Novel Ultrafast Annealing Processes for Performance Improvement in Advanced Nano Scaled Partially Depleted SOI-MOSFETs

R. Illgen[a], S. Flachowsky[a], T. Herrmann[a], T. Feudel[b], J. Höntschel[b], M. Horstmann[b], W. Klix[a], and R. Stenzel[a]

[a] Department of Electrical Engineering,
University of Applied Sciences Dresden
Friedrich-List-Platz 1, D-01069 Dresden, Germany
ralf.illgen@et.htw-dresden.de

[b] AMD Saxony LLC & Co. KG,
Wilschdorfer Landstrasse 101, D 01109 Dresden, Germany

Abstract

With the need to reduce vertical and lateral device dimensions, ultra fast annealing technologies either with or without prior conventional rapid thermal annealing has recently attracted attention. There are many advantages to this technology including high electrical activation, excellent lateral abruptness, controlled and limited dopant diffusion and the ability to engineer the extended defects remaining from the ion implantation. But there are also many challenges such as local and global wafer stress and difficulty in process integration.
This paper will provide an overview of currently used ultra fast annealing technologies and there compatibility with new materials such as embedded Silicon-Germanium with and without prior conventional rapid thermal annealing in advanced nano scaled partially depleted SOI-MOSFETs.

1. Introduction

In a standard process with a conventional rapid thermal annealing (RTA) stress techniques like embedded SiGe (eSiGe) and dual stress overlayers are standard features for advanced CMOS technologies to improve device performance [1]. Unfortunately, such an annealing scheme does not meet the 32 nm node requirements due to thermal diffusion and solid solubility limitations. To solve the problem, high temperature less-diffusive ultra fast annealing (UFA) technologies such as flash lamp annealing (FLA) [2] and non-melt laser spike annealing (LSA) [3] have been intensively investigated.

These new annealing technologies offer improved dopant activation for source-drain and gate polysilicon regions over conventional RTA techniques. This results in reduced source-drain resistance and polysilicon depletion for a process that has the advantage of almost no additional diffusion when added to a RTA process. For further scaling of the device dimensions the RTA temperature could be reduced or even could be replaced which lead to a diffusionless transistor architecture whose dimensions are only defined by the as-implanted profiles.

However, there are concerns that temperatures of 1300 °C and above cause reliability issues even if applied for milliseconds. This might become a critical impediment for further gate oxide scaling or might even require thicker gate oxides, in contradiction to device scaling theory and practice. Other problems are the possible relaxation of strained cap layers or eSiGe layers and dopant deactivation for cap layers deposited at low temperatures after UFA.

2. Ultra fast annealing technologies

In current generation RTA systems, junction annealing subjects the entire wafer to temperatures between 1000 °C - 1100 °C for seconds. This temperature range is too low to achieve the desired dopant activation due to solid solubility limits. In addition, considerable dopant diffusion takes place in this timeframe. The latest RTA tools attain about 2 s time exposures near the peak temperature of 1100 °C. This combination of time and temperature still does not meet the simultaneous requirements of improved activation and reduced thermal budget of the 32 nm node.

In a conventional RTA process the surface heat-up rates are slower than heat conduction through the wafer. Therefore, the wafer surface and bulk are in thermal equilibrium and the heat loss during ramp down occurs by heat conduction to the chuck and by radiation. The special features of the novel annealing schemes are the surface heating and the substrate selfcooling. Surface heating means that the surface heat-up rate is faster than heat conduction because the time constant of UFA technologies (about 1 ms) is much shorter than the thermal time constant of the wafer (10 ms - 20 ms). As a result the wafer bulk stays cool while only the surface heats up. The fast cooling is achieved since the bulk of the wafer acts as a heat sink removing heat from the top layer much more efficiently and faster than can be accomplished with the active cooling on a RTA system. Hence, very short annealing times in the millisecond range can be achieved. Fig. 1 shows a comparison of temperature-time profiles of a conventional RTA process and UFA technologies.

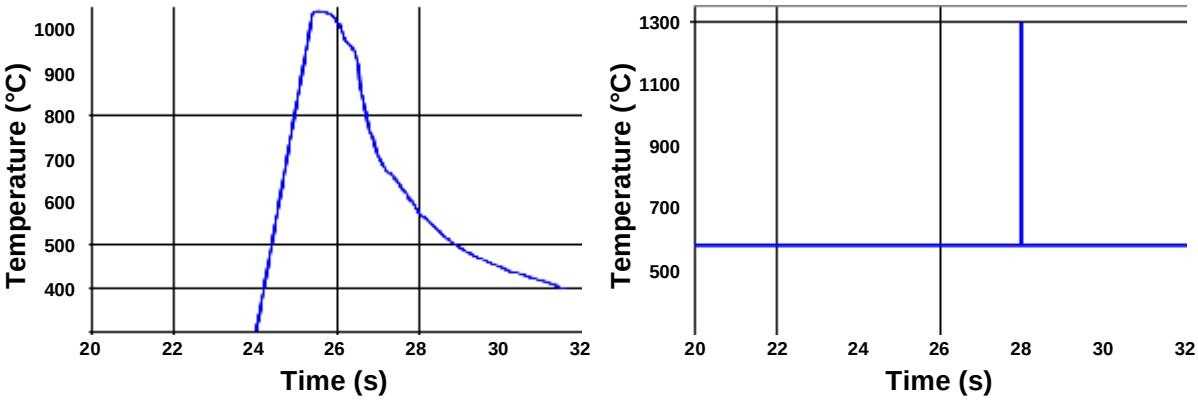

Fig. 1: Temperature-time profile of a conventional RTA (a) and UFA (b) process.

