Modeling of GIDL–Assisted Erase in 3–D NAND Flash Memory Arrays and Its Employment in NOR Flash–Based Spiking Neural Networks

Gerardo Malavena

Abstract Since the very first introduction of three-dimensional (3–D) vertical-channel (VC) NAND Flash memory arrays, gate-induced drain leakage (GIDL) current has been suggested as a solution to increase the string channel potential to trigger the erase operation. Thanks to that erase scheme, the memory array can be built directly on the top of a n^+ plate, without requiring any p-doped region to contact the string channel and therefore allowing to simplify the manufacturing process and increase the array integration density. For those reasons, the understanding of the physical phenomena occurring in the string when GIDL is triggered is important for the proper design of the cell structure and of the voltage waveforms adopted during erase. Even though a detailed comprehension of the GIDL phenomenology can be achieved by means of technology computer-aided design (TCAD) simulations, they are usually time and resource consuming, especially when realistic string structures with many word-lines (WLs) are considered. In this chapter, an analysis of the GIDL-assisted erase in 3–D VC NAND memory arrays is presented. First, the evolution of the string potential and GIDL current during erase is investigated by means of TCAD simulations; then, a compact model able to reproduce both the string dynamics and the threshold voltage transients with reduced computational effort is presented. The developed compact model is proven to be a valuable tool for the optimization of the array performance during erase assisted by GIDL. Then, the idea of taking advantage of GIDL for the erase operation is exported to the context of spiking neural networks (SNNs) based on NOR Flash memory arrays, which require operational schemes that allow single-cell selectivity during both cell program and cell erase. To overcome the block erase typical of NOR Flash memory arrays based on Fowler-Nordheim tunneling, a new erase scheme that triggers GIDL in the NOR Flash cell and exploits hot-hole injection (HHI) at its drain side to accomplish the erase operation is presented. Using that scheme, spike-timing dependent plasticity (STDP) is implemented in a mainstream NOR Flash array and array learning is successfully demonstrated in a prototype SNN. The achieved results represent an important step for the development of large-scale neuromorphic systems based on mature and reliable memory technologies.

G. Malavena (✉)
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy
e-mail: gerardo.malavena@polimi.it

L. Piroddi (ed.), *Special Topics in Information Technology*,
PoliMI SpringerBriefs, https://doi.org/10.1007/978-3-030-85918-3_4

Keywords 3–D NAND Flash memory arrays · NOR Flash memory arrays · GIDL · Neuromorphic computing · Spiking Neural Networks · Artificial Neural Networks

1 Introduction

Since their very first introduction, the performance improvement of Flash memory technologies was long achieved thanks to an uninterrupted scaling process that led to a NAND Flash cell feature size as small as 14 nm in 2015 [1]. However, as the size of the single memory cell was shrinked down to decananometer dimensions, some fundamental issues related to the increasingly complex fabrication techniques and to inherent physical limitations due to the discrete nature of charge and matter emerged, undermining both the manufacturing and the proper operation of Flash memory arrays [2, 3]. For this reason, an alternative integration paradigm has been adopted to break the classical trade-off between single-cell area and array storage density, consisting in stacking many layers of memory cells along the direction orthogonal to the wafer plane. Figure 1a, shows a schematic of a possible implementation of 3-D NAND Flash memory arrays, featuring vertical polycrystalline silicon (poly-Si) channels, contacted at the bit-line (BL) and source-line (SL) sides by n^+ regions. At the intersection between each poly-Si channel and a horizontal word-line (WL) plane, a gate-all-around (GAA) memory cell with *macaroni* structure is formed, as schematically depicted in Fig. 1b, with the oxide-nitride-oxide (ONO) stack adopted to store charge in the middle layer. Due to the lack of any p-doped regions to access the string channel in such architecture, the poly-Si channels cannot be contacted similarly to the case of planar NAND technologies. While this feature does not affect the read and program operations, the employment of a novel voltage scheme is required to increase the channel potential and trigger the emission of electrons from the storage layer or the injection of holes into it during erase. To this purpose, the voltage scheme displayed in Fig. 1c is adopted. A positive voltage ramp is applied to the BL and SL of the string while keeping to ground the WLs and the selector gates (SGs); the strong electric fields at the inner edge of the SGs are large enough to trigger the generation of electron/hole pairs by band-to-band tunneling (BTBT) [4, 5]. While electrons are swept towards the BL/SL contacts, giving rise to the so-called GIDL current, the BTBT-generated holes are directed towards the center of the string, where they accumulate and contribute to increase the channel potential. In this framework, Sect. 2 is devoted to the study of the GIDL-assisted erase in 3–D NAND Flash memory arrays by TCAD simulations first (Sect. 2.1), and, then, to the development of a compact model able to predict the string behavior and the threshold-voltage V_T evolution during erase (Sect. 2.2). All the presented results are from [6] and [7].

