Sensitivity-based Investigation of Threshold Voltage Variability in 32-nm Flash Memory Cells and MOSFETs

Valentina Bonfiglio[1], *Giuseppe Iannaccone*[1,2]

[1]Dipartimento di Ingegneria dell'Informazione and [2]SEED Centre, PUSL, University di Pisa.

Email: {valentina.bonfiglio, g.iannaccone}@iet.unipi.it

***Abstract*— We investigate variability of a 32 nm flash memory cell and of 32 nm MOSFETs with a methodology based on sensitivity analysis performed with a limited number of TCAD simulations. We show that - as far as the standard deviation of the threshold voltage is concerned - our method provides results in very good agreement with those from three-dimensional atomistic statistical simulations, with a computational burden that is orders of magnitude smaller. We show that the proposed approach is a powerful tool to understand the role of the main variability sources and to explore the device design parameter space.**

INTRODUCTION

Our intention in this paper is to assert that sensitivity analysis coupled with appropriate modelling can provide qualitative and quantitative information on variability in nanoscale Flash memories and in nanoscale MOSFETs.

Let us stress the fact that variability is even more critical for flash memory cells than for transistors. Indeed, volatile memory fabrication processes undergo even more aggressive scaling than CMOS technology for logic applications, as a means to increase bit density in response to the evolving demands of multimedia applications and mass storage. This exacerbates the device variability issue, which is especially acute in the case of multi-bit cells, where only few tens of electrons in the floating gate can separate two different logic levels [1].

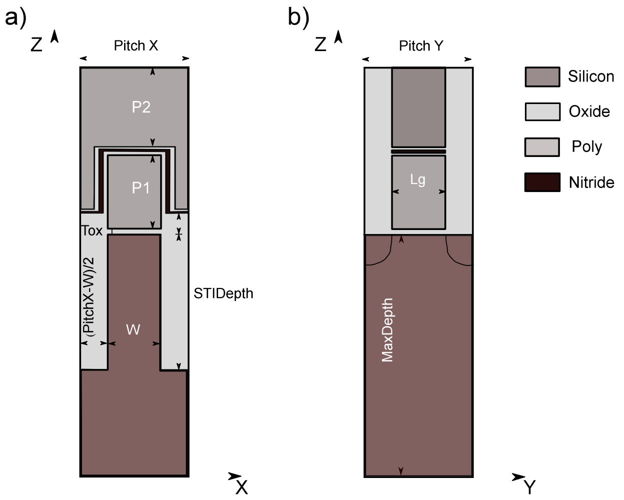
The problem is particularly severe because floating gate cells must be designed and characterized for more than eight standard deviations, and therefore the second order moment of the probability distribution is hardly sufficient [2]. Though this last issue is not addressed in the present work, we are aware that such consideration should inform any discussion on device variability.

In recent years, simulation studies of variability have typically relied on 3D atomistic statistical simulations, such as [3]. However, here we show that a recently proposed method based on TCAD-enabled sensitivity analysis [4-5], allows us to accurately compute the standard deviation of electrical quantities - such as the threshold voltage – with a reduction in computational cost of more than two orders of magnitude.

In the framework of the ENIAC Joint Undertaking MODERN project [6], we have considered a template device structure for a 32 nm CMOS flash memory cell, for which variability assessments based on three-dimensional atomistic statistical simulations and the impedance field method have been published [3]. We analyse the impact of variability sources such as random dopant distribution (RDD) [7], line-edge roughness (LER), line-width roughness (LWR), [8-9] interface trapped charge (ITC) [10], oxide thickness fluctuations (OTF) [11]. We also consider variability in 32 nm MOSFETs fabricated by STMicroelectronics, for which results from experiments and atomistic statistical simulations have been published [12].

The template device structure on the Flash Memory is illustrated in Figure 1. It is a simplified polysilicon floating gate device with dimensions typical of a 32 nm technology (indicated in the table in Figure 1), generated at the crossing point of two orthogonal lines of width 32 nm. Control gate and floating gate consist of polysilicon and are separated by an ONO (oxide-nitride-oxide) layer of 4-3-5 nm. The tunnel oxide thickness is 8 nm. Substrate is boron doped (2 x 10^{18} cm^{-3}), and arsenic doping of source and drain is symmetric with a maximum of 10^{20} cm^{-3}, Gaussian shape, and junction depth of 25 nm. Additional details are available in Ref. [3].