Flash lamp annealing uses a pre-heat step to heat the entire wafer surface up to 300 °C - 600 °C prior to the "flash" of a bank of xenon arc flash lamps (Fig. 2). The substrate preheating suppresses the thermal stress and it becomes possible to heat up the surface of a wafer up to more than 1000 °C without wafer breakage. The flash duration ranges from several hundred microseconds to tens of milliseconds, and heats the wafer surface up to near 1250 °C - 1300 °C. The spectra of xenon arc flash lamp cover the wavelength from UV to visible light. For silicon, interband absorption occurs for wavelengths shorter than 1.1 µm, and for that reason this is the dominant radiation absorption mechanism for flash lamp based technologies (Fig. 3).

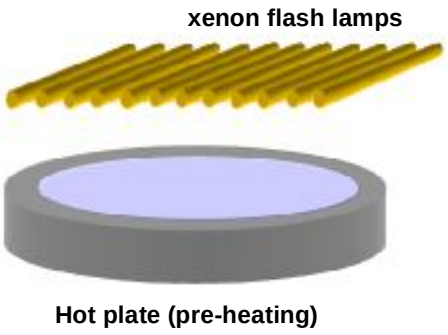

Fig. 2: Configuration of a xenon flash lamp annealing system [2].

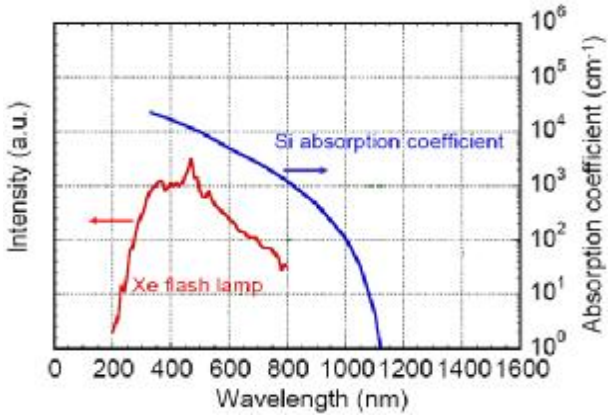

Fig. 3: Emission spectra of xenon flash lamp with optical absorption spectrum of Si [2].

A typical temperature-time profile of FLA shows Fig. 4(a). This new annealing technique can reduce the annealing time to the millisecond range as shown in Fig. 4(b).

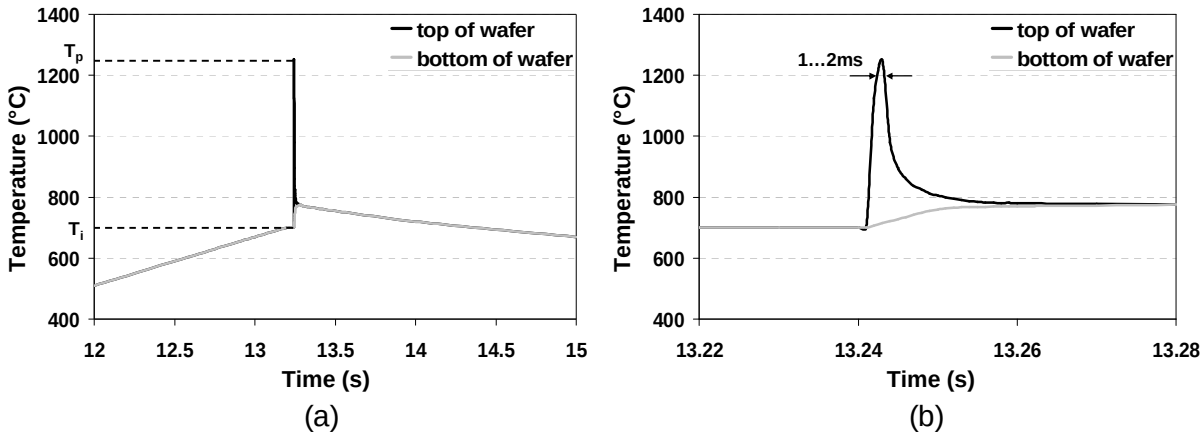

Fig. 4: Temperature-time profile of FLA.

Laser spike annealing uses a long-wave CO_2 laser radiation source at a wavelength of 10.6 µm. LSA equipment utilizes a p-polarized CO_2 laser beam incident at Brewster's angle to minimize emissivity deviation [4]. The wafer is sitting on a heated chuck mounted on a X-Y stage that is moved to scan the whole wafer under the laser beam spot (Fig. 5). The dwell time is simple defined as the time for beam spot to pass through a point on the wafer.

