Investigation of gate oxide traps effect on NAND flash memory by TCAD simulation*

He-Kun Zhang(章合坤)[1], Xuan Tian(田璇)[2], Jun-Peng He(何俊鹏)[1], Zhe Song(宋哲)[2], Qian-Qian Yu(蔚倩倩)[2], Liang Li(李靓)[2], Ming Li(李明)[1], Lian-Cheng Zhao(赵连城)[1], and Li-Ming Gao(高立明)[1,†]

[1] *School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*
[2] *SanDisk Info Tech Shanghai, Shanghai* 200241, *China*

(Received 15 October 2019; revised manuscript received 30 December 2019; accepted manuscript online 9 January 2020)

The effects of gate oxide traps on gate leakage current and device performance of metal–oxide–nitride–oxide–silicon (MONOS)-structured NAND flash memory are investigated through Sentaurus TCAD. The trap-assisted tunneling (TAT) model is implemented to simulate the leakage current of MONOS-structured memory cell. In this study, trap position, trap density, and trap energy are systematically analyzed for ascertaining their influences on gate leakage current, program/erase speed, and data retention properties. The results show that the traps in blocking layer significantly enhance the gate leakage current and also facilitates the cell program/erase. Trap density $\sim 10^{18}$ cm^{-3} and trap energy ~ 1 eV in blocking layer can considerably improve cell program/erase speed without deteriorating data retention. The result conduces to understanding the role of gate oxide traps in cell degradation of MONOS-structured NAND flash memory.

Keywords: NAND flash reliability, gate oxide traps, trap-assisted tunneling, TCAD simulation

PACS: 85.25.Hv, 85.40.–e, 85.35.Gv **DOI:** 10.1088/1674-1056/ab695f

1. Introduction

NAND flash has become the mainstream of data storage due to its high density, low cost, and much lower latency than hard disk drive (HDD). The NAND flash has successfully transferred from two-dimension (2D) to three-dimension (3D), in order to keep reducing cost for one bit. However, the cell reliability becomes more challenging due to smaller cell size and complicated 3D NAND process.

Most of NAND flash use metal–oxide–nitride–oxide–silicon (MONOS) multiple layer as the memory cell, in whichthe high-k block oxide is placed between gate and charge trapped layer (CTL), and a very thin bandgap engineered oxide on the top of channel silicon serves as a tunneling layer. Electrons are injected into or rejected out of CTL through Fowler–Nordheim (FN) tunneling due to high electric field in tunneling layer.[1,2] The high electric field used for progra/erase cycles creates traps or defects in oxide layer, leading to cell degradation and causing reliability issues.[3,4] The effects of traps in aluminum oxide on the fast V_t shift, cell program/erase (P/E) operations, and data retention properties of TANOS flash memory have been investigated experimentally and analyticaly.[5,6] Data retention phenomena of NAND flash device relating to temperature, program pattern, and bake time have been studied by TCAD simulation.[7] However, the influence of traps on MONOS-structured memory cell by comprehensively considering the trap position, trap density, and trap energy has not yet been fully discussed, especially by taking advantage of TCAD device modeling.

In this work, Sentaurus TCAD is adopted to simulate the memory characteristics of MONOS-structured NAND flash by placing hole traps in oxide layers. The gate leakage current (I_g) is more intensely affected by the traps in blocking layer rather than tunneling layer. Besides, the influence of trap density (N_t) and trap energy (E_t) on P/E speed and data retention properties are investigated by applying TAT model to the current transport. A specific trap profile ($N_t \sim 10^{18}$ cm^{-3} and $E_t = 1$ eV) shows a 10-times higher program speed with good data retention properties.

2. Experimental methods

The device simulation and characterizations were performed with Sentaurus device. Figure 1 shows the diagram of MONOS cell, where a SiO_2/SiON/SiO_2 multilayer is composed of the tunneling layer, SiO_2 layer between Si_3N_4 CTL and metal gate working as a blocking layer. The metal gate was made up of tungsten and a thin TiN layer, which is used to prevent tungsten from diffusing into oxide. The FN tunneling and direct tunneling were taken into account for electron tunneling during P/E operation. The current transport through the MONOS layer for gate leakage was simulated with the TAT model.[8,9] To include these mechanisms, several models were activated in the TCAD. Shockley–Read–Hall (SRH) recombination model was used to simulate recombination through deep defect levels in the gap;[10,11] mobility mod-

*Project supported by the SanDisk Info Tech Shanghai, China and the Institute of Microelectronic Materials & Technology, School of Materials Science and Engineering, Shanghai Jiao Tong University, China.

