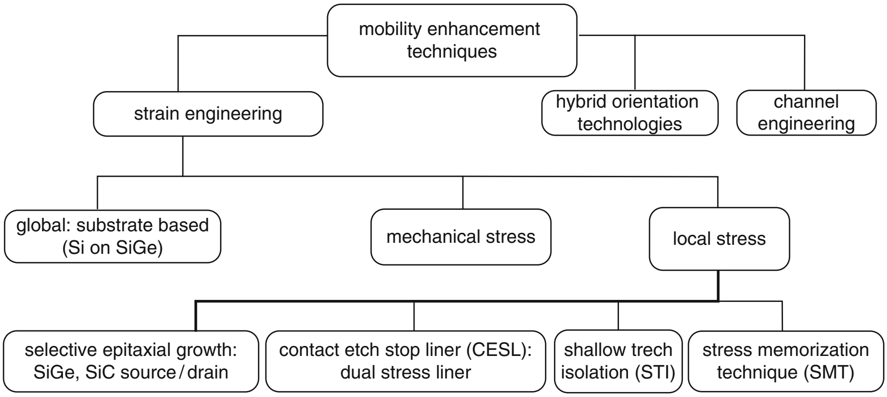
mobility enhancement techniques
strain engineering
hybrid orientation technologies
channel engineering
global: substrate based (Si on SiGe)
mechanical stress
local stress
selective epitaxial growth: SiGe, SiC source/drain
contact etch stop liner (CESL): dual stress liner
shallow trech isolation (STI)
stress memorization technique (SMT)

Fig. 2.1 Classification of stress techniques. Mechanical stress is used in laboratories

MOSFET size and architecture dramatically. Several techniques to deliver strain which require only little change in the process flow have been developed. This allows to integrate strained silicon into the manufacturing process at low additional production costs.

The influence of strain on transport in semiconductors has been a research topic for over half a century. Already in the beginning of the 1950s it was discovered that stress may influence the intrinsic silicon mobility [23,63]. To explain the effect, Herring and Vogt [26] have generalized the deformation potential theory initially proposed by Bardeen and Shockley [7] to describe the coupling between electrons and acoustic waves in solids and to express the relaxation times via the effective mass and deformation potentials. They have shown that the electron mobility change is due to repopulation between the valleys and reduced inter-valley scattering. Both effects are caused by stress-induced energy shifts which, depending on the stress condition, lead to the lifting of degeneracy of the six equivalent valleys. This interpretation of the mobility enhancement is often used to explain the mobility enhancement due to uniaxial stress as well, although, as we will show below, it is valid only for uniaxial stress in [001] direction, or, equivalently, for a biaxially, or inplane stressed sample. The effective mass change appears in [110] stressed samples, as was first demonstrated by Hensel [25] 1965 but since then well forgotten. Only recently [73] the Hensel-Hasegawa-Nakayma model of the conduction band was used to model the mobility enhancement in uniaxially stressed MOSFETs with technologically relevant [110] channel direction [72].

The stress-induced valence band shifts and warping are essential to understand the hole mobility modification. The **k**·**p**-based model [40] with a Hamiltonian including strain [8] has been a reliable and inexpensive method to address the stress-induced valence band modification since 1963 [24], which is successfully used nowadays to describe the subband structure in inversion layers [66].

The transport properties of strained silicon can be reasonably well predicted by piezoresistance coefficients for small stress values. However, the value of piezoresistances depends on the parameters like doping level or temperature and should be measured for each sample. Another problem is that the bulk values of the piezoresistances may not be used to predict the behavior of MOSFETs with confined carriers in the surface layer where the piezoresistance depends on the effective field as well.

Until the beginning of 1990 stressed silicon was studied by the physics community, but remained relatively unexplored for engineering applications [18]. In the pioneering work by Welser in 1992 it was demonstrated that an n-MOSFET with a channel built out of biaxially stressed silicon possesses nearly a 70% higher mobility [77]. In 1993 an increase of hole mobility in a p-MOSFET was reported [45,46]. The biaxial stress in silicon was achieved by growing the silicon layer on SiGe substrate. The drive current enhancement in pMOSFETs as a function of germanium concentration was investigated in [56], while short-channel n-MOSFETs were studied in [55]. History and the current status of the technology based on biaxially strained silicon, SiGe, and germanium channel MOSFETs is discussed in detail in a recent review [38].

