Scalability Study of Floating Body Memory Cells

Andreas Schenk
Integrated Systems Laboratory
ETH Zurich
Gloriastr. 35, CH-8092 Zürich, Switzerland
Email: schenk@iis.ee.ethz.ch

***Abstract*—This paper presents a TCAD study on the scalability of impact-ionization based floating-body memory cells to fully-depleted short-channel devices. Only the energy-balance transport model allows for transient simulations of realistic voltage wave forms. To attain qualified predictions, impact ionization rates were calibrated by full-band Monte Carlo simulations. Inclusion of band-to-band tunneling is crucial, although junction profiles were optimized for minimal gate-induced drain leakage. The memory effect is explored in detail for partially-depleted FETs, stressing the difference between two operation modi: the steady-state avalanche mode which makes use of the soft breakdown of the body-drain junction, and the bipolar mode where the collector current of the parasitic bipolar decays in time due to the loss of stored holes in the base. It is shown that in fully-depleted ultra-thin body FETs the common kink effect is absent, instead a similar floating body effect is found in the sub-threshold range which is however suppressed by unavoidable and predominating band-to-band tunneling. Although a large amount of excess holes can be created during WRITE1 and hold relatively steady in the body, this charge is always lost when switching to READ1. At sub-threshold gate voltages, the READ1 current becomes self-determined by band-to-band tunneling, which also fixes the READ0 current to the same value. Above threshold, stored holes are either swept out to the source or their electrostatic impact is negligible compared to the injected charge from the source. No wave form could be found that results in a READ1/READ0 programming window.**

I. Introduction

Floating body memories (FBMs) use a single transistor (1T) and no capacitor bit-cell [1]. They can be built with a few modifications of standard SOI logic processes and have been claimed to be as scalable as CMOS [2]. FBMs of the first generation make use of the V_{T}-shift due to the floating body (FB) effect, second-generation devices have been described as sensing the READ1/READ0 (R1/R0) ratio of the collector current of the parasitic bipolar transistor in the two charge states of the body (= base). Device simulation of FBMs is a challenge: the e-h pair generation process, either impact ionization (II) or band-to-band tunneling (B2BT), is nonlocal and far from equilibrium, defect-assisted tunneling (DAT) plays a role for generation-recombination, the used voltage wave forms involve fast transients, and the FB together with the high generation rates cause severe convergence problems. In an energy-balance (EB or "hydro") transport framework [3] which is the only option for transient simulations, as Monte Carlo (MC) codes are still unable to treat FBMs self-consistently, only well-calibrated generation models can give qualified predictions of the scalability of FBMs.

II. Device description

The devices chosen for this study are two silicon-on-insulator (SOI) double-gate (DG) nFETs with different dimensions (Fig. 1). The first structure is partially depleted

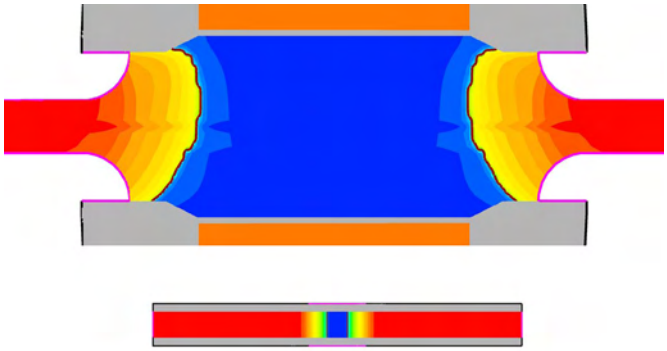

Fig. 1. Structure of the 110 nm (upper) and the 22 nm (lower) SOI DG nFETs plotted to scale.

(PD) and has a gate length of 110 nm, a body thickness of 75 nm, an equivalent oxide thickness (EOT) of 2.5 nm, a body doping of 1e18 cm^{-3}, and state-of-the-art source/drain (S/D) implants. The second device is fully depleted (FD) with a gate length of 22 nm, a thickness of the ultra-thin body (UTB) of 10 nm, an EOT of 1.1 nm, a body doping of 1.2e15 cm^{-3}, and with junction profiles tuned such that the OFF-current ITRS specification for LSTP devices (10 pA/m) [7], [8] is met. The latter profiles were further modified in [6] to minimize B2BT which is the origin for gate-induced drain leakage (GIDL).

