Stress-related Local Layout Effects in FinFET Technology and Device Design Sensitivity

Angelo Rossoni, Member IEEE, Tomasz Brozek, Senior Member IEEE, Sharad Saxena, Rajesh Khamankar, Luigi Colalongo, Member IEEE, Zsolt M. Kovacs-Vajna, Senior Member IEEE

***Abstract* — Transistor characteristics in advanced technology nodes are strongly impacted by devices design and process integration choices. Variation in the layout and pattern configuration in close proximity to the device often causes undesirable sensitivities known as Local Layout Effects (LLEs). One of the sensitivities is related to carrier mobility dependence on mechanical stress, modulated by device design and local/ global environment. In this paper we investigate the impact of stress, developed during FinFET device fabrication, on electrical characteristics of transistors manufactured in 7nm silicon FinFET technology. Two sources of stress modulation are studied: (i) active region isolation (Diffusion Break) (ii) Metal Gate extension outside of the fins of the transistor. A 3D TCAD process model of a FinFET device was created and calibrated using electrical characteristics measured on foundry fabricated silicon wafers. The model was then applied to simulate mechanical stress in transistors with various design attributes for Diffusion Breaks (Single vs. Double Diffusion Break) and Gate Cuts, following by modeling of electrical characteristics. Very good agreement between simulations and measured silicon data has been obtained for PMOS and NMOS FinFET transistors. This work demonstrates that the layout sensitivity in discussed design cases can be explained by modulation of the mechanical stress and that the model can be used to predict successfully the stress distributions and their impact on electrical characteristics of FinFET devices. It can be applied to assist designers and technologists with Design-Technology Co-optimization, design rule and PDK development, and process optimization for best performance and reduced variability.**

***Index Terms* — FinFET, transistor, 7nm node, silicon technology, electrical characteristics, local layout effects, TCAD, simulation, modeling, diffusion break, gate cut, mechanical stress**

I. INTRODUCTION

THE progress in VLSI scaling is enabled by advancements in patterning techniques, material engineering, and innovations in device architecture. Smaller geometrical dimensions increase the transistor sensitivity to the layout and the interaction between neighbor devices and isolation between them. Traditional isolation scheme, like Shallow Trench Isolation (STI) introduces discontinuity of active region and sensitivity to geometrical dimensions and materials used for STI fill. Similarly, patterning (lithography and etch) and its fidelity in reproducing intended layouts becomes very sensitive to proximity effects due to light interaction and etch loading between dense and sparse features. The proximity effects due to various factors have been known as Local Layout Effects (LLEs) and were identified as a significant source of device variability [1][2]. There are multiple sources of LLEs and most often they are driven by a boundary (discontinuity) of some layer or material in the close proximity to the device. Among most known LLEs are:

- Active area - shape and distance to neighbor active regions [3][4] – this includes the dimensions and spaces along the transistor channel and perpendicular to that direction. Gate width variability control and introduction of Replacement Metal Gate (RMG) process caused a restriction of gate pattern design and unidirectional orientation. Consequently, active regions and channel directions also became unidirectional, which simplified the design and helped to introduce stress as mobility boosters.
- Gate extension beyond the edge of the channel has a strong impact on off-state leakage control in traditional planar MOS transistors. Additional sensitivity is introduced by P-N boundary of Poly-silicon gate doping or the boundary between work function metals of PMOS and NMOS in case of the metal gate [5]. For RMG integration the gate metal stack also modulates the mechanical stress in the channel of the transistor, and the placement of the gate line end may impact the device performance [6][7].
- Well proximity and P-N boundary, the effects which becomes stronger when a border between the regions of the opposite doping type is very close to the edge of the

This paragraph of the first footnote will contain the date on which you submitted your paper for review, which is populated by IEEE.

Corresponding author: Angelo Rossoni.

Angelo Rossoni, Tomasz Brozek, Sharad Saxena, and Rajesh Khamankar are with PDF Solutions, Santa Clara, CA, USA (first author's email: angelo.rossoni@pdf.com),

Angelo Rossoni, Luigi Colalongo and Zsolt M. Kovacs-Vajna are with University of Brescia, Department of Information Engineering, Brescia, Italy.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

transistor, with possible "contamination" by opposite type of dopants (from implant scattering or out-diffusion), happening either in the channel or in the gate of the device.
- Device type neighborhood – the doping boundary proximity in case of the threshold voltage type and boundaries between various regions, since adjustment implants or work function metals can introduce dopant gradients, cross-diffusion, and channel doping variability.
- Contact placement sensitivity is a new layout effect, especially for technologies with high contact resistance, due to current crowding effects, but also due to possible modulation of stress effects in strained dielectric layers surrounding the transistor.

Those effects become more pronounced in 3-D structures like FinFET, with a fin-shape active region surrounded by a 3-D gate structure composed of multi-metal layers. Many of those effects can occur simultaneously and trigger multiple interactions which affect transistors in a complex way. For example, an isolated PMOS transistor can be impacted by a P-N boundary surrounding it from all sides, and an additional effect can arise due to density-driven volume change of the epitaxially grown SiGe (acting as a stressor and mobility booster in PMOS devices). Therefore, it is important to study each effect separately to understand device sensitivities and relative magnitude of each of the effects, but also their combination and interaction.

The LLE effects have a large impact on device variability in logic and analog designs. The logic designs are typically built out of standard cell library, which have pre-defined geometrical rules: gate length and pitch, number of fins, isolation width, gate extension, etc. Those rules limit the number of possible designs, but their possible combinations still lead to significant variability. In ideal case, such variability caused by LLEs should be predictable and included in device models in PDK (Process Design Kit). The problem becomes more complex in the case of analog or mixed mode circuitry, where designers use less restrictive rules or try to re-use functional blocks previously designed for different technologies or previous process generations. Characterization of transistors in such blocks can be very difficult and very often transistor parameters differ significantly from their expected characteristics. Availability of calibrated models capable of predicting the impact of various LLEs across a wide range of design attributes for different device types would be very helpful to both device designers and process integrators. The former ones could understand the trade-offs between the design style choices and consequences to device performance, and the latter ones could use the models to guide them through optimization of process integration and material innovation. Finally, layout optimization could be used for product design improvement as a DFM (Design for Manufacturing) tool for reduced variability and improved yields.