†Corresponding author. E-mail: liming.gao@sjtu.edu.cn

els including doping-dependent mobility degradation model and high field mobility saturation model were used to simulate the scattering of the carriers by charged impurity ions and the carrier drift velocity in high electric fields;[12–15] Poole–Frenkel model was used for the explanation of transport effects in dielectrics;[16–18] trap model was used to simulate single-energy level trap in Si_3N_4 layer; tunneling models including direct tunneling model and FN tunneling model were used to simulate main tunneling effects of this device and explain gate leakage current, *etc*.[19–21]

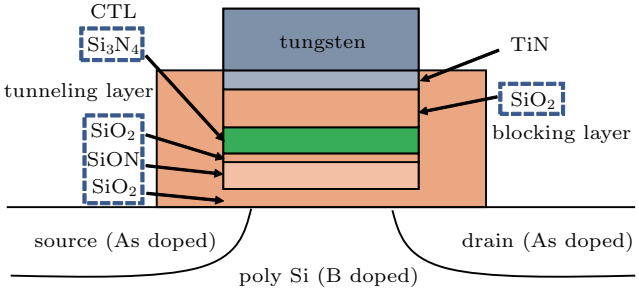

Fig. 1. Diagram of multilayer structure of MONOS cell.

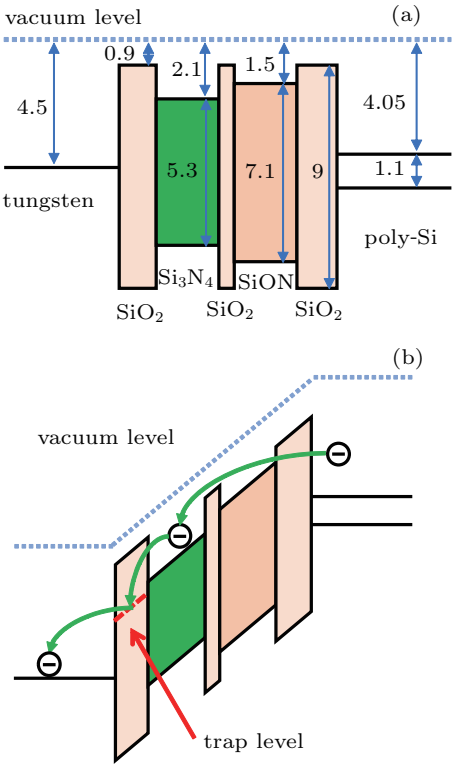

Fig. 2. Band diagram of MONOS layer under different bias conditions (energy in units of eV) at (a) thermal equilibrium state and (b) program state.

The band diagram of MONOS layer is shown in Fig. 2. In TAT model, when forward bias for program was used, the traps in blocking oxide can serve as an intermediate site for electrons tunneling to the gate. Tunneling current was therefore strongly enhanced as the barrier near the trap decreased due to Poole–Frenkel effect.[22] Since there could exist traps in both blocking oxide and tunneling oxide, the TAT process of MONOS layer became more complex than the MOS structure. In this simulation, traps were placed in blocking layer and tunneling layer respectively with N_t ranging from 1×10^{16} cm^{-3} to 1×10^{21} cm^{-3} based on previous flash memory studies.[23–26] The studies have been reported on TAT model in flash memory with $E_t = 4$ eV and on oxide trap influence with E_t from 0 eV to 4 eV.[27–29] Thus, trap energy (E_t) ranged from 0 eV to 4 eV under conduction band in this study. The program and erase speed were investigated by comparing trapped charge generation speed and dynamic V_t shift during P/E operation. The data retention properties at different trap profiles were also evaluated for device reliability.

Generally, higher temperature enhances gate leakage current and accelerates oxide breakdown because trap activation energy decreases as temperature increases.[30,31] Since this simulation mainly focus on the influence of trap concentration and trap energy, it was performed at room temperature (300 K).

3. Results and discussion

3.1. Leakage in program process

In order to study the influence of oxide traps on gate leakage, traps are placed in blocking or tunneling layer with trap energy of 2 eV. The gate leakage–gate bias (I_g–V_g) curves are obtained by ramping up V_g from 0 V to 27 V in 10^{-4} s as shown in Figs. 3(a) and 3(b). It is obvious that the traps in

