Characterization and physical modeling of endurance in embedded non-volatile memory technology

Davide Garetto*†, Alban Zaka‡, Jean-Philippe Manceau*, Denis Rideau‡, Erwan Dornel*
William F. Clark*, Alexandre Schmid†, Hervé Jaouen‡, and Yusuf Leblebici†
*IBM France - 850, rue Jean Monnet, Crolles, France - Email: david.garetto@fr.ibm.com
†Ecole Polytechnique Fédérale de Lausanne - Lausanne, Switzerland
‡STMicroelectronics - Crolles, France

***Abstract*— Transient and endurance mechanisms in high-performance embedded non-volatile memory flash devices are investigated in detail. An extraction methodology combining measurements on equivalent transistors and flash cells is proposed to discriminate the effects of defects on program/erase (P/E) efficiencies and on DC characteristics. A semi-analytical multiphonon-assisted charge trapping model is used to investigate the role and the impact of trapped charges on channel hot-electron injection and Fowler-Nordheim efficiencies, threshold voltage variations and endurance characteristics.**

***Index Terms*—flash endurance, interface traps, multiphonon trapping model**

I. Introduction and objectives

The improvements of program and erase (P/E) efficiencies in embedded non–volatile random-access memory (eNVRAM) flash cells in deep-submicron technologies are leading to significant challenges to be faced to preserve device endurance and retention. Indeed, aggressive and degrading techniques are used for injecting/removing the stored charge. Nevertheless, the channel hot elecron injection (CHEI) mechanism remains the preferred approach for embedded applications where programming speed is a priority. Therefore, understanding the physical aspects underlying these processes is required by technology development engineers, optimizing the technology process, as well as IC designers, focusing on worst-case analysis after device aging. Measurement extraction and modeling techniques are thus important to understand the role of key process parameters on transient performance and endurance.

Compact and analytical models of eNVRAM devices used in industry do not usually consider the physical mechanisms involved during the P/E of a flash memory cell. In particular, only read conditions are analyzed using simplified MOS models, where the threshold voltage V_{th} of the device is changed according to the state of the cell (programmed/erased) [1]. This excessive simplification represents a major limitation for IC designers, requiring an accurate estimation of the programming current and of the influence of soft-programming disturbs on the stored charge. Additionally, analytical models for simulating flash transient mechanisms are usually decoupled from the DC model of the device and oversimplified to facilitate parameter extraction, but compromising scalability [2]. These models are not suitable for worst-case analysis after device aging, and a direct estimation of the degradation of the characteristics is difficult to achieve.

In this work, an extraction methodology to decouple the effects of device degradation on endurance has been validated using a fully comprehensive model for flash devices with transient and device aging capabilities. The semi-analytical model is based on a multiphonon-assisted charge trapping approach and it has been used for the physical understanding of endurance. The endurance extraction methodology based on transient measurements is described. Its application to different process splits has been shown and correlation with process variation is illustrated. The generation and the effects of interface traps on transient mechanisms are investigated using the proposed model.

II. Device integration and characterization

Small memory matrices of flash devices in NOR configuration have been integrated and characterized in a high performance 65nm derivative technology, using process variations on tunnel oxide (TOX) formation and implant of the low-doped drain (LDD) region. The considered structure is a single Flash cell with terminals D (drain), S (source), B (bulk), C (control gate) and F (floating gate).

DC, transient and endurance characteristics are measured; Fowler-Nordheim (FN) tunneling and CHEI have been adopted to erase and program the device, respectively. A progressive sequence of P/E and read operations is used to measure the V_{th} dynamics during the operation. Additionally, programming by FN is considered for degradation studies. In this work, V_{th} corresponds to the control gate voltage V_{CB} needed to achieve a cell current of 8μA with a drain voltage of 0.7V.