III. Simulation approaches

The carrier temperature-driven II rate (see [3]) was calibrated for both devices by full-band MC simulations with the in-house package *SimnIC* [4] at different bias conditions. Although in an EB simulation the shape of the II rate, i.e. the location of its maximum and its spatial distribution, can never be perfectly harmonized with the MC result, the fit of the integral II rate can be assumed to reproduce at least the total amount of generated holes correctly. This is demonstrated in Figs. 2 and 3 for the 22 nm FD UTB FET. The distribution of the II rate in EB is much sharper and the dark space effect ($\approx$8 nm) is almost absent compared to MC. In the EB simulation one can only change the peak height of the II rate, but not the position and shape of the II rate profile.

978-1-4244-3947-8/09/$25.00 ©2009 IEEE

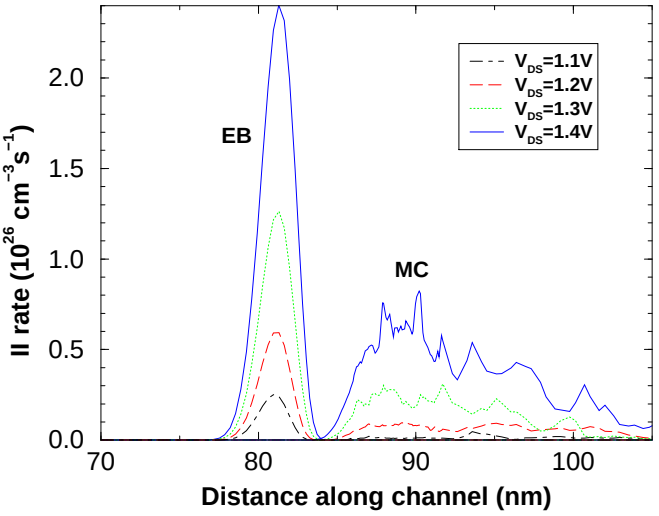

Fig. 2. Spatial distributions of the II rate along the channel in the EB and MC simulations at $V_{GS} = 0.5$ V. The MC profiles are a snap-shot (not averaged over time).

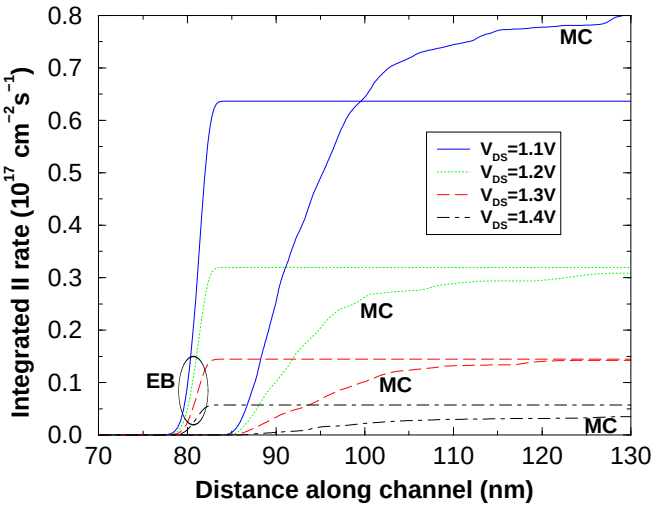

Fig. 3. Integrated spatial distributions of the II rate from Fig. 2 along the channel.

The calibration was done by integrating the II rate along the channel in the middle of the body. Increasing the parameter "*b*" in the exponent of the van Overstraeten model [3] by a factor 1.15, gives a reasonable match of the integrated II rates (see Fig. 3). Phonon-assisted non-local B2BT was modeled according to the microscopic theory of [5] with the same calibrated parameters as in [6]. DC and transient simulations were performed with *Sentaurus-Device* [3] both without and with DAT, varying the SRH lifetimes, the fixed oxide charge, and velocity overshoot (details not presented here).