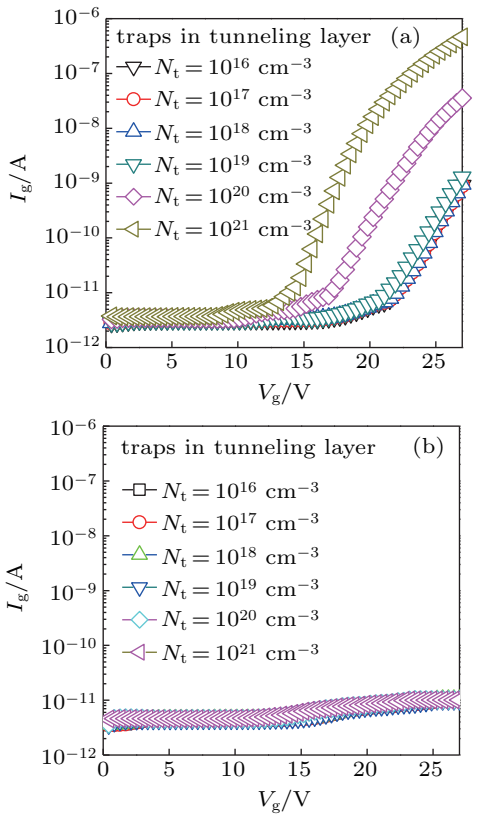

Fig. 3. Gate leakage currents varying with gate bias for (a) different trap densities in blocking layer at 2 eV and (b) different trap densities in tunneling layer at 2 eV.

tunneling layer just slightly enhance I_g (smaller than 10^{-11} A even at high V_g), while the traps in blocking layer dramatically increase I_g after N_t exceeds 1×10^{19} cm^{-3}. In the case of N_t below 1×10^{19} cm^{-3}, the maximum values of I_g for the traps in blocking layer are still about 100 times larger than those for the traps in tunneling layer. The influence of trap energy on gate leakage is shown in Fig. 4. It shows that I_g decreases for deeper traps, because the tunneling-related capture rate has inversely exponential relationship with trap energy as reported previously.[32–35]

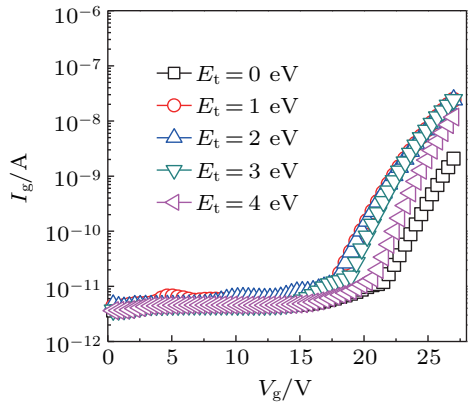

Fig. 4. Gate leakage currents varying with gate bias for different trap energies ($N_t = 1 \times 10^{20}$ cm^{-3}).

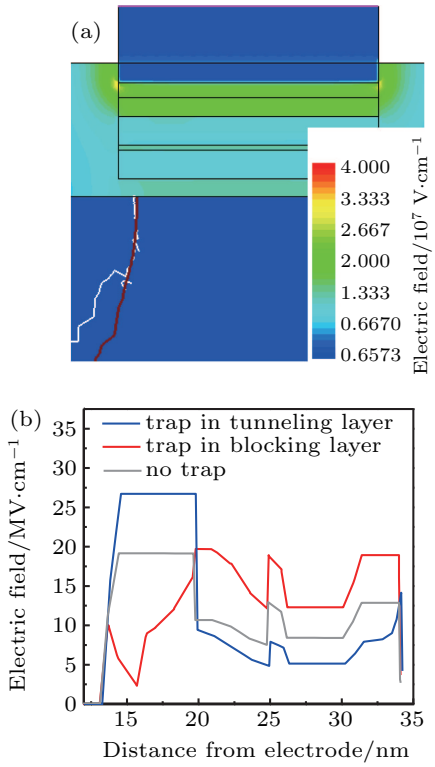

Fig. 5. (a) Electric field distribution of the whole device, and (b) plots of electric field across layers *versus* distance fron electrode.

To explain these results, the electric field distribution with different trap profiles is simulated. Figure 5(a) shows the electric field distribution of the whole device. By placing hole traps in blocking layer, electric field is redistributed as shown in Fig. 5(b). Specifically, in program, electric field decreases across blocking layer but increases in CTL and tunneling layer. An opposite result is observed by placing traps in tunneling layer. Figure 6(a) shows the influence of trap density on electric field distribution. Higher trap density in the blocking layer more strongly reduces the electric field on it. For different trap energyies it is apparent that the shallow traps ($E_t = 2$ eV) more significantly affect electric field than deep traps ($E_t = 4$ eV) as indicated in Fig. 6(b). According to the empirical trap-assisted tunneling model based on SILC measurements, the TAT current can be written as

$$J_{TAT} = CE_{diel} \exp(-q\phi_a / k_R T), \tag{1}$$

where ϕ_a is the activation energy, E_{diel} is the electric field in dielectric, and C is a constant. Clearly, I_g is directly proportional to E_{diel}. Apparently, the enhanced gate leakage does not result from the strong electric field, but possibly is caused by lower activation energy for TAT tunneling at higher trap density or smaller trap energy.