On the other hand, NOR Flash memory cells have never been scaled beyond the feature size of 40 nm as research efforts for embedded applications have been focused on different technologies, such as phase-change memories [9, 10]. Despite this, in the last years NOR Flash memory arrays attracted some interest also for their

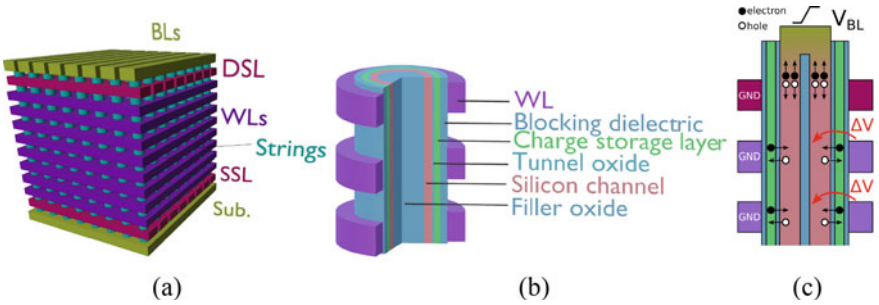

Fig. 1 **a** Schematic of a VC 3–D NAND Flash memory array and **b** of a GAA memory cell [1] (from [8] under CC BY 4.0 license). **c** Schematic of the string condition during a GIDL-assisted erase operation (a section of the string close to the BL is shown)

employment in the implementation of spiking neural networks (SNNs), representing a promising solution to outclass conventional CMOS systems based on the Von-Neumann architecture in problems dealing with unstructured data, such as image recognition and classification [11]. A mandatory condition to be met by memory arrays employed in SNNs is the possibility to tune the threshold voltage (V_T) of each cell independently of the others in both directions, meaning that single-cell selectivity not only during program and but also during erase operation is needed, with the block erase typical of Flash technologies clearly representing an obstacle for neuromorphic applications. To overcome this issue, some works have suggested design adjustments either to the cell or to the array level, with the drawback, however, of a larger array area occupancy and more complex manufacturing process [12–15]. Taking inspiration from the GIDL-assisted erase employed in 3–D VC NAND Flash memory arrays investigated in Sect. 2, the idea of moving from the classical erase scheme based on Fowler-Nordheim (FN) tunneling [16] to a novel single-cell selective one that exploits BTBT-generated HHI at the drain side is presented in Sect. 3. Exploiting such scheme, the operation of a SNN based on the STDP learning rule [17–19] exploiting a mainstream NOR Flash memory array with no modification either to the cell or to the array design is successfully demonstrated. The results presented in Sect. 3 are from [20–23].

