Advanced Annealing Strategies for the 32 nm node

C. Kampen[1], A. Martinez-Limia[1], P. Pichler[1,2], A. Burenkov[1], J. Lorenz[1], and H. Ryssel[2,1]
1: Fraunhofer Institute of Integrated Systems and Device Technology (IISB), Schottkystrasse 10, 91058 Erlangen, Germany
Phone: +49 9131 761 224, Fax: +49 9131 761 212, Email:christian.kampen@iisb.fraunhofer.de
2: Chair of Electron Devices, University of Erlangen-Nuremberg, 91058 Erlangen, Germany

***Abstract*—In this work, the influences of advanced annealing schemes, spike and flash annealing and combinations of them, on the electrical behavior of modern FD SOI MOSFETs have been investigated by numerical simulations. Process simulations have been performed for comparing the two-dimensional diffusion behavior of the dopants under the different annealing schemes. Device simulations have been performed for making conclusions about how the different annealing schemes are influencing the static and dynamic behavior of modern CMOS devices.**

I. Introduction

Modern CMOS devices require very high active dopant concentrations to improve their dynamic behavior via high drive currents and small access resistances [1], [2]. On the other hand, short channel effects and fringing capacitances should be reduced by controlling the lateral diffusion below the gate stack. To comply with both requirements, advanced annealing techniques are required. Optimization of devices at the cutting edge of technology by coupled numerical process and device simulation critically depends on the ability of the TCAD models available to predict implantation, diffusion, and activation quantitatively.

In this work, four different annealing methods have been investigated to assess their advantages concerning the static and dynamic transistor behavior, short channel effect (SCE) and contact resistances (R_{sd}). Improved diffusion simulation models have been used [3], [4], that we have recently implemented in SentaurusProcess [5].

The models have been calibrated for a wide range of annealing processes, from low temperatures (700 °C for boron and 650 °C for arsenic) via isothermal soak annealing series (750 °C–1000 °C), spike annealing (SpA) with peak temperatures in the range of 950 °C–1050 °C, and flash annealing (FlA) with peak temperatures from 1200 °C to 1300 °C. Experimental results of advanced annealing schemes (multiple flash and combinations of flash and spike annealing) have been included in the calibrations as well.

II. Process Flow

Ultra thin body fully depleted silicon on insulator (UTB FD SOI) MOSFETs have been simulated in this work. The physical gate length has been set to 21 nm and the effective gate oxide thickness (EOT) to 1 nm. The silicon body thickness t_{body} has been set to 7 nm, while the buried oxide (BOX) thickness has been chosen to be 20 nm. To reduce the drain induced barrier lowering (DIBL), a heavy doping of the ground plane of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ has been used [6], while the channel region has been kept completely un-doped for carrier mobility improvement.

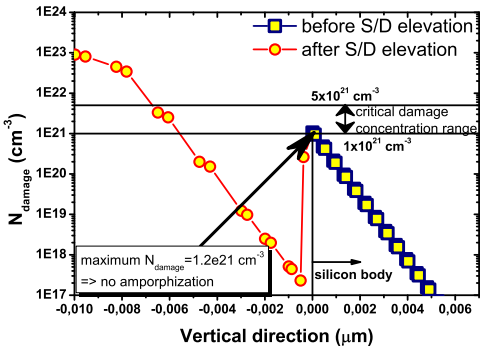

Fig. 1. Simulated damage concentration after the first implantation step (squares) and the second (circles). Amorphization is suppressed before the source/drain elevation as the damage concentration does not exceed $0.1 \times 5 \times 10^{22}\,\mathrm{cm}^{-3}$

