Performance Predictions for Scaled Process-induced Strained-Si CMOS

G Ranganayakulu and C K Maiti
Department of Electronics and ECE, IIT Kharagpur, Kharagpur – 721302, India

Abstract: Device and circuit simulations using process/physics-based Technology CAD tools are done to project the scaled CMOS speed-performance enhancement that can be expected from process-induced strained-Si CMOS.

Key Words: Process induced stress, Dual Stress Liners, Cutoff Frequency.

Introduction

For over four decades now, the semiconductor industry has relied largely on shrinking transistor geometries for improvements in both circuit performance and density, resulting in a consistent reduction of cost per function for each new technology node. CMOS scaling using conventional means such as, voltage and geometry have reached their limits. The quest to satisfy the low-power and low-leakage requirements of portable/mobile consumer electronics is driving efforts to improve transistor performance. As gate lengths approach sub-45 nm dimensions and gate oxides approach 1 nm, scaling becomes more challenging, and new material and device structures are required to overcome the fundamental physical limitations imposed by traditional semiconductor materials. Continuation of CMOS downscaling is now being attempted mainly by stress-induced carrier mobility enhancement. In general, in Si MOSFETs on <100> substrate with <110> channel direction, if uniaxial mechanical stress is applied along the channel direction, electron mobility is enhanced by tensile stress while hole mobility is enhanced by compressive stress [1].

Process-induced stress has been mainly employed due to its cost effectiveness. Fig. 1 illustrates possible directions of uniaxial stresses and their effects on N and PMOS. Different mechanical stresses can be introduced through the substrate and during the fabrication process. For NFET, tensile stressed liner has been used to induce tensile localized strain to improve electron mobility. For PFET, compressive localized strain has been induced by compressive stress line. Dual Stress Liner (DSL) has also been a cost effective technique to enhance both NFET and PFET simultaneously by selectively applying tensile stress liner at NFET and compressive stress liner at PFET [1]. The integration of DSL has been demonstrated for high performance CMOS. Recently stress proximity technique SPT for DSL has been successfully demonstrated by removing spacer between stress liner and poly gate to maximize stress proximity. The integration of cost effective technique DSL has been demonstrated.

Technology CAD process and device simulation tools play a critical role in advanced technology development by giving insight into the relationships between processing choices and nanoscale device performance that cannot be obtained from physical metrology tools alone. TCAD makes its greatest impact when a detailed understanding of the underlying physical mechanisms is tightly coupled within the technology development cycle so that physical insights feed directly into technology direction. In this paper, an attempt has been made to possible performance prediction of scaled strained-Si CMOS devices.

Contact etch-stop stress liner

Mechanical stress can be transferred to the channel through Si active area and poly gate if a permanent stressed liner is deposited on the device. Tensile liner will enhance NMOS and compressive liner will enhance PMOS mobility, respectively. The integration process is described below. After silicide
formation (either cobalt or nickel silicide), a highly stress liner (either tensile or compressive), such as PECVD (Plasma Enhanced Chemical Vapor Deposition) or RTCVD (Rapid Thermal CVD) nitride, is uniformly deposited over the wafer. The drive current dependence on stress is through the change in film thickness and material [2]. The remaining steps follow a conventional process flow, including interlayer dielectric and contact formation. Thicker nitride capping layer can increase the stress level. However, this will impact the conformity of gap fill itself, subsequent interlayer dielectric (ILD) gap fill, and contact opening process.

If one single liner is used, one drawback of this approach is that the device of the opposite type will be degraded. One of the possible solutions is to use high dose of Ge implantation to damage the nitride film and relax its stress. For example, a highly tensile stress nitride film is deposited over both N and PMOS, and Ge is selectively implanted into PMOS region to relax the nitride stress. However, the subsequent processes need to have low thermal budget to prevent the stress stability of the liner.

The above process does not result in both NMOS and PMOS to their highest performance. In order to achieve ultimate CMOS performance, two different types of stress liners should be applied to NMOS and PMOS, accordingly. For example, a highly tensile nitride liner is deposited first. It is selectively etched on the PMOS regions. Next, a highly compressive nitride liner is deposited, and this film is also selectively etched on the NMOS region. This process can be applied in the reverse order.

**Process Simulation**

The main process steps used in process simulation is similar to those reported process flow for the fabrication of a sub-45nm MOSFET from reference [6] and is summarized as follows:

1. Active patterning and threshold voltage implant
2. Gate oxidation
3. In-situ N⁺ poly gate deposition and patterning
4. NMOS S/D extension implantation
5. Spacer formation
6. S/D implantation & Annealing
7. Salicidation
8. Tensile nitride deposition
9. Removal of tensile nitride on PFETs
10. Compressive nitride deposition
11. Removal of compressive nitride on NFETs
12. Contact formation

Taurus Process [5] is a process simulator for one-, two-, and three-dimensional (1D, 2D, and 3D) structures and simulates the fabrication steps used to manufacture semiconductor devices. Simulation capabilities are focused on front-end processes such as ion implantation, activation, and annealing including oxidation and silicidation. The simulator allows the specification of arbitrary initial geometries. The simulation of etching and deposition is restricted to simple geometric operations where the resulting shape can be derived from the initial structure and the process description. Taurus Process uses a viscoelastic model to simulate stress evolution in the structure. By default, stresses are computed only during oxidation and silicidation steps and in material regions other than the silicon substrate.