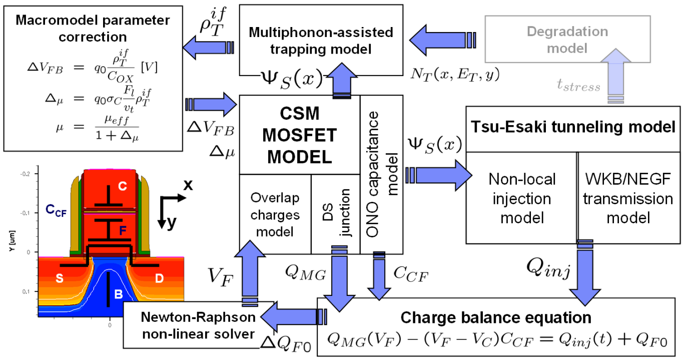
Endurance characteristics are obtained through the measurement of the V_{th} of the cell in erased and programmed states (V_{th}^E and V_{th}^P, respectively) after a high number of P/E cycles. Dummy equivalent-transistors are characterized in DC using the configuration also valid for flash; these devices are investigated applying the same electrical conditions that the flash experiences. Eventually, the approximated floating-gate voltage V_F has been calculated from the control gate bias V_{CB} using the proposed model and applied to the gate of the device to stress the TOX.

III. Model description and validation

Figure 1 shows an overview of the modeling blocks implemented in the semi-analytical model for simulating flash devices. This method is based on a charge sheet analytical model (CSM [3]) for DC analysis, the Tsu-Esaki numerical approach for tunneling mechanisms [4], and a physical non-local model for CHEI [5]. A numerical multiphonon-assisted

978-1-4577-0226-6/11/$26.00 ©2011 IEEE

trapping model inspired on a rigorous quantum mechanical approach [6] is included to simulate the effects of interface traps on the electrostatics [7].



Macromodel parameter correction
$\Delta V_{FB} = q_0 \frac{\rho_T^{if}}{C_{OX}}$ [V]
$\Delta_\mu = q_0 \sigma_C \frac{F_l}{v_t} \rho_T^{if}$
$\mu = \frac{\mu_{eff}}{1+\Delta_\mu}$
ρ_T^{if}
Multiphonon-assisted trapping model
Degradation model
$N_T(x, E_T, y)$
t_{stress}
$\Psi_S(x)$
ΔV_{FB}
$\Delta\mu$
CSM MOSFET MODEL
Overlap charges model
DS junction
ONO capacitance model
Tsu-Esaki tunneling model
Non-local injection model
WKB/NEGF transmission model
C_{CF}
C
F
S
D
B
x
y
Y [um]
-0.2
-0.1
0
0.1
0
V_F
Q_{MG}
C_{CF}
Q_{inj}
Newton-Raphson non-linear solver
ΔQ_{F0}
Charge balance equation
$Q_{MG}(V_F) - (V_F - V_C)C_{CF} = Q_{inj}(t) + Q_{F0}$

Fig. 1. Building blocks of the proposed semi-analytical model for flash devices. The model is built around a charge sheet analytical model (CSM) [3] including: a fully scalable physical compact model for C_{CF} for 3D fringing effects [8], a charge balance equation solver to calculate the floating gate voltage V_F [9], a non-local model for CHEI [5] and the Tsu-Esaki model for the calculation of the injected/tunneling charge Q_{inj} [4], a multiphonon-assisted numerical trapping model following [7]. The latter is adopted to determine the total trapped charge density ρ_T^{if} at the tunnel oxide interface with a multiphonon-assisted approach [7].

Transient mechanisms in flash devices have been studied after fabrication to show the model prediction in both erase and program nominal conditions. Figure 2(a) compares simulation results with measurements of the dynamics of the erase by FN tunneling. A sequence of erase pulses (V_{CB} from -16.5V to -18V) is applied to the control gate and the V_{th} of the cell is measured after each pulse. The same operation has been also performed for programming the cell using CHEI (drain pulse V_D = 4.2 V; t_r = 50ns; t_f = 50ns; t_{PW} = 200ns - control gate at a constant V_{CB} = 7V and 8V) (Figure 2(b)).

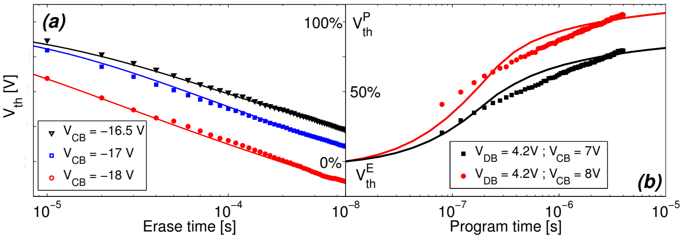


Fig. 2. Simulation results (lines) compared to transient measurements (symbols). In (a) the threshold voltage of the device is decreased by applying a progressive series of erase pulses of magnitude ranging from -16.5V to -18V, and sensing the V_{th} after each pulse. In (b), V_{th} increases upon the application of program pulses on the drain terminal. Both models are scalable and offer good bias dependency.

