Università di Pisa

Analytical model for the extraction of the trapped charge distribution in memories based on discrete storage nodes programmed via CHE injection

Luca Perniola
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, Università di Pisa and IMEP-ENSERG, Grenoble
Giuseppe Iannaccone
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, Università di Pisa
Barbara De Salvo
Commissariat à l'Energie Atomique–Laboratory of Electronics, Technology, and Instrumentation, Grenoble
Gérard Ghibaudo
Institut de Microélectronique Electromagnétisme et Photonique–Centre National de la Recherche Scientifique/Institut National Polytechnique de Grenoble
G. Molas
Commissariat à l'Energie Atomique–Laboratory of Electronics, Technology, and Instrumentation, Grenoble
Cosimo Gerardi
Central Research and Development, STMicroelectronics, Catania
S. Deleonibus
Commissariat à l'Energie Atomique–Laboratory of Electronics, Technology, and Instrumentation, Grenoble

L. Perniola, G. Iannaccone, B. De Salvo, G. Ghibaudo, G.Molas, C. Gerardi, S. Deleonibus, *Experimental and theoretical analysis of scaling issues in dual-bit discrete trap non-volatile memories* , International Electron Device Meeting 2005, pp. 877-880, Washington DC 2005.

Experimental and theoretical analysis of scaling issues in dual-bit discrete trap non-volatile memories

L. Perniola[1, 2, 3], G. Iannaccone[1] , B. De Salvo [2], G. Ghibaudo [3], G. Molas[2], C. Gerardi[4], S. Deleonibus[2]

(1)Università di Pisa, Via Caruso 16, 56122 Pisa, Tel. +33 438 786497, Fax +33 438 789456, Italy, luca.perniola@cea.fr
(2)CEA-LETI 17 Av. Des Martyrs, 38054 Grenoble, France (3)IMEP-CNRS/INPG Grenoble, France (4)STMicroelectronics Catane, Italy.

Abstract

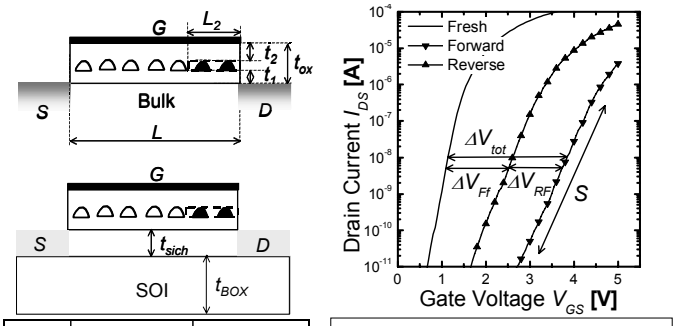
Here we present an experimental and theoretical analysis of dual-bit DT-NVMs. In particular data retention experiments on bulk and SOI silicon nanocrystal memory devices (Fig. 1) and their interpretation through a surface potential based model are shown (4). Our model is then exploited to investigate the main issues posed by dual-bit reading, when the dimensions of bulk and SOI devices are scaled down. We present two different reading schemes for a scaled device and we show that dual-bit performance of DT-NVMs, charged on both sides, is preserved even when the two pockets of charge coalesce. Finally, we conclude that both bulk and SOI dual-bit architectures are promising for memory cells with gate lengths down to 30-50 nm.

Introduction

Discrete trap (DT) non-volatile memories (NVMs), such as NROM, SONOS, and nanocrystal memories provide the particularly attractive opportunity of storing two bits per cell (1)-(3), which accelerates the effective scaling of NVM cells, and allows to effectively cope with the difficulties in scaling down the gate stack of traditional non-volatile memories. Nevertheless the physics of dual-bit devices is still not completely revealed, and in particular the electrical behavior during data retention and its impact on cell scaling are today subjects of interest in the research community.

Essentials of our model

We have developed an analytical model, thoroughly explained elsewhere (4), which is based on the computation of the surface potential of non-uniformly charged cells (Fig. 1). With this model we are able to express the relevant experimentally accessible parameters of DT-NVMs, after a channel hot electron (CHE) writing, (i.e. the programming windows or the subthreshold slope highlighted in Fig. 1), in terms of synthetic quantities describing the pocket of trapped charge (i.e. the "effective" charged length L_2 and the "effective" density of injected charge per unit area Q, see Fig.1).



