Overview of modeling approaches for scaled non volatile memories

Daniele Ielmini
Dipartimento di Elettronica e Informazione
Politecnico di Milano and IU.NET
p.zza Leonardo da Vinci 32, 20133 Milano – Italy
ielmini@elet.polimi.it

Abstract—**The success of non-volatile memory (NVM) in the last two decades have relied on the relentless downscaling of MOS-based Flash devices. Entering the 2X nm technology, the physical scaling of Flash memories will have to face potential reliability and feasibility showstoppers. To understand the ultimate scaling limits of the Flash technology and to evaluate alternative memory concepts, physically-based models able to describe small-size effects and non-silicon-based materials are required. This work reviews recent developments and open issues of Flash and post-Flash modeling, discussing physical mechanisms in reliability and operation and highlighting the challenges in the understanding of non-silicon active materials for phase change memory (PCM) and resistive-switching memory (RRAM) devices.**

Keywords-non-volatile memory; Flash memory; charge-trap memory; phase-change memory (PCM); resistive-switching memory (RRAM); physical modeling; reliability modeling; electro-thermal simulation

I. INTRODUCTION

Non volatile memory has been a strong market and technology driver for the micro-nanoelectronic industry in the last two decades. This is mainly due to the broad diffusion of portable electronic devices, such as cell-phones, laptops, digital cameras and multimedia players. The huge success of Flash memories has been driven and enabled by the remarkable scaling of these devices, which allowed a 50,000x cost-per-bit reduction from 1.5 µm (1987) to 65 nm (2007) technology nodes [1] along a scaling roadmap which has been surpassing Moore's law in terms of nodes per year [2]. Tomorrow, the strong push to further size and cost downscaling will follow mainly two alternative paths, one aimed at overcoming the inevitable scaling limits of planar Flash by 3D structures, such as FinFET and stacked/vertical Flash approaches exploiting charge trapping. The second path embraces alternative concepts, such as phase-change memory (PCM) and resistive-switching memory (RRAM), which may allow better scaling behavior beyond 20 nm, although still presenting some open issues in the physical understanding and modeling.

The purpose of this work is to review the recent developments and challenges in the modeling of advanced non volatile memories, considering both Flash devices and the so-called post-Flash technologies (charge-trap, PCM and RRAM). The physical mechanisms for programming and reliability will be described, and the scaling perspectives of each technology will be discussed.

This work was supported in part by EU (project Emma FP6 – 033751), MIUR (project FIRB-RBIP06YSJJ) and Intel (Grant 34523).

II. FLASH MEMORY MODELING

The Flash memory relies on charge storage in a floating gate (FG), thus a comprehensive modeling of the memory operation and reliability requires that the tunneling/hot-carrier injection mechanisms and the low-field oxide-leakage effects are fully understood. Program/erase processes are efficiently described by quantum-mechanical tunneling models and full-band Monte Carlo and analytical models for hot-carrier injection [3]. Reliability aspects, however, pose the most intriguing and challenging questions for modeling, at both the single-cell and array levels. Fig. 1 highlights the main Flash reliability issues below the 40 nm technology node. These include (i) oxide traps, contributing threshold-voltage (V_T) instability issues such as stress-induced leakage current (SILC) [4, 5], trapping/detrapping [6] and random telegraph-signal noise (RTN) [7, 8], (ii) edge field-enhancement and discrete-dopant effects, which collaborate with oxide-trap effects to enhance the V_T spread in the array, (iii) few-electron effects [9] and (iv) cell-cell electrostatic coupling in the array [10]. Physical understanding of all these issues is essential to develop models for designing error correction code (ECC) and wear-leveling methods [6]. The modeling approaches for each

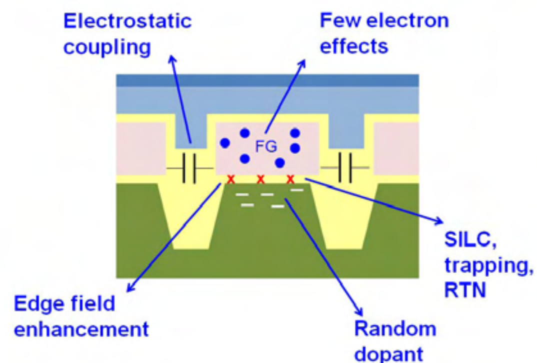
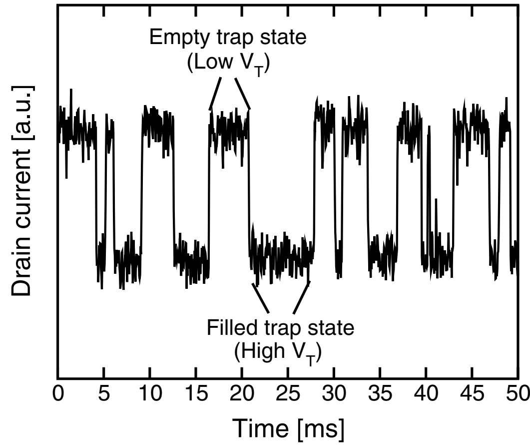


Fig. 1. Sketch for modeling issues in scaled Flash memory cells (FG = floating gate, STI = shallow trench isolation, W = channel width).