By now the industry has adopted several technologies to introduce strain in the Si channel of MOSFETs. The key challenge is to make the technology compatible with the CMOS manufacturing process flow. For uniaxial stress the integration was successfully achieved [9, 20, 35, 64]. This is why uniaxial stress first introduced in [19,33,61] is currently employed by the silicon industry. Uniaxial stress results in a smaller threshold voltage shift [69] and higher mobility enhancement [66]. Modern stress techniques are compatible with the multi-gate architectures [30–32, 67] and were recently integrated with high-k dielectrics and metal gates [15,79].

Although many strain technologies were developed and introduced up to now, they can be conveniently divided into two distinct categories: global techniques where stress is introduced into the whole wafer, and local techniques, where stress is delivered to each transistor separately and independently (Fig. 2.1). Local stress is usually introduced during the process of MOSFET fabrication and is sometimes called process-induced stress.

Stress must be beneficial for the transport boost in both n- and p-type channels. It turns out that to get the performance improvement n-MOSFETs should be stretched, while p-MOSFETs must be compressed. Obviously, the global stress technique cannot provide the current improvement for both n- and p-MOSFETs. Therefore, industry uses local stress techniques, although biaxially stressed Si can also be used to increase mobility of n-type transistors [16]. We begin with biaxially stressed Si on SiGe technology.

2.3 Global Strain Techniques and Substrate Engineering

High quality silicon wafers are the primary elements used in chip manufacturing. Due to growing needs for channels with improved transport properties and rapidly increasing expertise in synthesizing new materials with enhanced electrical, mechanical, or chemical characteristics several ways to engineer silicon wafers were recently explored. This results in a substrate with unique properties which cannot be achieved by using silicon alone. At the beginning of 1990 the system of a silicon layer grown on a thick SiGe virtual substrate attracted attention to enhanced mobility in strained silicon [77, 78]. The lattice constant of relaxed SiGe is slightly larger than the one in relaxed silicon. Thus, a thin silicon film grown epitaxially on top of a SiGe substrate becomes tensely strained due to the lattice mismatch between silicon and SiGe. Because of the lattice symmetry of silicon a (001) silicon film is equally elongated along [100] and [010] axes which results in biaxial strain. This type of strain is introduced globally through the whole wafer. Biaxial strain results in the conduction band modification which finally leads to improved electron transport. The drive current is increased by up to 25% in sub-100 nm strained silicon MOSFETs [54]. Global stress techniques are not restricted to standard bulk CMOS technology. Thanks to layer transfer and wafer bonding global stress is successfully integrated into SOI wafers. Recently, the performance enhancement in a 60 nm gate length n-MOSFET with an ultra-thin strained silicon layer grown on a SiGe substrate on insulator was demonstrated [21, 58].

Current enhancement alone is not sufficient for a technology to go into mass production. The new technology must be economically competitive [57] and deliver benefits exceeding production costs. Regardless of the proven electron current enhancement in biaxially strained silicon, the presence of the SiGe layer in a substrate introduces several challenges for process integration. One problem is that the SiGe layer induces a high density of defects in strained silicon [18]. The diffusion of Ge atoms into the strained silicon film reduces the thermal budget window. Due to the lower thermal conductivity of SiGe device self-heating may become a problem, especially in the SiGe on insulator structures. Finally, the diffusion rate of dopant atoms (boron, arsenic) is significantly different from that of silicon [74].

Several alternative approaches to introduce biaxial strain in silicon without SiGe layer were proposed. In the “strained silicon directly on insulator” technology the SiGe layer is eliminated before transistor fabrication. This technology delivers a 25% drive current enhancement while avoiding the difficulties of SiGe process integration.

Another back-end technique introduces strain into an already processed wafer. In this approach the wafer is mechanically stressed, after it was thinned down and put onto a polymer film. After that the wafer can safely be bonded to a final substrate. The advantage of the method is that it allows to introduce uniaxial as well as biaxial strain according to the mechanical deformation, and a 100% performance enhancement has been demonstrated [1, 53], however, yield and reliability issues have so far prevented the technique from being used in IC manufacturing.

As it was already pointed out, global stress techniques are able to provide only one type of strain through the whole wafer. However, n- and p-type channels are affected by strain alternatively: an in-plane biaxial tensile strain is beneficial for n-MOS but detrimental for a p-MOS, and vice versa. We briefly review local strain techniques delivering a particular stress to each MOSFET.