II. Stress- related LLEs in FinFET transistors

In this study we will focus on the LLEs related to mechanical stress. The effect of stress on the electrical characteristics of silicon devices is well understood and studied. Early research works focused on piezoelectric effects, but recent transistor developers introduced stress as a mobility enhancement factor and transistor performance booster. A qualitative impact of the stress on the carrier mobility [8] are summarized in Table I.

TABLE I
Longitudinal Channel Mobility sensitivity

	Mobility in the channel oriented <110>	
Stress Direction	**Electron mobility**	**Hole mobility**
Vertical <100>	Strong dependence Increases with Compressive Stress	Medium dependence Increases with Tensile Stress
Lateral <110>	Weak dependence Increases with Tensile Stress	Medium dependence Increases with Tensile Stress
Longitudinal <110>	Medium dependence Increases with Tensile Stress	Strong dependence Increases with Compressive Stress

The dependencies outlined in Table I were used to guide stress engineering activities to help improve transistor performance, first by employing the stressor layer on top of the transistor, and then by integrating it with the transistor structure itself. A good example here is a strained Silicon-Germanium (SiGe) epitaxially grown Source / Drain (S/D) region, which was used as a stressor for PMOS transistors [9].

Starting from 90nm technology node, stress engineering has been a strong contributing factor of VLSI technology [10]. However, the integration schemes and isolation techniques used in device fabrication influence the stress experienced by a transistor, which can impact device performance in unpredictable and sometimes undesirable ways.

Therefore, it is critical to understand and quantify these sensitivities and determine how significant the LLE impact on device performance is.

A. Device Isolation and Active Diffusion Break

CMOS planar technology generally relies on the Shallow Trench Isolation (STI) to isolate active areas of the transistors, and this remains the case for the FinFET technology as well, however the active area and channel regions of FinFETs are elevated above the STI region. To separate the channels of the devices along the fins, the fins have to be cut through their whole height. In advanced technology nodes, gate poly-Si pattern is very regular, "on-the-grid", to reduce variability. This requires an alignment of isolation edges with the poly-silicon lines. As the result, two alternative isolation configurations emerged [11]: Double Diffusion Break (DDB) and Single Diffusion Break (SDB). They are shown in Fig.1 (layout) and Fig.2 (3D simulated structures) and will be detailed below, in Section II.

For device isolation, DDB introduces an isolation region that spans between two neighbor poly-silicon lines, whereas SDB

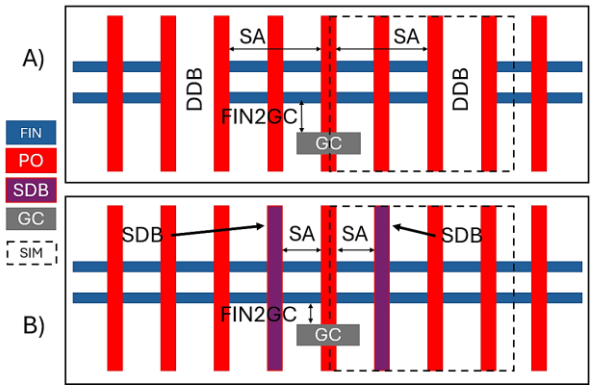

Fig. 1. Layout and model parameters of DDB (A) and SDB (B) test structures, and the area simulated by the model (dashed), showing fins (FIN), Poly/Gate (PO), Gate Cut (GC), distance FIN to PC (FIN2GC), distance channel to DB (SA)

uses the isolation region of much narrower width. In FinFET technologies SDB can be done with different integration approaches, with the fin cut performed either before or after the sacrificial gate poly formation [11]. The so-called Self-Aligned Single Diffusion Break (SA-SDB) has been a preferred choice across the industry as it helps to reduce the impact of the isolation on S/D SiGe growth and better control the fin cut location with respect to the gate.

Comparing the two isolation schemes, DDB consumes a significant amount of area between the active transistors, while the SDB allows to reduce part of STI area and improve transistor density. However, these two isolation constructs result in different stress modulation, particularly in PMOS devices, where the compressive stress is intentionally introduced by means of epitaxial SiGe in the Source/Drain regions to enhance hole mobility in the channel.

Understanding the impact of these Diffusion Break (DB) configurations on stress modulation is essential for optimizing device performance and achieving the desired electrical characteristics of transistors in advanced VLSI technology.

B. Poly/Gate continuity – Poly/Gate Cut and Cut location

In the early stages of VLSI, devices were relatively large. The precision control of the gate line ends was primarily achieved through lithography and patterning techniques. The focus was on integrating more transistors onto a single chip, moving towards higher density designs. As device dimensions shrank, the need for precise gate line end definition became more pronounced. The gate line end control is important in CMOS design, when the gates of neighbor PMOS and NMOS transistors need to be separated, and in order to reduce the space between the devices and avoid gate tip-to-tip shorts, a dedicated Gate Cut (GC) patterning step was introduced.

The transition from planar transistors to FinFETs brought about significant changes. FinFETs, with their three-dimensional structure, required more sophisticated GC techniques to ensure

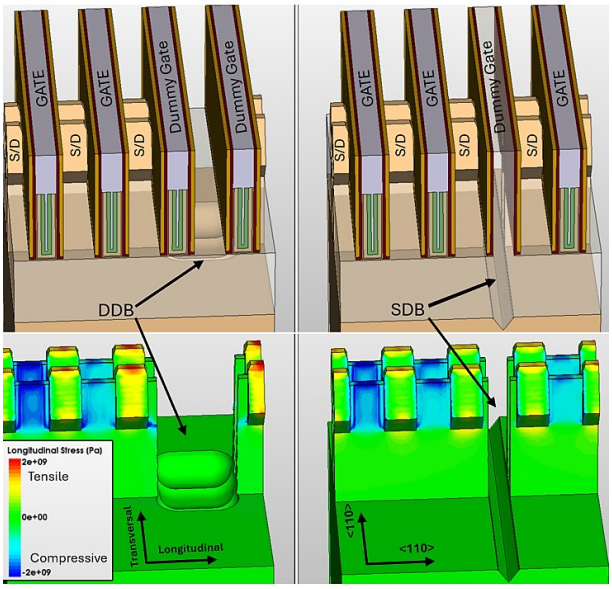

Fig. 2. DDB structure (left) and SDB structure (right). The top figures show the gate metal stacks, the gate spacers and the silicon at the end of the gate metal stack deposition process step. The bottom figures show the longitudinal stress on silicon for DDB (left) and SDB (right)

proper isolation and performance. The GC process became crucial for defining the fin structure and maintaining electrostatic control [12].