IV. Results

Starting point is the analysis of the 110 nm PD SOI FET. Fig. 4 shows its output characteristics for different gate voltages. The main effect of DAT is to smooth out the kink

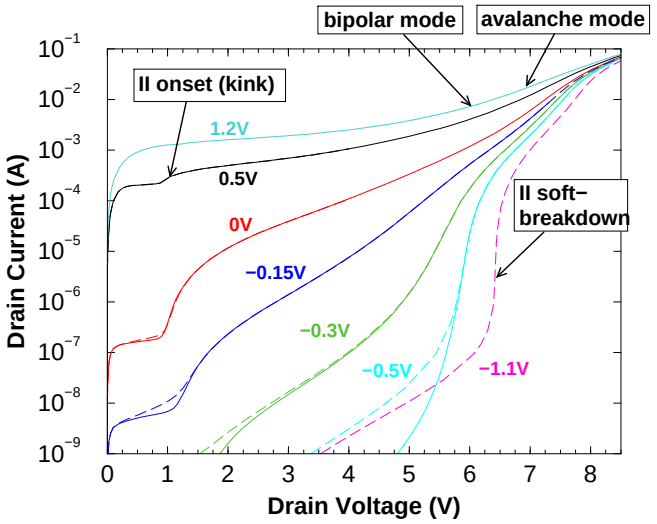

Fig. 4. DC output characteristics at indicated gate voltages of the 110 nm FET. Solid curves: without DAT, dashed curves: with DAT. T_L = 358 K.

which occurs for all V_{GS} at about the ionization threshold $V_{DS} \approx E_g$/e. For the understanding of the transient curves it is crucial to distinguish between two different V_{DS} regimes which are called "avalanche mode" and "bipolar mode" here. If during R1 the drain bias is higher than the soft-breakdown voltage (here about 6.5 V) , the strong II rate leads to a steady-state FB effect of the order of 0.2 V - 0.3 V which triggers a high and *constant* R1 collector current (source-drain current) of more than 0.1 mA/μm and, therefore, results in a large R1/R0 window (see the dashed curve in Fig. 5). The term "soft-breakdown" refers to the fact that in an SOI device the avalanche process is self-limiting because of the strong screening effect of the generated carriers. If V_{DS} during R1 is to the left of the soft-breakdown voltage, then the II rate is too weak to permanently compensate the loss of excess holes. As a consequence, the hole density and the collector current decay exponentially during R1 with a peak current much lower than in the avalanche mode (see the open circles in Fig. 5). In order to prove the bipolar mode, II was switched off between WRITE1 (W1) and R1 which results in the solid line in Fig. 5. Apart from a small difference due to the Early effect (channel length modulation), the same current turns out, even if V_{DS} is reduced to 2 V. At this bias, the bipolar transistor is still safely in saturation. It has to be mentioned that the relatively high breakdown voltage obtained in Fig. 4 can be significantly lowered by optimization of the junction profiles.

Now lets turn to the 22 nm FD UTB FET and analyze its potential for a FBM device. As can be seen from Fig. 6, the kink effect at $V_{DS} \approx E_g$/e is absent at all, instead a similar effect starts at $V_{DS} \approx$ 3 V. However, it is only visible with B2BT turned off and for gate voltages below threshold ($V_{GS} < 0.5$ V). The size of this FB effect decreases beyond a certain drain voltage (NDR-type behavior), while the II rate steadily increases until the current becomes a direct II current. (Here, "direct" means that it is not mediated by

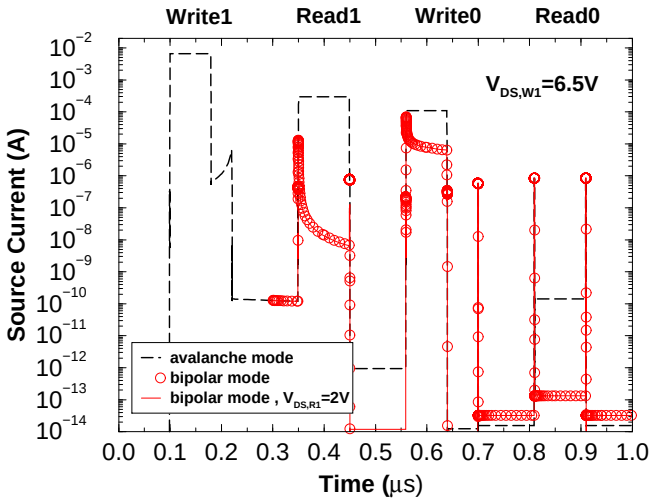

Fig. 5. Transient source current in the avalanche mode (dashed curve, $V_{\mathrm{DS,R1}} = 7\,\mathrm{V}$) and in the bipolar mode (symbols and solid curve, $V_{\mathrm{DS,R1}} = 6\,\mathrm{V}$). Lattice temperature $T_{\mathrm{L}} = 300\,\mathrm{K}$.

