SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2016, Vol. 59 122403:1–122403:10
doi: 10.1007/s11432-015-5475-7

A two-dimensional simulation method for investigating charge transport behavior in 3-D charge trapping memory

Zhiyuan LUN[1], Gang DU[1*], Kai ZHAO[2], Xiaoyan LIU[1] & Yi WANG[1]

[1]*Institute of Microelectronics, Peking University, Beijing* 100871*, China;*
[2]*School of Information and Communication, Beijing Information Science and Technology University, Beijing* 100101*, China*

Received August 11, 2015; accepted September 25, 2015; published online April 6, 2016

Abstract This work presents a self-consistent two-dimensional (2-D) simulation method with unified physical models for different operation regimes of charge trapping memory. The simulation carefully takes into consideration the tunneling process, charge trapping/de-trapping mechanisms, and 2-D drift-diffusion transport within the storage layer. A string of three memory cells has been simulated and evaluated for different gate stack compositions and temperatures. The simulator is able to describe the charge transport behavior along bitline and tunneling directions under different operations. Good agreement has been made with experimental data, which hence validates the implemented physical models and altogether confirms the simulation as a valuable tool for evaluating the characteristics of three-dimensional NAND flash memory.

Keywords charge trapping memory, semiconductor device modeling, 2-D charge transport, 3-D NAND flash, device modeling and simulation

Citation Lun Z Y, Du G, Zhao K, et al. A two-dimensional simulation method for investigating charge transport behavior in 3-D charge trapping memory. Sci China Inf Sci, 2016, 59(12): 122403, doi: 10.1007/s11432-015-5475-7

1 Introduction

With the intensive application of data storage units in the burgeoning contemporary electronic devices such as mobile phones, digital cameras and solid-state drives, the demand for high-density low-cost NAND flash memory has been continually on the rise. Under this trend, charge trapping (CT) flash memory technology has long been investigated and proposed as the most promising candidate within its domain due to its immunity to stress-induced leakage current and reduced coupling capacitance. As such, this technology has thus been recently attracting considerably more attention since it is ideally suitable for three-dimensional (3-D) vertical architectures which exploits the third dimension of the devices to fulfil the ever-increasing demand for bit cost [1–8]. 3-D CT NAND flash is also forecasted to continue the trend of scaling NAND flash below the 15 nm technology node [9]. These structures consist of gate-all-around memory cells piled up layer by layer, with channel runs vertically through multiple layers [1–3, 8]. However, the charge storage layer runs continuously along the memory string and thus cannot

* Corresponding author (email: gangdu@pku.edu.cn)

be interrupted due to deposition process restrictions. This constraint leads to the emergence of an additional charge leakage path in the memory cells, as the trapped charge can be transported out of the investigated cell towards adjacent cells on the same bitline (BL) [10]; such a condition is deemed to be a new disturbance for 3-D CT NAND architectures and therefore fosters a level of re-consideration regarding retention and reliability issues.

In recent years, several modeling approaches have been developed to describe CT flash memory. Most of them treat the CT devices in a one-dimensional way, only considering the direction along the charge tunneling path perpendicular to the gate stack (referred to as tunneling direction hereafter) in a single memory cell [11–13]. Previous two-dimensional (2-D) research assumed a simplified diffusion mechanism along BL direction in the storage layer [14, 15] or only focused on the modeling of retention behavior under high temperatures [16]. Therefore, a comprehensive 2-D simulation framework with consistent models in different operations is needed for the modeling of CT flash memory in 3-D NAND scenarios, especially with regard to charge transport mechanisms along BL direction.

Thus, in this study, on the basis of our previous work [17, 18], we present a comprehensive 2-D self-consistent simulation framework for investigating the charge transport behavior in 3-D CT NAND flash memory under different operation regimes with unified physical models. To this aim, three neighboring cells in a memory string are simulated via careful consideration of associated tunneling processes, charge trapping/de-trapping mechanisms, and 2-D drift-diffusion transport in the storage layer. After calibration of the simulator with experimental results, charge transport behaviors under room and elevated temperatures for different gate stack compositions were investigated. As such, this study's aim is to illustrate the influence of charge transport along BL and tunneling directions upon device characteristics and to evaluate associated disturbances from adjacent cells.

This paper is organized as follows: physical models and numerical treatment for the simulation are described in Section 2; in order to analyze the impacts of different parameters on charge transport within the storage layer, a simulation of retention behavior under different temperatures is evaluated in Section 3, which in addition, also addresses the various trends of threshold voltage shifts and dominant charge loss mechanisms during retention; final conclusion is drawn in Section 4.