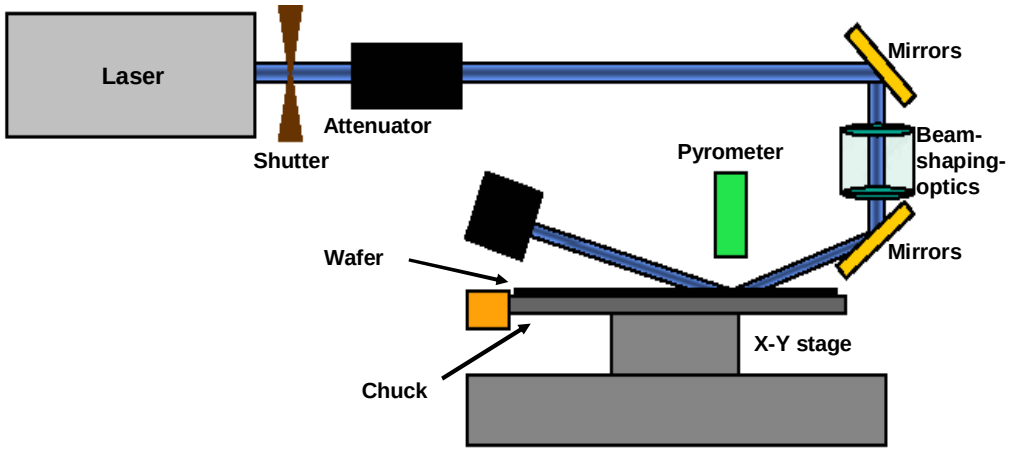

Fig. 5: Configuration of a laser spike annealing system [3].

The substrate of the entire wafer is pre-heated up to 300 °C - 600 °C before laser irradiation and only a small localized hot spot of the wafer surface will be heated up to 1350 °C by laser exposure, just below the silicon melting point (Fig. 6). For the mid-far IR light source that LSA uses, the principal absorption mechanism is free carrier absorption. In this case the density of free carriers in a semiconductor material can be easily changed by chuck heating.

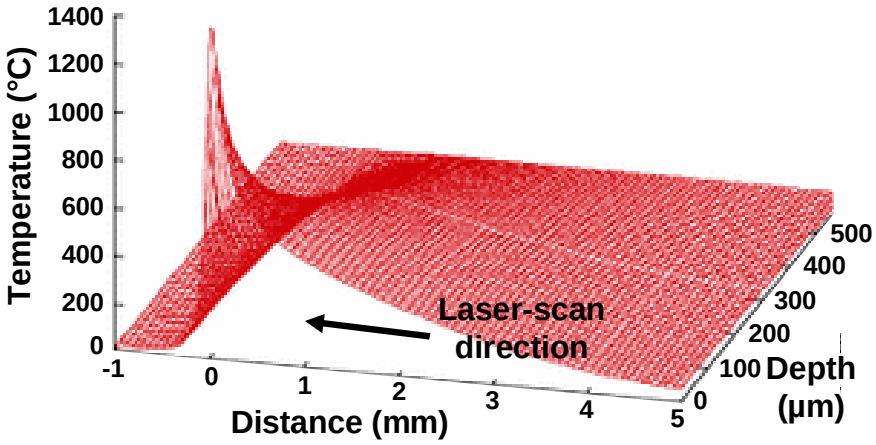

Fig. 6: Simulated temperature profile in a silicon substrate generated by a 100 µm-wide beam travelling at 500 mm/sec [3].

In such a temperature-time regime that UFA technologies use, dopant diffusion is negligible. Moreover, junctions with higher activation levels are achieved due to the higher solubility of dopants closer to the silicon melting point and the more rapid cooling rates. These approaches result therefore in an improvement over RTA in terms of junction depth and sheet resistance. Furthermore, processes that normally degrade device performance such as transient enhanced diffusion or boron penetration through the gate oxide are also minimized. Fig. 7 illustrates SIMS profiles for different annealing schemes. Using UFA no significant diffusion of the implantation profile can be observed in the SIMS profile whereas the dopant distribution after RTA treatment is much broader.

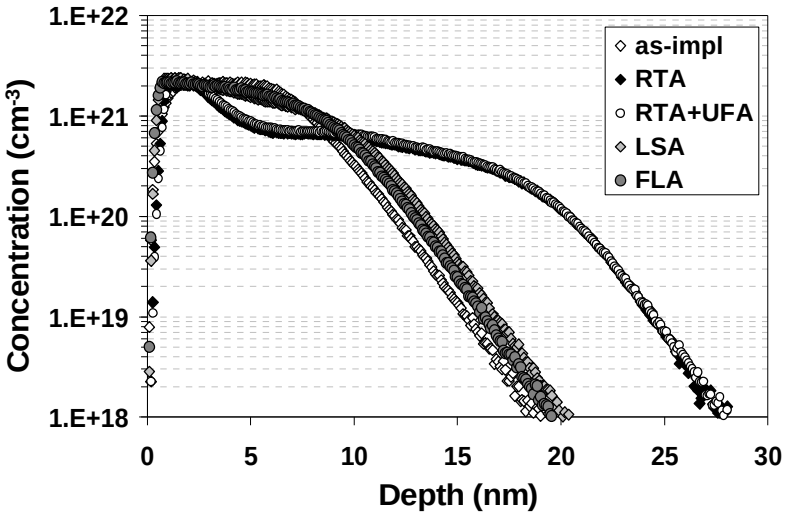

Fig. 7: SIMS profiles of arsenic implanted samples for different annealing schemes.