At advanced technology nodes, such as 7nm and 5nm, the GC process has become even more critical and can be performed at various stages of fabrication. Some technologies implement it immediately after sacrificial poly-Si gate patterning ("GC first"), others after the deposition of the nitride wall ("GC late"), and some delay it until after the deposition of the gate metal stack ("GC last"). These approaches have distinct characteristics and significantly impact the mechanical stress of the devices in different ways. Managing the stress introduced by the multi-metal stack is essential to maintain device reliability and performance.

The GC process involves etching to isolate individual transistor gates from continuous gate strips. Breaking the continuity of a metal gate (which in fact is a strained metal bar) results in

Parameter	Value
Fin pitch	30nm
Fin width	6nm
Fin height	40nm
Gate length	16nm
Poly pitch	50nm
Number of Fins	1, 2
Gate stack	TiN, TiAl, W

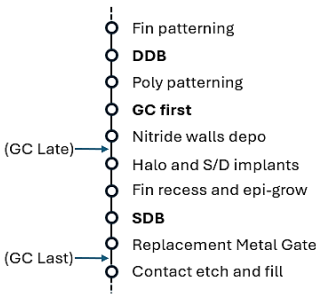

Fig. 3. Device structural parameters used in simulation (left). Process steps implemented in the model (right).

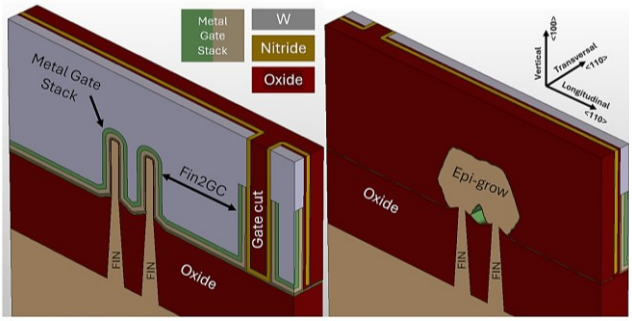

Fig. 4. Front (left) and back (right) of the 3D structure created by the process simulator and used as input for the device simulator to predict the electric characteristics

modification of the structure which can alter the mechanical stress profiles within the FinFET device, affecting the mechanical state of the device.

II. Methodology Description and Experiment

As a simulation tool we adopted Synopsys Sentaurus TCAD 3D (version U-2022.12) with a coherent model of a FinFET device based on a 7nm technology node. The stiffness parameters and the stress-to-strain conversion factors for the materials are standard and provided by the process simulator tool. The model included a target transistor surrounded by neighboring devices as in a CMOS logic block built with standard cells. The layouts were created with design attributes to represent cases of various diffusion breaks and gate line ending. The main device structural parameters used for simulations are shown in Fig. 3 (left). To collect experimental results, several test structures representing the layouts for such models were designed and manufactured in a 7nm FinFET technology to measure the electrical variability induced by LLEs and to validate the model. The test structures were designed as a part of the Characterization Vehicle® (CV) built with PDF Solutions characterization infrastructure [13].

The test structures used for this study were processed at a commercial foundry with 7nm node process capabilities. These structures help to study the impact of different DB designs and GC proximity on stress modulation and device performance. The model encompasses the main process steps (Fig. 3 right), from fin patterning to source/drain/gate (S/D/G) contact fill. The thermal budget and temperatures of key process steps are carefully integrated into the simulator to account for their impact on stress generation and relaxation.

The standard wafers used in manufacturing have a (100) orientation, the sidewall of the fins of FinFET transistors are (110) and the channels are oriented in <110> direction, those were also the settings used in our simulations. The insets in illustrations of Figs. 4 and 13 show the corresponding crystallographic indexes related to transistor channel/fin plane and orientation.

Fig. 1 shows two scenarios: (A) SDB and (B) DDB. Although shown separately, they can coexist in the same design. The figure also illustrates GC isolating the gate of the Device Under Test (DUT), with adjacent gates treated as dummy transistors. The dashed area represents the domain of both process and device simulations. Stress equilibrium equations require boundary conditions, assuming zero velocities of the particles in the direction perpendicular to the boundary planes. The resulting 3D structure is then mirrored relative to the left edge and electrically simulated.

Fig. 2 displays the resulting 3D structures of PMOS transistors generated by the process simulator at the end of the metal stack deposition process step (top), featuring the DDB (left) and SDB (right). The bottom images show the longitudinal stress distribution in the silicon for the two cases.

Fig. 4 shows a cross section of the FinFET structure with the gate cut. As mentioned in the introduction, the gate cut may introduce two LLEs: change in the mechanical stress and P-N boundary between PMOS and NMOS gates. Since this study focuses on the mechanical stress case, we simplified the structures and consider only the case where the modeled transistor is facing other transistors of the same type.

In this work, we simulated three orthogonal components of the mechanical stress in the channel of the transistors (longitudinal stress along the transistor channel and along the fin; transversal stress along the gate and perpendicular to the fin, and vertical stress perpendicular to the wafer surface). A good agreement between measured data and simulation has been achieved. It was decided not to include the diagonal stress terms in the analysis. Stress is studied as a function of the distance to both DB types and to the GC. The result of this simulation is then used to model the impact on transistor performance caused by the modulation of the carrier mobility in the channel.