2 GIDL–Assisted Erase in 3-D NAND Memory Arrays

2.1 Overview on String Dynamics

In order to investigate the erase operation in 3-D NAND Flash memory arrays when GIDL is triggered at the SGs, TCAD simulations were performed using a commercial device simulator (see [6] for more information about the simulation environment). As a starting point, no charge exchange between the channel and the storage layer

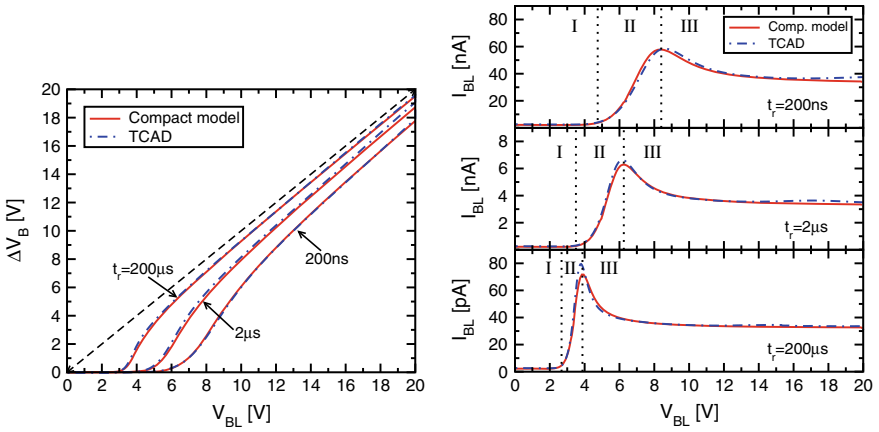

Fig. 2 **a** Comparison between the string potential V_B during the erase transient resulting from the TCAD simulations and that computed with the developed compact model. **b** Same as **a** but I_{BL} is shown. © 2018 IEEE [6]

is included. Simulation results are displayed in Fig. 2a and Fig. 2b, which report the variations of the string potential ΔV_B (the average value in the radial direction at the center of the string is considered) and the BL/SL currents $I_{BL} = I_{SL}$ during erase for different values of the BL/SL ramps rise time t_r. Results reveal that three different phases of the transient can be identified: I) $I_{BL} \approx$ constant and $\Delta V_B \approx 0$; II) I_{BL} increases steeply and the same does V_B, with rate larger than that of V_{BL}; III) I_{BL} reaches a peak and then saturates to a constant value while V_B continues to increase but at the same rate of V_{BL}. Figure 3 shows how the net charge density in the NAND string evolves during the erase transient. By comparing the former figure with Fig. 2a and Fig. 2b, it is easy to relate the charge distribution with the ΔV_B and I_{BL} transients: during phase I, BTBT-generated holes are a few and the string is approximately depleted of charge; during phase II, holes start to rule the string electrostatics, but they are confined in the central part of the string (that is, the SGs regions are still depleted); during phase III BTBT-generated holes spread also under the SGs, thus ruling their electrostatics.

2.2 *Compact Model*

Figure 4a shows the compact model developed to reproduce the results of the TCAD simulations. Holes distribution in the string is approximated to be uniform (orange region) over an equipotential region that extends from the center of the string to a distance ΔL within the channel of the SG, which is variable during the transient.

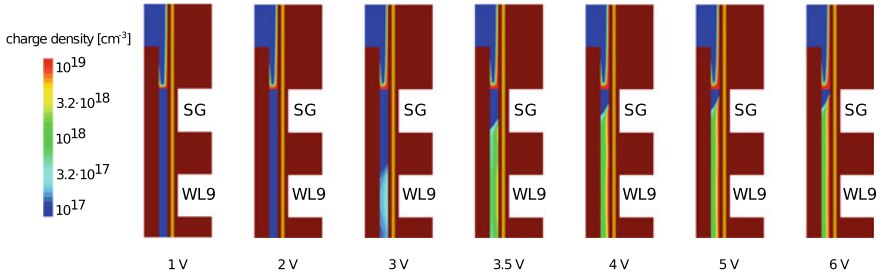

Fig. 3 TCAD results displaying the net charge density (normalized to the elementary charge) at different times during the same transients of Fig. 2a and Fig. 2b. Below each snapshot the V_{BL} at which it is taken is reported. Note that the net charge density in the central region of the string is due to holes, while that at the bottom of the n^+-doped regions is due to ionized donors. © 2018 IEEE [6]

