Extended Abstracts of the 2013 International Conference on Solid State Devices and Materials, Fukuoka, 2013, pp60-61

Effects of HfO_2 and Lanthanum Capping Layer Thickness on the Narrow Width Behavior of Gate First High-K and Metal Gate NMOS Transistors

Satya Siva Naresh[1], Nihar R. Mohapatra[1] and Pardeep Kumar Duhan[2]
[1] IIT-Gandhinagar, Chandkheda, 382424, Ahmedabad, India, Ph: +91-79-3245-5012, Email: nihar@iitgn.ac.in
[2] IIT-Bombay, Powai, 400076, Mumbai, India

Abstract - This paper discusses in detail the effects of Hafnium oxide (HfO_2) and Lanthanum (La) capping layer thickness on the performance of narrow and short NMOS transistors. It is shown that the threshold voltage of the NMOS transistors increase with decrease in channel width and this effect is enhanced for thicker HfO_2 and La capping layer. An empirical model is developed to understand and model this behavior.

1. Introduction: The use of higher K gate dielectrics with metal gates is acknowledged world over as the major change in the gate stack of MOS transistors during the last decade [1-3]. The newer gate stack coupled with the smaller geometries of modern MOS transistors give rise to many second order effects. One of them is the increase in threshold voltage (V_T) with decrease in the transistor width (W) for NMOS transistors. This behavior was attributed to the non-uniform distribution of fixed charges in the bulk of gate dielectric [4]. Note, the thickness optimization of HfO_2, interfacial layer (SiO_2) and La capping layer are popular methods to tune the V_T and adjust the long term reliability of high K metal gate MOS transistors. So, the effect of these thickness variations on the narrow width behavior of NMOS transistors needs to be studied. In this work, we discuss the effect of HfO_2 and La capping layer thickness on the narrow width behavior of NMOS transistors. The possible physical mechanisms responsible for this behavior are explained through detailed measurements. An empirical model is also proposed to explain the observed behavior and to incorporate the said effect in circuit simulations.

2. Experimental: The devices used in this work are fabricated using a 28-nm gate first CMOS technology with high-K dielectrics and metal gates. The gate dielectric stack is composed of HfO_2 (different thickness) and 8Å interfacial SiO_2 layer (T_{IL}). La is used as a capping layer between the high-K gate dielectric and TiN metal gate to provide band edge functionality. Note, all the devices used in this study are fabricated using the same process flow.

3. Results and Discussion: The measurements are performed on NMOS transistors (gate length (L) of 34nm and W varying from 500nm to 80nm) with different HfO_2 (T_{HfO2}) and La capping layer thickness ($T_{La\text{-}cap}$) as mentioned in Table I. The experiments are performed at room temperature and the V_T is extracted using maximum trans-conductance method.

Figs. 1(a) and (b) shows the V_T of NMOS transistors as a function of W for different T_{HfO2} and $T_{La\text{-}cap}$. As shown, V_T increases with reduction in W and the narrow width performance indicator, ΔV_T ($V_{T\ (W=80nm)}$-$V_{T\ (W=500nm)}$) increases with increase in T_{HfO2} and $T_{La\text{-}cap}$.

Table I: Device details

Device#	T_{IL} (Å)	T_{HfO2} (Å)	$T_{La\text{-}cap}$ (Å)	EOT (Å)
1	8	17	0	13.75
2	8	17	2	13.75
3	8	19	2	14.25
4	8	21	2	14.75
5	8	17	3	13.75
6	8	17	4	13.75
7	8	17	5	13.75

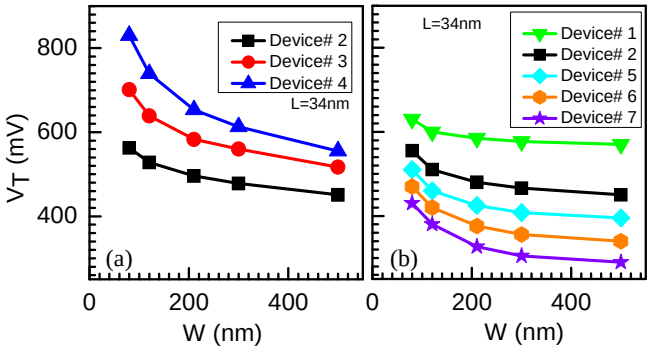

Fig. 1: V_T of NMOS transistors as a function of W (a) for different HfO2 thickness and (b) La capping layer thickness.