A. Model calibration

A specific 2-fin test structure was selected as a reference, characterized by a distant diffusion break (DB) located more than nine poly pitches away and the first GC at 200nm distance. To protect the confidentiality of the particular silicon process and the raw measurement data, we use only normalized values for performance modulation caused by stress LLEs. We are reporting relative changes in the transistor drive current in the

Group	Model	Parameters
Fermi-Dirac statistics	Fermi	---
High-K interface with silicon	RPS InterfaceCharge	--- 1e12 charge concentration
Mobility	IALmob BalMob HighFieldSaturation EffectiveIntrinsicDensity Recombination	PhononCombination, FullPhuMob KVM, Frensley GradQuasiFermi, OldSlotboom --- SRH, Band2Band
Strain effects and quantum models	MultiValley DeformationPotential DOS SubBand	ThinLayer, MLDA, kpDOS ekp, hkp minimum eMass, hMass Doping, EffectiveMass, Scattering(MLDA), RelChDir110

Fig. 5. List of models and their parameters used for simulations in this work

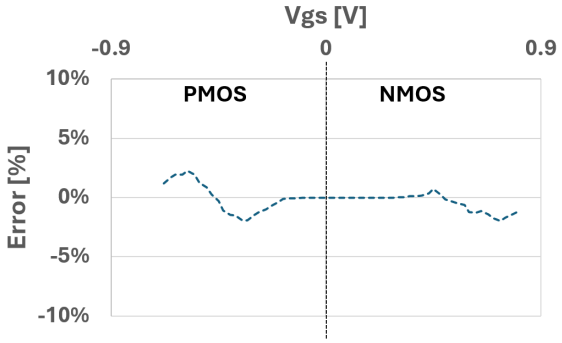

Fig. 6. Difference between simulation and average value of measured Drain current (in linear region VDS=0.05V) as a function of the gate bias VGS

linear regime, measured with the gate voltage at nominal supply voltage VG=0.8V and the drain voltage at VD=0.05V. Additional factors, like parasitic source and drain resistances add to the variability of the drive current in the saturation region, complicating the analysis of stress-induced LLE. Thus, the linear region current is used to detect stress-induced mobility changes.

The deposition temperatures of the materials are very important from stress management perspective. As the materials cool down from deposition to room temperature, the mismatch in thermal expansion rates can induce residual stresses, thereby impacting the performance of the devices.

The gate dielectric and metal stack used in our simulations is shown in Fig.3 (left), with HfO_2 deposited by Atomic Layer Deposition (ALD) at 250°C, TiN by ALD at 200°C, TiAl by Chemical Vapor Deposition (CVD) at 350°C, and Tungsten by CVD at 450°C [14][15][16].

The proposed model can effectively simulate all three types of GC processes discussed earlier. 7nm FinFET technologies use preferably the "GC first" approach and we also adopted it in this study. A comprehensive description of this process is provided in [17].

Adaptive meshes were used to guarantee good resolution even in the most critical cases. Given the presence of curvilinear geometries, attention was paid to the minimum angles of the

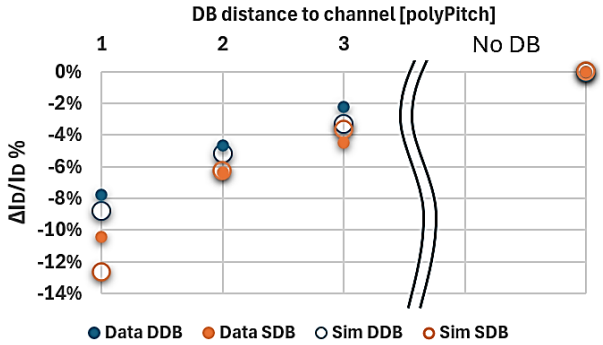

Fig. 7. Changes in PMOS Drain current (in linear region VDS=0.05V) as a function of the device gate distance to Diffusion Break

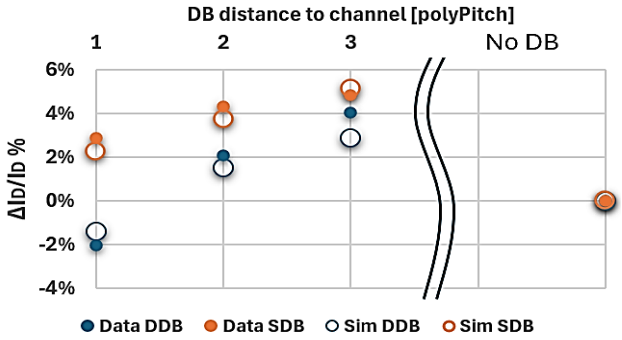

Fig. 8. Changes in NMOS Drain current (in linear region VDS=0.05V) as a function of the device gate distance to Diffusion Break

mesh to enhance convergence, and symmetry lines were defined to minimize the generation of artifacts both at the structural level and in the tensor fields.

The mesh of the channel of the DUT was always maintained at high resolution.

After the process simulation and before the device simulation, the mesh of the resulting 3D structure was regenerated to meet the minimum specifications of the physical models used. These models were chosen based on the fin dimensions, which necessitated the use of quantum models in addition to classical transport models [18][19].

The simulation was performed for each fin separately and the simulated results were combined to build electrical models of the 2-fin devices used for electrical characterization.

The device simulation was then calibrated using the reference test structure, whose electrical characteristics allowed

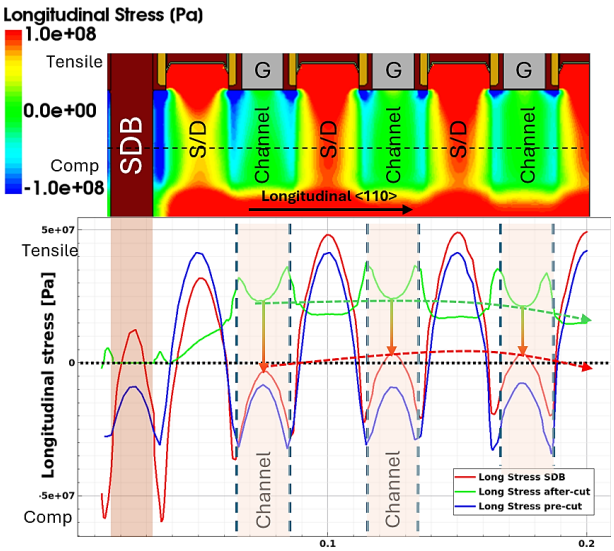

Fig. 9. Longitudinal cross section of the fins with longitudinal (Long) stress distribution (top) of NMOS devices. At the bottom, the graph of longitudinal stress of the cut line at different process steps: pre SDB etch (green), after SDB etch (blue) and SDB dielectric fill (red)