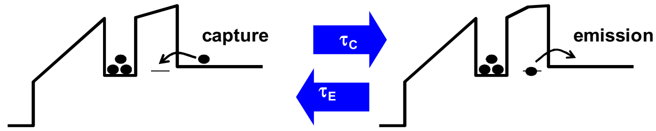


Fig. 2. Measured drain current I_D as a function of time, showing RTN (top, from [7]) and schematic for the capture/emission processes at the basis of V_T fluctuation (bottom).

of the reliability in Fig. 1 are reviewed in the following.

A. Oxide traps

Trap states in the dielectric layers (bottom dielectric, usually a thermal SiO_2 layer, and top dielectric, an ONO stack or high-k layer) contribute the most severe reliability issues in Flash memories. High-field stress induced by Fowler-Nordheim (FN) tunneling during program and erase (P/E) results in a generation of traps in the dielectric layers. Charge trapping in such states may then cause significant V_T shift after 10^5-10^6 P/E cycles, thus limiting the cycling endurance of the Flash memory [6]. Generated 'hard' traps in the bottom oxide may also contribute a trap-assisted-tunneling (TAT) leakage path, thus causing a partial charge-loss from the FG by SILC [6]. Fortunately, SILC affects a small minority of cells in the array, thanks to a careful control of P/E algorithms and of bottom-oxide quality and thickness. Also, it has been speculated that SILC may require the cooperation of at least two oxide traps, serving as an efficient percolation path for charge loss [4, 5]. The low percentage of leaky cells may thus be handle by ECC with a minimum amount of extra parity bits [6]. ECC design thus requires feasible SILC models predicting the probability of SILC within the array. Statistical SILC models have been developed, addressing both the physics of TAT leakage transport [4] and the statistical generation of SILC traps in the bottom dielectric [5, 11].

Oxide traps may also contribute to V_T shifts induced by charge detrapping from the dielectric layers: Electrons trapped at oxide states are released after the P/E pulse, resulting in a thermally-activated charge loss [12]. Although the V_T loss contributed by detrapping is lower than the SILC one, it affects most of the array cells, thus it needs to be carefully predicted for V_T window design, especially in multilevel cell (MLC).

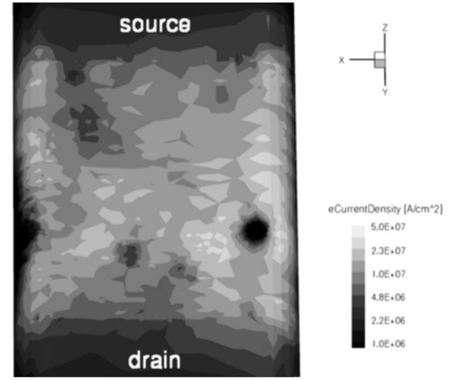


Fig. 3. Calculated current density map in a deca-nm Flash cell in presence of random dopant and edge field enhancement (from [8]).

Capture/emission processes at oxide traps can also lead to random fluctuations of the drain current I_D, as shown in Fig. 2: I_D randomly fluctuates between a high value (low V_T, empty trap state) and a low value (high V_T, filled trap state) as a result of alternated capture/emission of electrons at an oxide trap close to the substrate Fermi level (see bottom part of Fig. 2). RTN is generally stronger in Flash memories, as compared to MOS transistors at equal size, due to the larger oxide stack thickness, hence the smaller gate capacitance and the larger threshold voltage shift ΔV_T given by:

$$\Delta V_T = -\alpha \frac{q}{C_x} \quad (1)$$

where -q is the electron charge, C_x is the gate to trap capacitance (almost equal to the whole dielectric stack capacitance) and α is a parameter describing the impact of the trapped charge on channel conduction.