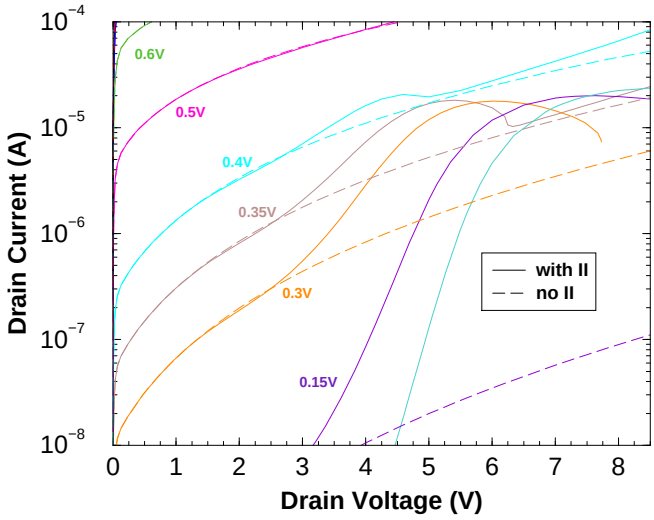

Fig. 6. DC output characteristics at indicated gate voltages of the 22 nm FD UTB FET with B2BT turned-off. Solid curves: with II, dashed curves: without II. $T_{\mathrm{L}} = 300\,\mathrm{K}$.

the FB effect.) The explanation of this behavior is involved. Below threshold, the number of injected electrons from the source is small and the II-generated holes can change the body potential significantly. The result is a strong FB effect, which however disappears at higher drain bias, when the stored holes are swept out to the source contact. It can be checked by inspecting the II rate that the latter continuously increases, but in the NDR branch this increase is greatly slowed down. The storage capability for the excess holes becomes worse with increasing gate voltage. Above threshold, the FB effect disappears, because the electrostatic impact of II-generated holes becomes negligible compared to the huge density of injected electrons from the source. However, the described FB effect is of no use as can be seen, if B2BT is now turned on.

Fig. 7 shows the output characteristics for the 22 nm FD UTB FET of Fig. 1 with optimized doping profile for the smallest possible GIDL, i.e. for the smallest B2BT rate [6]. This would

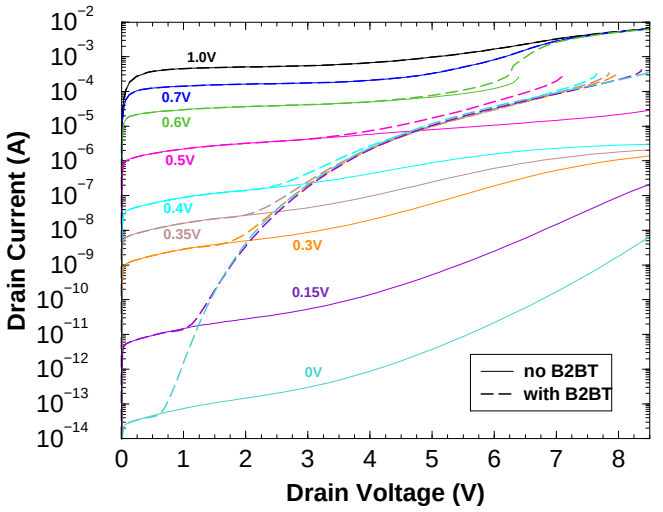

Fig. 7. DC output characteristics at indicated gate voltages of the 22 nm FD UTB FET. Solid curves: without B2BT, dashed curves: with B2BT. $T_{\mathrm{L}} = 300\,\mathrm{K}$.

be the best possible case for the scaled FBM. B2BT enhances the II soft-breakdown, but vanishes well above threshold, i.e. for $V_{\mathrm{GS}} > 0.7\,\mathrm{V}$. In the sub-threshold regime, all curves merge into a "B2BT envelope" which determines the unavoidable leakage and limits the choice of a reasonable V_{GS} for R1. It also predominates the II FB effect. As a consequence, the R1 current cannot be different from the R0 current as both are determined by B2BT, and there is no memory effect.