3. Experimental results

Firstly, the impact of the new annealing techniques on the suppression of gate depletion will be discussed. A 0.1 nm reduction in gate oxide thickness measured under inversion conditions was achieved due to reduced poly depletion (Fig. 8). An equivalent RTA step would require rather high peak temperatures or increased peak time. This would cause excessive diffusion of the source-drain extension implants. On the other hand, reducing the physical gate oxide thickness by 0.1 nm would result in an intolerable 4 times increase in gate leakage. The UFA step, however, caused only a 30 % increase in gate leakage due to increased electric field strength (Fig. 9).

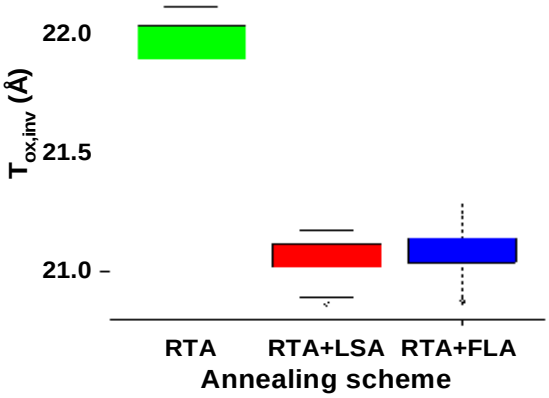

Fig. 8: Gate oxide thickness measured under inversion conditions ($T_{ox,inv}$) for different annealing schemes [5].

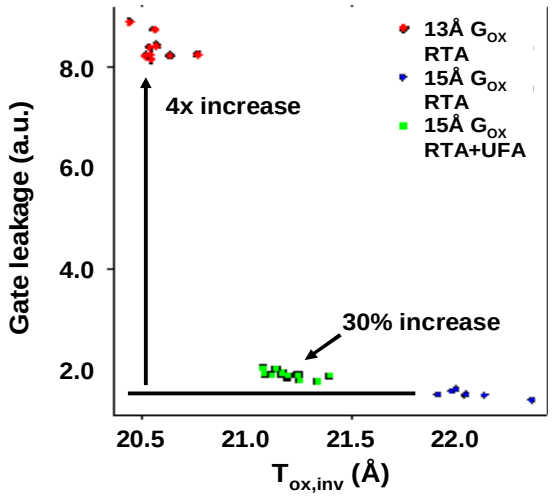

Fig. 9: Normalized gate leakage as a function of the electrically measured gate oxide thickness [5].

A critical problem with UFA techniques is the applied power density. Above a certain level a strong increase in gate leakage can be observed. For even higher power, the gates are physically damaged. We found the critical power density level in the range of an equivalent temperature of 1350 °C. Below that level long-term reliability issues were not observed for gate oxide thicknesses in the range of 0.9 nm to 1.5 nm. Thus, UFA techniques do not limit further gate oxide scaling.

Fig. 10 shows the impact of the additional UFA step on the sheet resistance of the source-drain extension regions dependent on RTA temperature for n- and p-MOSFET. Up to 30% decreased sheet resistance can be realized on n-MOSFET with an additional UFA step dependent on RTA temperature. On the p-MOSFET side, however, an improvement in sheet resistance can only be recognized at low RTA temperatures by the additional UFA.

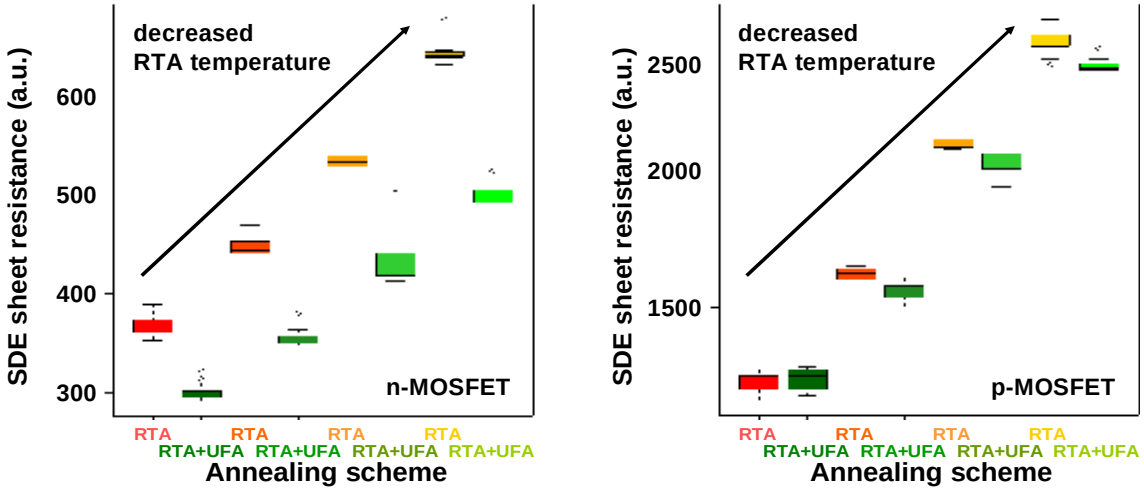

Fig. 10: Sheet resistance of the source-drain extension regions for different annealing schemes and RTA temperatures (left: n-MOSFET, right: p-MOSFET) [5].

Fig. 11 shows the $I_{D,sat}$-$I_{D,off}$ characteristics of n- and p-MOSFET devices annealed with and without UFA. A performance improvement can be achieved with UFA for both types of transistors. As a result from the former findings the performance enhancement on the n-MOSFET side is driven by higher carrier activation with non-diffusive features to reduce the parasitic source-drain resistance and reduced poly depletion, while the performance improvement on the p-MOSFET is primarily driven by the observed reduced poly depletion.