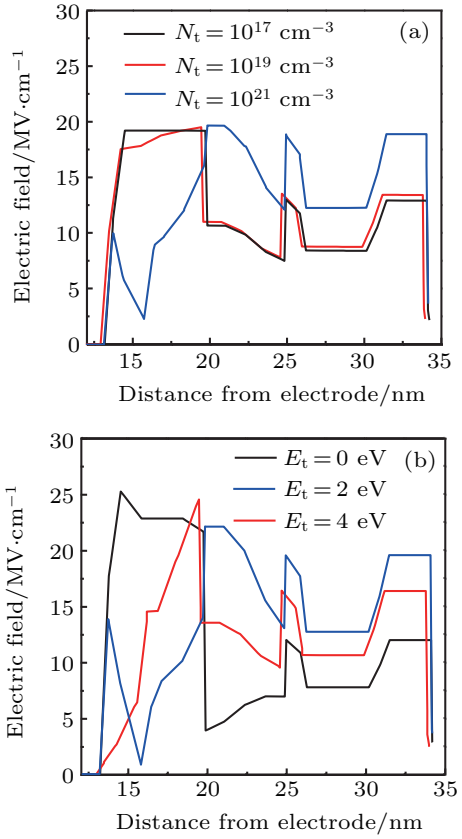

Fig. 6. Electric field profile across MONOS layer *versus* distance from electrode by placing hole traps with (a) different trap densities, and (b) different trap energies.

3.2. Program speed

The redistributed electric field could affect cell program as well. The enhanced program speed, which is contributed by TAT in blocking layer as mentioned above, is clearly demonstrated in Fig. 7(a) by measuring V_t shift as a function of programming time. When N_t increases to 1×10^{18} cm^{-3}, faster

V_t ramping up is observed. Comparing with the saturated V_t of non-trap cell, $N_t = 1 \times 10^{18}$ cm^{-3} saved 10× time to achieve the same V_t, as noted in Fig. 8(a). When trap density exceeds 1×10^{18} cm^{-3}, program speed also increases at the initial stage, however, V_t ramping up speed decreases later because leakage current overwhelms the tunneling current. When N_t increases to 1×10^{20} cm^{-3}, V_t intensely decreases after 1 μs due to higher TAT probability at higher N_t for electrons tunneling from CTL to gate, which is proved by the larger I_g in Fig. 3(a). The progress of trapped charge in CTL with gate bias is plotted in Fig. 7(b). The process is finished within 1 ms. Compared with non-trap cell, hole traps in blocking oxide can reduce the gate voltage for achieving the saturated trapped charge density. Specifically, 1×10^{20} cm^{-3} traps in blocking layer reduce the maximal V_t by about 4 V for a saturated charge density.

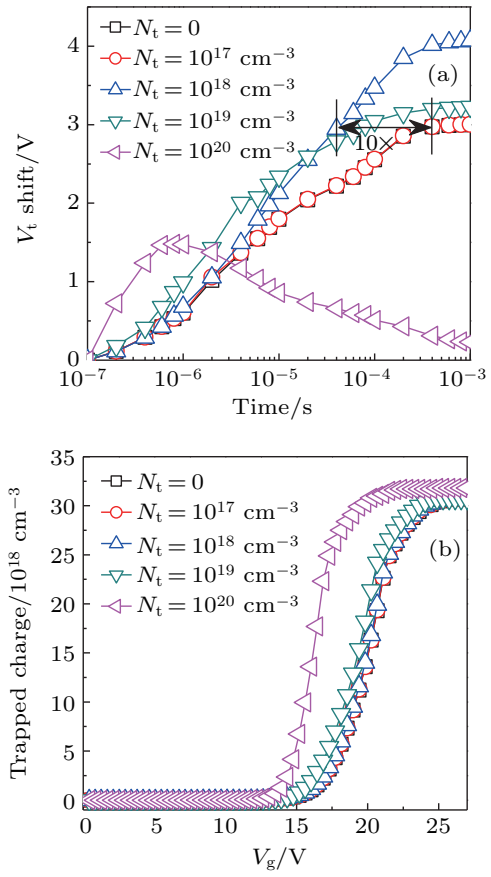

Fig. 7. (a) Plots of V_t shift *versus* program time at different trap densities, and (b) plots of trapped charge *versus* gate bias at different trap densities in program process.

Besides, the influence of trap energy on program speed is investigate, and the results are shown in Fig. 8(a). For various trap energy, $E_t = 1$ eV most strongly improved the program speed without V_t downshift at long program time. Again, we plot the progress of trapped charge in CTL with gate bias. An intermediate trap energy ($E_t = 1$ eV–3 eV) speeds up the program process by ∼ 0.8 V decrement for program voltage as shown in Fig. 8(b). From the above result, the trap energy does not enhance the program speed as intensely as trap density.