Device	Bulk	SOI
L / W	280 / 160 nm	100 / 20 nm
t_1 / t_2	5.5 / 8 nm	2.5 / 20 nm
D_{dot}	$3x10^{11}$ cm^{-2} 10^{12} cm^{-2}	$1.5x10^{12}$ cm^{-2}
t_{sich}	/	18 nm
t_{BOX}	/	400 nm
N_a	10^{18} cm^{-3} (p)	10^{18} cm^{-3} (n)

$L_2 << 2\text{-}3\ \lambda$ $\quad \Delta V_{RF} \approx \Delta V_{tot}$
$Q, L_2 \Leftrightarrow (S, \Delta V_{tot})_{exp}$

$L_2 > 2\text{-}3\ \lambda$ $\quad S_{rev} \approx S_{fresh}$
$Q, L_2 \Leftrightarrow (\Delta V_{RF}, \Delta V_{tot})_{exp}$

$L_2 > 4\text{-}5\ \lambda$ $\quad S_{rev} \approx S_{fresh}$, $\Delta V_{RF} \approx 0$

Fig. 1: **(left)** Bulk and SOI nanocrystal memory devices used for experiments. **(right)** The experimental quantities, that in our model provide information on the features of the pocket of injected charge L_2 and Q, are highlighted.

The electrical behavior of DT-NVMs is well known in the literature, but never explained through the surface potential approach. *We show that the dual-bit behavior is intrinsically due to Short Channel Effects (SCEs).* Indeed *dual-bit operation is possible only if the length of the pocket of charge L_2 is comparable to the characteristic length λ* (Fig. 2) which describes the degree of influence of the junctions on the channel potential. If this influence is drastically reduced, i.e. SCEs lowered, we do not see the two-bit typical behavior of DT-NVMs.

With this model we can analyze data retention behavior of DT-NVMs: The localised trapped charges can diffuse in the trapping layer (*charge diffusion*, QxL_2=const.) and/or can leak toward the channel (*charge loss*, thus L_2=const.). We show in Figs. 2-3 that the electrical behavior during data retention depends heavily on the *initial conditions of the pocket of charge,* feature never shown in the literature. With the dimensions specified in Fig. 1, our model predicts that ΔV_{Ff} raises during charge diffusion only if, immediately after the CHE write, $\Delta V_{Ff} \leq 1.2$ *V*. On the other hand, the charge loss process during data retention is apparent when ΔV_{RF}

keeps constant, only if initially $\Delta V_{tot} \geq 1.5\text{-}2\ V$.
Moreover from this model we argue that the threshold voltage V_{th} of a cell with both charged bits is higher than that of a cell with one charged bits *only if* $L_2 \approx L/2$.

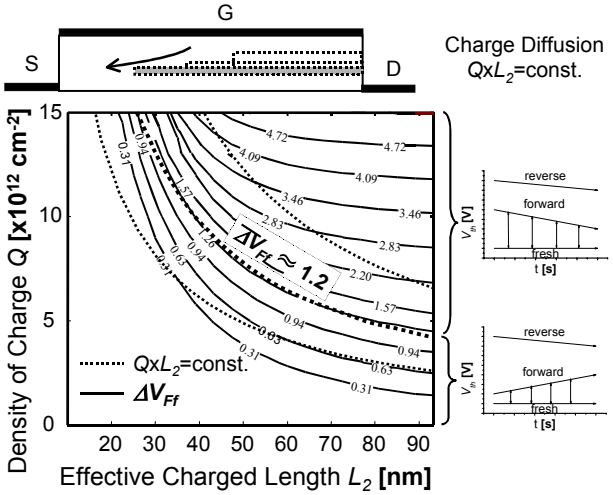


Fig. 2: Charge diffusion electrical behavior extracted by the model. Data retention results, macroscopically different, depend on the initial conditions of the pocket of charge.

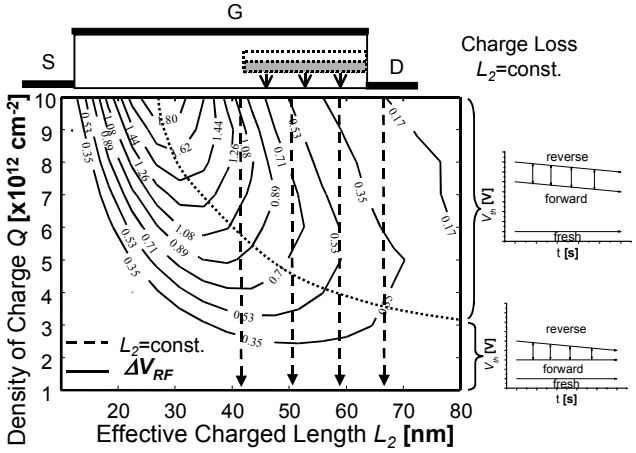


Fig. 3: Charge loss electrical behavior extracted by the model. As in Fig. 2, the data retention behavior depends on the initial condition of the pocket of charge.