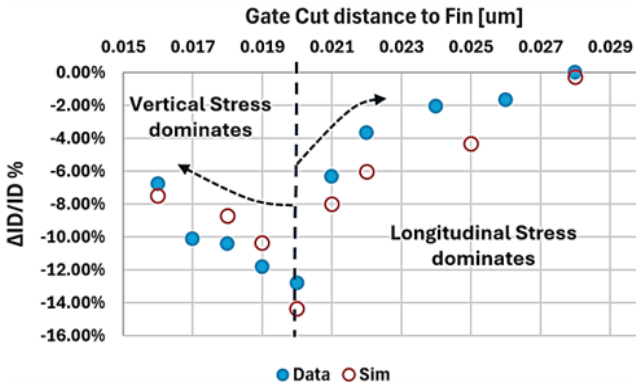

Fig. 10. Changes in PMOS Drain current (in linear region VDS=0.05V) as a function of the distance between the device outmost fin and the gate cut

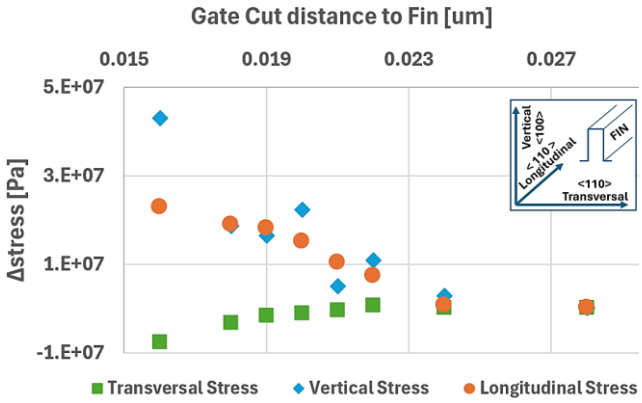

Fig. 11. Distribution of the stress variation (referenced to the case without gate cut) as a function of the distance to the Gate Cut for PMOS FinFET; three vector components of the stress are shown: longitudinal (orange circle), transversal (green square) and vertical (blue diamond)

to derive specific parameters for the models (showed in Fig.5). The difference of the final result and the measured parameters are comparable to the standard error of the available data set, reaching a maximum and an average deviation of 2.2% and 0.2% respectively (Fig.6).

B. Experimental data set

Test structures for 2-fin NFET and PFET transistors with SDB proximity, DDB proximity and GC proximity were used to characterize LLE effects in a 7nm FinFET technology. Each structure is specifically designed in a controlled environment to be responsive to a single LLE, simplifying the process of targeting each source of stress [1].

I-V curves were measured at room temperature (25˚C) for linear and saturation regimes and key device parameters (threshold voltage, drive current, subthreshold swing) were extracted for each transistor.

Every DUT is replicated eight times in a tightly clustered area for each die. Test structures for all devices are measured on every die on the wafer resulting in several hundreds of measurements for every transistor. Median values are used to represent electrical measurements of device parameters. This approach minimizes sensitivity to local and wafer spatial variations and generates a statistically robust data set. The standard deviation of measured parameters is less than 1.3% for the drain current in linear and saturation regions, and less than 1.6% for the threshold voltage and the subthreshold swing. This data set played a crucial role in both calibrating and validating the 3D TCAD model.

III. Results

A. Diffusion Break in PMOS devices

Fig. 7 shows the change in the PMOS drain current in linear region as a function of the distance from the Diffusion Break boundary. The plot shows both the simulated electrical performance (open circles marked “Sim”) and the electrical results measured on the silicon (closed circles marked “Data”) in DDB and SDB cases.

DDB is performed prior to the formation of the Source and Drain regions, which, in the case of PMOS, are rich in Germanium and generate the majority of the stress within the device. Conversely, the process steps that form SDB occur after the epitaxial growth (Fig. 3, right), and are thus carried out under conditions of high structural stress. It is inevitable that the fin cut in the SDB causes substantial relaxation, which is only partially contained by the nitride that constitutes the gate spacers. For this reason, the performance in the case of SDB is more sensitive to the distance of the cut compared to the DDB case.

As expected, the reduction in performance due to DB increases when the DB is closer to the device. The simulation may slightly overestimate the effect for the device closest to the DB. This is due to the lack of precise profile and depth of the etch performed to isolate the devices, which affects sensitivity in close proximity.

B. Diffusion Break in NMOS devices

The non-monotonic electrical characteristics of the NMOS device shown in Fig. 8 warrant a detailed discussion.

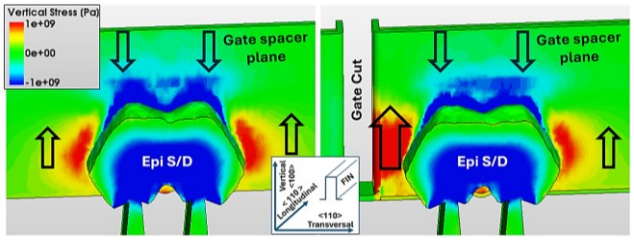

Fig. 12. Distribution of the vertical component of the stress arising from the interaction between epitaxial growth of Source and Drain and the surface of the gate spacer; left – continuous gate without cut; right – gate cut close to the fin (compressive stress is shown in blue and tensile in red)

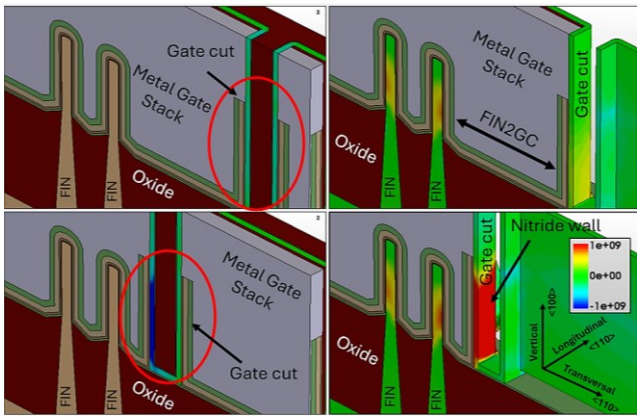

Fig. 13. Simulated longitudinal stress on nitride walls (left) and vertical stress on fins and nitride walls (right) when GC is far from fin (top) and close to fin (bottom) in PMOS

Fig. 9 (top) illustrates a longitudinal section of the fin, depicting the distribution of longitudinal stress in silicon. Fig. 9 (bottom) presents the graph of this stress across three consecutive process steps: before the SDB cut (blue line), after performing the cut (green line), and after filling the cut with the insulating dielectric (red line).