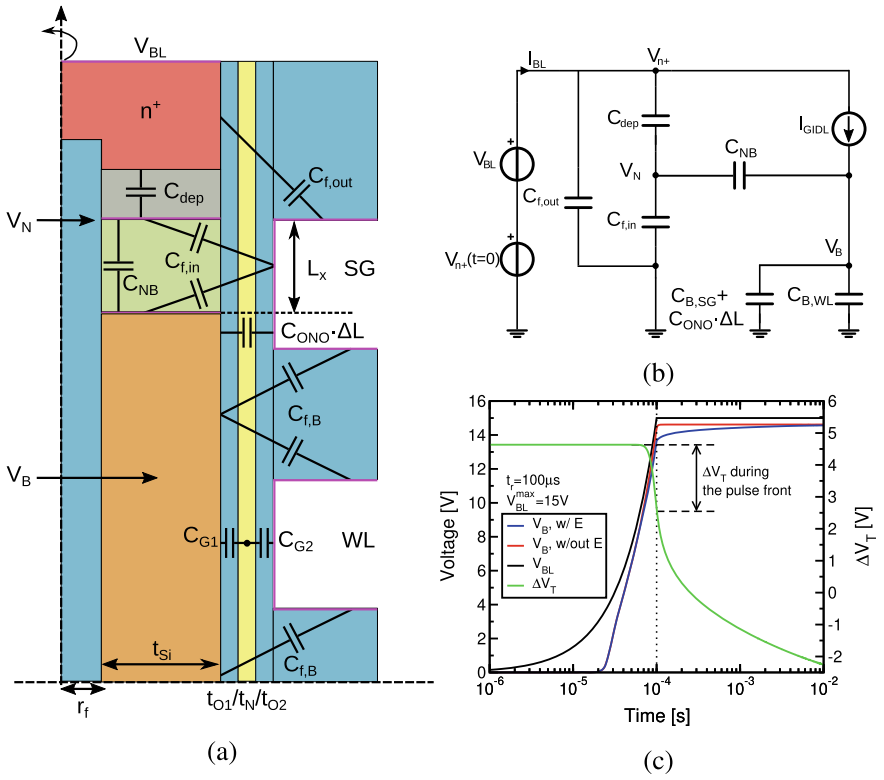

Fig. 4 **a** Schematic of the capacitive couplings considered in the developed compact model (only the upper region of the string close to the BL is showed, © 2019 Springer Nature [7]). **b** Schematic of the developed compact model. **c** Evolution of V_B and ΔV_T (with respect to the neutral V_T) during erase when also the charge exchange between the string channel and the nitride storage layer is included in the compact model of Fig. 4b. (© 2019 Springer Nature [7])

The electrostatics in the region of the SGs that is depleted of holes (L_x) is modeled through $C_{f,in}$ and C_{NB}; $C_{ONO} \cdot \Delta L$ is the remaining capacitive component between the bulk orange region and the longitudinal face of the SG; $C_{f,out}$ accounts for the fringing fields between the transverse face of the SG and the n^+ region while $C_{f,B}$ between the transverse face of the WLs and the central region of the string; C_{dep} simply models the variations of charge in the depleted portion of the n^+ region. Finally, the series between C_{G1} and C_{G2} represents the capacitance of the ONO stack, with the former calculated from the silicon/oxide interface to the middle of the nitride layer, and the latter from the middle of the nitride layer to the WL. The resulting compact model is shown in Fig. 4b, with the addition of the current generator I_{GIDL} that reproduces the GIDL current ($C_{B,SG}$ and $C_{B,WL}$ are the overall capacitances between the orange region and the SGs and WLs, respectively). Please refer to [6] and [7] for the calculation of I_{GIDL} and of all the capacitive contributions mentioned so far. Figure 2a and Fig. 2b show the V_B and I_{BL} transients computed with the compact circuit of Fig. 4b (red line). Model results nicely reproduce those from TCAD simulations confirming the validity of the developed compact model; refer to [6] for a similar analysis also for different string geometries and different electrical waveforms applied to the string contacts.