Data retention experiments

We performed electrical tests on bulk and SOI silicon nanocrystal devices (Fig. 1) with different dot densities D_{dot}. On bulk devices we performed a complete CHE write analysis, using several values of stressing parameters (Fig. 4). By extracting the charge distribution with our model, we deduce that by raising the stressing time, we inject an increasing amount of charge in the trapping medium, while raising the stressing gate voltage, we inject a smaller pocket of charge (Fig. 4). A lower saturation limit for L_2 seems to be achieved at $L_2 \approx 2t_{ox}$.

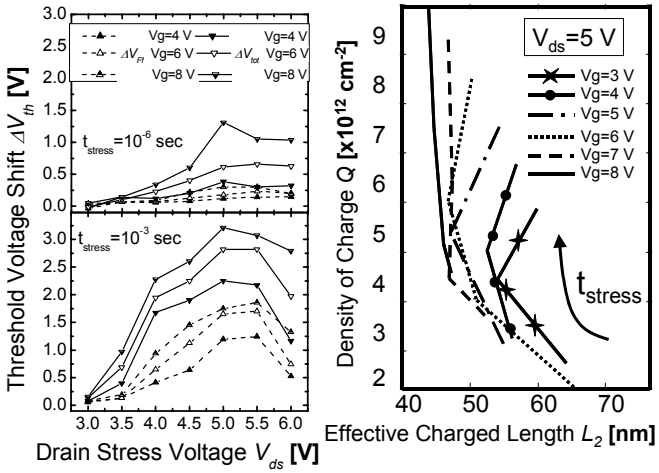


Fig. 4: **(left)** Experiments of CHE write on bulk devices, **(right)** fitted by our model, to extract L_2 and Q for the pocket of charge.

Several data retention experiments were performed on the wafers with bulk devices, at different temperatures (Fig. 5). With the help of our model, we found that at T=150 °C and T=200 °C data retention was affected by *charge diffusion* (less in the wafer with lower dot density), while at T=250 °C the competing effect of *charge loss* becomes evident (Fig.5). As it is clear from Fig. 5 and as highlighted in Ref. (5), the fact that ΔV_{Ff} raises during data retention is a sufficient condition to state the existence of a charge diffusion process. However, a raising ΔV_{Ff} is not necessary for charge diffusion: indeed in Fig. 6 we show a data retention at T=150°C, in charge diffusion regime, where ΔV_{Ff} is initially 2.06 V and then decreases during the experiment. This example stresses the importance of both the *initial conditions* of the injected pocket of charge, and of the *test temperature* as far as the electrical behavior during data retention is concerned.

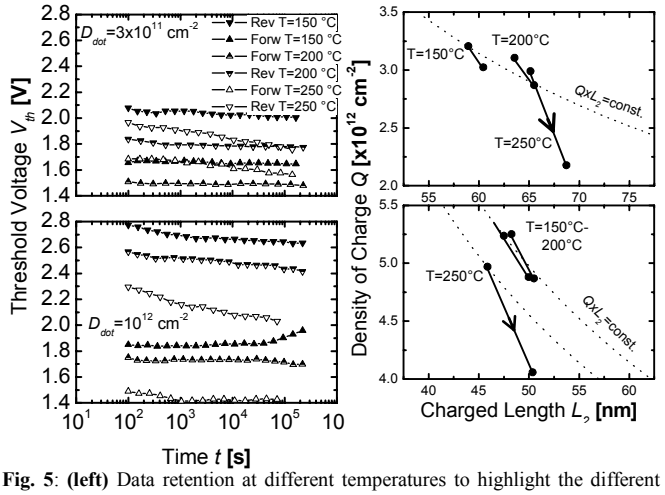


Fig. 5: **(left)** Data retention at different temperatures to highlight the different electrical behaviors. **(right)** Extraction of L_2 and Q from data, which tells that charge diffusion is apparent at T=150-200 °C; charge loss at T=250 °C.