The 32 nm MOSFETs considered in our simulations are designed and fabricated by STMicroelectronics: they are characterized by a retrograde channel doping profile with halo implants, high-K metal-gate stack with hafnium-based gate dielectric, TiN metal gate stack, and by the use of strain for mobility enhancement. Some more details can be found in Ref. [12]. Simulations have been performed using calibrated TCAD input files from STMicroelectronics.



Geometrical parameters	*Symbols and dimensions of layers*	
	Symbol	Value
Cell X dimension	Pitch X	64 nm
Cell Y dimension	Pitch Y	64 nm
Active Area Width	W	32 nm
Gate Length	Lg	32 nm
Silicon substrate thickness	MaxDepth	0.5 um
Isoltation Depth	STIDepth	0.2 um
Tunnel oxide thickness	Tox	8 nm
	StepH	20 nm
Poly1 thickness	P1	70 nm
Poly2 thichness	P2	100 nm
ONO bottom oxide	ONO_bot	4 nm
ONO nitride	ONO_nit	3 nm
ONO top oxide	ONO_top	5 nm
Junction depth	xj	25 nm

Figure 1: Device structure and geometrical parameters of the template 32 nm flash memory under investigation.

METHODOLOGY

The approach proposed is described in detail in [5]. First, all process and geometry variability causes are expressed in terms of a set of synthetic independent variability sources. Then, TCAD-based sensitivity analysis is used to evaluate the contribution to the dispersion of electrical parameters (e.g. the threshold voltage V_{th}) of each independent source. This step is based on the assumption that the effect of each source is sufficiently small that first-order linearization is applicable. Also in the case of the 32 nm Flash memory [3], the variance of the threshold voltage due to combined effect computed with 3D atomistic statistical simulations is shown to be very close to the sum of the variances due to individual effects, giving us confidence in the linear approximation.

As an example, let us consider the case of LER, considering the illustration in Fig. 2, where the 32 nm device is shown with the y axis running along the channel length direction, the z axis perpendicular to the device plane and the x axis running along the channel width.

We can translate line edge roughness in terms of the dispersion of the average position of both gate edges along the y axis (y_1 and y_2, where $\langle y_1 \rangle = 0$ and $\langle y_2 \rangle = L$). This basically means that the impact of line-edge roughness appears just in terms of gate length dispersion. For the sake of simplicity, we further assume that parameters y_1, y_2 are only affected by LER and are physically independent. In the case of very short gate length or narrow fin this assumption can be removed by just considering cross-correlation between y_1 and y_2. The average edge position is a random function $g(x)$ with zero mean value and Gaussian autocorrelation $r(d) \equiv \langle g(x) g(x+d) \rangle$ characterized by correlation length Λ_L and mean square amplitude Δ_L, i.e.:

$$r(d) = \Delta_L^2 e^{-\frac{d^2}{2\Lambda_L^2}} \qquad (1)$$

from which we can write the variance of g as

$$\sigma_g^2 \equiv \langle g^2 \rangle = \frac{1}{W^2} \left\langle \int_0^W g(x_1) dx_1 \cdot \int_0^W g(x_2) dx_2 \right\rangle. \qquad (2)$$

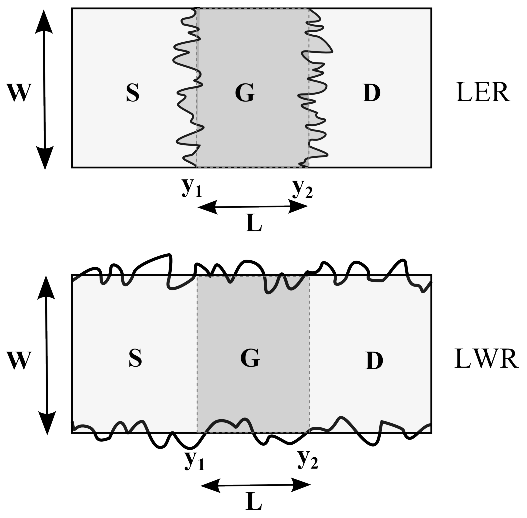


Figure 2: Illustration of the approach to the evaluation of line edge roughness (above) and line-width roughness (below).