IV. Degradation and Endurance Characterization

When cycling the device, the tunnel oxide is subject to FN electrical stress during erase and CHEI stress during program. Figure 3(a) shows the effects of electrical stress during cycling, inducing a modification of the V_{th} window, $W = V_{th}^P - V_{th}^E$. Both FN/FN and CHEI/FN P/E operations are characterized and reproduced using the described methodology. In the former case, the cell is programmed and erased by FN operation (program: pulse width 10ms; $V_{CB} = 18.9V$ - erase: pulse width 1ms - $V_{CB} = -17.65V$) and the V_{th} is measured after P/E operations. Two effects can be identified in this configuration: (a) both the V_{th}^E and V_{th}^P increase, (b) the increase of V_{th}^P is less pronounced than V_{th}^E (thus W decreases). In the latter case, the endurance characterization has been performed by cycling the cell with CHEI for programming and FN for erasing. Two phenomena are identified also in this case: (a) V_{th}^E increases due to the progressive filling of interface traps delaying inversion; (b) V_{th}^P initially decreases, inducing the closure of the window after moderate cycling, and then recovers.

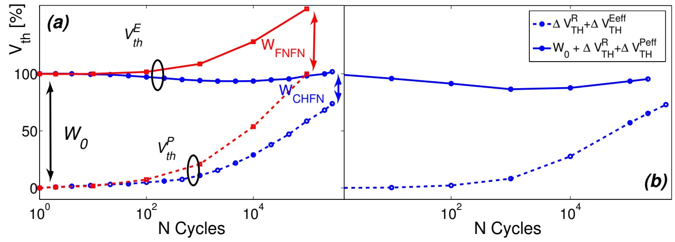


Fig. 3. Threshold voltage window $W = V_{th}^P - V_{th}^E$ vs. number of P/E cycles. In (a), direct V_{th} measurements after cycling using FN/FN and CHEI/FN operations for programming and erasing the cell, respectively, are shown. In (b), window extraction for CHEI/FN regime using the proposed methodology to decouple transient and DC degradation effects.

V. Physical Interpretation and Extraction

Border and interface defects traps attributed to the generation of sp^3 Si dangling bonds (P_b centers [10]) are created when the oxide layer is electrically stressed, such as during cell cycling. The amphoteric nature of P_b centers, being able to capture or emit an electron (+/0/-), cause performance reduction in CMOS devices, V_{th} shift, g_m reduction, subthreshold slope degradation, P/E efficiency reduction [11]. Additionally, fixed negative charges are stacked in the oxide layer causing an additional permanent shift of the characteristics. Figures 4(a-b) show the simulated and measured I_{DS}/V_C curves as cycling increases: the gradual filling of defects by electrons in degraded devices induces a negative charge close to the interface and retards inversion. As a consequence, V_{th} increases of a quantity ΔV_{th}^R

$$\Delta V_{th}^R(n_{cycles}) = V_{th}(t_R, n_{cycles}) - V_{th}(t_R, 0) \qquad (1)$$

where t_R is a time chosen such that the initial state of the device does not affect the dynamics; n_{cycles} is the total number of P/E cycles. Also the subthreshold slope is degraded and the g_m is reduced. The model reproduces well this behavior.

The influence of degradation on erase mechanisms is analyzed in Figure 5(a): ΔV_{th}^R affects the V_{th} vs. erase time measurements due to trap filling in inversion. This contribution taken at $t_R = 0.2$ms has been removed by vertically shifting all the curves (inset of Figure 5(b)).

The erase efficiency degradation, corresponding to the threshold voltage ΔV_{th}^{Eeff}, can be identified as the variation

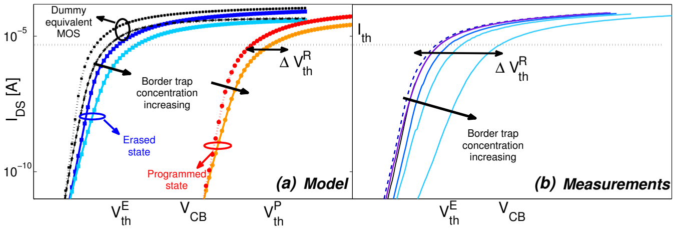


Fig. 4. I_{DS} vs. V_{CB} characteristics ((a) simulations; (b) measurements); in (a) the curves for both erase (blue) and programmed (red) states are shown. Dashed curves indicate fresh devices, while for solid curves a Gaussian distribution of traps is added in proximity of the TOX/substrate interface, affecting the sub-threshold slope and shifting the V_{th} of both the memory states of a voltage ΔV^R_{th}. Model results have been validated on DC measurements during cycling (b).

