A Figure of Merit for Flash Memory Multi-Layer Tunnel Dielectrics

B. Govoreanu[a,b], P. Blomme[a,b], M. Rosmeulen[a,b], J. Van Houdt[a] and K. De Meyer[a,b]

[a]STDI Division, IMEC Leuven
Kapeldreef 75, 3001 Leuven, Belgium

[b]ESAT, KU Leuven
Kasteelpark Arenberg 10, 3001 Leuven, Belgium

E-mail: Bogdan.Govoreanu@imec.be

Abstract

In this paper a figure of merit for evaluating the performance of multi-layer tunnel dielectrics for Flash memory applications is proposed. Further analysis provides an in-depth understanding of the multi-layer stacks and allows to select the most suitable stack for memory application.

1 Introduction

Scaling down the tunnel oxide of Flash memories faces major problems. The stress induced leakage current (SILC) strongly degrades the charge retention capability of the memory cell, once the tunnel oxide thickness is scaled down below 7-8 nm. Even in the absence of SILC, reducing the oxide thickness to 4-5 nm seems to be very difficult because of the direct tunneling current. Recent work [1] proposes to use high-k dielectric materials for replacing tunnel oxides and it has been shown [2] that by using a stack of several materials, enhancement of the Fowler-Nordheim tunneling is, in principle, possible.

In some cases, deposition of high-k dielectrics requires that a thin interfacial layer of SiO_2 exists between the silicon and the high-k dielectric. Hence, considering stacks with SiO_2 being one of the layers is of practical importance. This study shows that the existence of the oxide does not necessarily involve a decrease of the performance, and that careful analysis allows for selecting an optimal thickness of the oxide w.r.t. some specific performance indicators. For the sake of simplicity, in this paper only results of two-layer stacks are presented. They are compared with an oxide layer of the same electrical thickness (Equivalent Oxide Thickness - EOT). This choice has a major practical implication: since the capacitance of the stack is the same as that of the reference oxide layer, the capacitive coupling of the floating gate will remain unchanged and therefore the performance of the stack relative to that of the reference layer is a direct measure for the device performance improvement. Moreover, an easy translation of the time scale to the current density scale is possible, which simplifies the programming speed/retention time analysis.

2 Models

The tunnel current density was calculated using the independent electron approximation [3] and a WKB approach with Taylor expansion around the Fermi level in the injecting electrode, and taking into account the variation of the effective mass of the dielectrics across the stack. Although simplistic, this method gives results similar to the Airy implementation [4] and is still valid, at least within an order of magnitude, for the range of thicknesses used in Flash memory devices.

Fig.1 shows typical tunneling curves for the reference layer and for a stack of the same EOT. It is assumed that the dielectric constant of the reference layer (for this analysis, SiO_2 layer) equals the lowest dielectric constant within the stack. At high voltages, the energy barrier shifts downward so that only the first layer of the stack still shows a barrier to the tunneling particle. The electric field is the same in the reference layer and in the first layer of the stack and it therefore follows that in the high voltage range both I-V curves coincide. Maximum performance of the stack w.r.t. the reference layer is obtained when programming at a voltage V_P for which the current ratio J_{stack}/J_{ref} is maximum. Moreover, the retention performance of the stack is better as long as the maximum disturb voltage does not exceed the V_R value, where the stack and the reference curves cross. To get a narrow transition region between the two voltage levels, and therefore a steep programming curve, the ratio V_P/V_R should be relatively small.

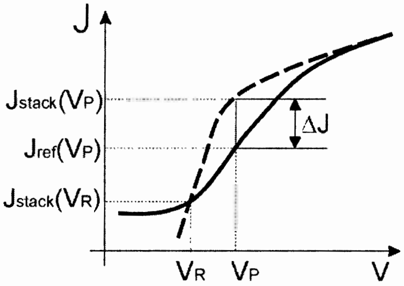

Fig.1. Tunneling curves for the reference layer (solid line) and for the stack of the same electrical thickness

The corresponding ratio of the currents $J_{stack}(V_P)/J_{stack}(V_R)$, should be of the order of several decades, according to the programming speed and charge retention requirements, expressed in current levels. The programming speed is related to the current level in a voltage range around V_P and is traded off for low voltage operation. *The voltage factor* $F_V = V_P / V_R$ *and the current factor* $F_J = \log(J_{stack}(V_P)/J_{ref}(V_P))$ *subject to the constraints* $J_{stack}(V_R) \leq J_{\max}$ *and* $J_{stack}(V_P) \geq J_{\min}$ *are proposed as a figure of merit for comparing stacks of identical EOT.* The current levels J_{max} and J_{min} should be fixed according to the retention and speed requirements, being related to the coupling capacitances of the cell.

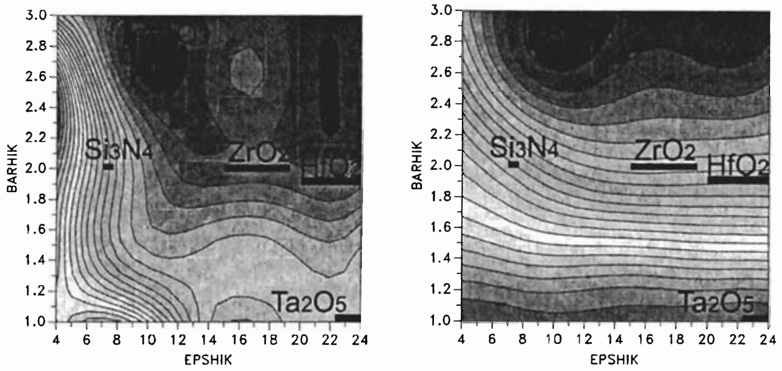

Fig.2. Contour plots of F_V (left) and F_J (right) as functions of k_{hk} (EPSHIK) and $\Phi_{B0,hk}$ (BARHIK) at $t_{ox,stack}$ = 3 nm, for 7 nm EOT.