2 Physical models and numerical approach

2.1 Basic equations

The cross section of a memory string in a 3-D CT NAND flash structure is demonstrated in Figure 1(a). In order to describe 2-D carrier transport and improve simulation efficiency, a string of three neighboring memory cells is simulated in this study, which takes into consideration the charge transport along the BL and tunneling directions in the storage layer as well as the interference between neighboring memory cells, as shown in Figure 1(b). A system of 2-D non-linear partial differential equations is solved self-consistently in the simulation domain. The equation system is as follows:

$$\nabla \cdot (\varepsilon \nabla \varphi) = q(n_f + n_t), \tag{1}$$

$$\frac{\partial n_f}{\partial t} = \frac{1}{q}\nabla \cdot \boldsymbol{J}_n - c_n n_f + e_n n_t, \tag{2}$$

$$\frac{\partial n_t}{\partial t} = c_n n_f + r_{\mathrm{BT}}(N_T - n_t) - (e_n + r_{\mathrm{TB}})n_t, \tag{3}$$

where q is the electronic charge; ε is the material dielectric constant; N_T, n_t, and n_f are the volume densities of traps, trapped carriers, and free carriers, respectively; c_n is the free-carrier capture coefficient; e_n is the trapped-carrier emissivity; r_{BT} and r_{TB} are the emission rates of band-to-trap (BT) tunnelling and trap-to-band (TB) tunnelling, respectively. The abovementioned equation system includes a Poisson equation for the gate stacks with electrostatic contributions of both free and trapped carriers, a carrier continuity equation for the charge storage layer with drift-diffusion transport scheme, and a trapped

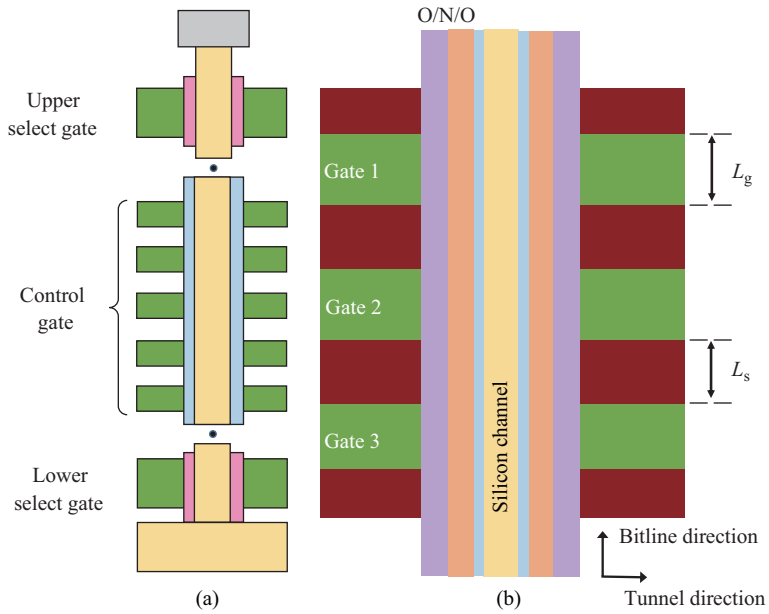

Figure 1 (Color online) (a) Cross section of 3-D NAND flash memory string; (b) simulation domain of CT NAND flash memory structure.

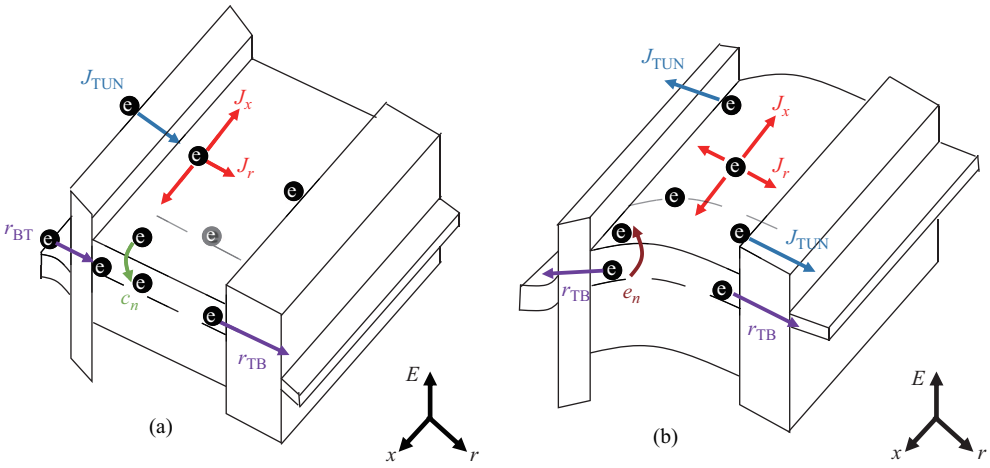

Figure 2 (Color online) Dominant physical mechanisms involved in (a) programming and (b) retention operation of the memory.

carrier conservation equation that accounts for trapping and de-trapping dynamics considering several mechanisms. In each simulation step, the threshold voltage shift resulting from the spatial distribution of both free and trapped carriers in the trapping layer is calculated.

2.2 Physical models

The dominant physical mechanisms implemented in the simulation are depicted in Figure 2, in which c_n is the coefficient for electron capture and e_n is the coefficient for electron emission. In addition to the emission process from trap sites to local conduction band, electron tunneling event from trap sites to conduction band of tunneling and blocking oxide (i.e. TB tunnelling) is also considered in the simulation, and is represented by r_{TB} in Figure 2.