Compressive stress is not beneficial for NMOS device performance (rather opposite), therefore a traditional approach with silicon epitaxy was used for source and drain regions of NMOS transistors. Because of that, the overall stress levels in NMOS channel are lower (on absolute scale) than in PMOS engineered FinFETs. The simulation shows that in continuous fin (no SDB in close neighborhood) of NMOS FinFET the stress is compressive, and its value is relatively small. The DB performed through the etch of a small fin segment causes a relaxation of the compressive stress and transition into a tensile state (Fig.9 blue to green). The stress profile in the S/D region becomes flat, lowering the central tensile peak and raising the lateral wings to the same level. The channels, which have a smaller silicon volume compared to the S/D regions, absorb this change from both sides, transitioning from a compressive to a U-shaped tensile stress.

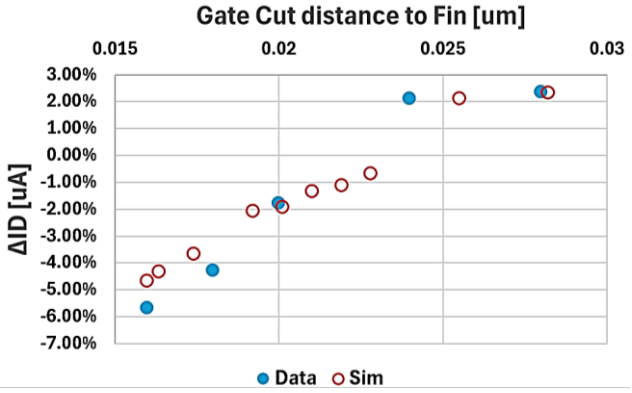

Fig.14 Changes in NMOS Drain current (in linear region VDS=0.05V) as a function of the distance between the device outmost fin and the gate cut

Following the SDB cut, the dielectric material fills the created gap. This dielectric fill exerts a compressive force on the adjacent fins, pushing back the stress towards its original state before the cut (Fig.9 green to red). Notably, even the maximum compressive stress induced by the dielectric fill is not able to completely revert the stress change induced by the cut.

The impact of both aforementioned process steps on stress gradually diminishes with increasing the distance from the DB (green and vertical arrows). Consequently, near the SDB, the compressive stress decreases as the distance from the SDB cut increases, due to the dominating stress gradient generated by the filling dielectric. This stress gradient positively impacts the performance of NMOS transistors, as reduced compressive stress enhances electron mobility.

For distances greater than four poly pitches, the relaxation due to the SDB cut starts to diminish, pushing back the longitudinal stress towards the original state before the cut. This stress gradient from tensile to compressive becomes the dominant component, thereby reversing its impact on mobility.

Fins that remain in an uncut line exhibit higher compressive stress compared to those that have undergone the SDB process. This elevated stress correlates with inferior performance metrics, underscoring the benefits of the SDB architecture in optimizing NMOS device functionality.

The performance of NMOS devices is characterized by stress variations resulting from individual process steps and the properties of the materials used, leading to a non-monotonic behavior. In contrast, in PMOS devices the longitudinal stress variation generated by SiGe dominates all other components, maintaining a monotonic performance curve and making the devices less sensitive to material properties used in DB process.

C. Gate Cut in PMOS devices

Fig. 10 shows the change in the PMOS drain current in linear region as a function of the distance from the gate cut. The plot

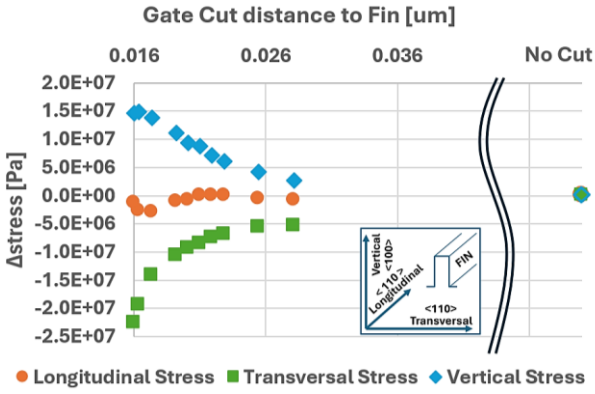

Fig. 15. Distribution of the stress variation (referenced to the case without gate cut) as a function of the distance to the gate cut for NMOS FinFET; three vector components of the stress are shown: longitudinal (orange circle), transversal (green square) and vertical (blue diamond)

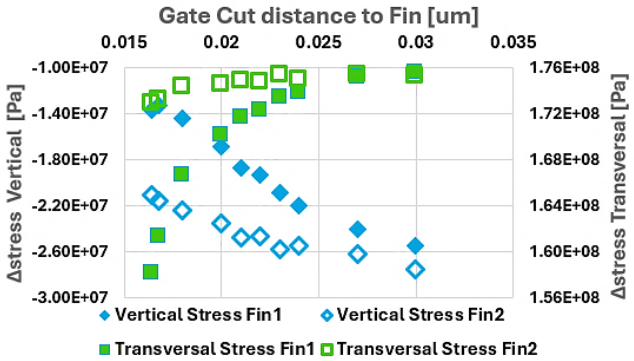

Fig. 16. Distribution of the vertical and transversal stress variations for the two fins of NMOS device as a function of the distance to the Gate Cut. Fin1 is the fin closer to GC

shows both the simulated electrical performance (open circles marked "Sim" and the electrical results measured on the silicon (closed circles marked "Data"). The electrical characteristics depicted in Fig.10 suggest the presence of two superimposing effects in this experiment. Fig.11 illustrates the simulated 3D stress component distributions within the device channel region. At shorter GC distances from the fin, the vertical stress enhancement on the mobility dominates [20], while at longer distances, the longitudinal stress effect prevails.

Fig. 12 presents the vertical stress in the nitride walls for devices at varying distances from the GC. GC far from the device (left), and very close to the device (right). The figure shows a simplified view of the epitaxial S/D grown next to the gate spacer. The stress developed along their interface contributes to the modulation of the mobility in the channel.

Fig. 13 illustrates how the gate cut and nitride walls contribute to the stress balance, affecting hole mobility and, consequently, PMOS transistor performance.