A. Implantation

Due to the un-doped channel, a well defined placement of the extensions is needed to suppress short channel effects. On the other hand, for high drive currents, a high active doping concentration in the extensions is required to minimize the sheet resistance. After the poly-silicon gate has been structured, a liner oxide layer of 3 nm has been deposited. Then, the extensions have been implanted. As the source/drain regions should be elevated by selective epitaxial growth, low non-amorphizing doses have been used for creating the extensions. For the nMOS, a dose of $1 \times 10^{14}\,\mathrm{cm}^{-2}$ of arsenic and an implantation energy of 1.3 keV have been used. After the first implantation, a damage concentration of $1.2 \times 10^{21}\,\mathrm{cm}^{-3}$ and below has been observed (Fig. 1) which does not lead to amorphization at room temperature. Therefore, the elevation of the source/drain regions by selective epitaxial growth (SEG) should be possible. For the pMOS, a dose of $1 \times 10^{14}\,\mathrm{cm}^{-2}$ of boron and an implantation energy of 0.2 keV has been used to create the extensions. After the extensions have been implanted, the nitride spacers have been structured and the source/drain regions have been elevated by 10 nm. After the source/drain elevation, a second implantation step has been performed to create the active source/drain regions. Here, an implantation energy of 1.5 keV and a dose of $1 \times 10^{15}\,\mathrm{cm}^{-2}$ of arsenic have been used for the nMOS, and 0.5 keV and a boron dose of $1 \times 10^{15}\,\mathrm{cm}^{-2}$ for the pMOS. The second implantation

978-1-4244-1753-7/08/$25.00 ©2008 IEEE

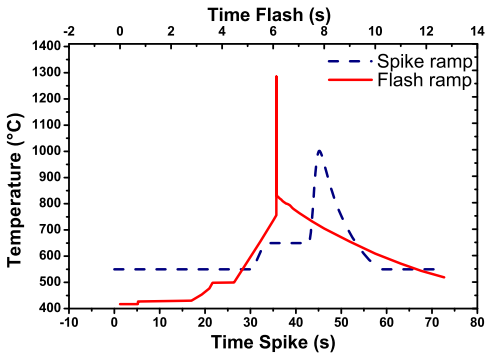

Fig. 2. Temperature profiles of the spike and flash annealing schemes

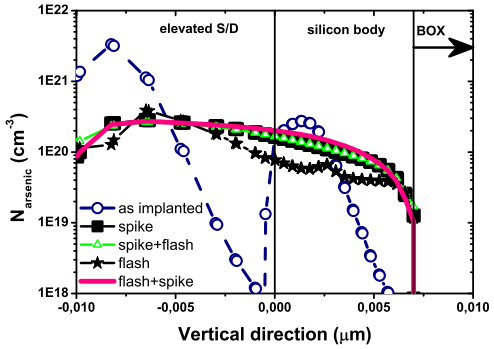

Fig. 3. Vertical active arsenic concentrations after different RTA methods

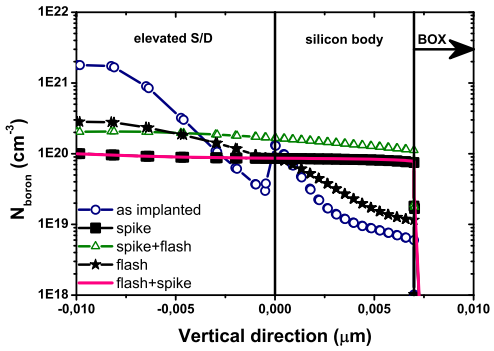

Fig. 4. Vertical active boron concentration after different RTA methods

step aims at reaching a high surface doping concentration after the annealing to reduce source/drain contact resistances.

III. Annealing

After the complete MOSFET structure has been created and implanted, four alternative rapid thermal annealing (RTA) methods have been applied: spike annealing (SpA), flash annealing (FlA), SpA followed by FlA, and FlA followed by SpA. The temperature profiles for the SpA and FlA that have been used in this work are displayed in Fig. 2. The spike and flash annealing have peak temperatures of 1000 °C and 1270 °C, respectively.

Looking at the vertical doping profiles of nMOS (Fig. 3) and pMOS (Fig.4) leads to the conclusion that the FlA results in a higher surface concentration than the SpA. An active surface concentration of $1.1 \times 10^{20}\,\mathrm{cm}^{-3}$ for arsenic and $2.8 \times 10^{20}\,\mathrm{cm}^{-3}$ for boron have been reached by the flash annealing, while the spike annealing results only in an active surface concentration of $8.7 \times 10^{19}\,\mathrm{cm}^{-3}$ for arsenic and $1.0 \times 10^{20}\,\mathrm{cm}^{-3}$ for boron. A dent in the flash annealed doping profile can be observed for arsenic (Fig.3 (stars)) below the maximum of the as-implanted (Fig.3 (circles)) profile. This is caused by precipitation of arsenic after high dose implantation. The short flash temperature peak fails to dissolve completely the arsenic precipitates.