In order to demonstrate this in more detail, proper voltage wave forms have to be found. As there is no systematic and straight-forward way to select the best voltage wave form, one has to probe different forms empirically based on the inspection of the DC output characteristics. (Note, that the DC output characteristics never apply to the case with stored excess holes.) Some general rules are obvious, e.g. during W1 a maximal number of excess holes should be created which requires both a high drain and gate voltage, whereas during R1 the gate voltage must be below the threshold voltage in order not to loose the stored holes immediately. Fig. 8 shows the wave form that is used for the following transient simulations. In Fig. 9 the transient behavior of the hole density is shown for two holding voltages, and Fig. 10 presents the hole distributions along the device at the different bias conditions. The very dense hole population created during W1 is relatively stable during HOLD between W1 and R1, provided the gate voltage remains well below threshold. A better holding condition ($V_{\mathrm{GS,H}}$ = -0.2 V instead of 0 V) increases the excess hole concentration before the start of R1 significantly, because of the deeper potential well responsible for the longitudinal confinement. When switching to the R1 condition, the hole density instantaneously takes the value determined by the "B2BT envelope", i.e. the R1 current does

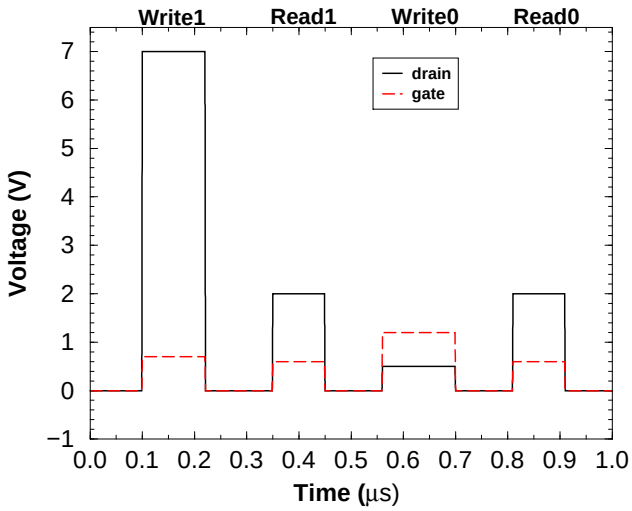

Fig. 8. Voltage wave form with $V_{\mathrm{GS,H}}$ = 0 V, $V_{\mathrm{DS,W1}}$ = 7 V for strong hole generation, and $V_{\mathrm{GS,R1}}$ = $V_{\mathrm{GS,R0}}$ = 0.6 V.

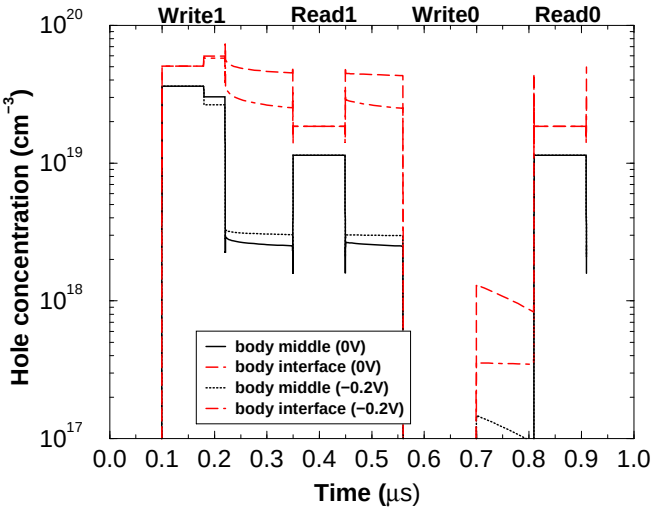

Fig. 9. Transient hole concentration for two holding voltages $V_{\mathrm{GS,H}}$ and two locations in the 22 nm FD UTB FET. T_{L} = 300 K.

not depend on the holding condition at all. Now, the hole profile is solely determined by the B2BT generation and the diffusion of the generated holes towards the source (see the curves labeled "R1=R0" in Fig. 10). During WRITE0 (W0) all excess holes are removed and the holding condition after W0 creates its own confined hole distribution with a density much smaller than after W1. However, when switching to the R0 voltages now (which are of course the same as during R1), exactly the same hole profile arises as during R1! Again, the current is a B2BT current which does not depend on the amount of stored holes.