Finally, in [7] the developed compact model was improved to account also for the variation in the cell V_T due to the emission of electrons from the nitride layer or by injection of holes into it; for compact modeling purposes, the net charge was assumed to be stored in the node between C_{G1} and C_{G2}. Figure 4c shows the evolution of V_T during the GIDL-assisted erase operation and the impact of the charge exchange between the silicon channel and the nitride layer on V_B.

3 NOR Flash–Based Spiking Neural Networks

Hardware neural networks (HNNs) are computing systems in which memory and computing units are not distinct entities exchanging data through a communication bus but rather they are distributed in a way that resembles the organization of synapses and neurons in the human brain [24]. A convenient way to implement HNNs consists in exploiting non-volatile memory arrays as synaptic arrays connecting adjacent layers of artificial neurons: each memory cell acts like a biological synapse, that is, an electrical connection of variable strength [11]. For example, Fig. 5a shows schematically a two layers NOR Flash-based HNN. The voltage signals coming from the presynaptic neurons (PRE) are applied to the WLs of the memory arrays; then, as result of the input signals and the state (V_T) of each memory cell, a current flows through each BL, corresponding to the output signals that are sent to the postsynaptic neurons (POST). In particular, each memory cell is operated in subthreshold regime [25, 26], in which the drain-to-source current I_{DS} can be expressed as a function of the WL voltage V_{WL} as $I_{DS} = I_0 \cdot \exp\left[\alpha_G \left(V_{WL} - V_T^{ref}\right)/(mkT)\right] \cdot \exp\left[\alpha_G \Delta V_T/(mkT)\right]$, where ΔV_T is the cell V_T shift from a reference condition V_T^{ref} (see [23] for the remaining

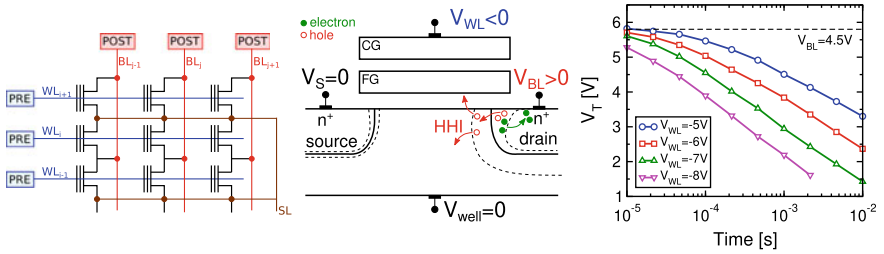

Fig. 5 **a** Schematic of a NOR Flash memory array used as a synaptic arrays connecting two layers of neurons. **b** Schematic a NOR Flash cell subjected to the bias scheme used to trigger HHI and **c** the resulting experimental V_T erase transients (© 2018 IEEE [23])

parameters). In the previous equation the scaling factor $w = \exp[\alpha_G \Delta V_T/(mkT)]$, which is a function of V_T but not of V_{WL}, plays the role of the synaptic weight and the remaining one corresponds to the input signal. A HNN specializes its behavior to perform a well defined task after a *learning* phase, during which the weights of all the memory cells are tuned according to specific *learning algorithms* or *learning rules*. Spiking Neural Networks (SNNs) are particular HNNs for which learning is carried out according to biologically inspired learning rules without external supervision, such as STDP; they take their name from the integrate-and-fire behavior of the artificial neurons, delivering asynchronous spikes during network operation [11, 27].