The spike annealing, on the other hand, causes more redistribution of the dopants, and, therefore, a lower active surface concentration, as well for arsenic Fig.3 (squares), as in the case of boron Fig.4 (squares). Comparing the spike annealed and flash annealed profiles to each other leads to the idea that a combination of spike followed by flash annealing would lead to a favorable behavior of the final MOSFETs. In Fig. 3 and 4 it can be seen that the SpA followed by FlA generates a high active surface concentration, which reduces the contact resistances, and a smooth shape of the doping profile, which decreases the parasitic sheet resistances.

Another important issue in CMOS device modeling is the placement of the extensions. Especially, if un-doped channels are used, the diffusion below the gate stack can only be well controlled by using RTA annealing schemes to prevent short channel effects. Furthermore, high active extension concentrations are required to reduce the sheet resistances at the channel interface.

Fig. 5 and 6 display the lateral active doping concentrations 1 nm below the gate-oxide after the four different annealing schemes. For the nMOS (Fig. 5), the flash annealed profile, as well as the FlA followed by SpA, seem to be the most promissing annealing schemes. As the active doping concentration nearly follows the as-implanted profile, short channel effects should be suppressed by using the FlA and the SpA-FlA combination, The spike annealed profile and the SpA followed by FlA profile, on the other hand, are distributed completely below the gate-stack, with a concentration of above $1 \times 10^{18}\,\mathrm{cm}^{-3}$. This might result in a stronger short channel effect, compared to the FlA and the SpA-FlA combination.

In case of the pMOS (Fig. 6), the SpA followed by FlA seems to be the best choice, as the active concentration below the spacers is high and the decay of the active concentration is very smooth. This behavior should result in low sheet resistances and therefore in high conductivity.

To investigate how the different RTA methods influence the electrical transistor performance, numerical device simulations have to be done. Thereby, qualitative conclusions on the static and dynamic behavior of the MOSFETs can be drawn.

IV. Device Simulation

The device simulations have been done by using the Synopsys software SentaurusDevice [5]. For the drift-diffusion simulation, several standard models have been used [5]. Quantum-

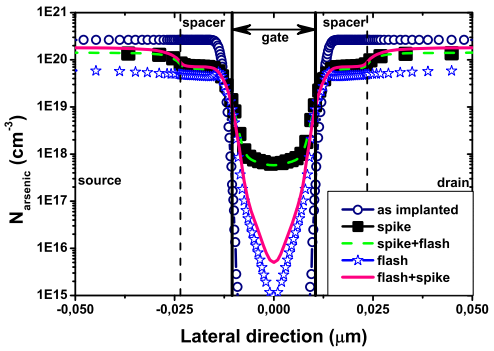

Fig. 5. Lateral active arsenic concentration after different RTA methods: 1 nm below the gate-oxide

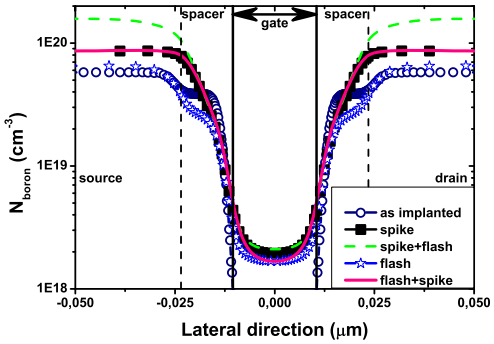

Fig. 6. Lateral active boron concentration after different RTA methods: 1 nm below the gate-oxide

mechanical depletion at the silicon-gate-oxide-interface has been taken into account by using the modified local density approximation (MLDA) [7], and quasi-ballistic carrier transport in un-doped channels has been simulated by using the approximation of Bude [8]. Parasitic contact resistances have been taken into account by the doping dependent resistance model [5]. To reduce the contact resistances, which depend on the contact area [9], the contact lengths have been chosen to be four times the physical gate-length.

A gate voltage V_{gate} of 1.0 V has been applied to a midgap workfunction gate electrode, while a drain voltage V_{drain} of 0.05 V has been applied for the low-field and $V_{drain} = 1.0$ V for the high-field case. The transfer characteristics have been calculated by a DC analysis, while the C_{gate}-V_{gate} characteristic of the gate capacitance has been been achieved by an AC small signal analysis.