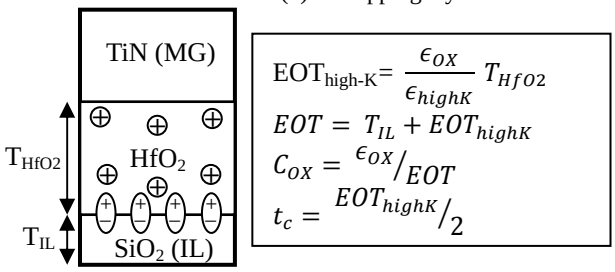

Fig. 2: High-K dielectric and metal gate stack with various kinds of charges that will contribute to the EWF of the gate electrode.

The increase in V_T with the reduction in W for the gate first high-K and metal gate NMOS transistors has been previously reported [4]. This behavior was attributed to the reduction in the positively charged oxygen vacancies and high-K/SiO_2 interface dipoles at the edges of the gate due to oxygen diffusion from ambient and La out-diffusion during post-gate high temperature process steps. Note, these positively charged oxygen vacancies increase linearly with increase in T_{HfO2} thereby increasing the annihilation of these charges at the gate edges due to oxygen diffusion from the ambient. The high-K/SiO_2 interface dipoles also increase with increase in $T_{La\text{-}cap}$ thereby increasing the possibility of La out-diffusion at the gate edges. Both these factors explain the increase in ΔV_T with increase in T_{HfO2} and $T_{La\text{-}cap}$. Although the above statement qualitatively explains the narrow width behavior of the high-K metal gate

NMOS transistors, a closed form V_T expression is necessary to model the said effect. Fig. 2 shows the high-K dielectric and metal gate stack with various kinds of charges that will contribute to the effective work function (EWF) of the gate electrode and thereby V_T of the NMOS transistor. Note, the positively charged oxygen vacancies are uniformly distributed across the bulk of HfO_2 and so could be modeled as a thin sheet of charge, Q_F (C/cm^2) at a distance t_c from the TiN/HfO_2 interface. The EWF in this case can be written as,

$$EWF = \varphi_{MS} - \left(\frac{EOT_{highK}}{2EOT}\right)\frac{Q_F}{C_{OX}} - \left(\frac{EOT_{highK}}{EOT}\right)\frac{Q_D}{C_{OX}} \quad (1)$$

The first term φ_{MS} is the metal-semiconductor work function difference. EOT_{highK} is the effective oxide thickness of the HfO_2 layer and EOT is the effective oxide thickness of the complete gate stack. C_{OX} is the gate capacitance per unit area and Q_D is the dipole charge (C/cm^2).

As discussed before, due to oxygen diffusion from the ambient and La out-diffusion, there is a reduction in Q_F and Q_D at the edges of the gate. The magnitude of the lost charges peak at the edges of the gate and therefore could be modeled as a 2-D exponential function shown below.

$$\Delta Q_F(x,y) = Q_1\left(e^{\frac{-x}{A_1}} + e^{\frac{-(W-x)}{A_1}}\right)\left(e^{\frac{-y}{B_1}} + e^{\frac{-(L-y)}{B_1}}\right) \quad (2)$$

$$\Delta Q_D(x,y) = Q_2\left(e^{\frac{-x}{A_2}} + e^{\frac{-(W-x)}{A_2}}\right)\left(e^{\frac{-y}{B_2}} + e^{\frac{-(L-y)}{B_2}}\right) \quad (3)$$

Here, Q_1 and Q_2 (C/cm^2) are the constant amplitudes of the exponential function. A_1, A_2 and B_1, B_2 are the decaying constants along W and L respectively. From this distribution the average lost charge (ΔQ_F, ΔQ_D) can be calculated by integrating eqns. (2) and (3) along the L and W and dividing it by the total gate area.

$$\Delta Q_F = \frac{4Q_1A_1B_1(1-e^{\frac{-W}{A_1}})(1-e^{\frac{-L}{B_1}})}{WL} \quad (4)$$

$$\Delta Q_D = \frac{4Q_2A_2B_2(1-e^{\frac{-W}{A_2}})(1-e^{\frac{-L}{B_2}})}{WL} \quad (5)$$

The modified EWF and the V_T taking care of the lost charges could now be written as,

$$EWF = EWF_0 + \left(\frac{EOT_{highK}}{2EOT}\right)\frac{\Delta Q_F}{C_{OX}} + \left(\frac{EOT_{highK}}{EOT}\right)\frac{\Delta Q_D}{C_{OX}} \quad (6)$$

$$V_T = V_{T0} + \left(\frac{EOT_{highK}}{2EOT}\right)\frac{\Delta Q_F}{C_{OX}} + \left(\frac{EOT_{highK}}{EOT}\right)\frac{\Delta Q_D}{C_{OX}} \quad (7)$$

