第27卷 第7期 半 导 体 学 报 Vol.27 No.7
2006年7月 CHINESE JOURNAL OF SEMICONDUCTORS July,2006

Modeling of Gate Tunneling Current for Nanoscale MOSFETs with High-k Gate Stacks*

Wang Wei[1,†], Sun Jianping[2], and Gu Ning[1]

(1 State Laboratory of Bioelectronics, Jiangsu Laboratory for Biomaterials and Devices, Southeast University, Nanjing 210096, China)
(2 Department of Electrical Engineering and Computer Science, University of Michigan, USA)

Abstract: A quantum model based on solutions to the Schrödinger-Poisson equations is developed to investigate the device behavior related to gate tunneling current for nanoscale MOSFETs with high-k gate stacks. This model can model various MOS device structures with combinations of high-k dielectric materials and multilayer gate stacks, revealing quantum effects on the device performance. Comparisons are made for gate current behavior between nMOSFET and pMOSFET high-k gate stack structures. The results presented are consistent with experimental data, whereas a new finding for an optimum nitrogen content in HfSiON gate dielectric requires further experimental verifications.

Key words: high-k; gate current; quantum model
EEACC: 2570A
CLC number: TN386 **Document code:** A **Article ID:** 0253-4177(2006)07-1170-07

1 Introduction

The on-going research on materials, device structures and assembly methods for microelectronic devices with feature sizes smaller than 100nm carries the goal to keep Moore's law in force beyond limits of current microchip technology. To understand the device behavior and provide design aids for sub-0.1μm devices, numerous studies have been conducted on the modeling of the gate current and gate capacitance of nanoscale MOSFETs. Control and reduction of gate leakage current bear practical significance for low power CMOS circuits and to alleviate the scaling limits. High-*k* gate stack structures as candidates to replace silicon dioxide layer for nanoscale MOSFETs have received great attention due to their promise in reduction of gate current and standby power consumption[1~5]. A number of high-*k* gate stack structures have been studied experimentally and theoretically, with most of the modeling approaches based on the WKB approximation. The WKB approaches have gained popularity recently due to their simplicity and less numerical work. However, they have inherent shortcomings in that some physics aspects are parameterized, that it is unable to evaluate gate current and capacitance simultaneously. In addition, in principle, it cannot model combinations of various high-*k* stack dielectrics with interfacial layers properly. Therefore, most approaches existed are insufficient to carry out the tasks for significant engineering of high-*k* stack structures and silicon dielectric interface, which are required due to complications associated with applicability and compatibility of the materials and processing of the high-*k* stacks with silicon technologies. Furthermore, most modeling work performed and published so far is mainly focused on nMOSFETs with high-*k* gate stacks. For pMOSFETs, to our knowledge, only modeling work published was presented in Ref. [6], in which the WKB approach was also employed.

In this work, a unified formulation is employed to model nanoscale n and pMOSFETs with high-*k* stacks as their gate dielectrics, aimed at the gate current reduction and evaluation of other relevant device properties. When the device feature sizes

* Project supported by the National Natural Science Foundation of China (Nos. 60371037, 20573019)
† Corresponding author. Email: wangwei@njupt.edu.cn
Received 24 January 2006, revised manuscript received 29 March 2006
©2006 Chinese Institute of Electronics

reach nanoscale dimensions, a gate dielectric thickness of less than 2.0nm would be required for nanoscale MOS devices with power supply voltages of 1.0 to 1.5V. The simultaneous requirement to keep the same or greater gate capacitance and lower gate current entails the same or greater equivalent oxide thickness (EOT) as well as a thicker physical insulating layer for reduction of gate tunneling current. The same or larger gate capacitance is necessary for good gate control over the channel current and reduction of the short channel effects.