of slope in the latter plot. A quantifiable estimation of this degradation is given by:

$$\Delta V^{Eeff}_{th}(n_{cycl}) = V_{th}(t_E, n_{cycl}) - \Delta V^R_{th}(n_{cycl}) - V_{th}(t_E, 0) \tag{2}$$

which represents the threshold voltage that cannot be restored when erasing the degraded cells for a given erase time t_E =1ms. The erase efficiency degradation is thus very limited with respect to the effect of filled traps ΔV^R_{th} on the electrostatics. Indeed, the ΔV^R_{th} vs. cycles curve on Figure 5(b) shows that the apparent erase performance degradation, i.e. the V^E_{th} increase in Figure 3(a), only corresponds to a change of the device electrostatics due to traps filling.

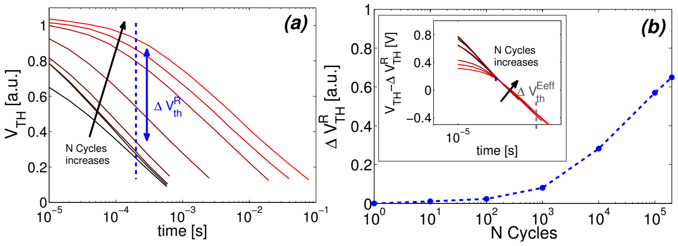


Fig. 5. Extraction methodology to determine ΔV^R_{th} and ΔV^{Eeff}_{th} from erase transient measurements. (a) Raw transient measurements performed after a different number of cycles. Due to the contribution of traps filling, the initial state of the cell is not the same and thus measurements need to be aligned by subtracting $\Delta V^R_{th}(n_{cycles})$ (inset of (b)). $\Delta V^R_{th}(n_{cycles})$ can also be extracted from DC measurements in Figure 4(b). The inset shows how erase efficiency is minimally affected by degradation (identical slope of the curves). (b) ΔV^R_{th} vs. n_{cycles} following the trend of the window baseline.

The same extraction has been performed in program conditions (Figure 6). In this case, the cell has been previously over-erased to exhibit the full transient dynamics. Figure 6(a) shows raw transient measurements before data processing; in the inset of (b), the ΔV^R_{th} contribution is removed so that the initial charge on the floating gate be the same for all the stress/cycling conditions. In the inset of (a), the $V_{th} - \Delta V^R_{th}$ is plotted as a function of the effective program time, measured starting from the instant where the threshold voltage is V^{Eeff}_{th}. The program efficiency degradation can be identified in the latter plot as the threshold voltage variation ΔV^{Peff}_{th} that cannot be restored after a given arbitrary program time $t_P = 4\mu$s:

$$\Delta V^{Peff}_{th}(n_{cycles}) = V_{th}(t_P, n_{cycles}) - \Delta V^R_{th}(n_{cycles}) - V_{th}(t_P, 0) \tag{3}$$

The program efficiency is thus sensibly affected by device degradation after stress and plays an important role on the window dynamics.

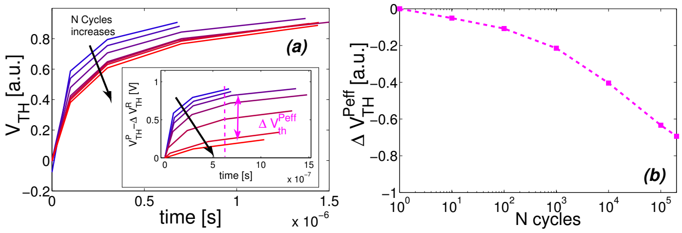


Fig. 6. Extraction methodology to determine ΔV^{Peff}_{th} from program transient measurements of V_{th} vs. programming time. (a) Raw transient program measurements performed after a different number of cycles. In this case normalization has to be performed removing the contribution ΔV^R_{th} of filled traps (inset). (b) ΔV^{Peff}_{th} vs. n_{cycles} representing the degradation of program efficiency.

