Extended Abstracts of the 2020 International Conference on Solid State Devices and Materials, VIRTUAL conference, 2020, pp59-60

Millisecond Post Deposition Annealing for Improving the EOT and D_{it} in TiN/HfO_2/SiO_2/Si Gate Stacks using Flash Lamp Annealing

H. Kawarazaki[1], A. Ueda[1], S. Kato[1], K. Izumi[2] and Y. Nara[2]
[1] SCREEN Semiconductor Solutions Co., Ltd., 480-1 Takamiya-cho, Hikone, Shiga 522-0292 Japan
[2] University of Hyogo, 2167, Shosha, Himeji, Hyogo 671-2280 Japan
Phone: +81-749-24-8460 ext. 8306112, E-mail: kawarazaki @screen.co.jp

Abstract

We demonstrate improvements in the EOT and D_{it} of TiN/HfO_2/SiO_2/Si gate stacks using post deposition flash lamp annealing. We show that densification of the dielectrics leads to EOT scaling. Also, we show that the assist heating temperature is a key parameter for improving D_{it}. These results suggest this method has the potential to improve gate stacks while maintaining a low thermal budget.

1. Introduction

Thermal budget management is a crucial challenge for gate stacks. Unlike traditional thermal oxide processes (typically ~1200ºC), current mainstream high-k last integration limits thermal process (typically ~900ºC) and BTI reliability is concern[1]. Furthermore, the latest roadmap reports the introduction of SiGe/Ge channels and 3D stacking for the future sub 5nm node [2], which implies there will be more severe limitations on the thermal budget. Therefore, with the development of devices and materials, low thermal budget processes for gate stacks are becoming increasingly necessary. We have proposed flash lamp annealing (FLA) for high-k post deposition annealing (PDA) with annealing times of the millisecond order [3,4,5]. To achieve lower thermal budgets, we have evaluated the assist temperature (Ta) dependency of industrially used HfO_2/SiO_2/Si gate stacks, and we discuss their electrical properties based on the physical analysis.

2. Experimental

Fig. 1 shows the process flow and conditions used in this work. After pre-cleaning, HfO_2 layers were deposited on the wafers by ALD. Then PDA was applied at various Ta and peak temperatures (Tp), with or without flash using a LA-3100 (SCREEN Semiconductor Solutions Co., Ltd). All PDAs were conducted in a N_2 or NH_3 ambient. The FLA pulse was fixed to 1.4ms. After gate patterning, C-V and I-V measurements were conducted. The measured C-V curves were modeled by NCSU's CVC program [6]. The interface state density (D_{it}) was calculated by the conductance method at room temperature.

3. Results and Discussion

First, we discuss the impact of PDA using FLA on the capacitance of the gate stack. As shown in the CV curves (Fig. 2), the capacitance increases after FLA. Fig. 3 shows XTEM images before and after FLA which indicates the capacitance increase is due to thinning of the gate stack. Note that both the HfO_2 thickness and SiO_2 thickness decrease and that the HfO_2 layer seems to be crystalized after FLA. The XRD spectrum, shown in Fig. 4 also shows crystallization peaks and reveals that the non-monoclinic phase is dominant in the FLA samples while the monoclinic phase is dominant in the RTA samples which may be the reason why the FLA samples have higher capacitance than the RTA samples. This unique result for FLA can be attributed to the high cooling rate, which prevents transition from the non-monoclinic phase to the monoclinic phase [7]. As shown in Fig. 5, the oxide peak in the XPS Si2p spectrum seems to have shifted to a slightly higher binding energy after PDA. It can be inferred that the sub-oxide decreases after PDA. As seen in the ATR-FTIR spectrum of SiO_2/Si-sub structure (Fig. 6) the longitudinal-optical (LO) phonon peak due to Si-O-Si asymmetric stretching has shifted towards a higher wavenumber after FLA. This implies a shortening of the mean Si-O bond length indicating densification of the SiO_2 layer. Therefore, densification of both HfO_2 and SiO_2 increases the capacitance leading to EOT scaling by FLA.