However, the behavior of the longitudinal stress is different than vertical stress. The nitride spacer layer around the GC region creates a fulcrum which absorbs part of the compressive longitudinal stress generated by the Source/Drain SiGe. The reduction of compressive longitudinal stress in the channel results in a reduction in hole mobility. However, this effect is less pronounced compared to the impact of vertical stress in the case of fin architecture with (110) channel orientation, as reported in Table I.

At a larger distance from GC, the longitudinal compressive stress increases, while the vertical component of the stress becomes constant. The new stress state leads to an overall increment in hole mobility as the GC distance increases. Transversal stress appears to have a negligible effect on hole mobility Table I.

D. Gate Cut in NMOS devices

Unlike PMOS, the NMOS transistors exhibit a monotonic dependence of the device performance on the distance to GC (Fig. 14). This is a consequence of the strong transversal and vertical stresses that impact channel mobility in a synergistic way. The impact of GC on the longitudinal component is negligible.

The variation of the transversal and vertical stress is driven by the stress from the gate metal stack. The metal stack is fabricated through multiple steps, with each layer deposited at different temperatures ranging from 200°C to 450°C, followed by cooling. Each of these thermal cycles generates a distinct stress profile within the device. Ultimately, the device will experience vertical compressive stress along the channel surface and transverse tensile stress perpendicular to the fin sidewall. These stresses degrade electron mobility, thereby reducing NMOS performance.

Fig. 15 shows that both vertical and transversal stress components intensify as GC distance to the fin is reduced, degrading the performance [20], as illustrated in Fig. 14.

E. Fin shielding effect

Most of the FinFET transistors use at least two fins to provide the required performance and reduce variability of a single-fin channel. Obviously, the stress experienced by a fin will depend on the distance from this fin to the gate cut.

In this work we studied two-fin transistors. For Diffusion Break effects, the stress variations are identical for both fins, since the isolation impact is the same for both by construction. This gives the same stress and performance modulation regardless of the number of fins in the device.

However, in case of the single-sided gate cut, the structure is asymmetrical, and the two fins of the transistor have different sensitivities to the gate cut.

For NMOS devices, the primary stress components affecting their performance are vertical and transverse, both of which are mainly influenced by the morphology of the gate metal stack. The fin closest to the gate cut is more affected due to the significant reduction in the volume filled by the gate metal stack between the fin and the cut. Conversely, the second fin does not experience any significant changes induced by the gate cut.

Fig.16 illustrates the variation of vertical and transverse stress in NMOS devices for both fins separately. It can be observed that the fin farther from GC is minimally affected by stress variations.

PMOS devices, on the other hand, are sensitive to longitudinal and vertical stress components, which are also influenced by the interaction between epi-grow and the nitride wall.

The profiles of the vertical stress illustrated in Fig.13 (right) show the change between stress inside the outer fin for the case with GC compared to continuous gate. The effect is much stronger for the fin closer to GC.

Due to this shielding effect, the impact on the device performance decreases for the devices with larger fin number.

F. Effect comparison and discussion

Table II summarizes the variation range of performance in the LLE investigated in this work for NMOS and PMOS devices.

In all instances, NMOS devices demonstrate lower sensitivity to LLE compared to their counterpart PMOS devices. The SiGe

Source/Drain compressive stressors in PMOS FinFETs induce substantial stress, therefore even small modifications in the three-dimensional structure can lead to significant variations, impacting PMOS performance in more substantial way.

TABLE II
PERFORMANCE VARIABILITY CAUSED BY STRESS-LLEs

LLE	NMOS	PMOS
Single Diff Break	+2% to +4%	-12%
Double Diff Break	-2% to +4%	-8%
Gate Cut	-5% to +2%	-13%

The NMOS SDB shows the lowest sensitivity among all analyzed effects. Lower intrinsic stress, compared to PMOS, and competition of two opposing effects generated by two separate process steps makes the device relatively insensitive to SDB.

The findings provide valuable insights into optimizing device architecture for improved performance and reduced variability by carefully managing stress components through diffusion breaks and gate cuts. Although not all stress model predictions were precise and accurate, they correctly captured the qualitative dependencies of device performance, including the non-monotonic and device type specific behaviors.

The present work also provides a robust framework for studying the impact of changes in device geometry, process recipes, and material properties. These efforts aim to enhance the understanding of material behavior and optimize the design for improved performance and reliability.

IV. CONCLUSION

Strain engineering is a key component of modern transistor, critical to achieve the desired performance. This work studies stress dependence and modulation due to diffusion break and gate cut in FinFET technology. We performed device stress and electrical performance simulation using a 3D Sentaurus TCAD process model in application to a 7nm FinFET technology.

Very good agreement between simulations and measured silicon data has proved that the model can be used to predict stress distributions for both PMOS and NMOS transistors under a variety of isolation environments and predict an impact on electrical characteristics of FinFET devices.

The magnitude of performance variation introduced by the stress-related LLE's in PMOS are as high as 10-12% in case of SDB and 8% for DDB. The performance variation in NMOS devices is smaller (less than 5%) due to lower stress levels. We experimentally observed a non-monotonic response of NMOS transistor performance in the proximity of the Diffusion Break which was correctly captured by the model and attributed to competing stress changes of opposite signs originating from two separate process steps.

For the LLE related to the Gate Cut, the magnitude of the performance reduction is up to 13% in PMOS, and 5% in NMOS. PMOS devices show a complex interaction between vertical and longitudinal stresses, resulting in a non-monotonic performance behavior when GC is in close proximity to the device.

Through this work we showed the importance of proper characterization and capturing of LLE effects in device models. The findings provide valuable insights into optimizing device architecture for improved performance by carefully managing stress components through diffusion breaks and gate cuts. The present work also provides a robust framework for studying the impact of device geometry, process, and material choices. Furthermore, the models can be used for layout optimization and for product design improvement as a DFM tool for reduced variability and improved yields.

ACKNOWLEDGEMENTS

The authors would like to thank the CV Design, Test, and Characterization teams at PDF Solutions for their contribution and help to generate and collect the data used in this work.

PDF Solutions' affiliated authors also appreciate the managerial support of Michael Yu, Howard Read, and PK Mozumder.

