Detailed Physical Simulation of Program Disturb Mechanisms in Sub-50 nm NAND Flash Memory Strings

C. D. Nguyen*, A. Kuligk*, M. I. Vexler*†, M. Klawitter*, V. Beyer‡, T. Melde§, M. Czernohorsky‡, and B. Meinerzhagen*
*BST, TU Braunschweig, 38023 Braunschweig, Germany
Email: a.kuligk@tu-bs.de
†Power Electronics Division, Ioffe Physical-Technical Institute, 194021 St. Petersburg, Russia
‡Fraunhofer Center Nanoelectronic Technology (CNT), 01099 Dresden, Germany
§NaMLAb gGmbH, TU Dresden, 01187 Dresden, Germany

***Abstract*— The hot electron induced mechanism disturbing the stored information in inhibited bit lines during the programming of nonvolatile memories with NAND architecture is studied in detail using a new dedicated advanced physical simulation scheme for the first time.**

I. Introduction

One of the most important issues for NAND flash memory technologies is scalability. Due to their improved scaling perspectives resulting from less cell to cell coupling TANOS ($TaN/Al_2O_3/Si_3N_4/SiO_2/Si$) memory cells are discussed as promising devices for replacing floating gate technology [1]. However, the scalability of such memories may still be limited by disturb mechanisms [2].

In Fig. 1 the scheme of a standard NAND memory is shown. The programming of the cell in the active word line (WL) and the active bit line (BL) is intended. However, a cell in the same WL but in a non selected (inhibited) BL may also change its stored charge due to the high WL voltage. This undesired programming is referred to as program disturb. In order to suppress the programming of non selected cells along the WL the channel potential of the inhibited BLs is boosted during the programming step. Typical applied voltages are given in Fig. 1.

In the present work, the physical origin of the program disturb in inhibited TANOS BL strings is analyzed and simulations for a quantitative characterization of the corresponding processes are performed. Furthermore, options to influence the program disturb are studied.

II. Origin of Program Disturb

The key leakage mechanism determining the upper limit and the initial transient decay of the boosted channel potential of the non selected BLs during the programming step is band-to-band tunneling (BBT), which occurs mainly close to the silicon surface near the select transistors [3].

In Fig. 2(a) the processes causing the program disturb are shown. The BBT occurs predominantly close to the silicon surface, where the electric field is maximal. Due to the large

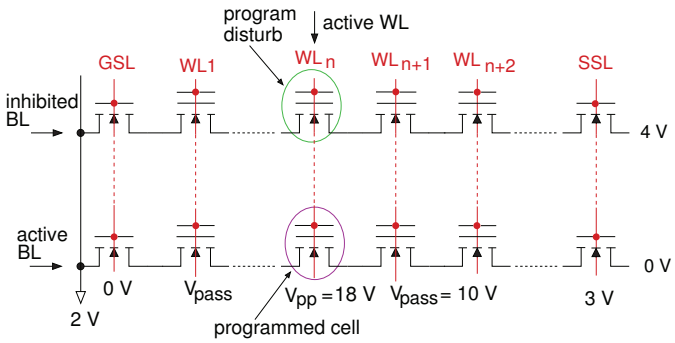

Fig. 1. Schematic of NAND strings with bias conditions during programming and the location of the program disturb.

potential difference between the channel and the substrate in inhibited BL the holes generated by BBT are accelerated towards the substrate. Close to the bottom boundary of the substrate space charge region the accelerated holes cause impact ionization (II). Due to the same large potential difference the electrons generated by this hole initiated impact ionization are accelerated towards the silicon surface. These accelerated electrons are hot and the main reason for the program disturb in the inhibited BLs during the programming step.

In addition to holes electrons are also generated due to BBT. In contrast to the holes, these electrons stay close to the surface and don't get hot.

In order to quantitatively estimate these processes and to control them it is necessary to know where the BBT occurs and how its location can be influenced. Typically, WL1 close to the GSL shows the largest program disturb. Therefore, our study concentrates on this situation. For the other WLs the processes are in principle the same.

III. Simulation Details

A. Simulation Setup

A simplified TANOS NAND string consisting of GSL (Ground Select Line), SSL (String Select Line), and three

978-1-4244-7700-5/10/$26.00 ©2010 IEEE

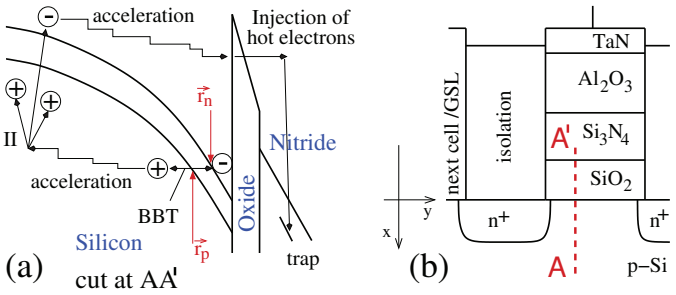

Fig. 2. (a) Schematic diagram of the processes due to BBT. (b) Schematic of TANOS device.