The free-electron capture coefficient is modeled by the Schockley-Read-Hall theory [19]:

$$c_n = \sigma_n v_n^{\mathrm{th}} (N_T - n_t), \tag{4}$$

where σ_n is the electron capture cross section and $\upsilon_n^{\rm th}$ is the electron thermal velocity. The free-electron capture coefficient is proportional to the volume density of the unoccupied traps. The process of electron emission from trap sites to the charge storage layers local conduction band is described by Poole-Frenkel emission [20], which is calculated as follows:

$$e_n = f_{\rm PF} \exp\left[-\frac{q(E_T - \beta\xi^{1/2})}{kT}\right], \tag{5}$$

$$\beta = \sqrt{\frac{q^3}{\pi\varepsilon}}, \tag{6}$$

where $f_{\rm PF}$ is the attempt-to-escape frequency for the Poole-Frenkel process; ξ is the local electric field; and ε is the high-frequency dielectric constant for material of the charge storage layer.

Finally, TB tunneling is modeled by

$$r_{\rm TB} = f_{\rm TB} \cdot T_p, \tag{7}$$

where $f_{\rm TB}$ is the attempt-to-escape frequency for TB tunneling and T_p is the tunneling coefficient computed by the Wentzel-Kramers-Brillouin approximation [21], which is related to the locations and energy levels of trapped electrons.

2.3 Numerical approach

In order to self-consistently solve the equation system for obtaining the time evolution of threshold voltage as well as the distribution of internal physical properties at each time step, several numerical approaches have been accordingly adopted within the simulation. For example, the five-point-stencil central difference method was implemented for the 2-D Poisson equation. To afford greater accuracy with fewer mesh nodes and larger time steps, the Scharfetter-Gummel scheme [22] is used for the discretization of current density in the 2-D carrier continuity equation:

$$J_{e'} = \frac{qD_{e'}}{L_e}\left[n_e B\left(\frac{\varphi_e - \varphi_c}{V_t}\right) - n_c B\left(\frac{\varphi_c - \varphi_e}{V_t}\right)\right], \tag{8}$$

where the subscript e refers to the midpoint between the center and east points in the mesh stencil; D is the electron diffusion coefficient; L is the length between two points; n is the electron density; φ is the potential of the mesh point; V_t is the thermal potential; and $B(x)$ is the Bernoulli function [22] which is defined as follows:

$$B(x) = \frac{x}{e^x - 1}. \tag{9}$$

Additionally, time stability is ensured by the implementation of implicit numerical method. Hence, ultimately, solution for the subject equation system is attained by employing the flowchart methodology depicted in Figure 3.

3 Results and discussion

3.1 Calibration of the simulator

To calibrate the developed simulator and determine certain parameters adopted in the models, simulation results are compared with measurement data for CT flash memory. Figure 4(a) shows the reproduction of experimental programming results [11] for a memory device with gate stacks composed of 4.5/6/10 nm $SiO_2/Si_3N_4/Al_2O_3$. For the simulated memory cells, sharp interfaces are assumed between the layers, and the influence of interfaces on device characteristics is not included in the simulation. A uniform trap spatial distribution in the trapping layer is assumed, with a trap density $N_t = 7 \times 10^{19}$ cm^{-3} and a trap cross section $\sigma_n = 1 \times 10^{-14}$ cm^2. The parameters used in the simulation are summarized in Table 1. Using a consistent set of material parameters, strong agreement is achieved with experimental retention results under a wide variety of temperatures [23], as shown in Figure 4(b). The gate stack composition for the retention measurements is 4.5/7/12 nm $SiO_2/Si_3N_4/Al_2O_3$. The programming and retention transients are measured on different samples within the same fabrication process [11, 23].

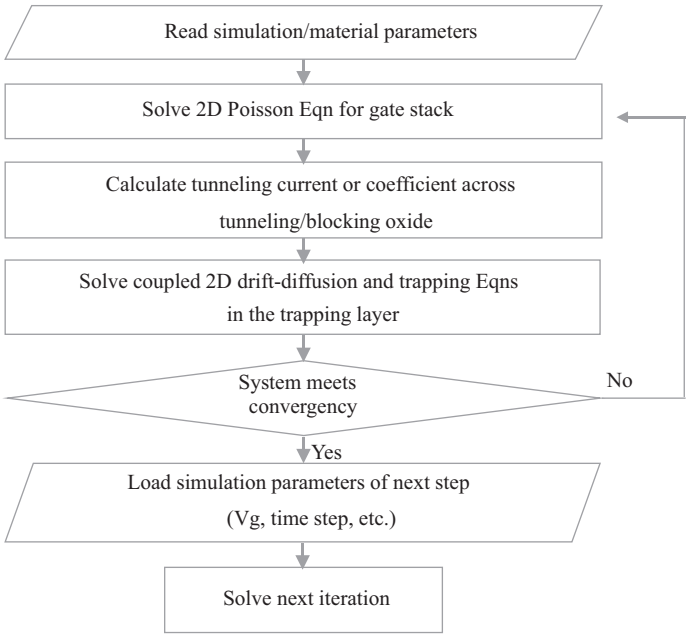

Figure 3 Simulation flowchart for the coupled equation system.