2 Method of calculation

The present model is based on self-consistent solutions of the Schrödinger-Poisson equations previously developed for modeling quantum-tunneling devices[7,8], with modifications made for nanoscale MOSFETs in the *y* direction perpendicular to the dielectric layers. The gate current components of the thermionic emission, FN tunneling, and direct tunneling through the oxide barrier are evaluated as a whole using a traveling wave calculation, hereafter referred to as J_{3D}, while the tunneling component from the inversion layer quantum well is evaluated by a transmission calculation, hereafter referred to as J_{2D}, both obtained from the solutions of the Schrödinger equation, self-consistently with its potential term determined by the Poisson equation. The calculated electron wave functions relevant to the current transport are shown in Fig. 1 to illustrate the current transport components originated from the 3D electrodes and 2D quantum well in the inversion layer. As an example of the 2D wave functions,

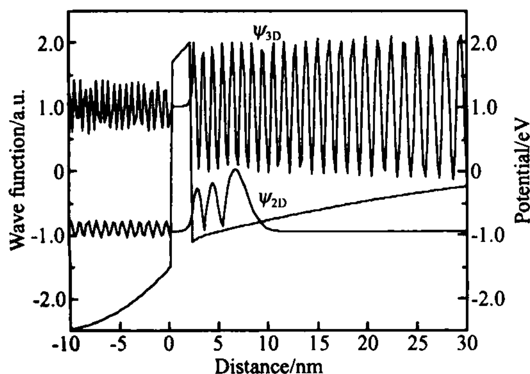

Fig. 1 Electron wave functions from the 3D and 2D electron states Only the third quasi-bound state in the inversion layer is shown.

Ψ_{2D}, only the wavefunction at the third level of the quasi-bound states is shown.

We first calculate the self-consistent potential and charge distributions in the polysilicon and substrate regions from the coupled Poisson equation and Fermi distribution function. The electrostatic potential across the device in the y direction is determined by the doping profile and mobile charges, as governed by the Poisson equation.

$$\frac{\partial^2 V(y)}{\partial y^2} = -\frac{q}{\varepsilon(y)}[N_D(y) - N_A(y) - n(y) + p(y)] \tag{1}$$

where $\varepsilon(y)$ is the dielectric constant of individual high-k stack layers, n(y) is the electron density, and p(y) is the hole density in the quasi-equilibrium contact regions in terms of local Fermi levels,

$$n(y) = N_C \frac{2}{\sqrt{\pi}} F_{1/2}\frac{E_F(y) - E_C(y)}{k_B T} \tag{2}$$

$$p(y) = N_V \frac{2}{\sqrt{\pi}} F_{1/2}\frac{E_V(y) - E_F(y)}{k_B T} \tag{3}$$

where N_C and N_V are the effective densities of states in the conduction band and valence band, respectively, and

$$F_{1/2}(\xi) = \int_0^\infty \frac{x^{1/2}}{1 + \exp(x - \xi)} dx \tag{4}$$

is the Fermi-Dirac integral function. Note that Fermi levels $E_F(y)$ are assumed to be constants, respectively, in the poly-silicon and silicon substrate regions and that conduction band and valence band edge are related to the potential by[7,8]

$$E_C(y) = E_C(\pm\infty) - q[V(x) - V(\pm\infty)] \tag{5}$$

$$E_V(y) = E_C(\pm\infty) - q[V(x) - V(\pm\infty)] - E_g \tag{6}$$

where E_g is the energy bandgap of silicon. The Fermi integrals in Eqs. (2) and (3) can be expanded by rapidly converging power series in terms of the ratios $r_1 = n(y)/N_C$ and $r_2 = p(y)/N_V$, respectively, based on the Joyce-Dixon approximation[7,8]:

$$\frac{E_F(y) - E_C(y)}{k_B T} - \lg r_1 = \sum_{k=1}^{N} A(k) r_1^k \tag{7}$$

$$\frac{E_V(y) - E_F(y)}{k_B T} - \lg r_2 = \sum_{k=1}^{N} A(k) r_2^k \tag{8}$$

with the coefficient A(k) determined using a least-squares fit technique. In each y point, Eqs. (1), (7), and (8) lead to three coupled nonlinear equations with three unknows, i.e., the electron density, hole density and potential, which are then

solved using the Newton iterative method with a sparse matrix technique. For the charges inside the quantum well of the inversion layer formed under positive gate bias voltages and for the gate current components through and/or above the ultrathin oxide barrier, quantum calculations are performed by directly solving the Schrödinger equation:

$$-\frac{\hbar^2}{2}\times\frac{\partial}{\partial y}\left(\frac{1}{m^*(y)}\times\frac{\partial\psi(y)}{\partial y}\right)+E_C(y)\psi(y)=E\psi(y) \quad (9)$$