The physical interpretation of the P/E efficiency degradation is illustrated in Figure 7, where a Poisson-Schroedinger simulator including a multiphonon-assisted trapping model [7] has been used to evaluate the band structure, the electrostatics and the tunneling current through TOX. Figure 7 shows the difference of band structure and gate current, in inversion/programming regime ($V_{CB} \geq 10$V) for an erased cell when acceptor like (0/-) traps are taken into account. The presence of trapped charges delays inversion and strongly reduces the tunneling current, causing program efficiency degradation in both FN and CHEI regimes. As the CHEI current is also dependent on the channel current I_{DS}, the CHEI efficiency is more sensitive to degradation.

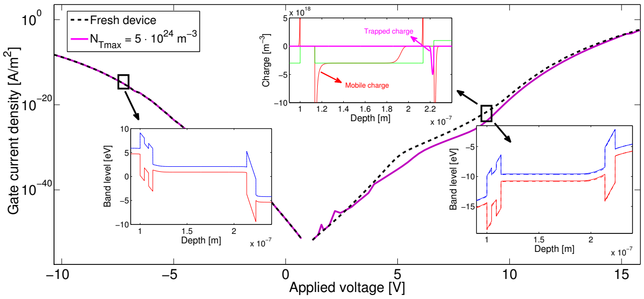


Fig. 7. Gate leakage current of flash devices evaluated with an advanced PS simulator and band diagrams in accumulation/erase regime (left inset) and inversion/program regime (right inset). The charge distribution in inversion is also plotted near the floating gate region and the trapped charge in proximity of the SiO$_2$/Si interface can be identified.

The semi-analytical model has been used to confirm experimental results of P/E performance degradation. Figure 8 shows the effects of the trap concentration in the oxide for FN and CHEI program transient operations, respectively. Finally, 3D TCAD simulations have been performed taking into account realistic dopant profiles in the cell using a commercial modeling tool [12], to study degradation phenomena and identify the defect localization. e^- tunneling in FN regime is concentrated on the floating-gate area overlapping with the substrate side wall near the overetched region of the STI edge (*divot* - Figure 9(a)), while CHE injection mainly occurs from the Lightly Doped Drain region (LDD - Figure 9(b)).

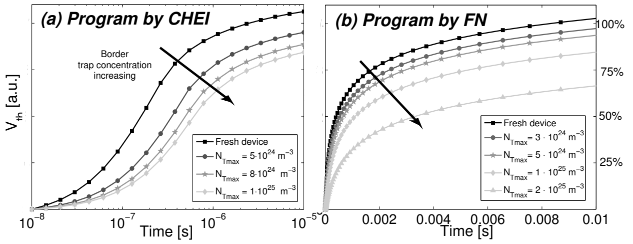


Fig. 8. Transient simulations showing the effects of device degradation on CHEI (a) and FN (b) program efficiencies for different trap concentrations. In both cases, uniform Gaussian trap distributions along the channel are placed at the $\mathrm{Si/SiO_2}$ interface. In (a), the gain is reduced due to the presence of interface traps decreasing the channel current, the electron distribution and consequently the injection current. In (b), trap filling causes inversion to be delayed, reducing the tunneling current.

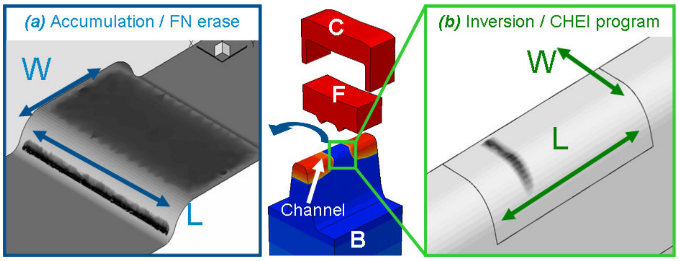


Fig. 9. 3D TCAD simulations to identify the points subject to high electrical stress during in accumulation/erase (a) and inversion/program (b). In the former case the current flows through the floating gate divot region. In the latter case, the current is concentrated in the LDD region where electrical field and injection are higher. Darker areas identify regions with higher current densities.

The proposed extraction method can be applied to the physical understanding of endurance characteristics: several process splits are considered, showing a large spread in W. Figures 10(a-b) group the results of process variations where the LDD implant characteristics are varied, while Figures 10(c-d) show the results of devices, where the oxide formation process is varied. The window W variations and the baseline trends for both the subsets are presented in Figure 10(b-d). Figure 11 shows the correlation between the V_{th} of the dummy device (V_{th}^{TREQ}) after 100s stress in both FN and CHEI and the cell window after 200k cycles. The tendencies evidence both the correlation between LDD splits and CHEI degradation, and TOX splits and erase efficiency.