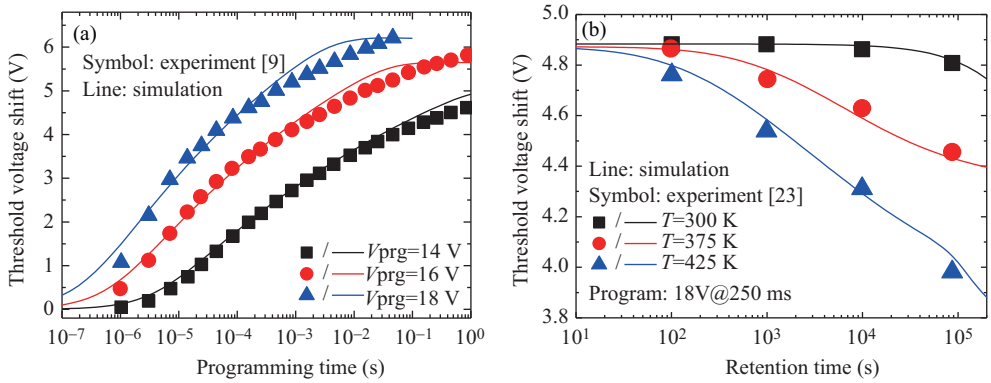

Figure 4 (Color online) Comparisons between simulation results and experimental data for (a) programming and (b) retention characteristics of CT flash memory.

Table 1 Main simulation parameters

Parameter	Symbol	Value
SiO_2 effective mass	m_{oxide}	0.40 m_0
Si_3N_4 effective mass	m_{SiN}	0.50 m_0
Si_3N_4 trap cross section	σ_n	1×10^{-14} cm^2
Si_3N_4 trap density	N_T	7×10^{19} cm^{-3}
Si_3N_4 electron mobility	μ_n	0.01 cm^2/V/s [13]
Si_3N_4 TB tunneling frequency	f_{TB}	1×10^{11} Hz
Si_3N_4 trap energy	E_T	1.5 eV [13]
Si_3N_4 PF frequency	f_{PF}	1×10^{12} Hz

3.2 Retention characteristics

In order to demonstrate major charge loss mechanisms in retention operations under room temperature and to observe their associated dependencies on material parameters, simulation of a memory device with a tunneling oxide of medium thickness (O/N/O composition: 4/5/5 nm $SiO_2/Si_3N_4/SiO_2$) was performed. Figure 5 shows the distribution of trapped electron density after a retention time of 10^8 s.

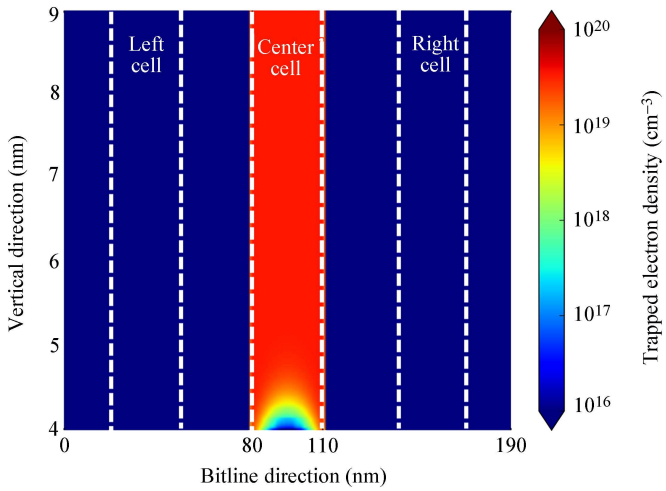

Figure 5 (Color online) Distribution of trapped electron density under room temperature after 10^8 s. (Note that the simulation domain in Figure 1(b) is rotated by 90 ° to demonstrate the distribution of free and trapped electron densities in order to clearly show the charge transport along BL and tunneling directions).

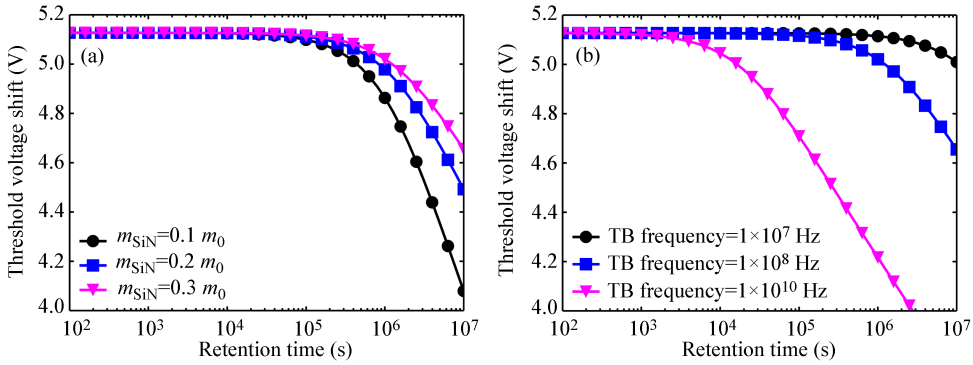

Figure 6 (Color online) Simulated retention behavior with different (a) effective electron tunneling masses and (b) attempt-to-escape frequencies for TB tunneling.