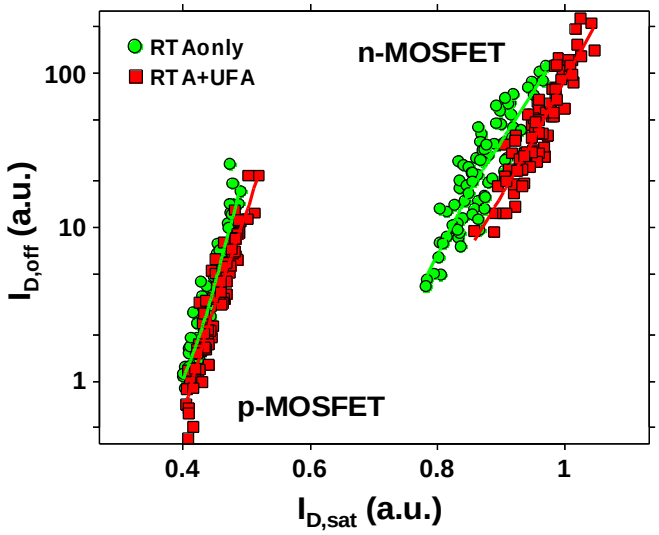

Fig. 11: $I_{D,sat}$-$I_{D,off}$ characteristics of n- and p-MOSFET devices annealed with and without UFA.

Apart from the performance gain, there are several questions about compatibility with other process steps. Low-temperature layer deposition steps are always applied after final activation during contact formation and backend processing, typically having temperatures between 350 °C and 500 °C for a few minutes. There are speculations that these temperature treatments can cause dopant deactivation.

But as shown in Fig. 12 there are no performance differences between various types of overlayers deposited at different temperatures on top of the transistor. The results are shown for the p-MOSFET. For the n-MOSFET the $I_{D,sat}$-$I_{D,off}$ characteristics behave likewise.

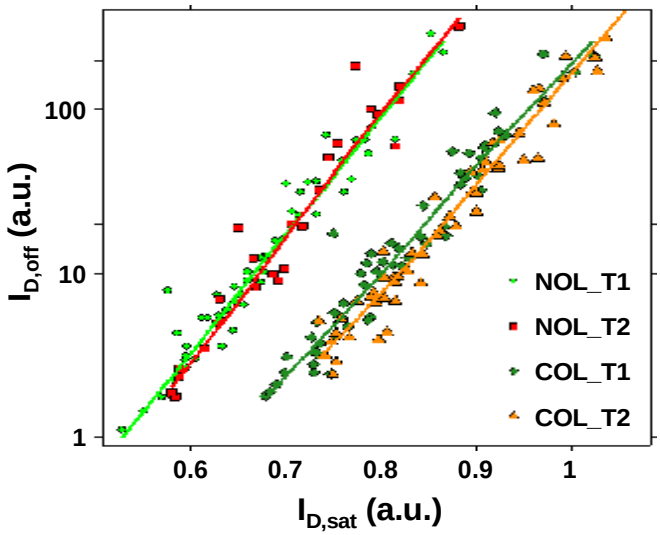

Fig. 12: $I_{D,sat}$-$I_{D,off}$ characteristics of p-MOSFET devices with neutral (N) and compressive strained (C) overlayers (OL) deposited at different temperatures (T) on top of the transistor (T1 < T2) [5].

A very efficient method to improve the performance of the p-MOSFET device is the implementation of eSiGe into the source-drain areas. Depending on the depth of the cavity, the fill high and the proximity of the eSiGe to the channel region, a high performance gain can be achieved by enhancing the hole mobility with compressive strain. But this compressive strain might be relaxed during subsequent high-temperature processing.

The new annealing techniques must therefore also be compatible with the eSiGe stress technique to prevent dislocation formation and strain relaxation. Another potential problem is the lower melting point of Si-Ge dependent on the germanium content when compared to silicon. For that reason the equivalent UFA temperature had to be reduced by 25 K in order to avoid damaging the active device areas.

As shown in Fig. 13 there are no strain relaxation effects on p-MOSFET device performance visible. On the contrary, despite the reduced UFA temperature we obtained the same performance improvement by adding the UFA step to the process flow either with or without eSiGe. As a result the new UFA technologies could be successfully implemented in a standard RTA process since the 65 nm device technology to further improve state-of-the-art CMOS transistors.

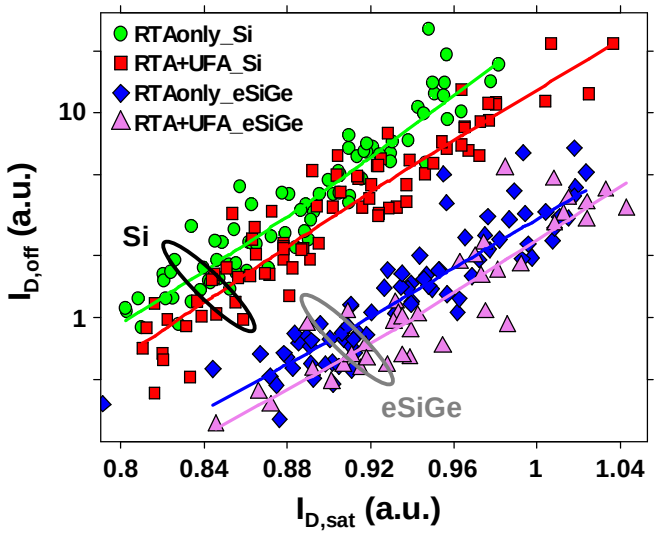

Fig. 13: $I_{D,sat}$-$I_{D,off}$ characteristics of p-MOSFET devices with and without eSiGe annealed with and without UFA.