V. Discussion

Fig. 7 displays the transconductance g_m characteristics of the pMOSFET after the four RTA methods. By using the spike followed by the FlA we observe the highest values for g_m as the annealing scheme induces a high active surface doping concentration and a smooth decay of the concentration into the depth. Furthermore, due to the high active extension concentration that results from the SpA followed by FlA ((dashed line) Fig. 6), the sheet resistance below the gate-stack is lowered. On the other hand, using only spike annealing leads to a very low conductivity, due to the lower surface concentration. Comparable results have been achieved for the nMOSFET.

After the transconductances have been calculated from the transfer characteristics, the source/drain contact resistances have been extracted from the low-field mobility by using the y-method [10]. For this purpose, the transfer behavior has been simulated again without using contact resistances and the low-field mobility μ_0 has been calculated. The source/drain contact resistances can then be calculated by using equation (1).

$$R_{sd} = \frac{\Theta_1 - \Theta_0}{C_{ox}\mu_0 \frac{W}{L}} \quad (1)$$

Here, Θ_0 is the mobility degradation factor that had been calculated from the low-field case without using contact resistances. Θ_1 specifies the mobility degradation factor for the case of including contact resistances, C_{ox} denotes the gate-oxide capacitance, μ_0 the low-field mobility, W the width of the transistor (1 μm), and L the physical gate-length. Fig. 8 displays the calculated contact resistances in dependence on the gate-voltage. For the nMOSFET as well as for the pMOSFET, the FlA-SpA combination induces the highest contact resistivity of nearly 1600 Ω-μm, which is 800 Ω-μm per contact. While using SpA followed by FlA or FlA alone leads to a great reduction of the source/drain contact resistivity, for nMOS as well as for pMOS.

The great influence of the contact resistances on the dynamic behavior of the MOSFETs can be observed in Fig. 9. Here, the $C_{gate}V_{DD}/I_{on}$ of the nMOSFET at fixed leakage current values is displayed for the four annealing schemes. As expected, using the SpA and the FlA followed by SpA leads to larger switching delays compared to the FlA and the SpA-FlA combination. Here again, the parasitic source/drain contact resistances are the most affecting parameter that slows down the switching speed, as the drive currents are reduced. To demonstrate how the source/drain contact resistances influence the dynamic behavior of the MOSFET, the switching speed characteristics for the spike annealing and for SpA followed by FlA have been plotted in addition without taking contact resistances into account. Both of the characteristics without contact resistances nearly fulfill the requirements of the international technology roadmap for semiconductors (ITRS) defined for the 32 nm node. However, by taking contact resistances into account, none of the simulation results would fulfill the requirements. For the spike annealing, a performance loss of nearly 65 % has been caused by the contact resistances, while it is only 30 % for the spike-flash combination. Overall, using the spike-flash combination annealing leads to the fastest switching speed in case of the nMOSFET.

Finally, the short channel behavior following the four RTA methods has been investigated (Fig. 10). Although a heavy ground plane doping [6] and a relatively thin body thickness have been used, a strong short channel effect is observed for each annealing scheme. Due to the fact that the flash annealing leads to a slightly higher threshold voltage than the other annealing schemes, the SCE is slightly lower than for the other

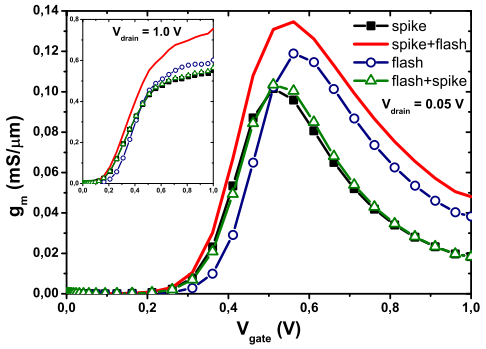

Fig. 7. Transconductance of the pMOSFET for different RTA schemes: low-field V_d=50mV, high-field V_d=1.0V

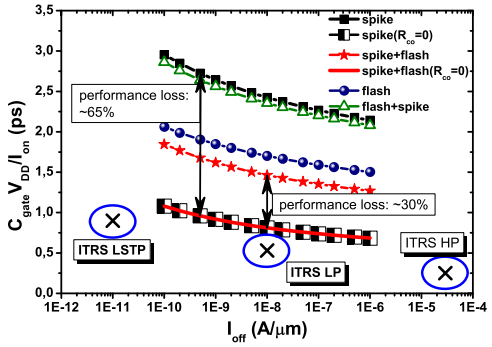

Fig. 9. Propagation delay $C_{gate}V_{DD}/I_{on}$ of the nMOSFET for fixed off-current values