An alternative possibility for improving scaling

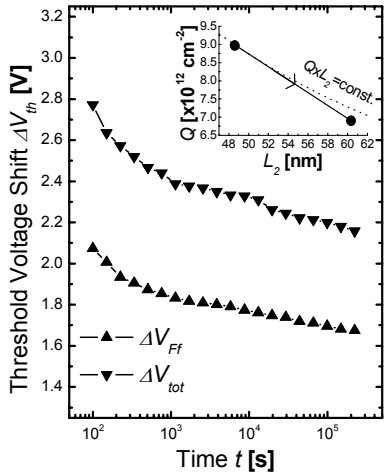


Fig. 6: Data retention at T=150 °C. If ΔV_{Ff}>1.2 V, during the charge diffusion ΔV_{Ff} lowers, contrarily to data retention, at the same temperature, highlighted in Fig 5.

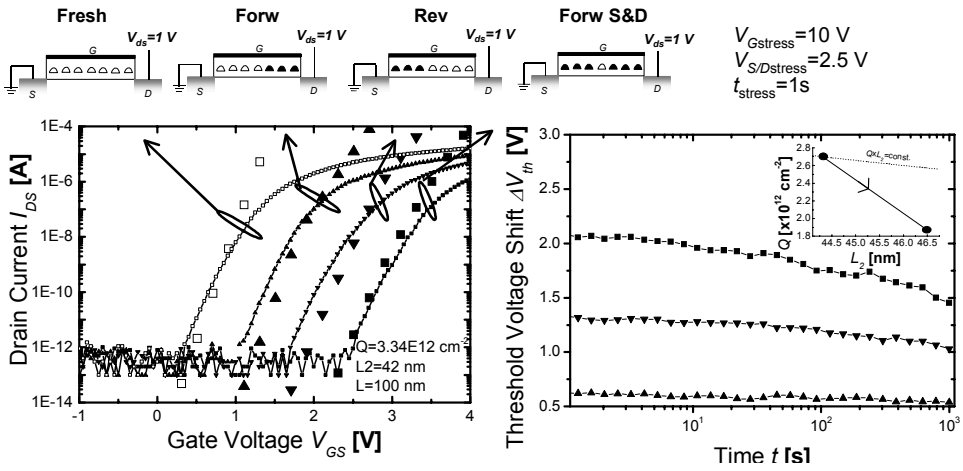


Fig. 7 : Transfer characteristics of a scaled SOI device charged consecutively on drain, source and on both sides with the same stressing conditions. Four clear states are apparent also if the two pockets of charge are very close to one another. The data retention at room temperature is poor, due to the very thin t_1. Indeed from our model (see the inset), we see that data retention is affected mostly by charge loss.

perspectives could be represented by SOI architectures. In Fig. 7 we analyzed the $I_D(V_{GS})$ of a scaled SOI cell with L=100 nm, charged consecutively via the CHE procedure near the drain, near the source, and near both drain and source. Four states are apparent and with the help of our model, we estimate that L_2=42 nm, which means that the two charge pockets are not well separated when both sides are charged. Nevertheless dual-bit performance remains.

Dual-bit performance analysis

Herewith we show the performance of dual-bit DT-NVMs as a function of cell parameters, as extrapolated from our model. We would like to stress that some cell configurations were simulated with a commercial TCAD tool (6) to validate the analytical model. We found that the error between our model and the numerical results was lower than 20%.

From Fig. 8, we see that the top oxide thickness t_2 has a larger impact than the bottom oxide thickness t_1 on the programming windows, as it is in the case of continuous floating gate Flash devices.

A limitation to the scaling of the cell length L can be represented in bulk devices by the punchthrough between drain and source. If we keep the doping profile uniform, to overcome this issue, we are obliged to raise the doping level of the bulk. In Fig. 9 we see that for N_a=3x10^{17}cm^{-3} we can reduce L down to 170 nm, and only with N_a=3x10^{18}cm^{-3} we are able go down to L=50 nm. However we see that the programming windows do not depend on the cell length, as far as $L \gg L_2$.

In Fig. 10 we stress the importance of SCEs on dual bit performance. As it appears from the figure, raising the level of the doping in the bulk (reducing the SCEs), ΔV_{RF} lowers: the asymmetry of the threshold voltages disappears and the device starts to behave as a one-bit NVM. Understanding the electrical behavior of a scaled DT-NVM device is of paramount importance to decide the suitable reading scheme. In Fig. 11 if $L_2 \ll L$, with one sense level we can read the information of each bit, because the reading is sensitive to the charging state of the bit near the low-voltage junction.

