Available online at www.sciencedirect.com
SCIENCE DIRECT®
Solid-State Electronics 49 (2005) 695–701
SOLID-STATE ELECTRONICS
www.elsevier.com/locate/sse

Scaling CMOS: Finding the gate stack with the lowest leakage current

Thomas Kauerauf [a,b,*], Bogdan Govoreanu [a,b], Robin Degraeve [a], Guido Groeseneken [a,b], Herman Maes [a,b]

[a] IMEC, Kapeldreef 75, B-3001 Leuven, Belgium
[b] KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Received 29 June 2004

The review of this paper was arranged by Prof. S. Cristoloveanu

Abstract

In order to reduce the gate leakage current, high-k gate dielectrics are expected to replace SiO_2 in future CMOS generations. Many of these novel dielectrics are stacks of a thin SiO_2 and a high-k layer. We present a theoretical study that aims at identifying the combination of the stack architecture and high-k material with the lowest leakage current. In the first part of this work the leakage current through high-k double stacks with various thicknesses, materials and gate electrodes is calculated assuming tunneling only. We discuss the difference between gate and substrate injection and show quantitatively the impact of interfacial layer thickness, barrier height, k-value and the work function of the gate material on the tunneling current. In the second part the material properties are no longer considered to be independent and with the universal relation between k-value and barrier height introduced, we are able to identify what material suits best all requirements. The leakage current is calculated for different EOTs and we demonstrate that if all dielectrics follow this universal relation, for a fixed thickness there exists one material, which gives a minimum gate leakage current. It is concluded that even for sub 1 nm EOT devices the k-value has not to exceed ~25 and ZrO_2 or HfO_2 come closest to the ideal high-k if only leakage current issues are considered.

Keywords: High-k; Gate dielectric; Leakage current; k-Value; Barrier height

1. Introduction

To increase the transistor performance and to fulfill the CMOS scaling laws, every new technology node requires a reduction of the gate oxide thickness. In the past this was achieved by simply depositing a thinner SiO_2 dielectric and more recently SiON has been introduced as gate dielectric. In the latest most advanced high-performance devices, oxides with a thickness down to 1.4 nm are used in production. In the following years this scaling will continue and in Fig. 1 an overview of some important parameters according to the ITRS (The International Technology Roadmap for Semiconductors) roadmap [1] is given.

A major penalty of reduced oxide thickness is the increase in gate leakage currents. At operating voltage the stack is in the direct tunneling regime and the current increases exponentially with decreasing oxide thickness. For sub-1.5 nm Equivalent Oxide Thickness (EOT) this leads to enormous current densities and high power dissipation. Especially for mobile application this is unacceptable.

Presently, high-k gate dielectrics are aggressively developed to replace SiO_2/SiON in order to reduce gate leakage currents with a physically thicker layer while still scaling the EOT in future CMOS generations.

* Corresponding author. Address: IMEC, Kapeldreef 75, B-3001 Leuven, Belgium.
E-mail address: kauerauf@imec.be (T. Kauerauf).

0038-1101/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sse.2005.01.018

Year	2004	2005	2006	2007	2010	2013	2016
Generation	90 nm			65 nm	45 nm	32 nm	22 nm
MPU, L_g	37 nm	32 nm	28 nm	25 nm	18 nm	13 nm	9 nm
MPU, EOT	0.9–1.4	0.8–1.3	0.7–1.2	0.6–1.1	0.5–0.8	0.4–0.6	0.4–0.5
MPU, V_{dd}	1.0 V	0.9 V	0.9 V	0.7 V	0.6 V	0.5 V	0.4 V
LOP, L_g	53 nm	45 nm	37 nm	32 nm	22 nm	16 nm	11 nm
LOP, EOT	1.4–1.8	1.2–1.6	1.1–1.5	1.0–1.4	0.8–1.2	0.7–1.1	0.6–1.0
LOP, V_{dd}	1.1 V	1.0 V	1.0 V	0.9 V	0.8 V	0.7 V	0.6 V
LSTP, EOT	1.8–2.2	1.6–2.0	1.4–1.8	1.2–1.6	0.9–1.3	0.8–1.2	0.7–1.1
LSTP, V_{dd}	1.2 V	1.2 V	1.2 V	1.1 V	1.0 V	0.9 V	0.9 V

Fig. 1. Excerpt taken from the ITRS showing the trends in gate length, operating voltage and oxide thickness. Especially for high performance applications (MPU), the scaling is very aggressive.