If we compute (2) considering (1) and the definition of *r* we find:

$$\sigma_{LER}^{2} = \frac{2\Delta_{L}^{2}\Lambda_{L}}{W^{2}}\left[\Lambda_{L}\left(e^{-\frac{W^{2}}{2\Lambda_{L}^{2}}}-1\right)+\sqrt{\frac{\pi}{2}}W\operatorname{erf}\left(\frac{W}{\sqrt{2}\Lambda_{L}}\right)\right]. \qquad (3)$$

The variance of V_{th} due to line edge roughness is:

$$\sigma_{V_{th}LER}^{2} = \left(\frac{\partial V_{th}}{\partial y_{1}}\right)^{2}\sigma_{y_{1}}^{2}+\left(\frac{\partial V_{th}}{\partial y_{2}}\right)^{2}\sigma_{y_{2}}^{2} = 2\left(\frac{\partial V_{th}}{\partial L}\right)^{2}\sigma_{LER}^{2}, \qquad (4)$$

where y_{1}, y_{2} in (4) are the average gate edges indicated in Fig. 2. All required derivatives can be computed with TCAD sensitivity analysis as illustrated in Fig. 3 (left). The very same approach can be used for LWR.

In the case of OTF we must consider surface roughness with a two dimensional Gaussian autocorrelation

$$r(x_{a}, y_{a}, x_{b}, y_{b}) = \Delta_{S}^{2}\exp\left(-\frac{\left(x_{b}-x_{a}\right)^{2}+\left(y_{b}-y_{a}\right)^{2}}{2\Lambda_{S}{}^{2}}\right), \qquad (5)$$

characterized by correlation length Λ_{S} and mean square amplitude Δ_{S}, which corresponds to a variance of the average position of the interface:

$$\sigma_{OTF}^{2} = \frac{2\pi\Lambda_S^2\Delta_S^2}{L^2W^2}\left[L\cdot\operatorname{erf}\left(\frac{L}{\sqrt{2}\Lambda_S}\right)+\sqrt{\frac{2}{\pi}}\Lambda_S\left(e^{-\frac{L^2}{2\Lambda_S^2}}-1\right)\right] \times\left[W\operatorname{erf}\left(\frac{W}{\sqrt{2}\Lambda_S}\right)+\sqrt{\frac{2}{\pi}}\Lambda_S\left(e^{-\frac{W^2}{2\Lambda_S^2}}-1\right)\right] . \quad (6)$$

The variance of the threshold voltage due to OTF is therefore

$$\sigma_{V_{th}OTF}^{2} = \sum_{m}\left(\frac{\partial V_{th}}{\partial s_m}\right)^2\sigma_{OTF}^{2}, \quad (7)$$

where s_m are all positions of the interfaces between dielectric layers and between dielectric and conducting or semiconducting layers. Also in this case, all derivatives can be computed with TCAD simulations following the example of Figure 3 (right).

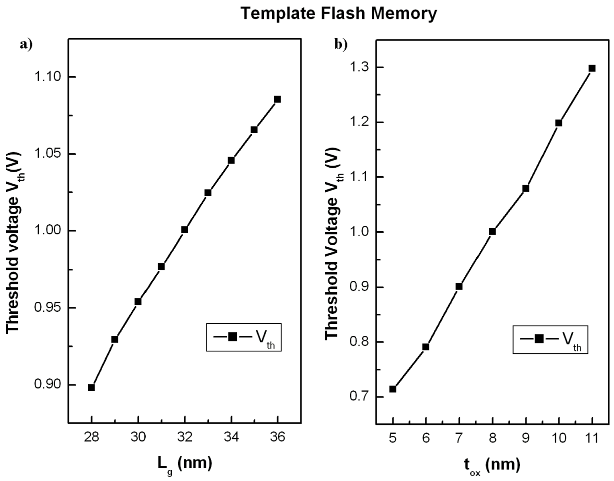


Figure 3 a) Threshold voltage as a function of gate length Lg and b) threshold voltage as a function of tunnel oxide thickness t_{ox} for the template Flash Memory as computed from TCAD simulations.