The initial distribution of trapped electrons in the charge trapping layer is assumed to be uniform, with a trapped charge density $n_{t,\mathrm{init}} = 3.5 \times 10^{19}$ cm^{-3}. It is shown that charge loss occurs only in a small region near the tunneling oxide, and there is barely any lateral transport of trapped charge along the BL direction. This is due to the fact that under room temperature the emission of trapped electrons is negligible, whereas the TB process through the tunneling oxide is the dominant charge loss mechanism.

Figure 6 shows the impact on the simulated retention behavior of two key parameters governing the TB component: the effective tunneling mass of the charge storage layer material (i.e. m_{SiN}, which affects the calculation of tunneling coefficients) and the electron attempt-to-escape frequency for TB tunneling (i.e. f_{TB}, which determines the frequency of the TB tunneling process). It can be observed that m_{SiN} mainly influences the slope of decline in threshold voltage shift with time evolution, while f_{TB} affects the retention time at which threshold voltage begins to notably decrease.

The simulation of retention performance under high temperatures is also of critical importance to a CT flash memory, because the higher the temperature a CT flash memory experiences during retention operations, the greater the acceleration of charge de-trapping mechanisms tend to be, which can ultimately lead to data bit failure event. To reduce the charge loss in tunneling-out components along tunneling directions and to illustrate the roles exclusively played by electron de-trapping and electron transport in the storage layer, a memory device with a thick tunneling oxide (O/N/O composition: 6/6/6 nm $SiO_2/Si_3N_4/SiO_2$) was simulated. Figure 7 demonstrates the distribution of trapped electrons and free electrons under 85 °C after a retention time of 10^8 s. A uniform distribution with an initial trapped

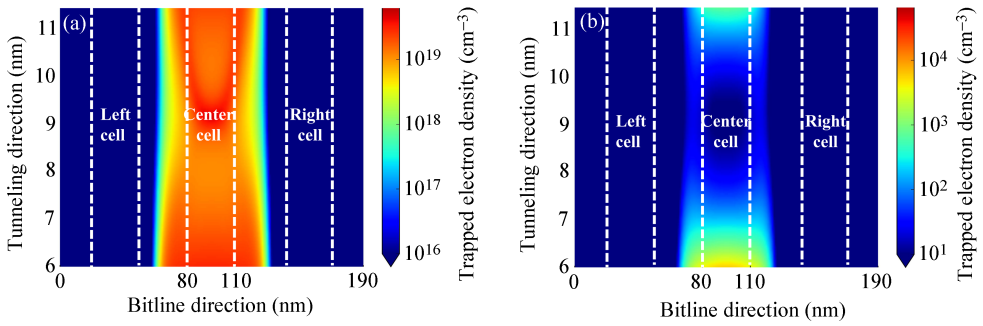

Figure 7 (Color online) Distribution of (a) trapped and (b) free electrons under 85 °C, after a retention time of 10^8 s.

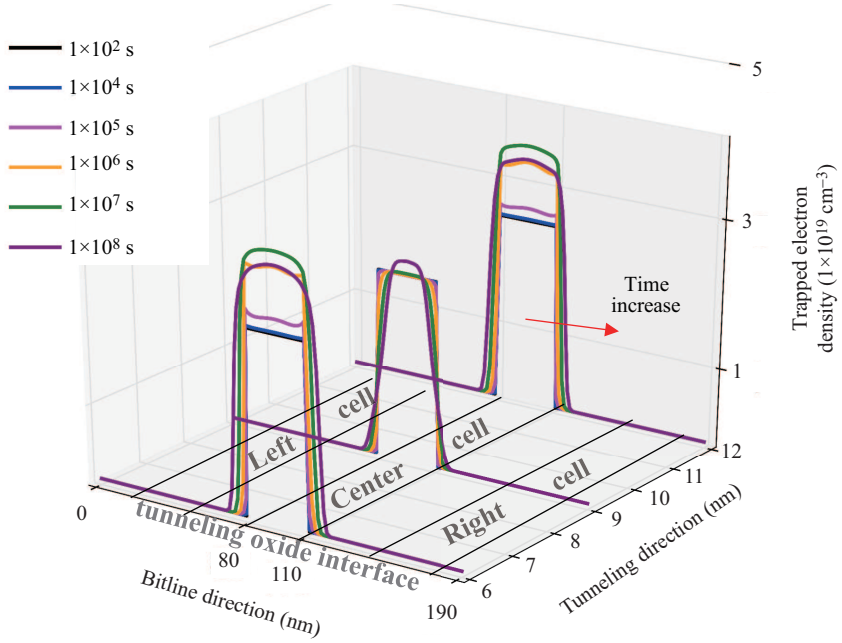

Figure 8 (Color online) Trapped electron distributions in the charge storage layer at different retention times along the BL direction at the tunneling oxide interface, through the middle of the trapping layer, and at the block oxide interface.

charge density $n_{t,\mathrm{init}} = 3.5 \times 10^{19}$ cm^{-3} was assumed at the start of the retention simulation.