In order to boost transistor performance, 32 nm CMOSFET devices continue to use mobility enhancement techniques such as tensile strained overlayers and stress memorization techniques for the n-MOSFET and compressive strained overlayers and eSiGe for the p-MOSFET. However, by reducing the poly pitch due to the current slow-down in device scaling the benefit from these stressors will also be reduced.

An example for this problem shows Fig. 14. Here the p-MOSFET drain-current enhancement from the compressive strained overlayer compared to neutral overlayer in dependency of the poly pitch is represented. With the switch from the 45 nm to the 32 nm technology node which comes along with a poly pitch reduction from 190 nm to 130 nm, the loss of the performance improvement only from this mobility enhancement technique for the p-MOSFET is almost 50 %.

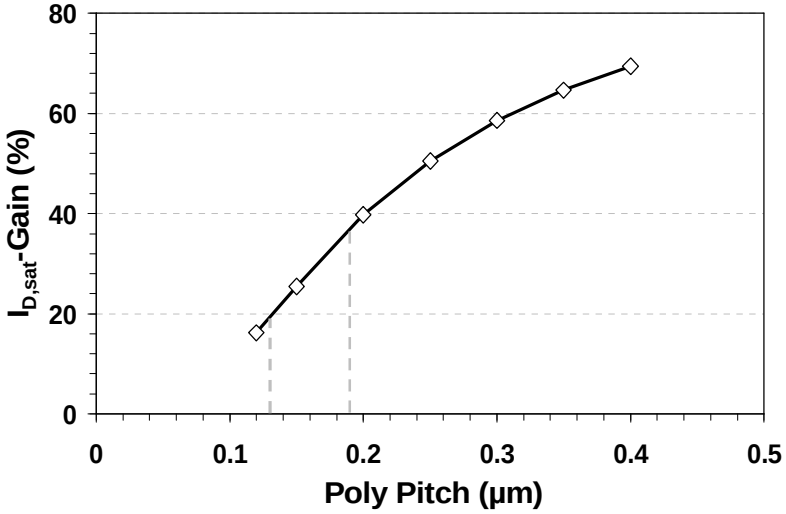

Fig. 14: Drain-current enhancement of compressive strained overlayers compared to neutral overlayers as a function of the poly pitch (obtained from TCAD simulations).

Increasing the stress value of the strained overlayers or increasing the germanium concentration in the eSiGe is one of the solutions to overcome the problem of reduced performance improvement with poly pitch reduction. But excessive stress value causes film crack or peeling and higher germanium concentration leads to crystal defect generation that loses not only the stress value but also the device functionality.

It is important to optimize the device structure to add the sufficient amount of the stress to the channel region. One possibility is a further reduction of device dimensions, especially the spacer width, to improve the impact from the strained overlayers. For that reason the RTA temperature should be reduced or even must be replaced with the new annealing techniques. These specifications lead to a diffusionless transistor architecture (UFAonly) whose dimensions are only defined by the as-implanted profiles.

One challenge of the integration of UFAonly is the complete redesign of the spacer layout as well as the junction implant parameters due to the diffusionless annealing. Fig. 15 shows the $I_{D,sat}$-$I_{D,off}$ characteristics of UFAonly n-MOSFET devices for different spacer widths. After optimization of the implant parameter for the halo, source-drain extension and deep source-drain areas the performance of the UFAonly n-MOSFET could be successive enhanced due to spacer width reduction. As a result the UFAonly device outperforms the standard transistor with RTA and UFA.

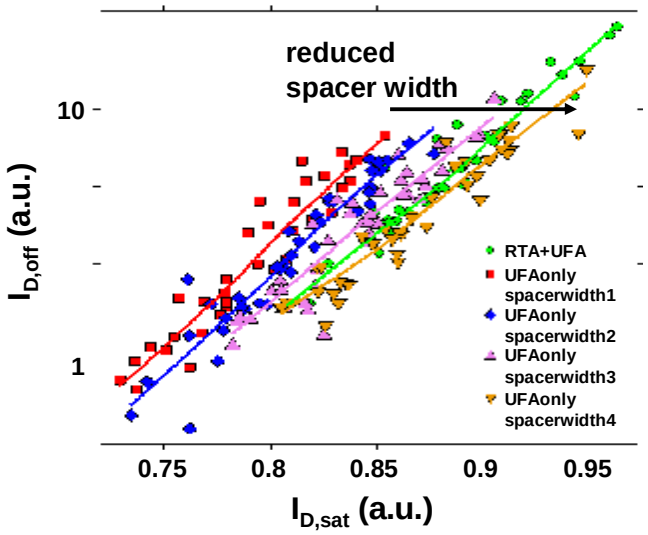

Fig. 15: $I_{D,sat}$-$I_{D,off}$ characteristics of UFAonly n-MOSFET devices for different spacer widths (n-MOSFET annealed with RTA+UFA also included as reference).