The starting material is a (100) Silicon wafer with an initial boron concentration of ~4.0x10¹⁷ cm⁻³. Channel doping is performed to adjust NMOS and PMOS Vₜ using ion implantation. To relieve the etch damage; a sacrificial oxide is removed before gate oxidation. 22 Å thermal oxide is grown and in-situ heavily doped N⁺ poly-silicon is deposited. After gate plasma etch, the source/drain extension implantation is simulated followed by annealing simulation. To simulate the stress history during the entire process flow and in the entire structure including the bulk region, one needs to activate model Keep-Stress-History. With this, the simulation will include thermal mismatch stress, lattice mismatch stress, stress relaxation during low temperature geometry changes, intrinsic stress during deposition, and stresses caused by material growth and volume expansion or shrinkage.

This model also activates the stress-strain relations in all the regions including the silicon substrate and does stress simulations. Spacer of silicon oxide is deposited and dry etched with final thickness of ~ 40 nm followed by original source/drain implantation. Thermal anneals above 1000°C are used for dopant activation. Dual Stress Liner (DSL) shown in Fig 4 is deposited as explained by the steps 9-12. After inter-layer dielectric deposition, W is used for metal contact plugging and Cu is used for interconnection.
Device Simulation

For deep submicron devices, quantum mechanical effects are becoming increasingly important. In particular, thinner oxides and higher substrate doping used in advanced technologies lead to high electric fields that can quantize electron and hole motion in the inversion layer. This phenomenon has an effect on:

- Threshold voltage
- CV characteristics
- Carrier distribution

Taurus Device [5] provides a wide range of quantum mechanical models for simulating the impact of carrier confinement on device operation. These models include the Van Dort model, the Modified Local Density Approximation (MLDA) model, and a Schrödinger equation solver and allow the user to trade-off speed and accuracy. Also the channel–gate oxide interface must be resolved to a very high level of accuracy. The MLDA is an alternative quantum mechanical model that is capable of calculating the confined carrier distributions that occur near Si/SiO₂ interfaces. It can be applied to both inversion and accumulation and simultaneously to electrons and holes. It is based on a rigorous extension of the local density approximation and provides a good compromise between accuracy and runtime.
Results and Discussion

The $I_d-V_d$ characteristics of NMOS and PMOS devices with DSL process compared with control samples are shown in Fig. 5. Significant drive current enhancement of both N/P-MOS is achieved with insertion of this DSL in standard CMOS flow.

Simulations are performed to study further scaling down to 32 nm gate lengths. Simulation results (id-vg characteristics) using synopsys taurusdevice tool [5] for mosfets with 20Å gate oxide thickness are shown in Fig. 6 for three different gate lengths.

![Fig. 6: Simulated $I_d-V_d$ for three different gate length 70nm, 45nm and 32nm MOSFETs](image)

![Fig. 7: Comparison of simulated transconductance for both NMOS and PMOS](image)
Fig. 7 shows the $g_m$ variation vs. gate bias for both NMOS and PMOS devices. It shows the transconductance improvement for 45 nm gate length transistors with DSL over the reference control transistor with same gate length. A 12.5% higher $g_m$ than reference transistor is obtained.

The focus of the ac analysis is set on the intrinsic device cutoff frequency, $f_T = \frac{g_m}{2\pi(C_{gs} + C_{gd})}$, where $C_{gs}$ and $C_{gd}$ are the gate-to-source and gate-to-drain capacitances, respectively. This analysis is performed to see the impact of stress on the RF performance. To get higher $f_T$, not only higher $g_m$ is needed, but also higher $g_m/C_{gs}$ ratio. In Fig. 8, the cutoff frequency $f_T$ is plotted against gate voltage. As expected a peak cutoff frequency of 411 GHz is observed for NMOS devices (with tensile cap layer) of gate length 22nm and 165 GHz is observed for PMOS devices (with compressive cap layer) of gate length 22nm. A 6% higher cutoff frequency, $f_T$ is obtained for NMOS with tensile stress than its control counter part.

Fig. 9 compares the peak $f_T$ values predicted by simulation with published experimental data for bulk NMOS and PMOS devices. Simulated $f_T$ values for planar bulk NMOS and PMOS devices reasonably fit the trend pattern of the experimental data [3, 4]. The DSL process technique is used to achieve higher values of the drive current and cutoff frequency for nanoscale MOSFETs.

**Conclusion**

Even when scaling reaches its limits, alternate technologies discussed will result in improved performance. Performance projections of scaled CMOS devices involving process-induced strain are made via technology CAD simulation.
Acknowledgement

The work has been supported by MCIT/DIT, N Delhi.

References