It can be concluded that the electrons are emitted from trap sites at high temperatures and then transported along BL and tunneling directions in the charge storage layer. In addition, the time-dependent evolutions of trapped electron densities along the BL direction in the charge storage layer at the tunneling oxide interface, through the middle of the trapping layer, and at the blocking interface, are plotted in Figure 8.

The impact of the attempt-to-escape frequency of Poole-Frenkel emission on the retention characteristic under 85 °C is shown in Figure 9. It can be seen that higher attempt-to-escape frequencies lead to an earlier decline of threshold voltage shifts.

The retention behavior of CT flash memory cells under 85 °C with channel lengths ranging from 20 nm to 50 nm, while maintaining a fixed space length of 30 nm, is shown in Figure 10. It is found that prior to 10^5 s, the time evolutions of threshold voltage shift in four distinct cases show similar tendencies. Figure 11(a) further illustrates the details of trapped charge redistributions along the tunneling direction through the center of the simulated cell (at $t < 10^5$ s), which can be explained as follows. The trapped electrons are emitted into the conduction band of the charge storage layer, where they become free electrons. These free electrons are driven by electrical fields and are ultimately transported to the interface of the tunneling/trapping and trapping/blocking layer, where they encounter a potential barrier and thus become trapped again. In contrast, after 10^5 s, memory cells with smaller gate lengths show

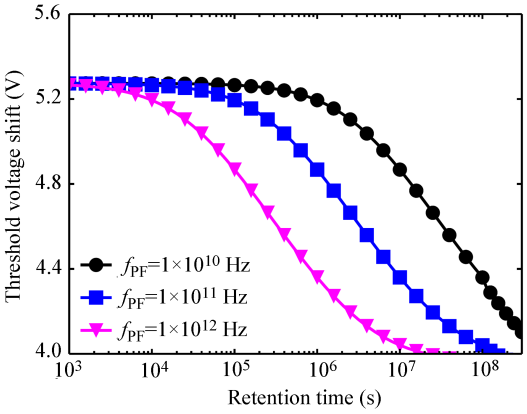

Figure 9 (Color online) Simulated retention behavior with different attempt-to-escape frequencies for Poole-Frenkel emission.

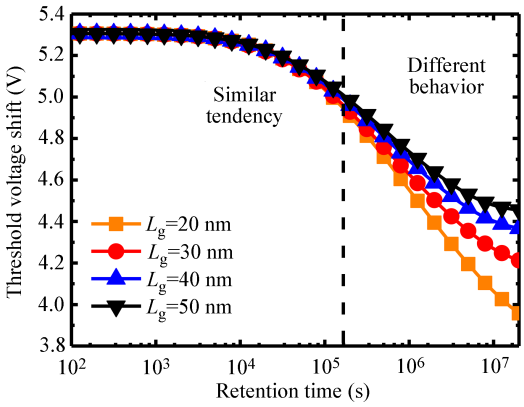

Figure 10 (Color online) Simulated retention behavior for memory cells with different gate lengths.

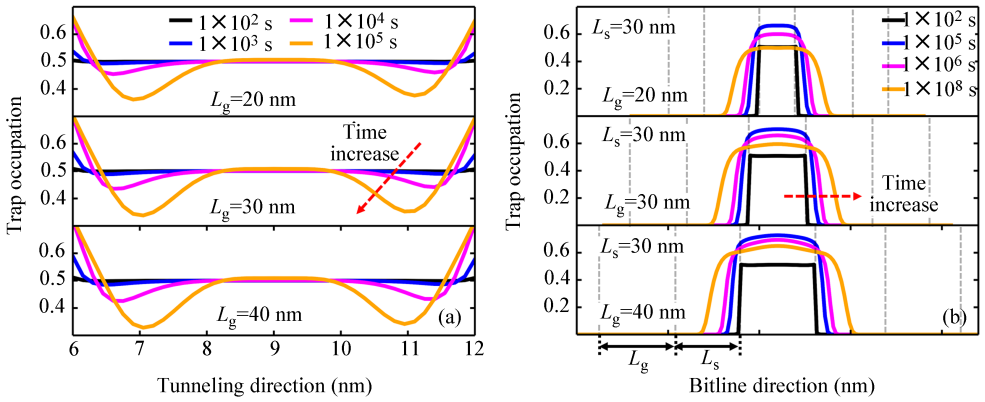

Figure 11 (Color online) Trap occupations along (a) the tunneling direction under the gate electrode and (b) the BL direction at the blocking oxide interface.

severe retention problems under high temperatures. The rationale behind this condition is depicted in Figure 11(b), where the lateral charge profile evolutions with retention times along the BL direction at the tunneling oxide interface are respectively plotted for varying channel lengths. Prior to the retention time of 10^5 s, the threshold voltage shift is determined by the trapped charge distribution along the tunneling direction under the gate. A similar trend of threshold voltage shift can be ascribed to similar trap occupation distributions along the tunneling direction. After a retention time of 10^5 s, lateral spreading of trapped charge begins to play a dominant role. Lateral charge spreading is more critical for memory cells with smaller channel lengths.