These observations are also confirmed by simulations using SYNOPSYS Sentaurus TCAD software [6]. Fig. 16 shows the simulated n-MOSFET drain-current enhancement and the horizontal average stress component as a function of the spacer width of devices with tensile strained overlayers.

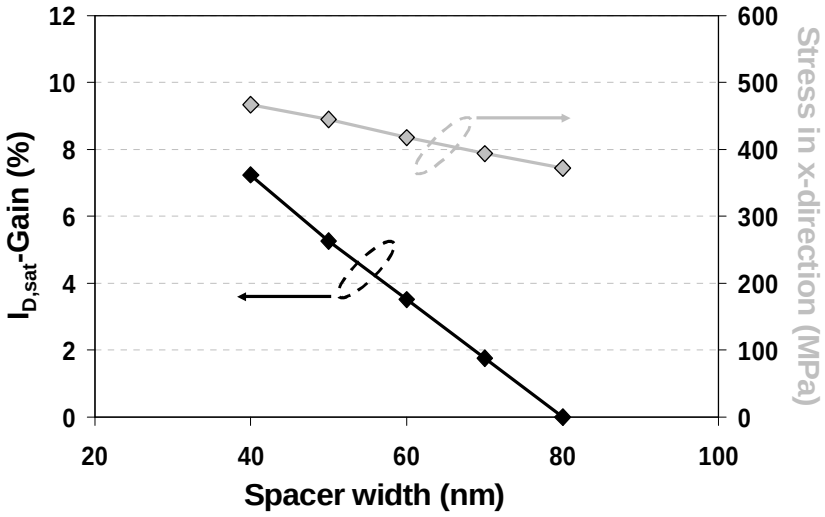

Fig. 16: Drain-current enhancement at constant $I_{D,off}$ and average stress (in channel direction 2 nm below the surface) as a function of the spacer width for n-MOSFET devices with tensile strained overlayers (obtained from TCAD simulations).

The horizontal stress component in the channel increases in absolute values with decreasing spacer width which leads to the drain-current benefit for smaller spacer widths. The reason of this performance improvement is the increased carrier mobility because the stress present in the channel at the Si/SiO_2-interface is predominantly responsible for the mobility enhancement.

A likewise drain-current benefit is also visible for n-MOSFET devices with neutral overlayers as shown in Fig. 17. Although the stress in the channel changes only slightly, the drain-current benefit increases for smaller spacer widths (but not as much as for tensile strained overlayers). The reason for this behaviour is the decrease of the parasitic source-drain resistance.

The n-MOSFET performance improvement with smaller spacer widths consists therefore of:

- a decreased channel resistance, caused by a higher stress and to the associated better mobility,
- and a decreased resistance of the parasitic source-drain extension areas.

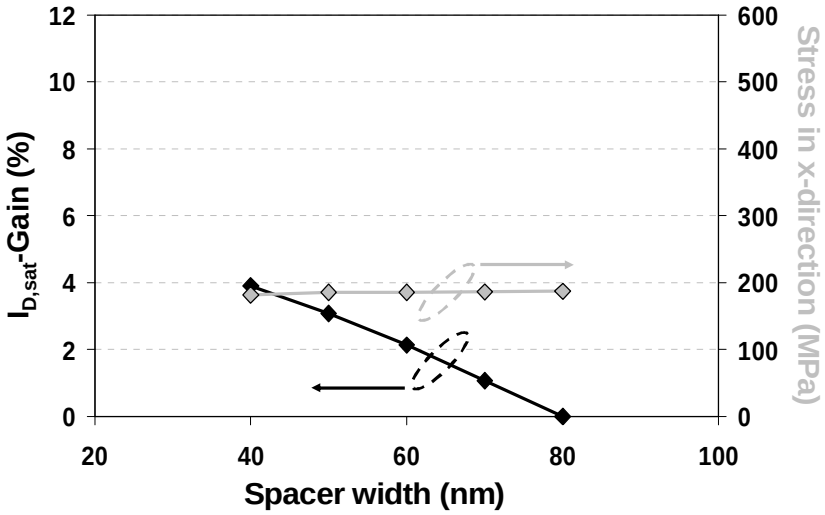

Fig. 17: Drain-current enhancement and average stress (in channel direction 2 nm below the surface) as a function of the spacer width for n-MOSFET devices with neutral overlayers (obtained from TCAD simulations).

Another challenge is the implementation of the eSiGe in the diffusionless transistor architecture. As shown in Fig. 18 the eSiGe enhances the p-MOSFET device performance dependent on eSiGe proximity to the channel in comparison to the unstrained case similar to the RTA+UFA process.