Secondly, we discuss the impact on D_{it}. Fig. 7 shows the Ta dependence of D_{it} with FLA. Interestingly, the use of flash improves D_{it} with Ta under 500ºC, while flash with Ta at 600ºC degrades D_{it}. To investigate this Ta dependence, ESR analysis was conducted (Fig.8). The P_{b0} centers decrease with low Ta. It is speculated that Ta under 500ºC prevents H desorption from the SiO_2/Si interface due to the low thermal budget while Ta above 600ºC causes H desorption after flash. Fig. 7 also shows that NH_3 further improves D_{it} compared with N_2. This NH_3 passivation mechanism is not clear and further investigation is needed. D_{it} seems to be correlated with the gate leakage current (Fig. 9). This suggests trap sites near the interface assist gate leakage. Fig. 10 shows that FLA improves both the EOT and the gate leakage current especially for low Ta. Fig. 11 shows V_{FB} as a function of Ta with FLA. V_{FB} is also an important parameter which determines the work function of the gate. There is no significant shift in V_{FB} after FLA processes with low Ta. Fig. 12 shows low Ta improves the stability of D_{it} with which improvement in the BTI can be expected.

4. Conclusions

PDA using FLA was proposed as a low thermal budget technique for gate stacks and was applied to the widely used HfO_2/SiO_2/Si gate stacks. We demonstrate a reduction in the EOT by FLA which is attributed to densification of the dielectrics. Furthermore, FLA improves D_{it}, especially with low Ta, which may be due to the decrease in P_{b0} centers. FLA with low Ta also decreases the gate leakage current while maintaining V_{FB} and improving the D_{it} stability. In conclusion PDA using FLA is a promising low thermal budget process for gate stacks.

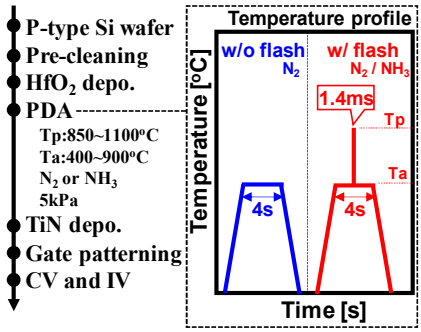

Fig. 1 Process flow and schematic image of temperature profile of PDA.

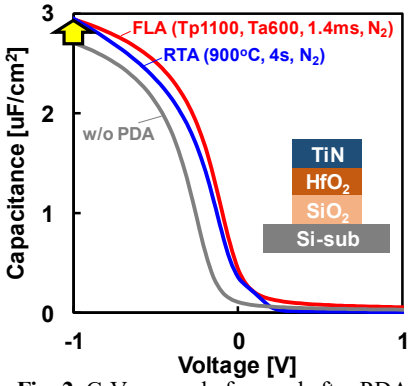

Fig. 2 C-V curves before and after PDA. PDA increases the capacitance.

References: [1] J. Franco et al., IEDM 2018, p.787. [2] IRDS 2020 [3] H. Kawarazaki et al., IIT, 2016, p287. [4] K. Shuto et al., IWDTF, 2019, p.48. [5] H. Kawarazaki et al., SISC, 2019, 4.2. [6] J. R. Hauser. NCSU's CVC [7] A. Toriumi, et al., IEDM 2019, p.338

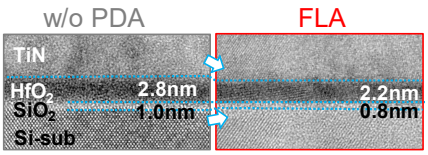

Fig. 3 XTEM of MOSCAPs. Thickness of the dielectrics decreases after FLA.

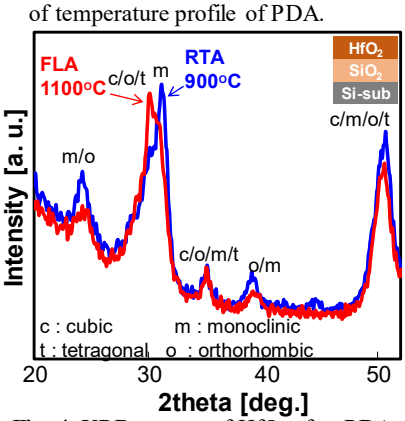

Fig. 4 XRD spectra of HfO_2 after PDA. Non-monoclinic phase is dominant in FLA.

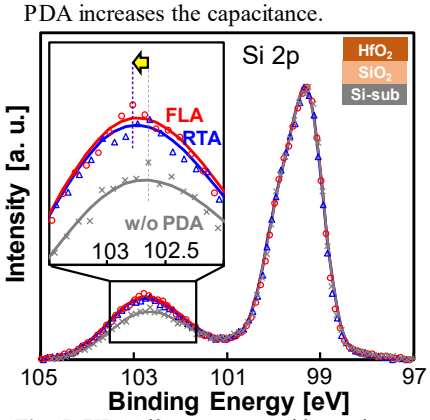

Fig. 5 XPS Si2p spectra. Oxide peak shifts to higher binding energy after PDA.