The self-consistent electron charge at the j-th energy subband in the i-th valley in the inversion layer is given by

$$n_{ij}(y)=\frac{m_i^* k_B T}{\pi\hbar^2}\ln\left[1+\exp\left(-\frac{E_{ij}-E_F}{k_B T}\right)\right]\times|\psi_{ij}(E_{ij},y)|^2 \quad (10)$$

where $\psi_{ij}(E_{ij},y)$ is the electron wavefunction at the j-th subband in the i-th valley from the solution of the Schrödinger equation in the inversion layer. The gate current evaluation is performed by the traveling wave calculations for the thermionic emission, Fowler-Nordheim (FN) tunneling, and direct tunneling through the oxide barrier, giving

$$J_{3D}=-q\hbar\sum_k W(k)\,\mathrm{Im}\left(\psi_k^*(y)\frac{1}{m^*(y)}\times\frac{\partial\psi_k(y)}{\partial y}\right) \quad (11)$$

where weighting function $W(k)$ is given by

$$W(k)=\frac{m^* k_B T}{2\pi^2\hbar^2}\ln\left(1+\frac{E_F-E_l}{k_B T}\right) \quad (12)$$

and $E_l=\hbar^2 k_l^2/2m^*$ is the longitudinal energy of the electron with wave-vector k_l.

For the electron tunneling current from the inversion layer into the oxide and gate electrode, a transmission calculation is performed. Based on the wave functions calculated in the inversion layer, the transmission is given by

$$T_{ij}=\frac{|C_{tr,ij}|^2}{|A_{in,ij}|^2}\times\frac{k_{tr,ij}}{k_{in,ij}}\times\frac{m_{in,i}^*}{m_{tr,i}^*} \quad (13)$$

where $C_{tr,ij}$ is the transmitted wave amplitude in the i-th valley and the j-th subband, and $A_{in,ij}$ the incident wave amplitude in the corresponding subband and valley. The 2D gate current component originated from the subbands in the inversion layers is then

$$J_{2D}=\sum_{i,j}J_{i,j}=q\sum_{i,j}n_{ij}T_{ij}f_{ij} \quad (14)$$

where T_{ij} is the electron transmission probability, $f_{ij}=\frac{E_{ij}}{j\pi\hbar}$ the electron impact frequency on the interface, n_{ij} the sheet electron density, and E_{ij} the quasi-bound state energies in the i-th valley and the j-th subband. The total gate current density is the sum of the 2D and 3D components.

In the present approach, we treat the thermionic emission, FN tunneling, and direct tunneling current components in a unified formulation. Most of other treatment to these current components involve complicated formulas and, in many cases, fitting parameters. We consider that the separation of the current transport components is a reflection of the historical development of the MOS scaling efforts toward nanoscale dimensions, and that, in principle, tunneling depends on the overall potential profile in the electron tunneling path, rather than a particular part of the potential profile. Therefore, the separation and delimitation of various gate current components may involve inconsistencies in the interfacing and integration of those components. In addition, the gradual potential barrier approximation inherent in the WKB approach may become improper for the sharp potential interfaces between the polysilicon and substrate for nanoscale MOSFETs. In our opinion, as the feature sizes approach nanoscale dimensions, the device structure in the direction of interest may well be treated as an integrated, open quantum system, and a fully self-consistent solution to the Schrödinger-Poisson equations is necessary and sufficient to model all the gate current transport components and obtain all the device terminal properties of interest.

An important feature of the present model lies in that the electrostatic potential $V(y)$ and the material parameters can be accurately specified at each y point in the multilayered stack structures, allowing evaluation of various schemes of combination having different stack materials and different layers. Moreover, the possible fixed charge and interface charge distributions in the stacked layers can also be accommodated in Poisson's equation, as long as these charge distributions are available. Therefore, the present approach enables us to study multilayer high-k stack structures of nanoscale MOSFETs, consisting of various dielectric materials for various configurations.