This leads to the following conclusion: Although a large amount of excess holes ($>$ 1e19 cm^{-3}) can be created by II during W1 and hold relatively steady in the body until the beginning of the R1 pulse, the excess charge decays

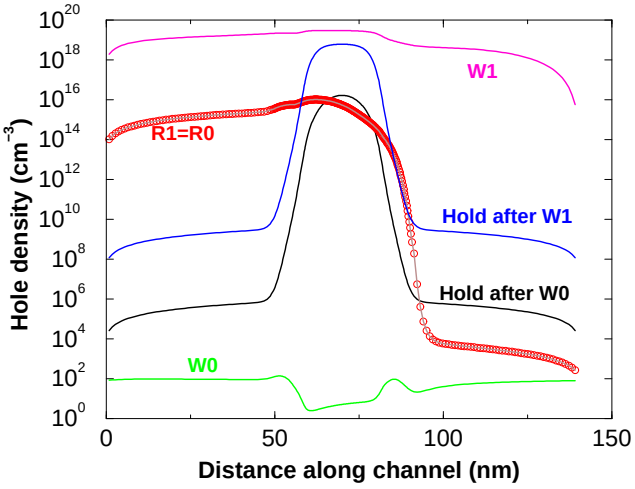

Fig. 10. Hole density profiles along the channel at a depth of 0.5 nm from the interface. The holding voltage was $V_{\mathrm{GS,H}}$ = 0 V.

instantaneously and the R1 current becomes self-determined by B2BT. A higher V_{GS} during R1 (above threshold) could increase the current above the "B2BT envelope", but then the stored holes are either swept out to the source or their electrostatic impact is negligible compared to the huge density of injected carriers from the source. It is, therefore, impossible to trigger a higher level of impact ionization by stored holes. This behavior is linked to the absence of the kink effect.

V. Conclusion

It seems challenging to scale FBMs based on impact ionization to fully-depleted short-channel FETs. A general reason is the absence of the kink effect which becomes predominated by unavoidable B2BT leakage. However, this statement has to be taken with great care because (i) TCAD simulations have only limited predictability, (ii) B2BT could be further reduced by a clever design, and (iii) only a limited number of voltage wave forms could be probed in this study.

Acknowledgment

The author thanks Dr. Simon Brugger (Stanford University) for *SimnIC* support and Dr. Cedric Bassin and Prof. Pierre Fazan (Innovative Silicon, Lausanne) for many valuable discussions.

References

[1] S. Okhonin, M. Nagoga, E. Carman, R. Beffa, and E. Faraoni, "New Generation of Z-RAM", IEDM, Dec. 2007, pp. 925 - 928.
[2] S. Okhonin, M. Nagoga, E. Carman, R. Beffa, and E. Faraoni, "New Generation of Z-RAM", IEDM, Dec. 2007, pp. 925 - 928.
[3] Synopsys Inc, Sentaurus Device User Guide, version Z-2007.12, Mountain View, California, (2007).
[4] SimnIC v1.06 User Manual, ETH Zürich, 2007.
[5] A. Schenk, "Physical Models for Semiconductor Device Simulation", Festkörperprobleme (Advances in Solid State Physics), vol. 36, pp. 245-263 (1996).
[6] A. Schenk, "GIDL Suppression by Optimization of Junction Profiles in 22nm DGSOI nFETs", Proc. EUROSOI, Göteborg, Sweden, Jan 19 - 21, 2009, pp. 31 - 32.
[7] EU-IST-4-026828 PULLNANO, http://www.pullnano.eu.
[8] http://www.itrs.net/reports.html.