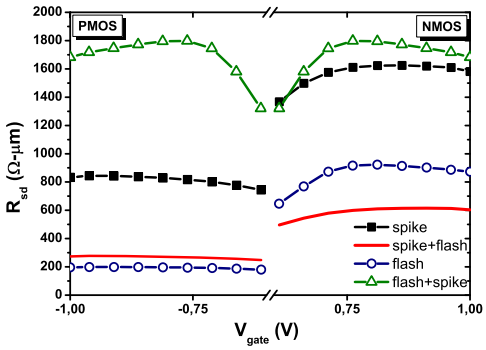

Fig. 8. Source/drain contact resistances; calculated by the y-method from the low-field mobility

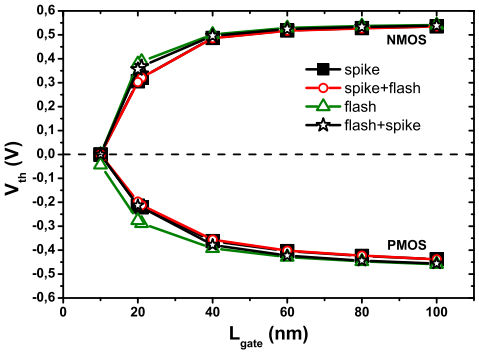

Fig. 10. Short channel behavior of the nMOSFET and the pMOSFET for different RTA schemes

three schemes. But, as it can be seen in Fig. 10, for each RTA method that has been used in this work, the threshold voltage is already zero at a gate-length of nearly 10 nm.

VI. Conclusion

The influences of advanced thermal annealing schemes, such as spike and flash annealing, as well as combinations of them, on the electrical behavior of ultra thin body fully depleted silicon on insulator have been investigated. In these simulations, improved activation and diffusion models for arsenic and boron have been used. It could be demonstrated that reducing the parasitic source/drain contact resistances by high active surface doping concentrations is a key issue for the next generation CMOS devices. Furthermore, high active doping concentrations and smooth concentration decays of the lateral diffusion profile are important, as they raise the conductivity. Thus, a spike plus flash annealing scheme has been found to be the most promissing candidate, as high active surface doping concentrations and highly doped extensions could be achieved. Finally, no significant differences in the short channel behavior have been found between the four advanced annealing schemes.

Acknowledgment

This work was funded in part by the European Union in the framework of the IST projects 027152 ATOMICS and 026828 PULLNANO.

References

[1] A. Burenkov, C. Kampen, E. Bär, J. Lorenz, and H. Ryssel, "Application-driven simulation of nanoscaled transistors and circuits," *Journal of Computational and Theoretical Nanoscience*, no. 6, pp. 1170–1182, 2008.

[2] C. Kampen, A. Burenkov, J. Lorenz, and H. Ryssel, "Alternative source/drain contact-pad architectures for contact resistance improvement in decanano-scaled CMOS devices," in *ULIS Conference*, 2008, pp. 179–182.

[3] J. Schermer, A. Martinez-Limia, P. Pichler, C. Zechner, W. Lerch, and S. Paul, "On a computationally efficient approach to boron-intersticial clustering," 2008, accepted to be published.

[4] A. Martinez-Limia, P. Pichler, C. Steen, S. Paul, and W. Lerch, "Modelling and simulation of advanced annealing processes," *Materials Science Forum*, vol. 279, pp. 573–574, 2008.

[5] *Sentaurus TCAD*, Release Z-2007.03 ed., Synopsys, Mountain View, CA, USA, 2007.

[6] T. Ernst and S. Cristoloveanu, "Buried oxide fringing capacitances: a new physical model and its implication on SOI device scaling and architecture," in *SOI Conference*, 1999, pp. 38–39.

[7] G. Paasch and H. Übensee, "A modified local density approximation: Electron density in inversion layers," *Physica Status Solidi (b)*, vol. 113, no. 1, pp. 165–178, 1982.

[8] J. Bude, "MOSFET modeling into the ballistic regime," in *SISPAD*, Seattle, WA, September 2000, pp. 23–26.

[9] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998.

[10] G. Ghibaudo, "New method for extraction of MOSFET parameters," *Electronic Letters, IEEE*, vol. 24, no. 9, pp. 543–545, April 1988.