As discussed earlier, lateral spreading of trapped charge and tunneling-out through a tunneling/blocking oxide are two major charge loss mechanisms in CT NAND flash structures. To compare and contrast the charge loss behavior of these two mechanisms, a simulation of the aforementioned memory cells with a tunneling oxide of medium thickness under varying temperatures was performed. The ratios of laterally spreading and tunneling-out electrons to the initial amount of total trapped charge under different temperatures are plotted in Figure 12(a) and (b), respectively.

The amount of laterally spread trapped charge also shows a dependence on the thickness of the charge storage layer, which is depicted in Figure 13, wherein simulation was performed under the assumption that for different cases, retention starts with the same trapped charge density.

To understand the influence of lateral charge spreading on neighboring cells, retention behavior after 10 years at 85 °C for a check-board programming pattern [16] was simulated, in which the monitored cell was a central erased cell without charge storage. It was assumed that the monitored cell was only

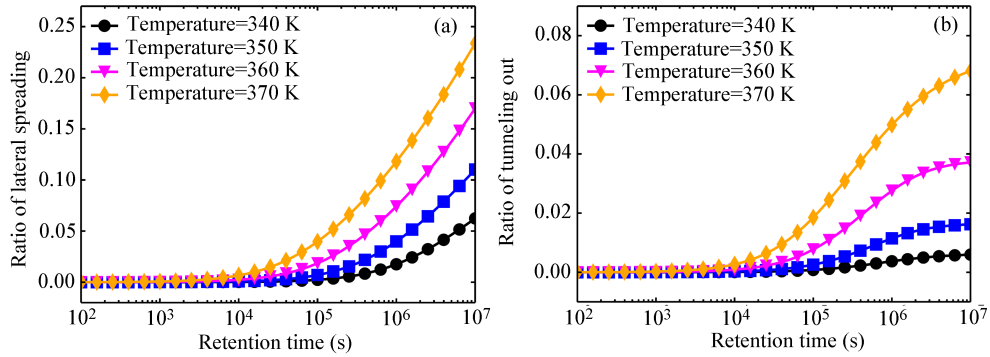

Figure 12 (Color online) The ratios of (a) laterally spreading and (b) tunneling-out electrons to the initial amount of total trapped electrons with evolution of retention under varying temperatures.

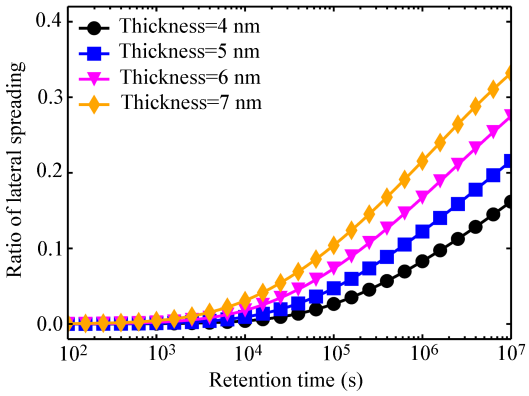

Figure 13 (Color online) The ratios of laterally spreading electrons to the initial amount of total trapped electrons versus retention time in memory cells with different trapping layer thicknesses.

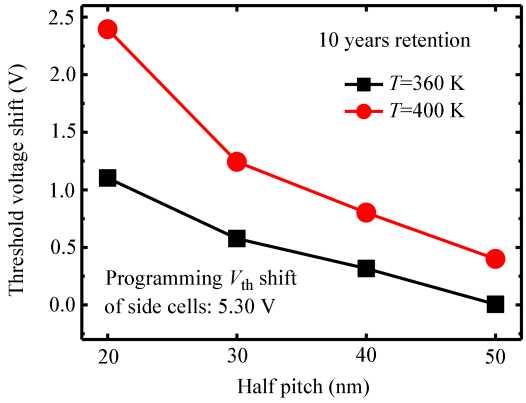

Figure 14 (Color online) Threshold voltage shift of central cell after 10 years for different cases of channel and spacer lengths. The initial threshold voltage shift of the side cells is 5.3 V. Note that in this figure, half pitch = gate length = spacer length.

influenced by two programmed neighboring cells in the same BL. Under high temperatures, since the transported electrons from the neighboring cells can reach the central cell region, the threshold voltage of the central cell correspondingly increases. The degree of this threshold voltage increase is correlated to the channel lengths of the cells and spacer lengths between two adjacent cells, as shown in Figure 14.

4 Conclusion

We have presented in this study the numerical implementation of a 2-D self-consistent simulator with unified physical models for different operation regimes of charge trapping memory, able to describe 2-D charge transport in charge storage layers. Moreover, the simulator takes into consideration various tunneling processes, charge trapping/de-trapping mechanisms, and 2-D drift-diffusion transport within the storage layer. A string of three memory cells was methodically simulated and evaluated for different gate stack compositions and temperatures. Good agreement has been made with the experimental results, thus validating the viability of the implemented physical models. The simulation results consistently demonstrate that the charge transport along BL and tunneling directions plays a vital role in 3-D charge trapping memory for a variety of applications. The TB process is found to be the major charge loss mechanism under room temperature, while charge redistribution along BL and tunneling directions dominates under high temperatures. Thus, in closing, the study's simulator is deemed a useful tool for accurately assessing 3-D NAND flash memory device characteristics.