2.4 Local Stress Techniques

Already in the 1990s it was found that certain process steps and IC elements appearing during wafer processing result in channel stressing and thus performance increase. Shallow trench isolation [41, 59], silicidation at the source and drain region [65], and formation of nitride contact-etch-stop layer [33, 61] were among earlier local stress techniques investigated (Fig. 2.2). Although the process induced stresses were moderate and could not provide sufficient drive current boost at the earlier stage, local techniques have certain advantages over global ones. Process-induced stress can be independently delivered to p- and n-MOSFETs guaranteeing the performance enhancement in both types of transistors. Additionally, stress can be introduced along three coordinate axes. This allows to optimize performance enhancement and costs, reduce the threshold voltage shifts [39], and improve integration into the process flow [36]. Importantly, the interest in stress technology was supported and motivated by industry needs to optimize the ratio of the performance to heat generation for the upcoming 90 nm technology node. Several process-induced local stress techniques, such as stressed nitride contact etch stop liner, stress memorization technique, selective epitaxial growth for embedded SiGe in the source and drain contacts, and stress from shallow trench isolation were introduced in mass production of integrated circuits.

In modern sub-100 nm technologies the transistor dimensions are so small that the mechanical stress induced by shallow trench isolation becomes important [9,75]. Stress can be induced both parallel and orthogonal to the channel lateral directions.

Another way to introduce compressive uniaxial stress into a p-channel is by filling the source and drain regions with SiGe [6,17,27,49,70,85]. For this purpose, the source and drain regions are etched out and a recess area is created. This recess is later filled by SiGe grown epitaxially in the source and drain regions [6, 49]. Alternatively, SiGe can also be grown on top of source and drain [12]. Depending on the

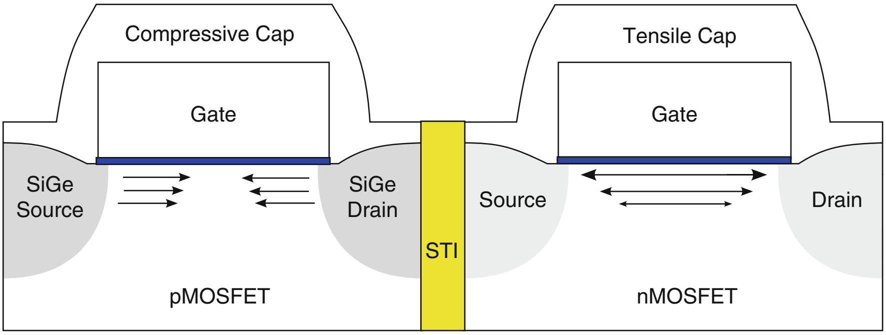

Fig. 2.2 Process-induced stressors employed by the semiconductor industry. Shallow-trench isolation, highly compressive and tensile capping layers, and compressive stress due to SiGe embedded in the source and drain regions are used in the CMOS process

thickness of the epitaxial $Si_{1-x}Ge_x$ and the Ge content x large uniaxial stress can be created using this method.

A part of the mechanical stress from a permanently stressed layer grown on top of a transistor can be transferred into the channel. The value of stress transferred depends on the thickness and the material properties of the liner [33]. To boost performance of an n-MOSFET a tensile cup layer is needed, while for a p-MOSFET the compressive layer is required. Thus, two different types of stress liners should be used to get performance enhancement in n-channel and p-channel MOSFETs simultaneously. Industry adopted a Dual Stress Liner (DSL) process, where a highly compressive nitride is deposited on top of the p-channel MOSFET, while a highly tensile nitride is deposited on top of the n-channel MOSFET. Silicon nitride (Si_3N_4) capping layers can produce both tensile and compressive strain depending on deposition conditions. In the fabrication process, a tensile silicon nitride layer is created by thermal chemical vapor deposition over the whole wafer. Parts of the layer are removed above p-MOSFETs by selective etching. After that a compressive Si_3N_4 layer is created by plasma-assisted chemical vapor deposition, followed by selective etching of the compressive layer above n-MOSFETs. Dual stress liners technology alone can improve the drive current by 11% in n-MOSFETs and by 20% in p-MOSFETs [60, 81].