3.1 Implementing STDP and Unsupervised Learning

Figure 5b shows a schematic of the erase scheme devised to overcome to achieve single-selectivity during erase to enable the adoption of mainstream NOR Flash memory arrays in SNNs. By applying simultaneously a positive V_{BL} and a negative V_{WL}, holes, generated by BTBT and accelerated by the horizontal electric field, become energetic enough to overcome the Si/SiO_2 energy barrier and to be injected into the cell floating-gate FG, leading to $\Delta V_T < 0$. Figure 5c displays the resulting V_T transients, measured for V_{BL} =4.5 V and different values of V_{WL}, confirming the feasibility of the suggested erase scheme.

Once single-cell selectivity during erase is achieved, it is possible to implement STDP in the NOR Flash array, according to which w variations of each memory cell must depend only on the timing between the presynaptic spike and the postsynaptic one (Δt). To that purpose, the voltage scheme displayed in Fig. 6a was devised, exploiting HHI for erase and the classically adopted channel hot-electron injection (CHEI) [28] for program. A presynaptic spike triggers a double-triangular WL pulse of duration t_{WL}; the postsynaptic spike, instead, results in the application of a rectangular pulse to the BL of amplitude equal to 4.5 V, duration much shorter than t_{WL} and delayed with respect to the fire event of $t_{WL}/2$. According to such scheme, if $\Delta t > 0$,

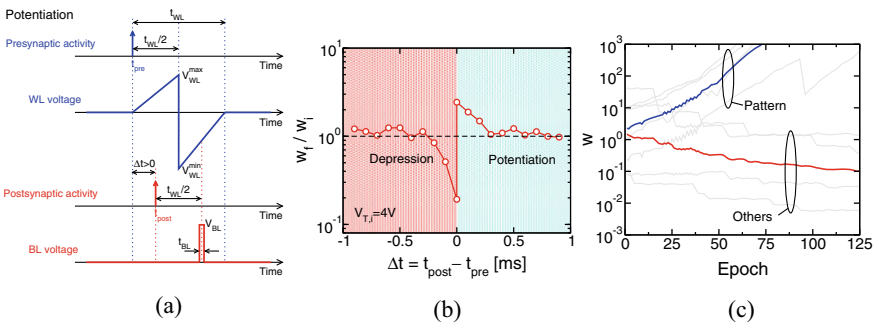

Fig. 6 **a** Voltage scheme suggested to implement STDP in a NOR Flash cell exploiting CHEI and HHI (the $\Delta t > 0$ case is shown) and **b** the resulting experimental STDP waveform. **c** Evolution during the learning phase of the weights of the memory cells in the prototype SNN; the blue curve is the average of the weights belonging to the input pattern, while the red one is the average of the remaining ones. © 2019 IEEE [21]

the BL pulse is applied in correspondence of a negative V_{WL}, thus triggering HHI that results in $\Delta V_T < 0$ ($\Delta w > 0$). In the opposite case, CHEI results in $\Delta V_T > 0$ ($\Delta w < 0$). The experimental STDP waveform resulting from the implementation of the scheme of Fig. 6a is reported in Fig. 6b, displaying that the final weight w_f after a fire event depends on Δt similarly to what observed on biological synapses [11, 17].

Finally, starting from the STDP scheme of Fig. 6a, a prototype SNN with 8 input signals e 1 output was implemented and tested in a pattern learning problem. The input pattern was encoded in the activity of the input neurons, meaning that neurons that are part of the pattern continuously deliver input spikes, otherwise their outputs are kept to ground. The input pattern is correctly *learned* by the SNN if the weights of the cells belonging to it are potentiated and the remaining ones are depressed, as demonstrated in Fig. 6c for the implemented SNN. Please refer to [21, 23] for a full discussion.

Besides, it is worth mentioning that when the employment of NOR Flash memory arrays in HNNs is considered, the impact of their non-idealities during V_T-tuning processes and their typical reliability issues must be carefully assessed. As a matter of example, in [22] the impact of program noise [29] and random telegraph noise [30, 31] on the performance of a neuromorphic digit classifier is investigated in detail. From the suggested analysis, also some quantitative criteria to determine how scaled NOR Flash cells can be when targeting neuromorphic applications are provided.