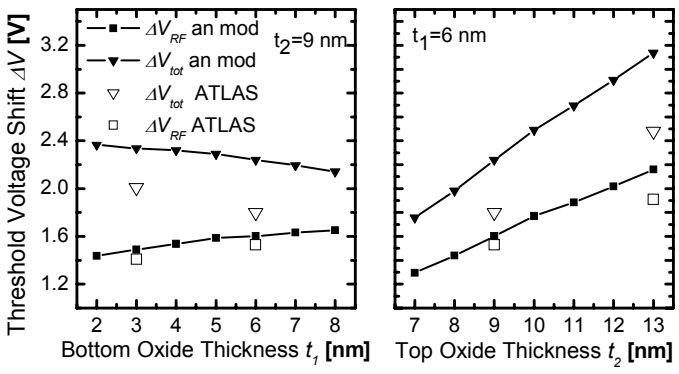


Fig. 8 : Dependence of the programming windows on the gate stack thickness as from our model and from numerical simulations.

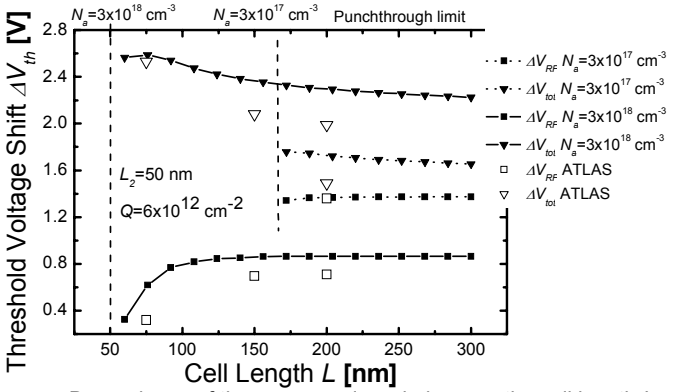


Fig. 9 : Dependence of the programming windows on the cell length L.

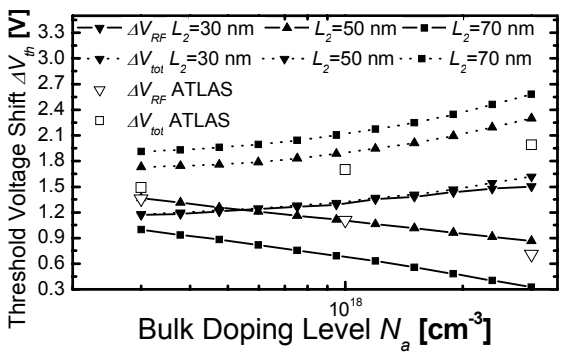


Fig. 10 : Dependence of the programming windows on N_a of the bulk, as from our model and from numerical simulations. Raising N_a (i.e. lowering SCEs), dual-bit performance is degraded ($\Delta V_{RF} \rightarrow 0$).

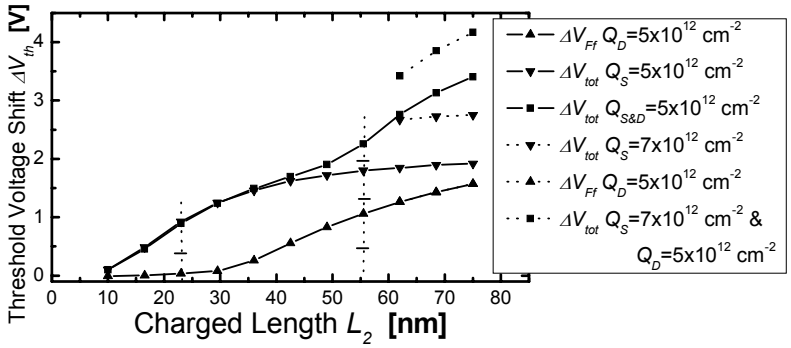


Fig. 11 : Dependence of the programming windows on the cell length L for a cell with L=100 nm. One sense level scheme is suitable for small pockets of charge, while three sense level scheme is suitable for pockets of charge close to one another.