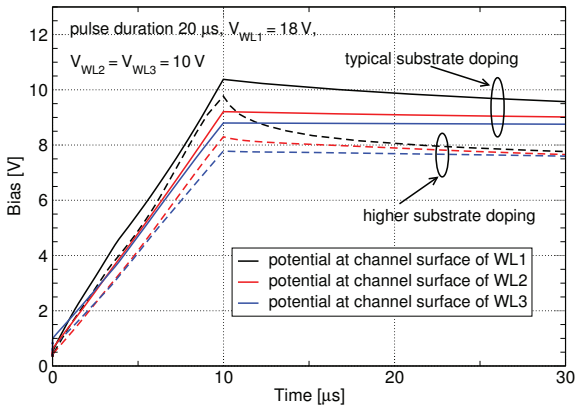

Fig. 3. Transient simulation results for the surface potential in the middle of the channel of different cells and both substrate doping levels.

cell transistors with lateral dimension usual for a 48 nm technology is considered (see Fig. 1). The distance between two neighboring WLs is 40 nm. For the distance between WL1 and GSL 85 nm is assumed in the simulations if no other distance is specified. A schematic of the considered TANOS device is shown in Fig. 2(b). ONA stack dimensions are 4.5 nm SiO_2, 6 nm Si_3N_4 and 12 nm Al_2O_3; 4.65 eV is used for the TaN work function. Standard doping levels for this technology generation are assumed except for the substrate doping for which boron concentrations of about $10^{17}\,\mathrm{cm}^{-3}$ (typical for 48 nm NAND strings, denoted by "typical") and $4 \cdot 10^{17}\,\mathrm{cm}^{-3}$ (refers to a 24 nm technology, denoted by "higher") are investigated.

To inhibit programming the string is first isolated by applying 4 V to the drain of SSL which causes SSL to operate in the off-state. Subsequently the WL gate voltages are ramped up from 0 V to their final values in 10 μs (see Fig. 3) in order to generate the inhibit potential in the BL channels. Finally, the gate voltages are kept constant for 20 μs programming time. Besides the ratio of the gate-to-channel and the channel-to-substrate capacitances the final value of the boosted channel potential depends on the dominant leakage mechanisms discharging these capacitancies which are BBT and II.

B. Nonlocal Modeling of BBT

The BBT is self-consistently calculated using a nonlocal model [4]. Within this approach, the BBT generation rate is proportional to the interband (Zener) elastic tunneling probability in silicon. It is assumed that an electron at the valence band edge located at the position $\vec{r}_p$ is tunneling directly (along a straight line) into the nearest position $\vec{r}_n$ within the conduction band edge having the same energy (see Fig. 2(a)).

Between $\vec{r}_n$ and $\vec{r}_p$ the potential drop equals E_g/q, where E_g is the band gap. Therefore, the band-to-band generation rate is obviously not a function of the local electric field. This is important because the topology of the field in a TANOS string is typically very steep and complicated and a constant field approximation is not valid. A similar nonlocal approach is used for the modeling of II in GALENE III [5].

C. Simulation Procedure

A complete programming pulse is simulated by the transient Drift-Diffusion model in GALENE III including self-consistently the nonlocal models for BBT and II, which have been extensively verified by comparison with experimental data [4], [5]. This specialized transient simulation algorithm including nonlocal models for BBT and II is a unique feature in GALENE III and does not exist in commercial TCAD software. The spatial distributions of electrostatic potential and BBT resulting after 20 μs are transferred to the full band Monte Carlo (MC) simulator ELWOMIS [6]. The MC simulations are post processor frozen field simulations including II [5] where the carriers are injected according to the spatial BBT distributions. Thus, the MC simulation can be restricted to the hot carrier region of the device. Based on these MC simulations the energy distribution function for electrons is evaluated at the Si surface.

The injection current density is calculated based on the energy distribution function. To calculate this current density both tunneling and thermal emission mechanisms are considered. Finally, the injected surface charge density is approximated by multiplying the injection current density by the total programming time.