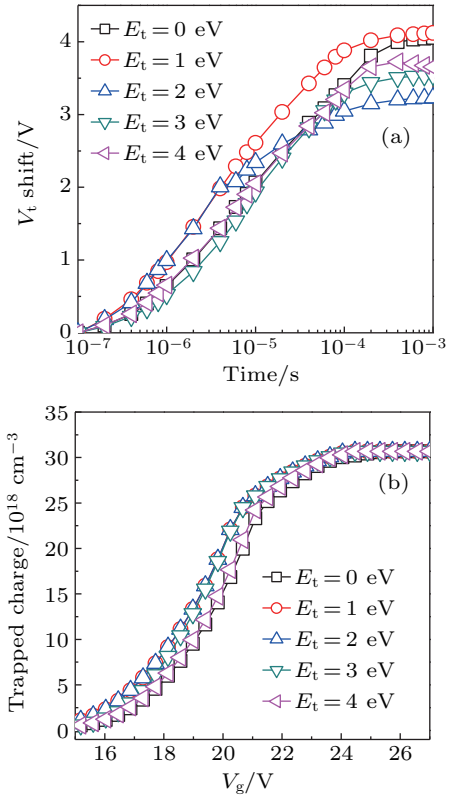

Fig. 8. (a) Plots of V_t shift *versus* program time at different trap energies and (b) plots of trapped charge *versus* gate bias at different trap energies.

3.3. Leakage in erase process

To investigate the erase process of the device with trap in blocking and tunneling layer, we apply voltage to gate from 0 V to −20 V with other settings being the same as those in program experiments. As we can see from Figs. 9(a) and 9(b), it is obvious that the traps in tunneling layer just slightly enhance I_g (smaller than 10^{-11} A even in high V_g), while the traps in blocking layer strongly increase I_g when V_g exceeds 12 V. After N_t exceeds 1×10^{19} cm^{-3}, I_g becomes larger. In the case of traps in tunneling layer, electrons at metal gate are blocked by blocking layer. Therefore, the TAT process in blocking layer is very weak, so I_g is below 10^{-11} A. In the case of traps in blocking layer, when V_g exceeds 12 V, the TAT and FN tunneling are enhanced, so a large number of electrons tunnel to CTL. Since the traps in blocking layer will enhance the electric field of tunneling layer, which is proved by Fig. 6(a), the tunneling current through the tunneling layer will also be intensified. Therefore, the traps in blocking layer can significantly increase the gate leakage current.

Again, we simulate the influence of trap energy ($E_t =$ 1 eV–4 eV) on I_g and the results are shown in Fig. 10(a). It is found that the shallow traps ($E_t = 1$ eV–2 eV) can significantly influence I_g, which indicates that shallow traps strongly reduce the activation energy for TAT, according to Eq. (1) or further confirmation, we plot the trapped charge density as a

function of V_{gate} as indicated in Fig. 10(b). This figure shows that the trapped charge increases dramatically after V_{gate} has exceeded about 11 V in the case of traps in blocking layer. When $N_t = 1 \times 10^{20}$ cm^{-3}, the trapped hole in CTL decreases compared with the scenario of lower trap density. This is because the enhanced TAT due to traps in blocking layer will make electrons from the gate easier to tunnel to CTL, which increases the possibility of electron–hole recombination, leading to lower trapped hole density in CTL. This result confirms the existence of trap assisted tunneling and its role in tunneling of erase process.

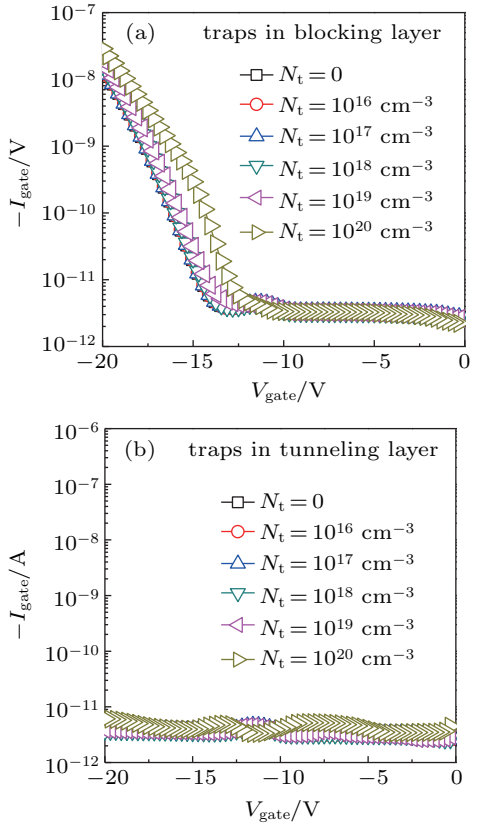

Fig. 9. Gate leakage currents varying with gate bias under different trap conditions in erase process: (a) different trap densities in blocking layer at 2 eV and (b) different trap densities in tunneling layer at 2 eV.