As the channel length is further reduced, the two-dimensional effects like the edge direct tunneling (EDT), drain-induced barrier lowering

(DIBL), and V_T roll-off become non-negligible. In the high drain bias, these problems cannot be described directly in 1D model. In this work, to solve the 2D problems such as EDT, we begin by using 2D Poisson equation and electron and hole current continuity equations to solve the charge and potential distribution throughout the device. Then, in y-direction (perpendicular to the dielectric layers), we solve the 1D Schrödinger-Poisson equations for each x position (assuming x-direction along the channel) using the calculated charge and potential distribution as input data. Using the gate current model given above, we can compute the tunneling current density for each x position, which we can integrate to obtain the total gate current.

3 Results and discussion

3.1 Comparison with experiment

Figure 2 shows a comparison between the calculated and measured gate tunneling current characteristics of nMOS and pMOS structures (EOT of

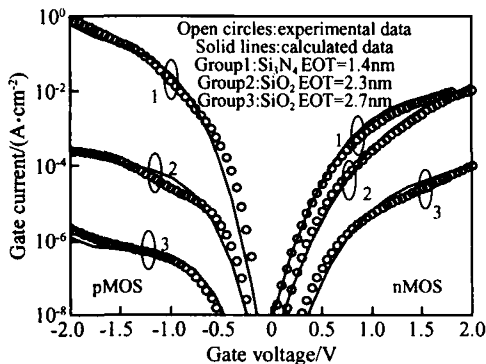

Fig. 2 Comparison of gate current for pMOS and nMOS with calculated data and experimental data The experimental data are taken from Refs. [9,10]. The electron and hole barrier height, effective mass, are known for SiO_2 (3.10eV/4.5eV, 0.40m_0/0.32m_0) and Si_3N_4 (2.10eV/1.90eV, 0.50m_0/0.40m_0)[10], and dielectric constant for SiO_2 (3.9) and Si_3N_4 (7.8)[6].

2.3 or 2.7nm for SiO_2, 1.4nm for Si_3N_4), with excellent agreement between the theory and experiment. The curves of Group 1 are for the Si_3N_4 cases, and the curves of Group 2, 3 are for the cases of SiO_2. It is apparent that the gate current decreases with increasing dielectric thickness for SiO_2. Furthermore, for the same SiO_2 film thickness, the gate current density for the n-MOS case is larger than that for the p-MOS case due to the lower conduction band offsets in the nMOS structures and higher valence band offsets in the pMOS structures[10]. However, for the cases of Si_3N_4, the gate current in pMOS is larger than that in nMOS due to the higher conduction band offset in nMOSFET for binary dielectric materials and the higher electron effective mass in the dielectric layer for nMOSFET, as explained in Ref. [10].

3.2 Effects of high-k dielectric thickness

For high-k stack structures consisting of a single high-k layer and a SiO_2 layer, HfO_2 stacks are taken as an example of gate current dependence on the thickness of the high-k layer with the same EOT of 1.3nm, as shown in Fig. 3. All the high-k stacks make a reduction of gate current as compared with SiO_2 gate dielectric layer, especially at lower gate voltages applicable for nanoscale

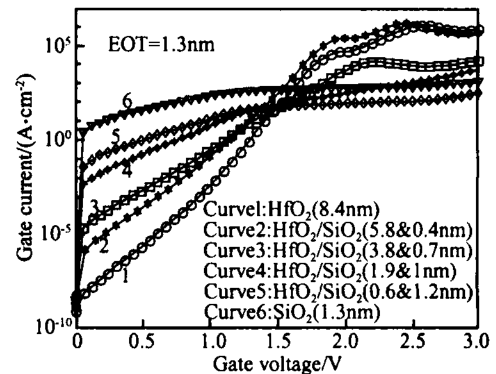

Fig. 3 Calculated gate current versus gate voltage with different physical thickness of the high-k dielectric with the same EOT (1.3nm) for HfO_2/SiO_2 gate stack structures The electron barrier height, effective mass, and dielectric constant for HfO_2 are 1.50eV[4], 0.20m_0[11], and 25[4], respectively.

