Abstract

In this paper, we investigate on the suitability of a spike anneal to reduce Boron diffusion while maintaining a high level of dopant activation on both poly and active area. We demonstrate that, playing on the spike anneal temperature, as well as ramp-up and ramp-down conditions, an optimum for this trade-off can be found. In particular, we demonstrate that a spike anneal at 1113°C, performed with moderate thermal ramps (75°C/s) gives similar activation but reduced Boron diffusion than our reference rapid thermal anneal (RTA). Based on this result, a 0.13 µm device architecture (gate length at 0.11 µm) for a multi-Vt platform and optimized for both digital and analogue operation is presented.

1. Introduction

The formation of ultra shallow junctions is one of the most challenging tasks for the coming CMOS generations. This is particularly true at pMOS side, because of the low atomic mass and the transient enhanced diffusion (TED) of Boron. One of the keys to reduce the Boron enhanced diffusion is to play on the activation step (rapid thermal anneal or RTA). Many recent publications reported on the reduction of Boron diffusion by using a short-time (spike) anneal [1-4]. However, it was also reported that the spike anneal may lead to a degradation in gate depletion and series resistance due to dopant deactivation [5]. In this paper, we demonstrate that the spike anneal temperature, as well as the thermal ramp-up and ramp-down conditions, governs the trade-off between Boron diffusion and dopant activation. In particular, we demonstrate that the peak temperature plays the prevalent role in this trade-off, while aggressive thermal ramps are not necessarily required to reduce Boron diffusion. This gives to our solution a high degree of manufacturability. An optimized 0.13 µm device architecture (with gate length at 0.11 µm), based on our optimization of the spike anneal conditions is presented. It is shown that the combination of a this optimized spike anneal with an ultra low energy (ULE) S/D extension implant allows a significant pocket dose lightening, resulting in a flat Vt(L) characteristic (no more Vt roll-up) and improved analogue operation.

2. Device fabrication

Once the trench isolation module is done, the wells are implanted and a 2 nm gate oxide is grown. The polysilicon is next deposited and predoped. After gate etching, S/D extensions are implanted with As at 2 keV at nMOS side and BF2 at 4 keV at pMOS side. A double pocket architecture is used in order to compensate for the short channel effect down to 0.1 µm: Indium + Boron at nMOS side, Arsenic + Phosphorus at pMOS side. The deep S/D implant is made of As at 30 keV at nMOS side and Boron at 3 keV at pMOS side. The splits are then performed: a reference RTA at 1025°C 15s is compared with various spike anneal conditions. Three conditions for peak temperature (1075°C, 1113°C, 1150°C) are crossed with two conditions for ramp-up and ramp-down (75°C/s vs 100°C/s), see Fig. 1.

3. Impact on pMOSFETs

The impact from the spike anneal conditions on Boron diffusion is illustrated by pMOS Vt(L) plots, see Fig. 2. It is clearly evidenced that the prevalent player is the peak temperature rather than the ramp-up or the ramp-down slope. Performing a spike anneal at 1150°C does not improve short channel effect immunity, compared to the reference RTA at 1025°C for 15s. In that case, the (too) high peak temperature drives the Boron diffusion. On the contrary, when the spike anneal is performed at 1075°C, the short channel effect is totally suppressed. Thereby, the spike anneal fully fulfills its
objective while the Boron diffusion is frozen. The spike anneal at 1113°C is in between. The benefit from the spike anneal concept is then partially taken. The short channel effect at pMOS side is also dependent on the ramps conditions. Ramping at 100°C/s reduces Boron enhanced diffusion. However, the gain is relatively thin when comparing with what is obtained when playing with the peak temperature. This means that “aggressive” ramps are not necessarily required to take most of the advantage of the spike anneal concept, especially regarding “cold” spike anneals.

Fig. 2: Vt(L) at pMOS side.

The square resistance for poly gate, deep S/D and S/D extensions are plotted as a function of the peak temperature, see Fig. 3. The trade-off between Boron diffusion and Boron activation is then clearly settled. Indeed, it is shown that the spike at 1075°C, which freezes Boron diffusion on one hand, gives poor Boron activation on the other hand, leading to high gate depletion and series resistances. On the contrary, the spike at 1150°C gives superior activation than the reference RTA. Thereby, it is very interesting to focus on the spike at 1113°C. For this intermediate temperature, the activation figures at pMOS side are in line with what is reached by the RTA at 1025°C for 15s. In other words, the spike at 1113°C appears as the best trade-off between Boron diffusion and activation.

Fig. 3: square resistance vs spike temperature

Tuning the peak temperature also introduces the well-known trade-off between P+ gate depletion, and Boron penetration into the gate oxide, see Figs. 4 and 5. However, it is noticeable that the spike at 1113°C gives less gate depletion than the reference RTA together with less Boron penetration. The spike at 1113°C is then not only superior to the reference RTA in terms of Boron diffusion freeze-out but also in terms of P+ gate activation vs Boron penetration.