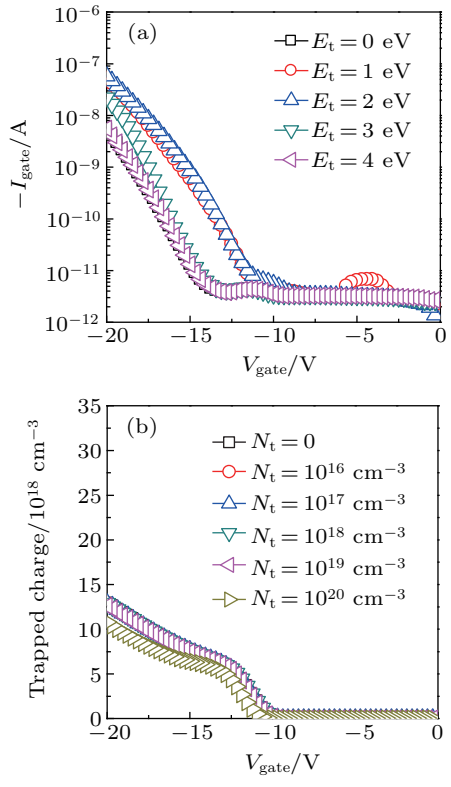

Fig. 10. (a) Gate leakage currents varying with gate bias in erase process at different trap energies, and (b) trapped charges changing with gate bias at different trap densities in erase process.

3.4. Erase speed

Also, erase speed is discussed. For erase process, we reduce the gate voltage from 0 V to −20 V and read the V_t of cell at $t = 10^{-7}$ s, 10^{-6} s, 10^{-5} s, 10^{-4} s, 10^{-3} s. Then we read the V_t of device by applying 0.5 V to the drain and increasing the gate voltage from −2 V to 4 V.

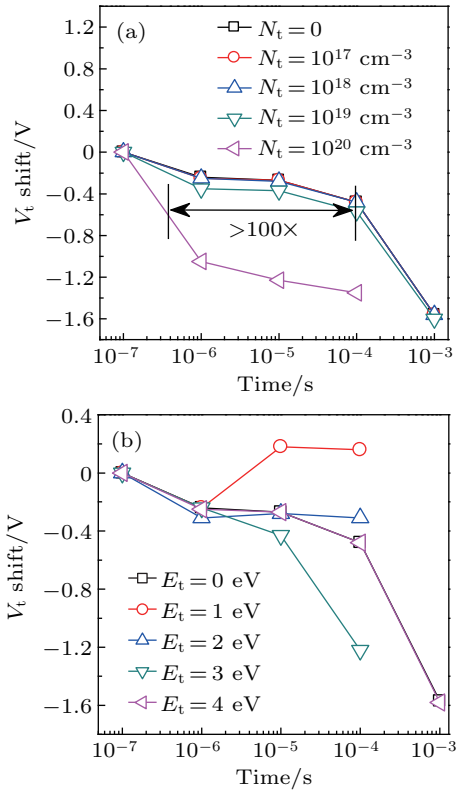

Fig. 11. The V_t shifts varying with erase time for (a) different trap densities and (b) different trap energies, respectively.

As figure 11(a) shows, when N_t increases to 1×10^{19} cm^{-3} faster, the V_t shifting down is observed. The erase speed, when N_t reaches to 1×10^{20} cm^{-3}, is extraordinarily high, more than 100 times that when $N_t = 1 \times 10^{18}$ cm^{-3}. As discussed in Fig. 6(a), traps in blocking layer enhance the electric field in tunneling layer, thus enhancing the erase speed. As erase time increases, the trapped holes become saturated and erase speed decreases in $t = 10^{-6}$ s–10^{-4} s. In Fig. 11(b), we show the influence of trap energy on erase speed. Generally, shallow traps ($E_t = 1$ eV–3 eV) can improve the erase speed when trap density is high ($N_t = 1 \times 10^{19}$ cm^{-3} in Fig. 11).

However, in the case of $E_t = 1$ eV, V_t shifting down stops and begins to recover and even rams up in $t = 10^{-6}$ s–10^{-4} s, resulting in degraded erase efficiency. This is because electrons tunneling from gate to CTL becomes much stronger than holes tunneling from channel to CTL. Therefore, more electrons accumulate in CTL and thus increase V_t. This indicates that the high trap density ($> \times 10^{19}$ cm^{-3}) at shallow trap energy (~ 1 eV) is fatal for erase operation due to large leakage current and low erase efficiency.

3.5. Data retention

Finally, the influence of gate oxide traps on data retention properties is investigated. We program the cell for 10^{-4} s in program retention experiment and erase the cell for 10^{-3} s in erase retention experiment. Then we read the V_t of device at $t = 10^1$ s–10^5 s by applying 0.5 V to drain and ramping up the gate voltage from -2 V to 4 V.