Several materials like Al_2O_3 and ZrO_2 were investigated, but recently HfO_2 [5] and particularly HfSiON [2–4] showed the most promising results. There the extracted k-values generally range between 12 and 18, depending on the Si and Hf content, respectively. These materials were chosen for processing compatibility reasons and especially HfSiON for the integration with poly-Si electrodes with a high thermal budget during the activation anneal. With the introduction of metal gates, several other candidate materials may become viable options.

Considering new gate dielectric materials, there exists a common misconception that higher k-values will automatically lead to lower leakage current, but leakage current is not only determined by the k-value. In this work, a systematic analysis of the impact of the interfacial layer thickness, capping layers, k-value, barrier height and gate work function is done in order to determine the optimal choice of gate dielectric from a leakage current point of view. It is shown that a strong asymmetry exists between gate and substrate injection, and that small changes of the dielectric properties may have a large impact on the leakage current.

2. Simulation model and the dual-layer stack

In this study, we restricted ourselves to only considering the leakage due to electron quantum-mechanical tunneling through defect-free high-k stacks. The calculations are therefore best-case and can only be reached with mature processes and in cases where chemical and atomic properties allow the fabrication of low-defect density material.

The tunneling current was calculated using the independent electron approximation [6,7], according to which the tunneling current is the energy integral of the tunneling probability multiplied by a factor accounting for the occupation probability and the density of states in the electrodes:

$$J = \frac{qm^*}{2\pi^2\hbar^3}\int_0^\infty \mathrm{d}E \int_0^E \mathrm{d}E_\parallel T(E)[f_1(E) - f_2(E)]$$

where q is the elementary charge, $\hbar$ is the reduced Planck constant, m^* the conduction band effective mass, T the tunneling probability, and f_1, f_2 are the Fermi–Dirac distribution functions in the source and destination electrodes, respectively. It is assumed that tunneling is an energy conserving process. If the parallel momentum conservation is neglected, the equation may be reduced to a single integral.

The tunneling probability was calculated using the Wentzel–Kramers–Brillouin (WKB) approximation by either numerical integration or using a Taylor expansion around the Fermi level in the injecting electrode. Variations of the dielectric constant, barrier height and effective mass across several dielectric stacks have been taken into account. This method gives qualitatively similar results as compared to the Airy implementation [7] for the bias and thickness range of interest in this work, and is therefore a useful approach for predicting the trends.

The applied bias V_g relates to the voltage drop across the stack V_{stack} through the potential balance equation:

$$V_g = \psi_s + V_{stack} + \varphi_{ms} - Q_{ox}/C_{ox}$$

$$V_g = \psi_s + V_{stack} + V_{FB}$$

where ψ_s is the surface potential, φ_{ms} is the metal–semiconductor work function and Q_{ox}/C_{ox} is the shift induced by the interface-equivalent oxide charge density, in this work set to 0. The band diagram of a MOS structure at flatband with a high-k dual layer stack is shown in Fig. 2, At applied bias, the potential distribution across the stack is calculated using basic electrostatics.

Most of the current high-k gate stacks consist of a thin interfacial SiO_2 or SiON layer, a thicker high-k layer and either a poly-Si or metal gate. Any applied

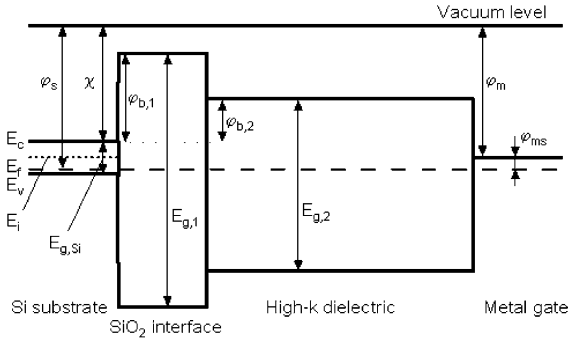

Fig. 2. The energy band diagram in flatband condition with E_c the Si conduction band, E_v the Si valence band, E_f the semiconductor Fermi level, E_i the intrinsic Fermi level, E_g the band gap, φ_s the semiconductor work function, φ_m the metal work function, $\varphi_{b,1}$ the potential barrier for the interface and $\varphi_{b,2}$ for the high-k, φ_{ms} the work function difference and χ the semiconductor electron affinity. The EOT of the stack is 1.6 nm with p-Si substrate, 1 nm interfacial SiO_2, 3 nm high-k dielectric (k = 20, $\varphi_{b,2}$ = 1.5 eV) and a mid gap metal gate electrode.