EWF_0 and V_{T0} are the effective work function and threshold voltage of wide NMOS transistors where the effect of lost charges is minimum. The equations (6) and (7) have 6 fitting parameters (Q_1, Q_2, A_1, A_2, B_1 and B_2). V_{T0} can be determined from the largest geometry NMOS transistor. The rest of the fitting parameters can be determined by curve fitting for more than one gate length simultaneously. Fig. 3 shows the comparison between the model and experimental V_T as a function of W for different T_{HfO2} and $T_{La\text{-}cap}$. As shown, the model successfully predicts the V_T within 2% for all geometries and for different T_{HfO2} and $T_{La\text{-}cap}$. The seven model parameters used to fit the measured data are shown in Table II. Fig.4 shows the simulated bulk fixed charge distribution in the dielectric stack for, a) large and small geometry transistors, b) different T_{HfO2} and c) different $T_{La\text{-}cap}$. As shown the fixed charge is maximum at the centre of the channel and reduces as we move towards the edges. This explains the V_T change with W. The fixed charge at the centre of the dielectric stack and the charge reduction at the gate edges are also higher for thicker T_{HfO2},$T_{La\text{-}cap}$ which explains the observed increase in narrow width effect for thicker T_{HfO2} and $T_{La\text{-}cap}$.

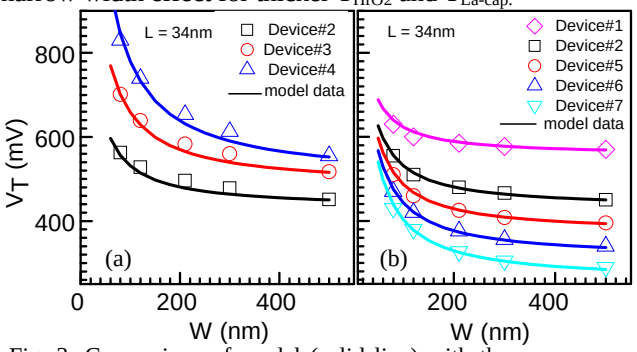

Fig. 3: Comparison of model (solid line) with the measurement data (points) for different (a) T_{HfO2} and (b) $T_{La\text{-}cap}$.

Table II: The empirical parameters for different devices

Device#	V_{T0} mV	Q_1 μC/cm^2	A_1 nm	B_1 nm	Q_2 μC/cm^2	A_2 nm	B_2 nm
1	555	6.0	13	13	0	0	0
2	430	6.0	13	13	1.02	16	16
3	480	7.0	17	17	1.02	16	16
4	500	7.8	21	21	1.02	16	16
5	370	6.0	13	13	1.50	16	16
6	310	6.0	13	13	1.99	16	16
7	255	6.0	13	13	2.44	16	16

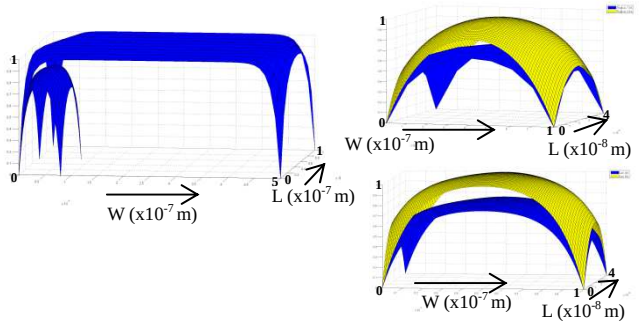

Fig. 4: Simulated bulk fixed charge distribution (normalized) in the gate dielectric for, a)large (500x100nm) and small (80x50nm) geometry transistors, b)different T_{HfO2} (blue-17A, yellow–21A) and c) different $T_{La\text{-}cap}$ (blue-0A, yellow–5A). Non-uniform distribution of bulk fixed charges in all the cases could be clearly seen.

4. Conclusions: It is shown that the V_T of the gate first high-K and metal gate NMOS transistor increases with decrease in W and this behavior is enhanced for larger T_{HfO2} and $T_{La\text{-}cap}$. The reason behind this behavior is attributed to the increased annihilation of positively charged oxygen vacancies and higher La out-diffusion at the gate edges. A closed form V_T expression is developed to model this behavior. The model can accurately predict the V_T of different device geometries and for a wide range of T_{HfO2} and $T_{La\text{-}cap}$.

Acknowledgement: Pardeep is thankful to DST for fellowship.

Refs:[1] Chudzik, p.194, Symp. VLSI Technology 2007, [2] Chen, p.88, Symp. VLSI Technology 2008, [3] Mistry, p.247, IEDM 2007, [4] Walke, p.2582, TED 2012.