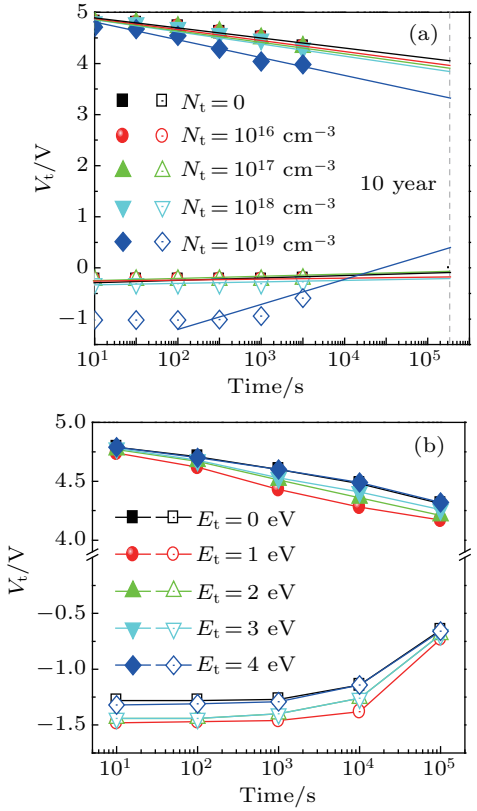

Fig. 12. Retention characteristics of MONOS cell with (a) different trap densities, and (b) different trap energies in blocking layer.

The V_t windows of various trap densities are compared with each other in Fig. 12(a). For trap densities $\leq 1 \times 10^{18}$ cm^{-3}, the initial V_t and final V_t window are barely affected. However, as N_t increases to 1×10^{19} cm^{-3}, data retention is notably deteriorated. Specifically, erased V_t has more V_t loss than programmed one, because positive traps in blocking layer enhance electrons tunneling from substrate to CTL as demonstrated in Fig. 5(b). From the above results, data retention at trap density $N_t < 1 \times 10^{18}$ cm^{-3} is comparable to non-trap cell. Using fitting curve and extrapolation method, we obtain V_t window ~ 5 V at $t = 0$ s and V_t window ~ 4 V after 10 years.

For $N_t = 1 \times 10^{19}$ cm^{-3} at different E_t values, $E_t = 1$ eV shows the worst data retention property. For deeper traps ($E_t \geq 2$ eV), data retention property is recovered. Besides, erased V_t with more V_t loss can also be found, especially after 10^{-4} s, which means that data retention characteristic turns worse. Deeper traps can improve data retention compared with shallow traps because electrons tunneling through block oxide are intensely prevented, which is confirmed by the reduced gate leakage in Fig. 4. From the figure, we obtain V_t window ~ 6 V at $t = 0$ s and V_t window ~ 5 V at $t = 10^5$ s.

4. Conclusions

In this work, the influences of trap position, trap density, and trap energy on device characteristics of MONOS-structured NAND flash are investigated through TCAD simulation. It is found that traps in blocking layer significantly increase the gate leakage in both program and erase process due to stronger TAT process. Besides, traps in blocking layer increase the program speed and the erase speed in a short period (less than 1 μs), but slows down cell program in long time range if trap density is over 1×10^{19} cm^{-3}. Furthermore, for trap density $\geq 1 \times 10^{19}$ cm^{-3}, data retention is obviously weakened, especially after 10^4 s. From the simulation results, trap in blocking layer with a density of $\sim 10^{18}$ cm^{-3} at ~ 1 eV can increase single cell program speed by 10 times and increase erase speed slightly, and retain a V_t window as large as 4 V after 10 years. The result conduces to understanding the role of gate oxide traps in cell degradation of MONOS-structured NAND flash memory.

References

[1] Cha S Y, Kim H J and Choi D J 2010 *3rd International Nanoelectronics Conference (INEC)*, January 3–8, Hong Kong, China, p. 1210
[2] Monzio Compagnoni C, Spinelli A S, Gusmeroli R, Lacaita A L, Beltrami S, Ghetti A and Visconti A 2007 *IEEE International Electron Devices Meeting*, December 10–12, Washington, USA, p. 165
[3] Albert F, Seol K S, Na J H, Hur S H, Choi J D and Kim K 2009 *IEEE International Electron Devices Meeting (IEDM)*, December 7–9, Baltimore, USA, p. 1
[4] Carmona M, Lopez L, Ogier J and Goguenheim D 2014 *IEEE International Integrated Reliability Workshop Final Report (IIRW)*, October 12–16, South Lake Tahoe, USA, p. 147
[5] Padovani A, Larcher L, Marca V D, Pavan P, Park H and Bersuker G 2011 *J. Appl. Phys.* **110** 014505
[6] Toledano-Luque M, Degraeve R, Zahid M B, Kaczer B, Kittl J, Jurczak M, Groeseneken G and Van Houdt J 2009 *IEEE International Electron Devices Meeting (IEDM)*, December 7–9, Baltimore, USA, p. 1
[7] Oh D, Lee B, Kwon E, Kim S, Cho G, Park S, Lee S and Hong S 2015 *IEEE International Memory Workshop (IMW)*, May 17–20, Monterey, USA, p. 1
[8] Lee K and Shin H 2017 *IEEE Trans. Dev. Mater. Reliab.* **17** 758
[9] Vianello E, Driussi F, Esseni D, Selmi L, Duuren M J V and Widdershoven F 2006 *European Solid-State Device Research Conference*, September 19–21, Montreux, Switzerland, p. 403
[10] Zimmerman W 1973 *Electron. Lett.* **9** 378
[11] Cuevas A, Stocks M, McDonald D, Kerr M and Samundsett C 1999 *IEEE Trans. Electron Dev.* **46** 2026