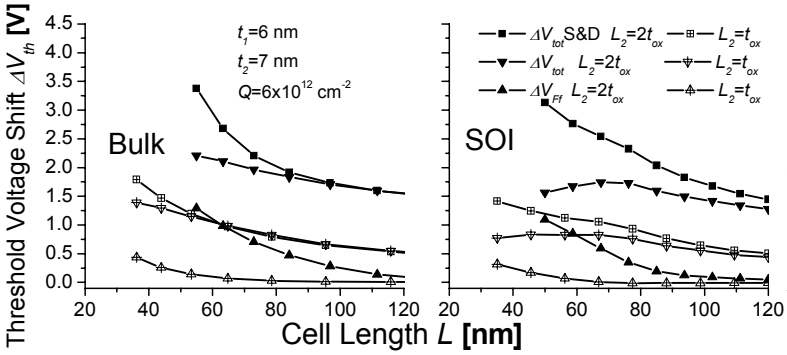


Fig. 12: Bulk and SOI dual-bit DT-NVMs scaling perspectives.

We assumed t_1=6 nm and t_2=7 nm, due to the stringent retention requirements for future cells. Concerning Q, after experiments shown in Fig. 4, we believe to adjust the Q value to our suits, changing the stressing time.

In the case of bulk devices, while scaling the cell length, we avoid punchthrough raising the value of the doping level of the bulk (N_a=7x10^{18} cm^{-3} for L=35 nm); while for SOI devices, scaling the cell length, we keep unaltered the aspect ratio of the cell, proportionally reducing the silicon film thickness (5 nm for L=35 nm).

Two limits for L_2 are investigated: L_2=2t_{ox} (as shown by Fig. 7), or L_2=t_{ox}, as demonstrated in Ref. (7) for NROM devices. *Both architectures maintain dual-bit behavior with gate lengths down to 30-50 nm.*

The charging state of the opposite bit is assured by swapping the voltages of source/drain, i.e. with bias reversal. When $L_2 \approx L/2$, the cell becomes sensitive to the 2 bits with no bias reversal, at the cost of three sense levels. If the coalescence of the charge pockets is significant, we have $\Delta V_{RF} \approx 0$, the four states coalesce in three, thus we loose the dual-bit behavior. Let us stress the fact that one possibility to obtain again four well separated charged states is represented by injecting different amounts of charge near the two junctions, as shown in Fig. 11.

Scaling perspectives

In Fig. 12, we show the behavior of the four V_{th}s with respect to the fresh V_{th}, when the cell length is scaled down, in the case of bulk and SOI devices. The SOI architecture seems to offer no clear advantage with respect to bulk devices, due to the fact that normally it keeps limited the SCEs, thus degrading the performance of DT-NVMs. We conclude that, in the case of $L_2 \approx t_{ox}$, we are able to reduce L down to 35 nm, using a three level sensing scheme and no bias reversal for both the bulk and SOI architecture.

Conclusion

Among the main results achieved in this work, we highlight the following conclusions:

- SCEs are vital for dual-bit behavior in DT-NVMs.
- The electrical behavior of DT-NVMs depend on the test temperature as well as on the initial conditions of the pocket of charge.
- Dual-bit performance is maintained in the case of ultra-scaled NVMs charged on both sides, with overlapping pockets of charge.
- Bulk and SOI architectures are both valuable candidates for future DT-NVMs.
- 2-bit information is possible in devices with gate lengths as short as L=35 nm.

References

(1) B.Eitan et al., "NROM: a novel localized trapping, 2-bit nonvolatile memory cell", *IEEE El. Dev. Lett.*, Vol. 21, No. 11, pp. 543-545, November 2000.

(2) A. Shappir et al., "The two-bit NROM reliability", *IEEE Trans. on Dev. Mat. and Rel.*, Vol. 4, No. 3, pp.397-403, September 2004.

(3) B. De Salvo et. al., "How far will silicon nanocrystal push the scaling limit of NVMs technologies ?", *Tech. Dig of IEDM*, pp. 597-600, 2003.

(4) L. Perniola et al., "Analytical model of the effects of a nonuniform distribution of stored charge on the electrical characteristics of discrete-trap non-volatile memories", *IEEE Trans. on Nanotech.*, Vol. 4, No. 3, pp. 360-368, May 2005.

(5) T. Sugizaki et al., "Novel multi-bit SONOS type Flash memory using a high-k charge trapping layer", *Symp. on VLSI Tech., Dig. of Tech. papers*, pp. 27-28, 2003.

(6) *SILVACO-ATLAS User's Manual*, Vol. I-II, SILVACO Int., Santa Clara CA, 1998.

(7) E. Lusky et al., "Investigation of Channel Hot Electron injection by localized charge-trapping nonvolatile memory devices", *IEEE Trans. on El. Dev.*, Vol. 51, No. 3, pp. 444-448, March 2004.