nMOSFETs, due to the additional physical width in the tunneling path for the lower energy electrons that contribute most of the gate current, as shown in Fig. 4. As expected, the physical thickness of the high-k layer plays a dominating role at the lower gate voltages. The difference can be as large as 4～5 orders of magnitude between the high and low HfO_2/SiO_2 thickness ratios for the same EOT, as observed in the Fig. 3. At higher gate bias voltages near and above the barrier height of the high-k dielectric layer, however, the gate current is dominated

by FN tunneling and thermal emission. Therefore the magnitude of gate current of the high-k stack structure and the SiO_2 structure becomes comparable, and that the former can be orders of magnitude larger than that of the SiO_2 structure at even higher gate voltage. In other words, at lower gate bias, the gate current is mainly affected by the dielectric constant of the high-k dielectric material utilized, while at higher bias, it is mainly determined by its barrier height.

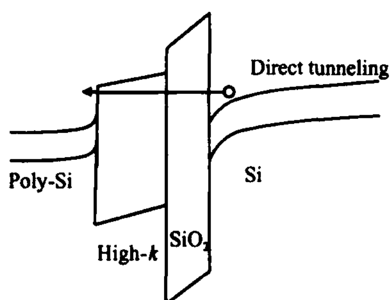

Fig. 4 Electron direct tunneling band diagram of high-k gate stack nanoscale MOS structure

Other types of high-k stack structures using $SiO_2/Ta_2O_5/SiO_2$ gate dielectrics[12] and adding Al_2O_3 capping layer[13] have been proposed to alleviate the interfacial structure problems[12]. The present modeling approach is applied to study of this type of multiple layered structures, with the material and structural parameters specified accurately in the individual stacked layers.

Recent studies[14,15] have shown that the direct tunneling between the source/drain extension and gate overlap region, i. e. the edge direct tunneling (EDT), dominates the off-state current, especially in very short channel devices. This results from the fact that the ratio of the gate overlap to the total channel length is larger for the short channel device than that of the long channel device.

Figure 5 shows the simulated gate current versus the length of overlap region for high-k stack structures. In the figure, we can see that the gate currents are higher with larger overlap length. This effect is due to the fact that the flat-band voltage for channel region is negative, whereas the flat-band voltage for the overlap region is almost zero, thus resulting in higher vertical electric field, and more leakage current density for the overlap region. It is also noted that the gate current is re-

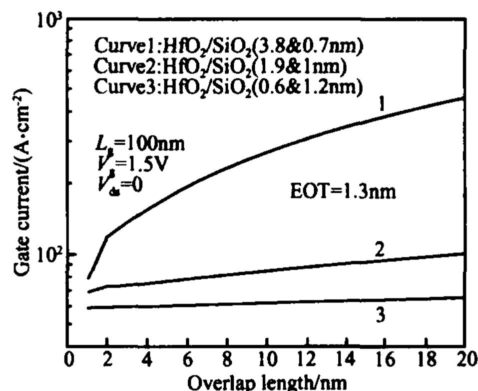

Fig. 5 Calculated gate current versus the length of overlap region with different physical thicknesses of the high-k dielectric with the same EOT (1.3nm) for HfO_2/SiO_2 gate stack structures

duced with increasing dielectric thickness of SiO_2 layer in the HfO_2/SiO_2 stacks. This is attributed to the interplay among various factors of the structure and material parameters, such as the oxide barrier height, electron effective masses, and dielectric constant. However, EDT current is mainly affected by its barrier height. As expected, with increasing the length of overlap region, the EDT effect results in higher vertical field. Thus, the thickness of SiO_2 plays a dominating role, due to SiO_2 layer with higher barrier in the tunneling path for the higher energy electrons of EDT.