[12] Masetti G, Severi M and Solmi S 1983 *IEEE Trans. Electron Dev.* **30** 764
[13] Arora N D, Hauser J R and Roulston D J 1982 *IEEE Trans. Electron Dev.* **29** 292
[14] Barnes J J, Lomax R J and Haddad G I 1976 *IEEE Trans. Electron Dev.* **23** 1042
[15] Meinerzhagen B and Engl W L 1988 *IEEE Trans. Electron Dev.* **35** 689
[16] Swain R, Jena K and Lenka T R 2016 *IEEE Trans. Electron Dev.* **63** 2346
[17] Lau W S, Wong O Y and Wong H 2013 *IEEE International Conference of Electron Devices and Solid-state Circuits*, June 3–5, Hong Kong, China, p. 1
[18] Southwick R G, Reed J, Buu C, Butler R, Bersuker G and Knowlton W B 2010 *IEEE Trans. Dev. Mater. Reliab.* **10** 201
[19] Masetti G, Severi M and Solmi S 1983 *IEEE Trans. Electron Dev.* **30** 764
[20] Schenk A and Heiser G 1997 *J. Appl. Phys.* **81** 7900
[21] Fossum J G, Mertens R P and Lee D S 1983 *Solid-State Electron.* **26** 569
[22] Swain R, Jena K and Lenka T R 2016 *IEEE Trans. Electron Dev.* **63** 2346
[23] Blauwe J, Houdt J V, Wellkens D, Groeseneken G and Maes H E 1998 *IEEE Trans. Electron Dev.* **45** 1745
[24] You J H, Kim H W, Kim D H, Kim T W and Lee K W 2011 *International Conference on Simulation of Semiconductor Processes and Devices*, September 8–10, Osaka, Japan, p. 199
[25] Yin Y, Wu J, Lan L, Li X and Wang Q 2013 *IEEE International Conference on Solid Dielectrics (ICSD)*, June 30–July 4, Bologna, Italy, p. 726
[26] Najam F, Yu Y S, Cho K H, Yeo K H, Kim D, Hwang J S, Kim S and Hwang S W 2013 *IEEE Trans. Electron Dev.* **60** 2457
[27] Kamohara S, Park D and Hu C 1998 *IEEE International Reliability Physics Symposium Proceedings. 36th Annual)*, March 31–April 2, Reno, USA, p. 57
[28] Yamada R, Mori Y, Okuyama Y, Yugami J, Nishimoto T and Kume H 2000 *IEEE International Reliability Physics Symposium Proceedings. 38th Annual*, April 10–13, San Jose, USA, p. 200
[29] Driussi F, Iob R, Esseni D, Selmi L, Schaijk R V and Widdershoven F 2004 *IEEE Trans. Electron Dev.* **51** 1570
[30] Suehle J S and Chaparala P 1997 *IEEE Trans. Electron Dev.* **44** 801
[31] Strong A W, Wu E Y, Vollertsen R P, Sune J, Rosa G L, Sullivan T D and Rauch S E 2009 *Reliability Wearout Mechanisms in Advanced CMOS Technologies* (Hoboken: John Wiley & Sons) p. 272
[32] Chaneliere C, Autran J L and Devine R A B 1999 *J. Appl. Phys.* **86** 480
[33] Sahhaf S, Degareve R, Roussel P J, Kauerauf T, Kaczer B and Groeseneken G 2007 *IEEE Int. Electron Devices Meeting Techn. Digest*, December 11–14, San Francisco, USA, p. 501
[34] Lo V L, Pey K L, Tung C H and Ang D S 2007 *Proc. 45th Int. Reliability Physics Sym. IEEE*, April 15–19, Phoenix, USA, p. 576
[35] Sakura T, Utsunomiya H, Kamakura Y and Taniguchi K 1998 *IEEE Int. Electron Devices Meeting Techn. Digest*, December 6–9, San Francisco, USA, p. 183