For random discrete dopants (RDD) [7] and interface trapped charge (ITC) [11], we adopt an approach based on a propagator with a very coarse granularity, which is in principle very close to the concept of impedance field method [13]. Let us note that for the 32 nm MOSFETs, as described in [6], we can simply perform 2D simulations, since the nominal device has translation symmetry in the *x* direction. On the other hand, for the Flash memory cells we have to perform 3D simulations, because the nominal structure is inherently three-dimensional.

For a given variation of doping concentration $\Delta N_A(x,y,z)$ with respect to the nominal value we can write the following expression for the variation of V_{th}:

$$\Delta V_{th} = \int K(x,y,z)\Delta N_A(x,y,z)dxdydz \tag{8}$$

where $K(x,y,z)$ has the role of a propagator. The expression requires the linearity assumption to hold.

To conveniently compute the propagator K, we can assume that K is a smooth function of x, y, and z, and move from the continuum to a discrete space, partitioning the active area in small boxes. Now we can write:

$$\Delta V_{th} = \sum_i \Delta V_{th_i} = \sum_i K_i \Delta N_i \tag{9}$$

The sum runs over all boxes, ΔN_i is the variation of the number of dopants in box i, and ΔV_{th_i} is the threshold voltage variation if only dopants in box i are varied.

In practice, we multiply doping in box i by a factor $(1+\alpha)$ and compute ΔV_{thi} with TCAD simulations. Therefore we have

$$\begin{aligned}\Delta N_i &= \alpha N_i \\ \Delta V_{th_i} &= \alpha K_i N_i\end{aligned}, \tag{10}$$

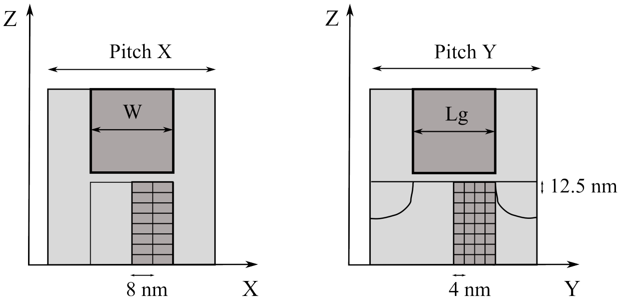
so that (9) becomes,

$$\Delta V_{th} = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha}\right)\alpha = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha}\right)\frac{\Delta N_i}{N_i} \quad . \tag{11}$$

If we finally assume that doping variations in different boxes are independent Poisson processes, we can write

$$\sigma^2_{V_{th}RDD} = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha}\right)^2 \frac{1}{N_i} = \sum_i \sigma^2_{V_{th}RDD}{}^{[i]}. \qquad (12)$$

The threshold voltage dispersion due to RDD only requires a single TCAD simulation for each box, and an integral of the doping profile in each box. To evaluate the most convenient level of granularity in device partitioning, we have made tests with different box sizes, as reported in the table in Figure 4.

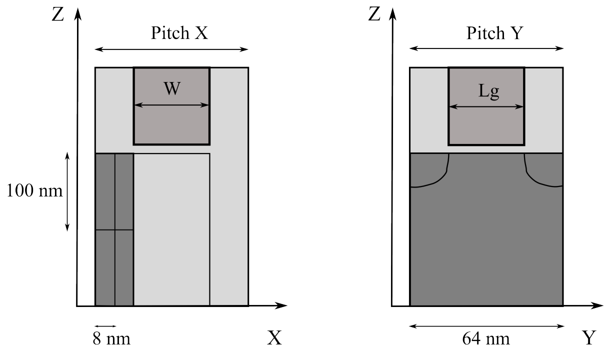


Box size (x, y, z) nm^3	8 x 8 x 25	8 x 4 x 25	4 x 8 x 25	8 x 8 x 12.5	8 x 4 x 12.5
$\sigma_{V_{th}}$(mV)	137	159	137	153	156

Figure 4 - Top: transversal (left) and longitudinal (right) device cross-sections for the assessment of the proper box partitioning. Bottom: computed standard deviation of the threshold voltage as a function of the box size for different choices of the partition.