3.3 Effects of nitrogen incorporation

It is reported that incorporation of nitrogen into hafnium silicate (HfSiO), hafnium aluminate (HfAlO) and HfO_2 greatly enhances the dielectric constant of silicates, suppresses dopant diffusion from gate poly-Si into the channel during high temperature annealing process, and increases crystallization temperature of the high-k stacks[16]. These properties are ascribed to the homogeneity of the bond structure in the film containing nitrogen through high temperature annealing. These improvements make these Hf-based materials more suitable for the CMOS process. The increase of nitrogen content leads to an increase of the dielectric constant but a decrease of the conduction band offset as well as the valence band offset. It is interesting to observe that, as the only exception among most high-k dielectrics, the conduction band offset with nitrogen incorporation may be larger than the valence band offset, such as in the case of HfSiON.

It is an advantage for the reduction of gate current in n-MOSFETs. For the pMOSFETs, however, such property would not be useful since the hole tunneling is dominant for the gate current.

In Fig. 6, the effects of nitrogen content on the gate current are shown for tunneling current through HfSiON dielectric film. Three different electron effective masses in between that of SiO_2, HfO_2, and Si_3N_4 were used in the calculations as the value of the electron and hole effective mass is not available, and such modeling calculations have

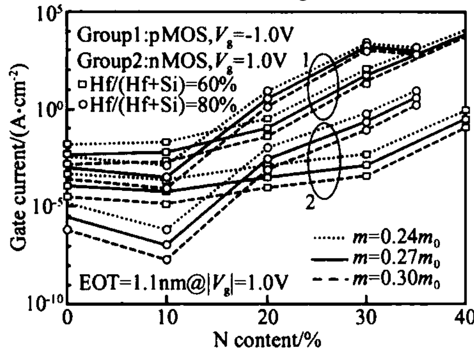

Fig. 6 Simulated tunneling current through HfSiON dielectric films of different nitrogen contents for nMOSFET and pMOSFET

not been found in the open literature. However, the values of dielectric constant and barrier height were taken from experimental work[17] with the N content up to 40 % for the curves of Hf/(Hf + Si) being 60 %, and up to 35 % for the rest of the curves[16]. The results here show that, for both cases, there is an optimum value of the N content (approximately 10 %) where minimum gate current may be achieved. This is attributed to the interplay between the dielectric constant and barrier height; the incorporation of N leads to an increase of the dielectric constant while maintaining essentially the same barrier height until the N content reaches 10 % where the barrier height start decreasing significantly. This optimum value of N content is a new finding from this work, to be verified experimentally. It is also noted that the gate current in the pMOSFETs is larger than that of the nMOSFETs with the same EOT and N content; this is because that, with the virtually same dielectric constant, the valence band offset in the pMOSFETs is smaller than the conduction band offset in the nMOSFETs.

4 Conclusion

Based on quantum modeling of electron tunneling, we have modeled the influence of quantum effects on the gate current of nanoscale p- and n-MOSFETs with high-*k* stack dielectric structures. The main feature of the present approach is that most of the effects of interest can be studied simultaneously and in a unified fashion, without having to use fitting parameters as done in approaches based on the WKB approximation. Moreover, since the material and structural parameters can be specified accurately at each structural point, various configurations of high-*k* stack materials and their combinations can be investigated properly. Therefore, the present approach can be used to provide guidelines for experimental studies of high-*k* stack structures for nanoscale MOSFETs. Our modeling results have demonstrated effects of dielectric film thickness, nitrogen incorporation, and various stack configuration, on the gate current, showing consistency with experimental results. A new finding has also been generated as a result of this work, indicating the optimum nitrogen content for nitrogen incorporation scheme. However, further experimental work is required to verify this result.

References

[1] Fan Y Y, Nieh R E, Lee J C, et al. Voltage and temperature dependent gate capacitance and current model: application to ZrO_2 MOSCAPs. IEEE Trans Electron Devices, 2002, 49(11): 1969

[2] Fan Y Y, Xiang Q, An J, et al. Impact of interfacial layer and transition region on gate current performance for high-*k* gate dielectric stack: its tradeoff with gate capacitance. IEEE Trans Electron Devices, 2003, 50(2): 433

[3] Kizilyalli I C, Huang R Y S, Roy P K. MOS transistors with stacked SiO_2-Ta_2O_5-SiO_2 gate dielectrics for giga-scale integration of CMOS technologies. IEEE Trans Electron Device Lett, 1998, 19(11): 423