Acknowledgements The authors would like to thank Prof. Tiao Lu from the Department of Scientific and Engineering Computing, School of Mathematical Sciences, Peking University, for a host of extremely meaningful discussions. This work was supported by National Natural Science Foundation of China (Grant No. 91230107), National Basic Research Program of China (973) (Grant No. 2013CBA01604), and National High Technology Research and Development Program of China (863) (Grant No. 2015AA016501).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Im J W, Jeong W P, Kim D H, et al. A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate. In: Proceedings of IEEE Solid-State Circuits Conference, San Francisco, 2015. 1–3
2 Katsumata R, Kito M, Fukuzumi, et al. Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices. In: Proceedings of Symposium on VLSI Technology, Kyoto, 2009. 136–137
3 Jang J, Kim H S, Cho W, et al. Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND flash memory. In: Proceedings of Symposium on VLSI Technology, Kyoto, 2009. 192–193
4 Kim J, Hong A J, Kim S M, et al. Novel vertical-stacked-array-transistor (VSAT) for ultra-high-density and cost-effective NAND flash memory devices and SSD (solid state drive). In: Proceedings of Symposium on VLSI Technology, Kyoto, 2009. 186–187
5 Choi E S, Yoo H S, Joo H S, et al. A novel 3D cell array architecture for terra-bit NAND flash memory. In: Proceedings of International Memory Workshop, Monterey, 2011. 1–4
6 Kar G S, van den Bosch S, Cacciato A, et al. Novel bi-Layer poly-silicon channel vertical flash cell for ultrahigh density 3D SONOS NAND technology. In: Proceedings of International Memory Workshop, Monterey, 2011. 1–4
7 Lu C Y, Lue H T, Chen Y C. State-of-the-art flash memory devices and post-flash emerging memories. Sci China Inf Sci, 2011, 54: 1039–1060
8 Tanaka H, Kido M, Yahashi K, et al. Bit cost scalable technology with punch and plug process for ultra high density flash memory. In: Proceedings of Symposium on VLSI Technology, Kyoto, 2007. 14–15
9 The International Technology Roadmap for Semiconductor (ITRS). 2013 Version. http://www.itrs2.net/2013-itrs.html
10 Han Y L, Huo Z L, Li X K, et al. Investigation of charge loss mechanism of thickness-scalable trapping layer by variable temperature Kelvin probe force microscopy. IEEE Electron Dev Lett, 2013, 34: 870–872
11 Padovani A, Larcher L, Heh D, et al. Modeling TANOS memory program transients to investigate charge-trapping dynamics. IEEE Electron Dev Lett, 2009, 30: 882–884
12 Vianello E, Driussi F, Palestri P, et al. Impact of the charge transport in the conduction band on the retention of Si-nitride based memories. In: Proceedings of Solid-State Device Research Conference, Edinburgh, 2008. 107–110
13 Mauri A, Amoroso S M, Compagnoni C M, et al. Comprehensive numerical simulation of threshold-voltage transients in nitride memories. Solid-State Electron, 2011, 56: 23–30
14 Maconi A, Arreghini A, Compagnoni C M, et al. Comprehensive investigation of the impact of lateral charge migration on retention performance of planar and 3D SONOS devices. Solid-State Electron, 2012. 74: 64–70
15 Maconi A, Arreghini A, Compagnoni C M, et al. Impact of lateral charge migration on the retention performance of planar and 3D SONOS devices. In: Proceedings of Solid-State Device Research Conference, Helsinki, 2011. 195–198
16 Park S, Choi S, Jeon K S, et al. A 3D simulation of the lateral charge spreading effect in charge trapping NAND flash memory. In: Proceedings of International Conference on Simulation of Semiconductor Processes and Devices, Denver, 2012. 376–379
17 Lun Z Y, Liu S H, Zhao K, et al. Two-dimensional self-consistent simulation on program/retention operation of charge trapping memory. In: Proceedings of International Workshop on Computational Electronics, Paris, 2014. 81–82
18 Lun Z Y, Liu S H, He Y, et al. Investigation of retention behavior for 3D charge trapping NAND flash memory by 2D self-consistent simulation. In: Proceedings of International Conference on Simulation of Semiconductor Processes and Devices, Yokohama, 2014. 141–144
19 Shockley W, Read W T. Statistics of the recombinations of holes and electrons. Phys Rev, 1952, 87: 835–842
20 Song Y C, Du G, Yang J F, et al. Influence of the Poole-Frenkel effect on programming and erasing in charge trapping memories. In: Proceedings of International Conference on Simulation of Semiconductor Processes and Devices, Vienna, 2007. 1–4
21 Gehring A, Selberherr S. Modeling of tunneling current and gate dielectric reliability for nonvolatile memory devices. IEEE Trans Device Mater Reliab, 2004, 4: 306–319
22 Scharfetter D L, Gummel H K. Large-signal analysis of a silicon Read diode oscillator. IEEE Trans Electron Dev, 1969, 16: 64–77
23 Padovani A, Larcher L, Heh D, et al. Temperature effects on metal-alumina-nitride-oxide-silicon memory operations. Appl Phys Lett, 2010, 96: 223505