We have evaluated that a partition of the three dimensional silicon body in 64 boxes of size $8 \times 4 \times 12.5$ nm^3 represents a good trade-off between computing time and accuracy. Considering that we can exploit the symmetry of the structure also along the transport direction at very low drain-to-source voltage, only sensitivities corresponding to 32 boxes must be computed with TCAD simulations.

For ITC, the situation is similar: we assume an average trap density of 5×10^{11} cm^{-2} and partition the tunnel oxide in tiles of $100 \times 8 \times 64$ nm^3, for a total of only four simulations, if the symmetry of the nominal structure is exploited. As can be seen in Figure 5, finer partitions do not lead to a different estimation of the threshold voltage dispersion.



Box size (*x*, *y*, *z*) nm^3	16 x 64 x 200	8 x 64 x 100	16 x 64 x 50	8 x 32 x 50	4 x 32 x 50
$\sigma_{V_{th}}$(mV)	27	59	56	59	59

Figure 5: Region partitioning in boxes $100 \times 8 \times 64$ nm^3 for the evaluation of propagators due to interface-trapped charge. Left: transversal cross section. Right: longitudinal cross section.

RESULTS

Let us first discuss results for the 32 nm flash memory. For the sake of comparison with atomistic simulations [3], we consider for LER and LWR a Gaussian autocorrelation with mean square amplitude $\Delta_L = 1.5$ nm and correlation length $\Lambda_L = 20$ nm. For OTF, we consider a Gaussian autocorrelation with mean square amplitude $\Delta_S = 0.2$ nm and correlation length $\Lambda_S = 18$ nm. For random dopants, as mentioned above we consider a partition of 32 boxes of size $8 \times 4 \times 12.5$ nm^3. For ITC, a partition of only four boxes is used.

The computed standard deviation of the threshold voltage due to each mechanism is compared in Table I with results obtained from 3D atomistic simulations on 1000 samples performed with GARAND [3]. The threshold voltage is defined with a current criterion of 100 nA for a drain-to-source voltage of 100 mV.

Considering that statistical simulations have been performed on ensembles of N=1000 devices, the mean square relative error on the estimated standard deviation of the threshold voltage is $(2N)^{-0.5}$, i.e., 2.2%: all terms lie within or very close to the error bars of statistical simulations.

TABLE 1 STANDARD DEVIATION OF THE THRESHOLD VOLTAGE OF THE 32NM FLASH MEMORY CELL DUE TO SEVERAL MECHANISMS COMPUTED WITH THE METHOD PROPOSED IN [5] AND WITH STATISTICAL SIMULATION IN [3].

σ_{Vth} (mV)	Our method [5]	Atomistic Sim. [3]
LER	46	48
LWR	28	26
OTF	14	14
RDD	156	144
ITC	59	67

As far as the 32 nm MOSFETs are concerned, we assume for LER a Gaussian autocorrelation with mean square amplitude $3\Delta_L = 4$ nm and correlation length $\Lambda_L = 30$ nm as in Ref. [12]. In order to evaluate the effect of RDD, we partition the active area in boxes of 4×4 nm^2. The effect of RDD on threshold voltage is limited to a very small area, as can be seen in the 2D color map of Fig. 6, where we plot the partial local contributions $\sigma^{2\;[i]}_{V_{th}RDD}$ of acceptor doping to the variance of the threshold voltage for the “regular V_{th}” (RVT) nMOSFET. It is pretty clear that a limited part of the active area has an actual impact on threshold voltage dispersion. We can focus only on the red region of the colour plot, and apply to it a fine partition for the accurate evaluation of the threshold voltage dispersion, neglecting contributions from other regions.