Si_3N_4 layers with more than 2.0 GPa tensile and 2.5 GPa compressive stress which introduce approximately 1.0 GPa stress in the MOSFET channel are routinely used in 65 nm process [4]. This technique is successfully combined with selective epitaxial growth for embedded SiGe in the source and drain contacts [43]. Thus, strain engineering techniques may not only be combined for the same transistor, but can be superimposed to yield even larger performance boost [27].

Residual channel stress may by preserved after removal of the nitride layer. This fact is exploited in the stress memorization technique [11, 27, 36, 48]. In a process using this technique, the conventional dopant activation spike anneal is performed after the deposition of a tensile stressor capping layer. This layer is subsequently removed before an eventual salicidation process. Even though the stressor nitride layer is removed from the final structure, the stress has been transferred from the nitride film to the channel during annealing and memorized by the re-crystallization of source, drain and the poly gate amorphized layers. Stress from the capping layer can be memorized in the channel. Stress is preserved in the channel even after the stressed layer is removed from the final structure providing a 15% improvement of the drive current in n-channel MOSFETs [10].

Process-induced local stress techniques depend strongly on device geometry and must be adjusted and optimized to maximize beneficial effects from stressors [17]. However, regardless of the challenges of local stressors integration into the manufacturing flow, local stress techniques have proven useful for industrial applications and promising for future technology nodes.

2.5 Advanced Stress Techniques

Stress was introduced into the fabrication process flow at the 90 nm technology node. Since then stress is a compulsory technique to get the MOSFET performance enhanced included in all technology nodes. Stress techniques were constantly improved and perfected through the 65 nm and 45 nm technology nodes in order to transfer more strain into the channel. The germanium concentration in the source/drain regions of p-MOSFETs was constantly increased from 17% at the 90 nm technology node to 23% at the 65 nm which resulted in a 60% increase of the channel strain. At the same time an enhanced process flow adopted for the Si_3Ni_4 capping layers increase the channel strain in n-MOSFETs by 80% [6]. Strain techniques are compatible with high-k dielectrics/metal gate technology and were successfully integrated in the process flow at the 45 nm technology node, resulting in the third generation strained silicon [43].

At the International Electron Devices Meeting in 2008 Intel has reported its second generation of high-k dielectrics/metal gate 32 nm transistors. The fourth generation of stress technology allowed to get approximately 14% in performance improvement [44] as compared to the 45 nm transistors. The technique allowed to build the largest SRAM with more than 1.9 billions transistors. Multiple stressors are combined to produce even higher strain in the channel. The fourth generation stress technology includes improved stress liners for both n- and p-MOSFETs. Compared to the 45 nm technology node where the dual stress liners with 1.5 GPa tensile and 2.8 GPa of compressive stress were used [43], capping layers with more than 2 GPa tensile and 3.5 GPa compressive stress are introduced for the 32 nm node. In combination with SiGe source/drain regions with high (approximately 30%) germanium concentration uniaxial stress of approximately 1.5 GPa is produced in the channel. The replacement metal gate, or gate last, process when the poly-silicon gate of a transistor is removed and later substituted by a metal gate allows to produce even more uniaxial compressive stress [5, 44], as demonstrated in Fig. 2.3. This allows to obtain the best drive currents of 1.55 mA/μm for n-MOSFETs and 1.21 mA/μm for p-MOSFETs reported for 32 nm technology node at the end of 2008 [44].

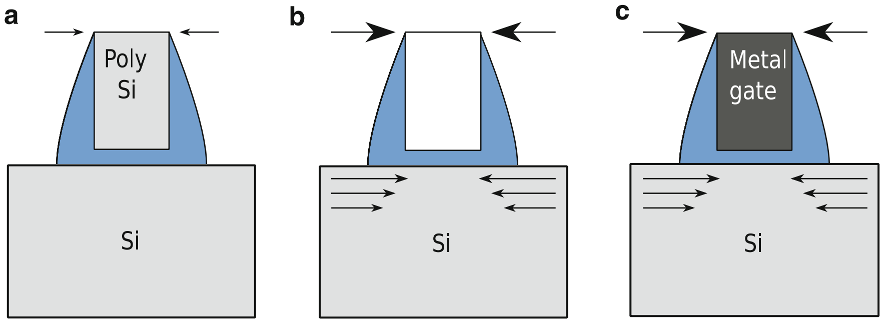

Fig. 2.3 Illustration of additional tensile strain introduced in the gate-last process [5, 44]