3 Results and Discussion

The dielectrics that are currently being investigated for SiO_2 replacement usually have lower energy barriers and higher dielectric constants (Tab. 1). Without loss of generality we limit ourselves here to a range from 4 to 24 for the dielectric constant and 1 to 3 eV for the barrier height relative to the bottom of the Si conduction band.

Material	SiO_2	Si_3N_4	Al_2O_3	ZrO_2	HfO_2	Ta_2O_5
Dielectric constant k[-]	3.9	7.5	9-10	15-20	20-25	~25
Barrier Height Φ_{B0}[eV]	3.15	2.0	~3.0	~2	~1.5	~1.0

Table 1: Dielectric constants and barrier heights of some dielectrics

Simulations were carried out for stacks of different EOT and showed similar results to those presented hereafter, where an EOT of 7 nm is considered. A multilevel design of experiment (DOE) approach was used. The independent variables are the dielectric constant of the high-k material k_{hk}, the barrier height of the high-k material ($\Phi_{B0,hk}$) and the physical thickness of the oxide in the stack ($t_{ox,stack}$). A statistical interpolation method [5] was used to construct the models. Contour plots for a fixed thickness of 3 nm oxide are shown in Fig. 2, where the black strips correspond to the considered dielectrics. From the analysis of the models, the following conclusions are drawn: (a) better F_V are obtained for the Al_2O_3, ZrO_2 and HfO_2 stacks; (b) F_J is, however, more favorable in case of either HfO_2 or Ta_2O_5 and it is also less sensitive to the dielectric constant; this is consistent with the fact that such barriers have less "tunneling area" as compared to the stacks with higher barrier height; (c) high programming currents indicate stacks with ZrO_2 or Al_2O_3 or layers whereas lower retention current is achieved for stacks with higher k and lower Φ_{B0} if the oxide is not very thin and for stacks with higher k and higher Φ_{B0} for very thin oxide layers. Overall, it can be derived that the SiO_2/ZrO_2 stack is among the most promising combinations. By fixing k_{hk} and ($\Phi_{B0,\,hk}$) an optimal thickness of the oxide layer can be found by minimizing the objective function $FO = \alpha(F_V - F_{V0})^2 + (1-\alpha)/F_J^2$, with α a parameter which trades off the voltage range operation for the programming speed. For the case of a stack of 5 nm EOT and assuming a transition region so that Fv = 2.0, the optimal thickness of the SiO_2 layer in the stack is 1.8 nm, if no requirement is placed on the current factor. However, maximizing the programming speed leads to an oxide thickness of 3 nm (Fig.3), if no constraint is placed on Fv.

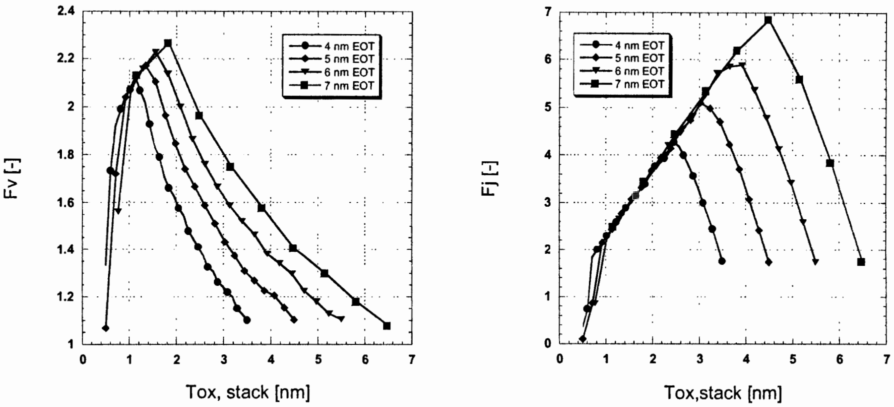

Fig. 3. Voltage and current factors for a SiO_2/ZrO_2 stack, as a function of the oxide thickness in the stack, at different equivalent oxide thicknesses

4 Conclusion

The tunneling current through multilayer stacks was compared to the current through a reference monolayer tunnel dielectric. This allowed for the definition of a figure of merit to be used for evaluating several stacks of a given EOT as replacement for the tunnel dielectric in Flash memories. SiO_2/ZrO_2 stack is proposed as one of the most promising candidates, which allows for higher speed or lower voltage programming.

Acknowledgement – The authors wish to thank Dr. M.Lorenzini, Dr. W.Magnus, Dr. W.Schoenmaker (IMEC Leuven) for useful discussions and comments.

References

[1] Manchanda L. High-K dielectrics for CMOS and Flash. Proc. ICSSDM, 1999, 150-151

[2] Likharev K.K. Layered tunnel barriers for nonvolatile memory devices. Appl Phys Lett 1998, 73(15):2137-2139

[3] Harrison W.A. Tunneling from an independent particle point of view. Phys Rev 1961, 123(1): 85-89

[4] Schenk A, Heiser G. Modeling and simulation of tunneling through ultra-thin gate dielectrics. J Appl Phys 1997, 81(12):7900-7908

[5] Sacks J, Welch W.J, Mitchell T.J, Winn H.P. Design and Analysis of Computer Experiments. Stat Science, 1989, 4:409-435.