VI. Conclusion and Perspectives

Transient and endurance characteristics in embedded flash memory devices have been characterized and modeled using a physical extraction methodology and a novel semi-analytical approach. The proposed model enables to study the effects of electrical stress and the role of border traps on the device electrostatics and transient characteristics. The role of traps and their impact on DC characteristics and P/E efficiency has been decoupled and the extraction methodology applied for endurance analysis. The model can be adopted by technology development teams, studying the impact of structure morphology and process variations, as well as by IC designers, engineering the decoding and voltage multiplying circuits to achieve the final product performances.

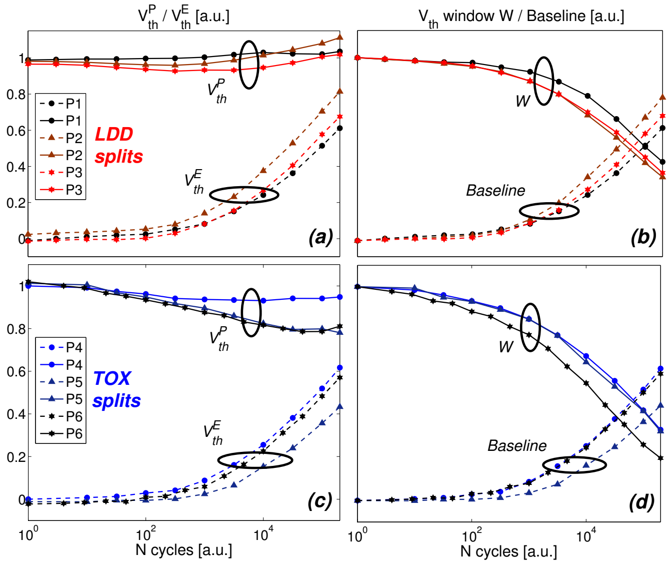


Fig. 10. Endurance characteristics for different process splits on oxide growth and LDD doping.

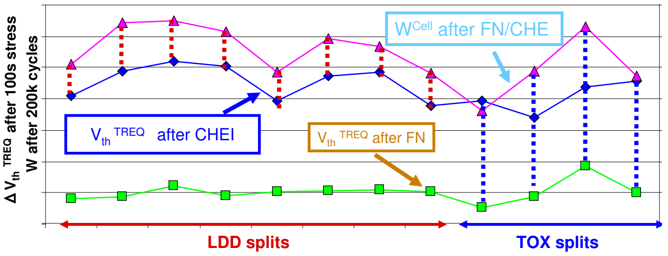


Fig. 11. Correlation between the threshold voltage V_{th} of stressed dummy devices (V_{th}^{TREQ}) and the flash cell window W after 200k cycles. Dummy devices have been stressed using both FN and CHEI stress. These results show how the oxide process variations influence V_{th}^{TREQ} after FN stress, while LDD process variations affect V_{th}^{TREQ} after CHEI stress. In both cases, the variation on V_{th}^{TREQ} is reflected on the window of the stressed flash device.

References

[1] A. Kolodny *et al.*, *Electron Devices, IEEE Transactions on*, vol. 33, no. 6, 835–844, 1986.

[2] A. Gehring *et al.*, *Journal of applied physics*, vol. 92, no. 10, 6019–6027, 2002.

[3] J. Brews, D. Kahng, Ed. Academic Press, 1981.

[4] R. Tsu *et al.*, *Applied Physics Letters*, vol. 22, 562, 1973.

[5] A. Zaka *et al.*, *To be published in proceedings of the ICMTS*, 2011.

[6] D. Garetto *et al.*, in *13th International Nanotech Conference and Expo 2010*, 2010.

[7] ——, in *Inproceeding for IWCE 2010 conference*, 2010.

[8] ——, in *Proceedings of ESSDERC 2009 - fringe poster session*, 2009.

[9] ——, in *Technical Proceedings Workshop on Compact Modeling (WCM)*, 2009.

[10] K. Brower, *Physical Review B*, vol. 42, no. 6, 3444–3453, 1990.

[11] D. Fleetwood *et al.*, *JAP*, vol. 73, no. 10, 5058–5074, 1993.

[12] Synopsys, *SProcess / SDevice*, release Z-2010.03.