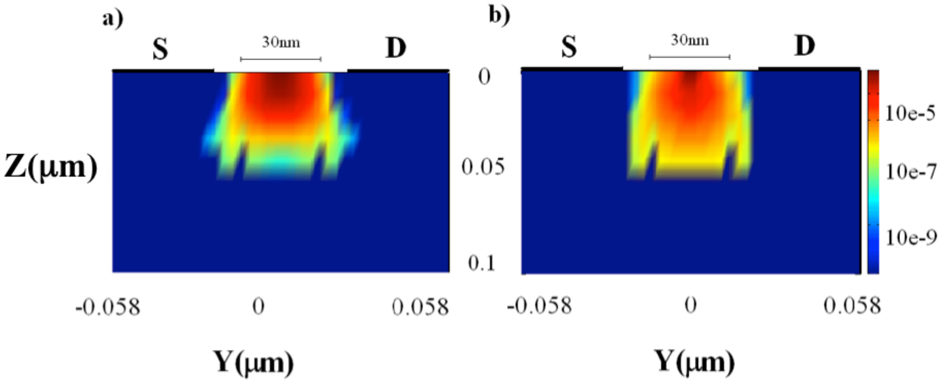


Figure 6: Colour map of the local contributions to the variance of the threshold voltage indicated with $\sigma^{2}_{V_{th}RDD}{}^{[i]}$ as a function of position: effect of the acceptor doping of the 32-nm RVT NMOSFET: a) for $V_{DS} = 50mV$ and b) for $V_{DS} = 1V$.

The same consideration holds for the 32 nm RVT pMOSFET, for which the $\sigma^{2}_{V_{th}RDD}{}^{[i]}$ due to donor doping is plotted as a function of positions in the colour maps of Fig. 7.

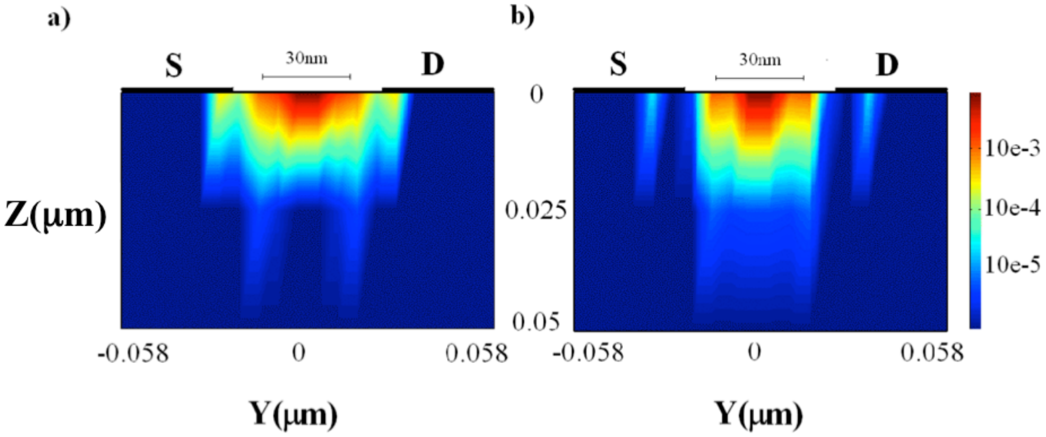


Figure 7: Colour map of the local contributions to the variance of the threshold voltage indicated with $\sigma^{2}_{V_{th}RDD}{}^{[i]}$ as a function of position: effect of donor doping of the 30-nm RVT PMOS: a) for $-V_{DS} = 50mV$ and b) for $-V_{DS} = 1V$.

The standard deviation computed with our method and from statistical simulations in Ref. [12] on an ensemble of 1000 devices is shown in Table II. As in the case of the memory, the standard deviation of the error on the computation of $\sigma_{V_{th}}$ with statistical simulations 2.2%, and the difference between the two methods is mostly within the error bars.

TABLE 2 STANDARD DEVIATION OF THE THRESHOLD VOLTAGE OF THE 32NM RVT NMOSFETS AND PMOSFETS DUE TO LER AND RDD COMPUTED WITH THE METHOD PROPOSED IN [3] AND WITH STATISTICAL SIMULATION FROM REF [12].

Device	V_{DS}(mV)	σ_{Vth} (mV)	Our method [5]	Atomistic Sim. [12]
32 nm nMOSFET	50 mV	RDD	43	44.4
		LER	13	12.7
	1V	RDD	48	49.8
		LER	25	24.9
32 nm pMOSFET	50 mV	RDD	46	42.3
		LER	10	12.8
	1V	RDD	58	54.4
		LER	29	33.3

CONCLUSION

We have proposed a methodology for the quantitative evaluation of the effects of the main mechanisms affecting threshold voltage variability, based on the careful identification of the main independent and relevant physical quantities. Our approach requires the calculation of partial derivatives of V_{th} with respect to device structure parameters, that can be obtained with a very limited number of TCAD simulations. We have shown that in all cases we are able to obtain results in good agreement with 3D atomistic statistical simulations [3] at a much smaller computational cost. We qualify this statement to the second order moment of the threshold voltage distribution, because the proposed approach does not provide information on the far tails of the distribution, which are important especially for large Flash memory arrays, and would require extension of the method to higher order terms. We also can count as a drawback of our proposed method the fact that it requires preliminary device inspection, in order to identify the independent synthetic parameters describing the variability sources.