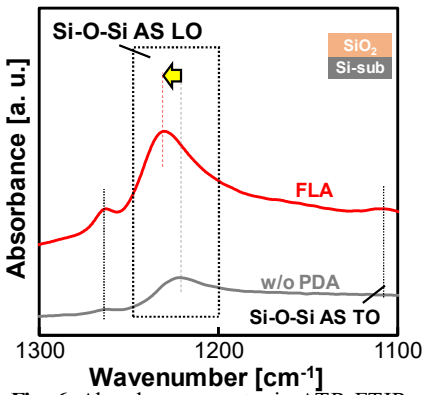

Fig. 6 Absorbance spectra in ATR-FTIR. LO peak shifts to higher wavelength.

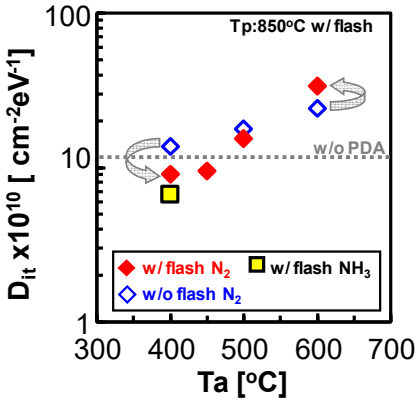

Fig. 7 D_{it} measured by conductance method at room temperature as a function of Ta. Flash improves Dit with Ta under 500°C.

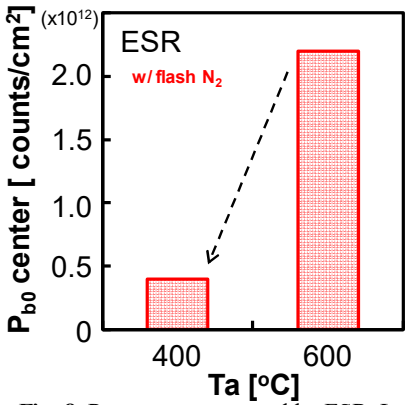

Fig. 8 P_{b0} centers measured by ESR. Low Ta decrease P_{b0} centers. Low Ta may prevents H desorption from SiO_2/Si-sub.

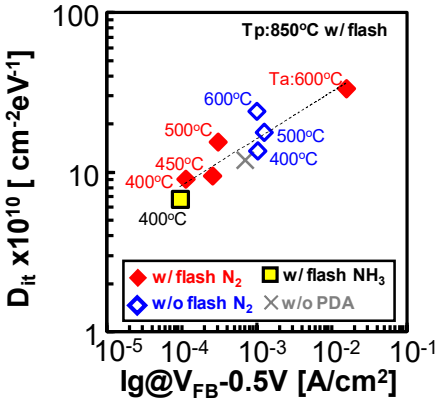

Fig. 9 D_{it} as a function of gate leakage. D_{it} seems to be correlates with the gate leakage.

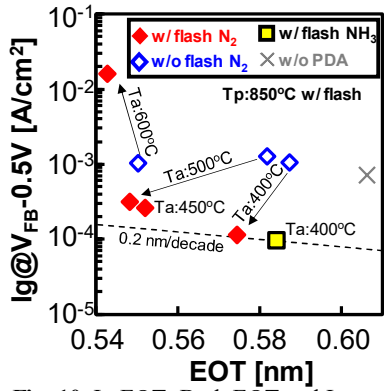

Fig. 10 Ig-EOT. Both EOT and Ig improve after FLA especially for low Ta.

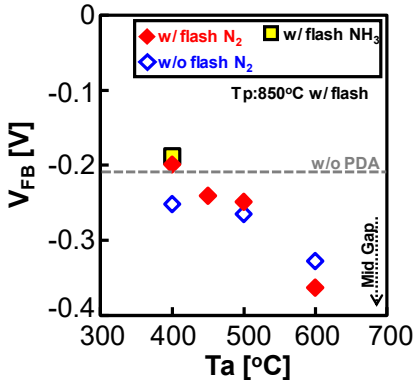

Fig. 11 V_{FB} as a function of Ta. V_{FB} shift is prevented with low Ta.

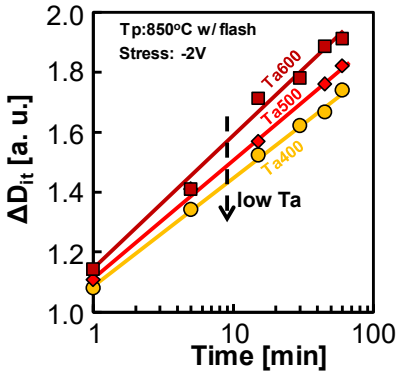

Fig. 12 D_{it} stability under -2V stress. Low Ta reduce D_{it} degradation.