4 Conclusions

In this chapter, the GIDL-assisted erase operation in 3–D NAND Flash memory arrays has been investigated by means of TCAD simulations and a compact model to reproduce the evolution of I_{BL}, V_B and the cell V_T has been presented. Thanks to its

simplicity and accuracy, the model represents a valuable tool for the optimization of the array performance during erase assisted by GIDL. Then, a similar erase scheme has been employed also in NOR Flash memory arrays, exploiting BTBT-generated HHI to enable single-cell selectivity during erase and allowing the adoption of mainstream NOR Flash memory arrays in SNNs without any modification either to the cell or to the array design. The presented results pave the way towards the development of neuromorphic systems based on cost-effective and highly-reliable memory arrays.

References

1. C. Monzio Compagnoni, A. Goda, A. Sottocornola Spinelli, P. Feeley, A.L. Lacaita, A. Visconti, Reviewing the evolution of the NAND flash technology. Proc. IEEE **105**(9), 1609–1633 (2017)
2. C. Monzio Compagnoni, A. Sottocornola Spinelli, Reliability of NAND Flash arrays: a review of what the 2-D-to-3-D transition meant. IEEE Trans. Electron Devices **66**(11), 4504–4516 (2019)
3. A. Sottocornola Spinelli, C. Monzio Compagnoni, A.L. Lacaita, Reliability of NAND Flash memories: planar cells and emerging issues in 3D devices. Computers **6**(2), 16 (2017)
4. T. Chan, J. Chen, P. Ko, C. Hu, The impact of gate-induced drain leakage current on MOSFET scaling, in *1987 International Electron Devices Meeting (IEDM)*. (IEEE, 1987), pp. 718–721
5. J. Fan, M. Li, X. Xu, Y. Yang, H. Xuan, R. Huang, Insight into gate-induced drain leakage in silicon nanowire transistors. IEEE Trans. Electron Devices **62**(1), 213–219 (2014)
6. G. Malavena, A.L. Lacaita, A. Sottocornola Spinelli, C. Monzio Compagnoni, Investigation and compact modeling of the time dynamics of the GIDL-assisted increase of the string potential in 3-D NAND Flash arrays. IEEE Trans. Electron Devices **65**(7), 2804–2811 (2018)
7. G. Malavena, A. Mannara, A.L. Lacaita, A. Sottocornola Spinelli, C. Monzio Compagnoni, Compact modeling of GIDL-assisted erase in 3-D NAND Flash strings. J. Comput. Electron. **18**(2), 561–568 (2019)
8. A. Sottocornola Spinelli, G. Malavena, A.L. Lacaita, C. Monzio Compagnoni, Random telegraph noise in 3-D NAND Flash memories. Micromachines **12**(6), 703–717 (2021)
9. S. Lai, Current status of the phase change memory and its future, in *2003 International Electron Devices Meeting (IEDM)* (IEEE, 2003), pp. 255–258
10. C. Monzio Compagnoni, , Gusmeroli, R., Sottocornola Spinelli, A., Ielmini, D., Lacaita, A.L., Visconti, A.: Present status and scaling challenges for the NOR Flash memory technology (Chap. 3), in *Solid State Electronics Research Advances*, ed. by S. Kobadze (Nova Science Publishers, Inc., 2009), pp. 101–134
11. G.W. Burr, R.M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola et al., Neuromorphic computing using non-volatile memory. Adv. Phys. X **2**(1), 89–124 (2017)
12. F.M. Bayat, X. Guo, H. Om'Mani, N. Do, K.K. Likharev, D.B. Strukov, Redesigning commercial floating-gate memory for analog computing applications, in *2015 IEEE International Symposium on Circuits and Systems*. (IEEE, 2015), pp. 1921–1924
13. C.H. Kim, S. Lee, S.Y. Woo, W.M. Kang, S. Lim, J.H. Bae, J. Kim, J.H. Lee, Demonstration of unsupervised learning with spike-timing-dependent plasticity using a TFT-type NOR Flash memory array. IEEE Trans. Electron Devices **65**(5), 1774–1780 (2018)
14. H. Kim, S. Hwang, J. Park, S. Yun, J.H. Lee, B.G. Park, Spiking neural network using synaptic transistors and neuron circuits for pattern recognition with noisy images. IEEE Electron Device Lett. **39**(4), 630–633 (2018)