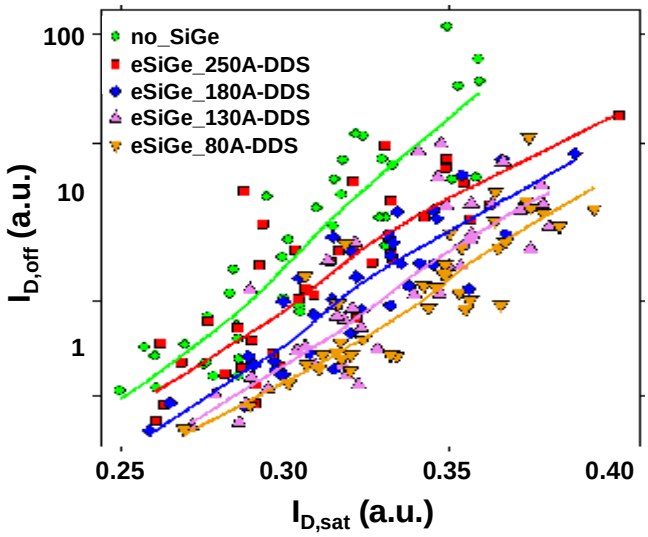

Fig. 18: $I_{D,sat}$-$I_{D,off}$ characteristics of UFAonly p-MOSFET devices with and without eSiGe and different eSiGe proximity.

Furthermore, as shown in Fig. 19 the same eSiGe to channel proximity trend for RTA+UFA and UFAonly can be observed, which is in excellent agreement with simulations. Both annealing strategies show only slightly different characteristics.

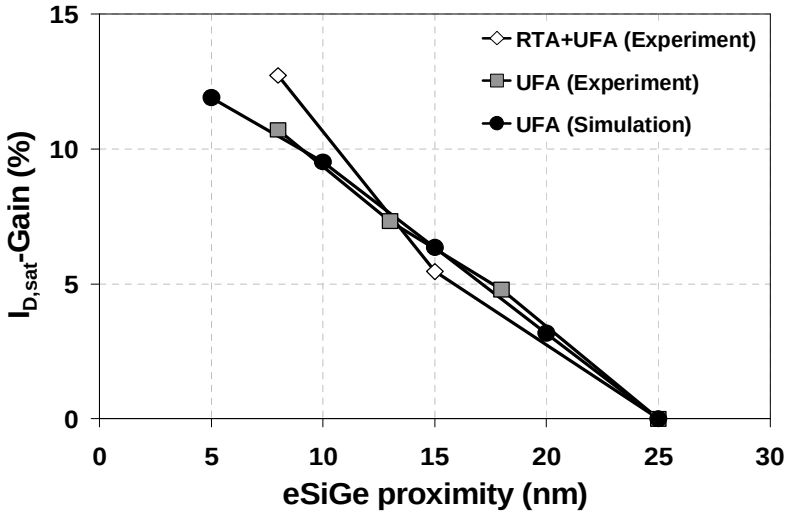

Fig. 19: Drain-current enhancement as a function of the eSiGe proximity for p-MOSFET devices for different annealing schemes. Also shown is a comparison to simulation data.

4. Conclusions

In this paper an overview of currently used ultra fast annealing technologies such as flash lamp annealing and non-melt laser spike annealing has been presented. It was shown that these new annealing strategies have been successfully implemented in a standard RTA process to further improve state-of-the-art CMOS transistors. A transistor performance improvement could be achieved due to reduced poly depletion and reduced source-drain resistance. The advanced annealing scheme does not cause gate oxide reliability issues provided that a certain power density limit is not exceeded. A full compatibility with embedded silicon-germanium source-drain stressor and no deactivation with subsequent low temperature backend processing steps for the ultra fast annealing technologies in addition to a standard RTA were demonstrated.

Furthermore, due to the current slow-down in device scaling, the conventional RTA must be replaced with the new annealing techniques and their non-diffusive features for a further reduction of device dimensions. Due to a smart junction engineering the diffusionless transistor architecture outperforms the standard transistor with RTA and ultra fast annealing. The embedded silicon-germanium source-drain stressor also enhances the p-MOSFET device performance similar to the RTA and ultra fast annealing process.

These results show the potential advantage of ultra-high temperature and non diffusive annealing that will be necessary for the next generation technology nodes.

5. Acknowledgement

This project was funded by the German Federal Ministry of Education and Research, registered under funding number 01M3167B. The author named in the publication bears responsibility for all published contents.

6. References

[1] M. Horstmann, A. Wei, T. Kammler, J. Höntschel, H. Bierstedt, et al.: *Integration and optimization of embedded-SiGe, compressive and tensile stressed liner films, and stress memorization in advanced SOI CMOS technologies*, IEEE International Electron Devices Meeting Technical Digest, pp. 233-236, December 2005

[2] T. Ito, K. Suguro, M. Tamura, T. Taniguchi, Y. Ushiku, et al.: *14nm-depth Low Resistance Boron Doped Extension by Optimized Flash Lamp Annealing*, IEEE International Symposium on Semiconductor Manufacturing, p. 19, 2002

[3] S. Talwar, D. Markle, and M. Thompson: *Junction scaling using lasers for thermal annealing*, Solid State Technology, Vol. 46, No. 7, p. 83, July 2003

[4] L.M. Feng, Y. Wang and D.A. Markle: *Minimizing pattern dependency in millisecond annealing*, Ext. Abstract the 6th International Workshop on Junction Technology, p. 25, 2006

[5] Th. Feudel, M. Horstmann, L. Herrmann, M. Herden, M. Gerhardt, et al.: *Process Integration Issues with Spike, Flash and Laser Anneal Implementation for 90 and 65 nm Technologies*, 14th IEEE International Conference on Advanced Thermal Processing of Semiconductors, 2006

[6] Sentaurus Process & Device User's Manual, Release Z-2007.12, Synopsys Inc., 2007