IV. Results

Fig. 3 shows the resulting transient surface potential in the middle of the channels of the cell transistors WL1 - WL3 for both doping levels. For the typical doping level the boosted channel potential is higher and decays slower when the word line voltages are constant. These effects are mainly due to the higher substrate doping which increases both the substrate capacitance as well as the BBT current, which is predominantly responsible for the charge leaking out of the substrate capacitance.

The spatial distributions shown in Figs. 4, 5, 6, and 8 refer to the situation 20 μs after the beginning of the ramp. The electric fields are shown in Fig. 4(a) for the typical doping and in Fig. 4(b) for the higher doping, respectively. Fig. 5(a) shows the BBT generation rate for holes for the typical doping. The electric field and BBT distribution for the typical doping are consistent with the gate induced drain leakage (GIDL) effect, since the electric field and the BBT maxima are both below the gate edge of the GSL device. However, for the higher doping

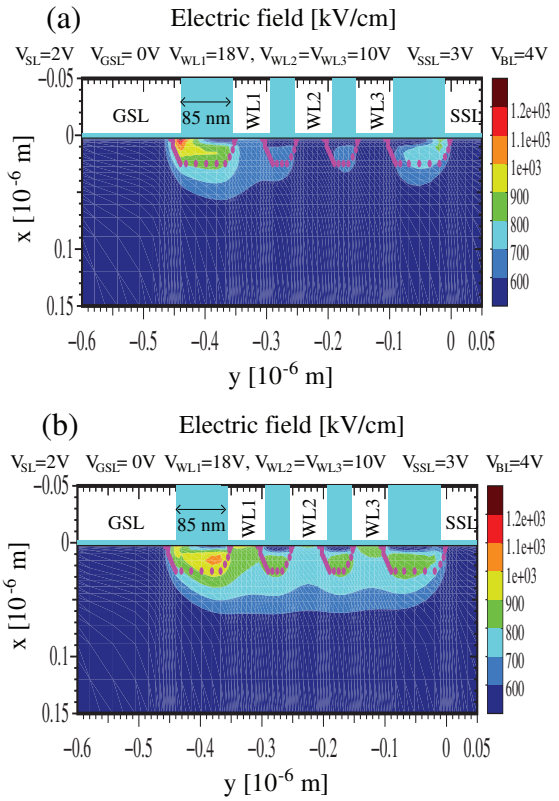

Fig. 4. Electric field (scale on the right [kV/cm]). (a) Typical substrate doping, (b) higher substrate doping.

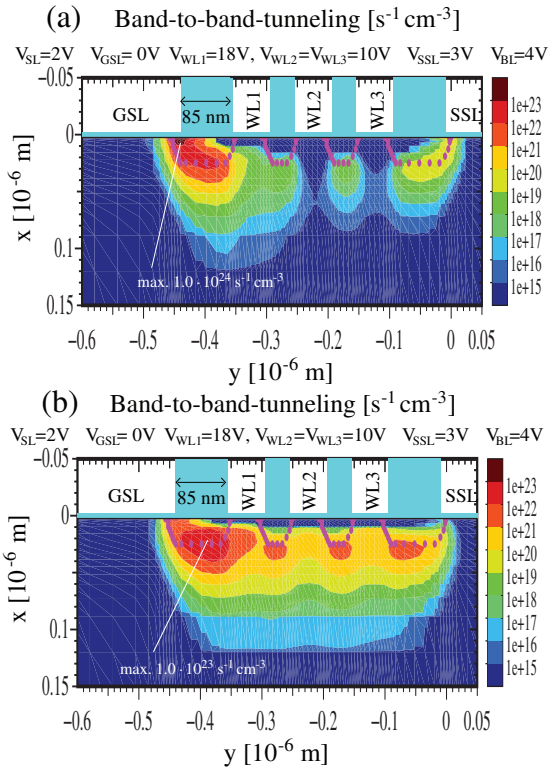

Fig. 5. Spatial distribution of hole generation due to BBT (scale on the right [$s^{-1}cm^{-3}$]). (a) Typical substrate doping, (b) higher substrate doping.