15. F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, K.K. Likharev, D.B. Strukov, High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays. IEEE Trans. Neural Netw. Learn. Syst. **29**(10), 4782–4790 (2017)
16. M. Lenzlinger, E. Snow, Fowler-Nordheim tunneling into thermally grown SiO_2. J. Appl. Phys. **40**(1), 278–283 (1969)
17. G. Bi, M. Poo, Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. **18**(24), 10464–10472 (1998)
18. D.O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, Chapman & Hall, 1949)
19. S. Lowel, W. Singer, Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. Science **255**(5041), 209–212 (1992)
20. G. Malavena, M. Filippi, A. Sottocornola Spinelli, C. Monzio Compagnoni, Unsupervised learning by spike-timing-dependent plasticity in a mainstream NOR Flash memory array-part I: cell operation. IEEE Trans. Electron Devices **66**(11), 4727–4732 (2019)
21. G. Malavena, M. Filippi, A. Sottocornola Spinelli, C. Monzio Compagnoni, Unsupervised learning by spike-timing-dependent plasticity in a mainstream NOR Flash memory array-Part II: array learning. IEEE Trans. Electron Devices **66**(11), 4733–4738 (2019)
22. G. Malavena, S. Petrò, A. Sottocornola Spinelli, C. Monzio Compagnoni, Impact of program accuracy and random telegraph noise on the performance of a NOR Flash-based neuromorphic classifier, in *ESSDERC 2019-49th European Solid-State Device Research Conference (ESSDERC)* (IEEE, 2019), pp. 122–125
23. G. Malavena, A. Sottocornola Spinelli, C. Monzio Compagnoni, Implementing spike-timing-dependent plasticity and unsupervised learning in a mainstream NOR Flash memory array, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 2.3.1–2.3.4
24. P. Sterling, S. Laughlin, *Principles of Neural Design* (MIT Press, s2015)
25. C. Diorio, P. Hasler, B.A. Minch, C.A. Mead, A floating-gate MOS learning array with locally computed weight updates. IEEE Trans. Electron Devices **44**(12), 2281–2289 (1997)
26. P.E. Hasler, C. Diorio, B.A. Minch, C. Mead, Single transistor learning synapses, in *Advances in Neural Information Processing Systems* (1995), pp. 817–824
27. W. Gerstner, W.M. Kistler, R. Naud, L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge University Press, 2014)
28. B. Eitan, D. Frohman-Bentchkowsky, Hot-electron injection into the oxide in *n*-channel MOS devices. IEEE Trans. Electron Devices **28**(3), 328–340 (1981)
29. C. Monzio Compagnoni, A. Sottocornola Spinelli, R. Gusmeroli, S. Beltrami, A. Ghetti, A. Visconti, Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics. IEEE Trans. Electron Devices **55**(10), 2695–2702 (2008)
30. A. Ghetti, C. Monzio Compagnoni, A. Sottocornola Spinelli, A. Visconti, Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer Flash memories. IEEE Trans. Electron Devices **56**(8), 1746–1752 (2009)
31. C. Monzio Compagnoni, R. Gusmeroli, A. Spinelli Sottocornola, A.L. Lacaita, M. Bonanomi, A. Visconti, Statistical model for random telegraph noise in Flash memories. IEEE Trans. Electron Devices **55**(1), 388–395 (2007)