REFERENCES

[1] S. Saxena *et al*., "Impact of layout at advanced technology nodes on the performance and variation of digital and analog figures of merit," *2013 IEEE International Electron Devices Meeting*, Washington, DC, USA, 2013, pp. 17.2.1-17.2.4, doi: 10.1109/IEDM.2013.6724646

[2] H. Xu *et al.*, "Impact Study of Layout-Dependent Effects Toward FinFET Combinational Standard Cell Optimization," *IEEE Trans. Circuits Syst. II*, vol. 70, no. 2, pp. 731–735, Feb. 2023, doi: 10.1109/TCSII.2022.3179382.

[3] M. G. Bardon *et al*., "Layout-induced stress effects in 14nm & 10nm FinFETs and their impact on performance," *2013 Symposium on VLSI Circuits*, Kyoto, Japan, 2013, pp. T114-T115.

[4] C. Lee *et al*., "Layout-induced stress effects on the performance and variation of FinFETs," in *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, Washington DC, USA: IEEE, Sep. 2015, pp. 369–372. doi: 10.1109/SISPAD.2015.7292336.

[5] F. Sato *et al*., "Process and local layout effect interaction on a high performance planar 20nm CMOS," *2013 Symposium on VLSI Circuits*, Kyoto, Japan, 2013, pp. T116-T117.

[6] P. Zhao *et al.*, "Influence of stress induced CT local layout effect (LLE) on 14nm FinFET," in *2017 Symposium on VLSI Technology*, Kyoto, Japan: IEEE, Jun. 2017, pp. T228–T229. doi: 10.23919/VLSIT.2017.7998182.

[7] P.-Y. Yang, "Effect of Gate-Line-End-Induced Stress and Its Impact on Device's Characteristics in FinFETs," *IEEE Electron Device Lett.*, vol. 37, no. 7, pp. 910–912, Jul. 2016, doi: 10.1109/LED.2016.2565901.

[8] W. T. Chiang, J. W. Pan, P. W. Liu, C. H. Tsai, C. T. Tsai, and G. H. Ma, "Strain Effects of Si and SiGe Channel on (100) and (110) Si Surfaces for Advanced CMOS Applications," in *2007 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, Hsinchu, Taiwan: IEEE, 2007, pp. 1–2. doi: 10.1109/VTSA.2007.378930.

[9] D. Esseni, F. Conzatti, M. De Michielis, N. Serra, P. Palestri, and L. Selmi, "Semi-classical transport modelling of CMOS transistors with arbitrary crystal orientations and strain engineering: (Review invited paper)," *J Comput Electron*, vol. 8, no. 3–4, pp. 209–224, Oct. 2009, doi: 10.1007/s10825-009-0284-0.

[10] C.-C. Lee, P.-C. Huang, and T.-P. Hsiang, "Interactive Lattice and Process-Stress Responses in the Sub-7 nm Germanium-Based Three-Dimensional Transistor Architecture of FinFET and Nanowire GAAFET," *IEEE Trans. Electron Devices*, vol. 69, no. 12, pp. 6552–6560, Dec. 2022, doi: 10.1109/TED.2022.3216227.

[11] K. Miyaguchi *et al.*, "Single and Double Diffusion Breaks in 14nm FinFET and Beyond," in *Extended Abstracts of the 2017 International Conference on*

Solid State Devices and Materials, Sendai International Center, Sendai Japan: The Japan Society of Applied Physics, Sep. 2017. doi: 10.7567/SSDM.2017.E-2-03.

[12] A. Greene *et al.*, "Gate-Cut-Last in RMG to Enable Gate Extension Scaling and Parasitic Capacitance Reduction," in *2019 Symposium on VLSI Technology*, Kyoto, Japan: IEEE, Jun. 2019, pp. T144–T145. doi: 10.23919/VLSIT.2019.8776493.

[13] C. Hess *et al.*, "Device Array Scribe Characterization Vehicle Test Chip for Ultra Fast Product Wafer Variability Monitoring," in *2007 IEEE International Conference on Microelectronic Test Structures*, Bunkyo-ku, Japan: IEEE, Mar. 2007, pp. 145–149. doi: 10.1109/ICMTS.2007.374472.

[14] R. Saad, M. Silberg, Y. Dafne, and Z. Lando, "Optimizing the Tungsten Deposition Process," *High Temperature Materials and Processes*, vol. 15, no. 3, pp. 217–222, Jul. 1996, doi: 10.1515/HTMP.1996.15.3.217.

[15] Y. J. Kim *et al.*, "The effects of process temperature on the work function modulation of ALD HfO2 MOS device with plasma enhanced ALD TiN metal gate using TDMAT precursor," *Microelectronic Engineering*, vol. 178, pp. 284–288, Jun. 2017, doi: 10.1016/j.mee.2017.05.023.

[16] J. Robertson and R. M. Wallace, "High-K materials and metal gates for CMOS applications," *Materials Science and Engineering: R: Reports*, vol. 88, pp. 1–41, Feb. 2015, doi: 10.1016/j.mser.2014.11.001.

[17] S. L. Fei, W. Q. Peng, Z. J. Hong, and C. Y. Shan, "Exploring Gate-Cut Patterning Approaches Using Simulation and Defect Modelling," in *2021 China Semiconductor Technology International Conference (CSTIC)*, Shanghai, China: IEEE, Mar. 2021, pp. 1–4. doi: 10.1109/CSTIC52283.2021.9461509.

[18] M. Wagner, M. Karner, and T. Grasser, "Quantum Correction Models for Modern Semiconductor Devices," International Workshop on the Physics of Semiconductor and Devices, New Delhi, 2005, pp. 458-461

[19] O. Penzin, G. Paasch, F. O. Heinz and L. Smith, "Extended Quantum Correction Model Applied to Six-Band k·p Valence Bands Near Silicon/Oxide Interfaces," in *IEEE Transactions on Electron Devices*, vol. 58, no. 6, pp. 1614-1619, June 2011, doi: 10.1109/TED.2011.2122264.

[20] S. Yang *et al.*, "10nm high performance mobile SoC design and technology co-developed for performance, power, and area scaling," in *2017 Symposium on VLSI Technology*, Kyoto, Japan: IEEE, Jun. 2017, pp. T70–T71. doi: 10.23919/VLSIT.2017.7998203.