Fig. 4: CV plot for pMOS, focus on depletion.

Fig. 5: CV plot for pMOS, focus on B penetration

4. Impact on nMOSFETs

No significant impact from the spike anneal conditions is found on Arsenic diffusion. This is clearly evidenced when looking at nMOS Vt(L) plots, see Fig. 6. The peak temperature does not drive the Arsenic diffusion. However, as for Boron, Arsenic activation is strongly dependent on the spike temperature, see Figs. 7 and 8. The spike at 1075°C gives a poor Arsenic activation, leading to high gate depletion and series resistance, while the spike at 1150°C gives superior activation than the reference RTA. The spike at 1113°C gives similar Arsenic activation than the reference RTA, for N+ poly gate, N+ S/D extensions and N+ deep S/D.
Fig. 6: V(t) at nMOS side

Fig. 7: square resistance vs spike temperature

Fig. 8: CV plot for nMOS, focus on depletion.

The influence of the peak temperature on Arsenic activation is also clearly visible when looking at Ion vs Ioff plots, see Fig. 9. Stronger gate depletion and series resistance are responsible for the Ion vs Ioff trade-off degradation with a spike anneal at 1075°C, while reducing gate depletion and series resistance significantly improve performance for a spike anneal at 1150°C. Both spike anneal at 1113°C and RTA at 1025°C 15s give similar performance, because of a similar Arsenic activation in both cases.

Fig. 9: Ion vs Ioff trade-off at nMOS side

Our observations demonstrate that the trade-off between diffusion and activation is not solved by the same spike anneal condition for Boron and Arsenic. The optimum is found to be at 1113°C for Boron, and at 1150°C for Arsenic. This is because Arsenic diffusion is not impacted by the spike anneal temperature, while its activation is enhanced for the hottest condition (no trade-off between diffusion and activation at nMOS side). For both case, it is shown that “slow” ramps (75°C/s) are sufficient to take most of the advantages from the spike anneal concept. This gives to this solution a high degree of manufacturability: the on-wafer uniformity is even slightly better with the spike anneal than with the reference RTA, see Fig. 10.

Fig. 10: Wafer map for electrical tox at –1.2V, pMOS (relevant of the temperature variation between the center and the edges of the wafer)

5. Device optimization

Integrating a spike anneal at 1113°C, we developed a 0.13 µm device architecture (gate length at 0.11 µm), for a multi-Vt platform, designed for both digital and analogue operation. This requires a specific work on pocket architecture, since digital and analogue operations need opposite pocket optimization [6-7]. This is made possible thanks to the use of a spike anneal,
combined with ULE S/D extension implant. Indeed, the better control of short channel effect at pMOS side, due to Boron diffusion freeze-out, may be traded for a lighter pocket dose, resulting in an higher voltage gain for long channel device. At nMOS side, the use of Boron (instead on Indium) to design a steep lateral pocket profile can also be reconsidered since the spike anneal is expected to freeze pocket lateral diffusion. This is illustrated by the \( V_t(L) \) and gain(L) plots drawn below, for pMOSFETs designed with various pocket condition. The control pMOSFET is activated with the reference RTA at 1025°C for 15s. In that case, because of the enhanced Boron diffusion, heavily doped pockets (made of Arsenic, implanted at \( 4.2 \times 10^{13} \) at.cm\(^{-3} \)) are mandatory to control the short channel effect on one side, and to target the right \( V_t \) value at \( L_g=0.11 \) µm on the other side. Using a spike anneal at 1113°C reduces Boron diffusion and allows a significant pocket dose reduction. In that case, the Arsenic dose is decreased down to \( 1.9 \times 10^{13} \) at.cm\(^{-3} \), while short channel effect are even better controlled; and while the target \( V_t \) for the nominal device is still reached, see Fig. 11. This approach is greatly helpful in terms of analogue operation, see Fig. 12. Because of the pocket dose lightning, the long channel voltage gain significantly increases, while the short channel gain is improved through short channel effect reduction.

6. Conclusions

In this paper, we demonstrated that, playing on the spike anneal temperature, an optimum for the trade-off between Boron diffusion and dopant activation can be found. We also demonstrate that the peak temperature does not influence Arsenic diffusion. Regarding CMOS integration, a spike anneal at 1113°C appears as the best condition, since the Boron diffusion is significantly reduced compared to the reference RTA, and while the dopant activation is similar at both sides. Moreover, we demonstrate that “aggressive” ramps are not necessarily required to take most of the advantage of the spike anneal concept. This gives to our solution a high degree of manufacturability. Based on this process step optimization, we proposed a device architecture for a 0.13 µm CMOS platform (gate length at 0.11 µm). The better control on short channel effect due to Boron diffusion freeze-out was traded for lowly doped pocket. By so doing, the analogue operation of the device was significantly improved, without altering digital performances.

7. References