[4] Zhao Y, White M H. Modeling of direct tunneling current through interfacial oxide and high-*k* gate stacks. Solid-State Electron, 2004, 48: 1801

[5] Zhang J, Yuan J S, Ma Y, et al. Design optimization of stacked layer dielectrics for minimum gate leakage currents. Solid-State Electron, 2000, 44: 2165

[6] Yu H Y, Hou Y T, Li M F, et al. Investigation of hole tunneling current through ultrathin oxynitride/oxide stack gate dielectrics for p-MOSFETs. IEEE Trans Electron Devices, 2002, 49(7): 1158

[7] Sun J P, Mains R K, Chen W L, et al. *C-V* and *I-V* characteristics of quantum well varactors. J Appl Phys, 1992, 72(9):

2340

[8] Sun J P, Haddad G I, Mazumder P, et al. Resonant tunneling diodes: models and properties. Proceedings of the IEEE, 1998, 86(4): 641

[9] Hou Y T, Li M F, Jin Y, et al. Direct tunneling hole currents through ultrathin gate oxides in metal-oxide-semiconductor devices. J Appl Phys, 2002, 91(1): 258

[10] Yeo Y C, Lu Q, Lee W C, et al. Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric. IEEE Electron Device Lett, 2000, 21(11): 540

[11] Hou Y T, Li M F, Yu H Y, et al. Modeling of tunneling current through HfO_2 and $(HfO_2)_x(Al_2O_3)_{1-x}$ gate stacks. IEEE Electron Device Lett, 2003, 24(2): 96

[12] Cassan E, Galdin S, Dollfus P, et al. Analysis of electron energy distribution function in ultra-thin gate oxide n-MOSFETs using Monte Carlo simulation for direct tunneling gate current calculation. Physica B, 1999, 272: 550

[13] Yang C W, Fang Y K, Chen C H, et al. Dramatic reduction of gate leakage current in 1.61nm HfO_2 high-k dielectric polysilicon gate with Al_2O_3 capping layer. Electron Lett, 2002, 38: 1223

[14] Yang K N, Huang H T, Chen M J, et al. Characterization and modeling of edge direct tunneling (EDT) leakage in ultrathin gate oxide MOSFETs. IEEE Trans Electron Devices, 2001, 48(6): 1159

[15] Choi C H, Nam K Y, Yu Z P, et al. Impact of gate direct tunneling current on circuit performance: a simulation study. IEEE Trans Electron Devices, 2001, 48(12): 2823

[16] Koyama M, Kaneko A, Ino T, et al. Effects of nitrogen in HfSiON gate dielectric on the electrical and thermal characteristics. IEDM Tech Dig, 2002: 849

[17] Koike M, Ino T, Kamimuta Y M, et al. effect of Hf-N bond on properties of thermally stable amorphous HfSiON and applicability of this material to sub-50nm technology node LSIs. IEDM Tech Dig, 2003: 107

高 k 栅介质纳米 MOSFET 栅电流模型*

王 伟[1,†] 孙建平[2] 顾 宁[1]

(1 东南大学生物电子学国家重点实验室, 江苏省生物材料与器件重点实验室, 南京 210096)
(2 美国密西根大学电气工程和计算机科学系, 美国)

摘要: 介绍了一种纳米 MOSFET(场效应管)栅电流的统一模型,该模型基于 Schrödinger-Poisson 方程自洽全量子数值解,特别适用于高 k 栅介质和多层高 k 栅介质纳米 MOSFET. 运用该方法计算了各种结构和材料高 k 介质的 MOSFET 栅极电流,并对 pMOSFET 和 nMOSFET 高 k 栅结构进行了分析比较. 模拟得出栅极电流与实验结果符合,而得出的优化氮含量有待实验证实.

关键词: 高 k; 栅电流; 量子模型

PACC: 7340G; 7340Q

中图分类号: TN386 **文献标识码**: A **文章编号**: 0253-4177(2006)07-1170-07

*国家自然科学基金资助项目(批准号:60371037,20573019)

†通信作者. Email: wangwei@njupt.edu.cn

2006-01-24 收到, 2006-03-29 定稿

©2006 中国电子学会