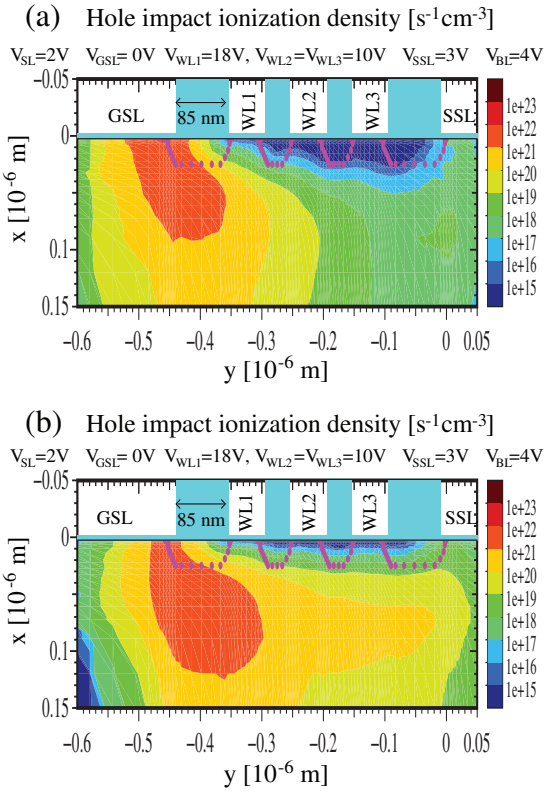

Fig. 6. Spatial distribution of hole initiated II density (scale on the right [$s^{-1}cm^{-3}$]). (a) Typical substrate doping, (b) higher substrate doping.

(Fig. 5(b)) the BBT maximum as well as the maximum of the electric field is between GSL and WL1 and cannot be explained by GIDL. This difference is mainly caused by the different substrate doping.

Fig. 6 compares the results for II for both doping levels. It can be seen that this hole initiated II is very important for the program disturb since it generates a lot of electrons deep in the bulk under WL1 where the potential is much lower compared to the channel so that these electrons can get hot when they are accelerated towards the channel. Though the boosted channel potential is about 2 V higher for the typical substrate doping level compared to the higher one, the II distribution is worse for the higher doping level, because more electrons are generated below WL1 due to the less favorable location of the BBT maximum. Consequently, the injection of hot electrons into the nitride layer of WL1 and its neighboring cells is much higher for the higher doping level (see Fig. 7) though the maximum BBT rate is smaller. A charge injection of more than 10^{12} cm^{-2} in the WL1 devices causes a significant program disturb. Moreover, the simulated disturb pattern shown in Fig. 7 for the typical doping agrees well with respective experimental results (e.g. [3]).

V. Suppression of the Program Disturb

The most obvious way to reduce the disturb is to increase the GSL-WL1 spacing [3], whereby the distance between the BBT maximum and WL1 may increase. In Fig. 8 the hole

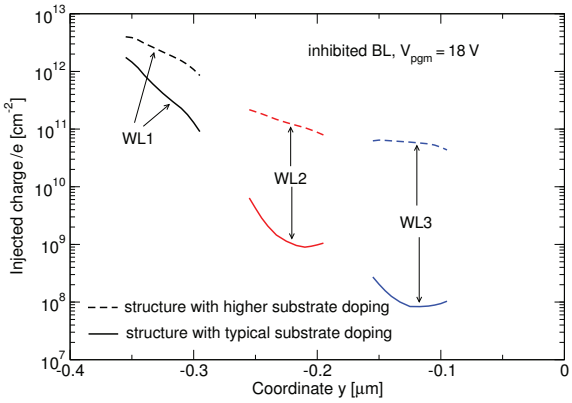

Fig. 7. Charge injection for both doping levels.

generation due to BBT is shown for NAND strings with 135 nm GSL-WL1 space. Comparing Fig. 8(a) and 8(b) it can be seen that this increased spacing is more efficient for the typical doping than for the higher doping level, because the maximum of the BBT rate stays at the gate edge at the drain of the GSL device for the typical doping and can still be interpreted as being consistent with the GIDL effect while this is not the case for the higher doping.

In Fig. 9 the injected charge is shown for three GSL-WL1 spacings (85 nm, 135 nm, and 185 nm). The increased spacing is obviously much more efficient for the suppression of program disturbs in case of the typical doping than for the higher doping. For the typical doping the 135 nm space reduces the injected charge in WL1 by more than two orders of magnitude. However, the reduction factor is only about 3 for the higher doping level.

VI. Conclusion

The crucial role of the spatial distribution of BBT for the program disturb effect in nonvolatile NAND memories has been demonstrated. Since higher substrate doping levels are unavoidable for scaled strings, this observation may have serious consequences in future technology generations.

Acknowledgment

This work has been partially supported by the European Commission under the FP7 research contract 214431 "GOSSAMER".