We assert that our approach has some advantages over statistical modelling, not only because it is orders of magnitude faster, but also because it represents a powerful tool for understanding the impact of individual factors and to efficiently explore the design space using tools already available and routinely used by semiconductor technology developers.

Acknowledgments

This work has been supported by the ENIAC JU project 12003 MODERN awarded to IUNET.

References

[1] A. Calderoni, P. Fantini, A. Ghetti, A. Marmiroli, "Vth fluctuations in nanoscale floating gate memories, Proc. SISPAD, Sept. 9-11, 2008, pp. 49-52.

[2] A. Spessot, A. Calderoni, P. Fantini, A. S. Spinelli, C. Monzio Compagnoni, F. Farina, A. L. Lacaita, and A. Marmiroli, “Variability effects on the VT distribution of nanoscale NAND Flash memories,” in Proc. IRPS, 2010, pp. 970–974.

[3] G. Roy, A. Ghetti, A. Benvenuti, A. Erlebach, and A. Asenov, “Comparative Simulation Study of the Different Sources of Statistical Variability in Contemporary Floating-Gate Nonvolatile Memory,” IEEE Transactions on Electron Devices, vol. 58, no. 12, pp. 4155-4163, Dec. 2011.

[4] V. Bonfiglio and G. Iannaccone, “Analytical and TCAD-supported Approach to Evaluate Intrinsic Process Variability in Nanoscale MOSFETs”, Proceeding ESSDERC 2009, pp. 193-196, Athens 2009.

[5] V. Bonfiglio and G. Iannaccone, “An Approach Based on Sensitivity Analysis for the Evaluation of Process Variability in Nanoscale MOSFETs,” IEEE Transactions on Electron Devices, vol. 58, no. 8, pp. 2266-2273, 2011.

[6] Deliverable D.2.2.3 of the ENIAC MODERN Project, 2011.(Project website: www.eniac-modern.org).

[7] H.-S. Wong, Y. Taur, “Three-dimensional “atomistic” simulation of discrete random dopant distribution effects in sub-0.1 mm MOSFETs”, Tech. Dig. IEDM 1993, pp. 705-708, 1993.

[8] A. Asenov, S. Kaya, and A. R. Brown, “Intrinsic parameter fluctuations in decananometre MOSFETs introduced by gate line edge roughness,” IEEE Trans. Electron Devices, vol. 50, no. 5, pp. 1254–1260, May 2003.

[9] Ji-Young Lee, Jangho Shin, Hyun-Woo Kim, Sang-Gyun Woo, Han-Ku Cho, Woo-Sung Han, and Joo-Tae Moon, “Effect of line-edge roughness (LER) and line-width roughness (LWR) on sub-100-nm device performance”, Proc. SPIE 5376, 426 (2004).

[10] A. Asenov, S. Kaya and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," IEEE Transactions on Electron Devices, vol. 49, pp. 112–119, 2002.

[11] C. L. Alexander, A. R. Brown, J. R. Watling, and A. Asenov, “Impact of single charge trapping in nano-MOSFETs—Electrostatics versus transport effects,” IEEE Trans. Nanotechnol., vol. 4, no. 3, pp. 339–344,May 2005.

[12] X. Wang, G. Roy, O. Saxod, A. Bajolet, A. Juge, A. Asenov, "Simulation Study of Dominant Statistical Variability Sources in 32-nm High- κ /Metal Gate CMOS," IEEE Electron Device Letters, vol.33, no.5, pp.643-645, May 2012.

[13] W. Shockley, J. A. Copeland, R. P. James, “The impedance field method of noise calculation in active semiconductor devices”, in Quantum Theory of Atoms, Molecules, and the Solid State, A tribute to John C. Slater. Edited by Per-Olov Loewdin. New York: Academic Press, 1966, p.537.