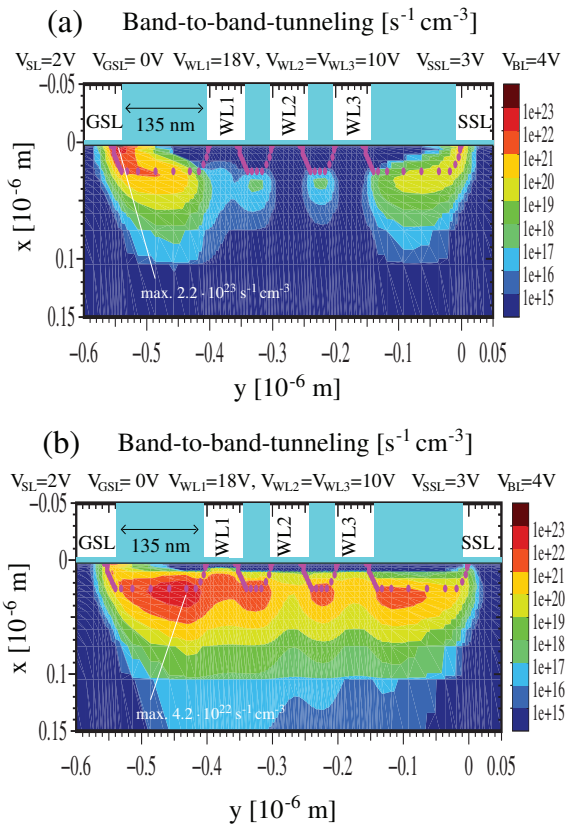

Fig. 8. Spatial distribution of hole generation due to BBT (scale on the right [$s^{-1}cm^{-3}$]). (a) Typical substrate doping, (b) higher substrate doping.

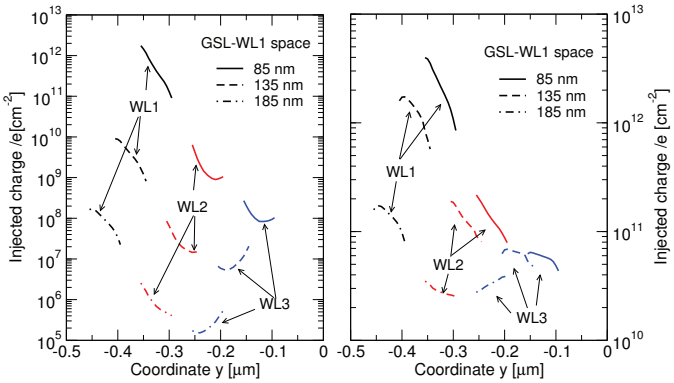

Fig. 9. Charge injection for the typical (left) and the higher (right) substrate doping level and different GSL-WL1 spacings.

References

[1] Y. Park, J. Choi, C. Kang, C. Lee, Y. Shin, B. Choi, J. Kim, S. Jeon, J. Sel, J. Park, K. Choi, T. Yoo, J. Sim, and K. Kim, "Highly Manufacturable 32Gb Multi-Level NAND Flash Memory with 0.0098 μm^2 Cell Size using TANOS(Si - Oxide - Al_2O_3 - TaN) Cell Technology," *IEDM Tech. Dig.*, pp. 29–32, 2006.

[2] C.-J. Tang, C. W. Li, T. Wang, S. H. Gu, P. C. Chen, Y. W. Chang, T. C. Lu, W. P. Lu, K. C. Chen, and C.-Y. Lu, "Characterization and Monte Carlo Analysis of Secondary Electrons Induced Program Disturb in a Buried Diffusion Bit-line SONOS Flash Memory," *IEDM Tech. Dig.*, pp. 173–176, 2007.

[3] J.-D. Lee, C.-K. Lee, M.-W. Lee, H.-S. Kim, K.-C. Park, and W.-S. Lee, "A New Programming Disturbance Phenomenon in NAND Flash Memory by Source/Drain Hot-Electrons Generated by GIDL Current," *NVSMW*, pp. 31–33, 2006.

[4] A. v. Schwerin, W. Bergner, and H. Jacobs, "Self-Consistent Simulation of Hot-Carrier Damage Enhanced Gate Induced Drain Leakage," *IEDM Tech. Dig.*, pp. 543–546, 1992.

[5] C. Jungemann, B. Meinerzhagen, S. Decker, S. Keith, S. Yamaguchi, and H. Goto, "Is Physically Sound and Predictive Modeling of NMOS Substrate Currents Possible?" *Solid-State Electron.*, vol. 42, no. 4, pp. 647–655, 1998.

[6] C. Jungemann and B. Meinerzhagen, *Hierarchical Device Simulation (The Monte-Carlo Perspective)*. Wien, New York: Springer-Verlag, 2003.