B. Random dopant and edge field enhancement

The amplitude of RTN V_T fluctuations has been shown to increase upon channel size downscaling: this can be explained by percolation effects through non-uniform V_T landscape in the randomly and discretely doped channel and by confinement effects of the current at the channel edges due to local field enhancements. Fig. 3 shows the calculated map of current density in a scaled-Flash channel [8]. The current density increases at the left/right channel edges and decreases at discrete dopant ions which are close enough to the channel, thus resulting in a non-uniform channel conduction. A fluctuating trap which is placed very close to the channel conduction bottleneck can dramatically affect I_D, resulting in giant RTN. TCAD physics-based simulations with atomistic doping and Monte Carlo placement of traps in the bottom oxide indicate that the exponential slope λ of the V_T distribution in the array obeys to $\lambda \propto W^{-1}L^{-0.5}$, where W and L are the channel width and length [8]. The larger dependence on W scaling is due to the increasing probability that the trap overlaps with the channel percolation path, thus quenching I_D [8, 13]. In addition to the strong size dependence, an increase of channel doping enhances RTN due to the increased concentration of dopants affecting channel conduction. Simulation results in Fig. 4 indicate a large enhancement of RTN effects beyond the 40 nm node, due to the strong geometry and doping dependence [8].

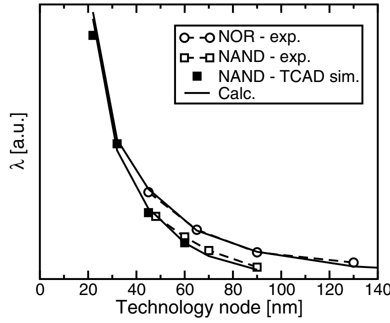


Fig. 4. Calculated exponential slope of V_T distribution λ, as a function of the technology node. The slope largely increases below 40 nm (from [8]).

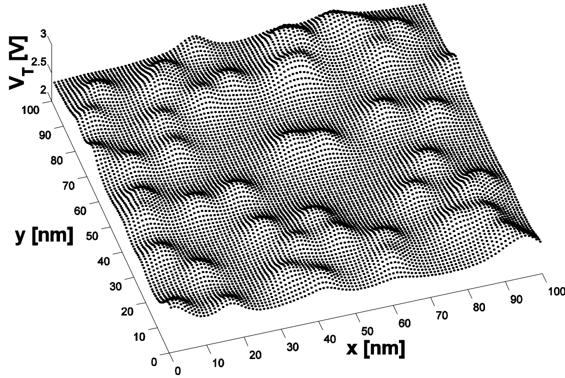


Fig. 5. Calculated V_T landscape within the channel of a charge trap memory with nanocrystal trapping layer (from [13]).

C. Few-electron and electrostatic coupling effects

In addition to V_T shift contributed by RTN, other sources of V_T spread should be considered. As the cell scales down, the number of electrons stored in the FG decreases at equal ΔV_T [14]. The stochastic nature of quantum-mechanical tunneling used for injecting charge into the FG degrades the control in the amount of injected charge, thus affecting the precision of MLC P/E operations [9]. Accurate modeling of V_T spread due to few-electron effects requires Monte Carlo and analytical techniques. A further source of ΔV_T instability comes from electrostatic coupling between FGs, which increases for decreasing pitch between cells in the array, thus contributing to V_T distribution width [10]. These effects require careful array-level electrostatic 3D modeling, which can provide valuable tools for developing P/E algorithms to minimize cell-cell interferences and improve V_T distributions.

III. Evolutionary Flash

A potential solution to SILC and electrostatic interference comes from the charge-trap memory [1]. This consists of a Flash memory where the polysilicon FG is replaced by a charge-trapping layer, usually SiN. The discrete nature of storage nodes prevents complete leakage of charge through SILC spots, while electrostatic coupling among adjacent cells is strongly inhibited due to the lack of significant capacitive coupling between thin trapping layers. The most popular charge trap memory features a high-k top dielectric and a metal gate to prevent erase saturation, see e.g. the TANOS concept with TaN gate, Al_2O_3 top dielectric, SiN trapping layer and thin (about 4 nm) bottom SiO_2 [15]. Charge injection, trapping, emission and vertical/lateral transport in the trapping/high-k layers are fundamental issues for predicting P/E and reliability characteristics. Many physical details (trap energy distribution, nitride transport and emission mechanisms and their field and temperature dependences) are still widely debated. The discrete nature of storage nodes makes the cell electrostatics more complicated than in Flash, due to the non-uniform distribution of channel charge. This is similar to the discrete-dopant effect in Fig. 3, and results in a percolative conduction in the cell channel [13]. Fig. 5 shows the calculated V_T landscape for a charge-trap memory with Si nanocrystals dispersed in a SiO_2 film. The resulting distribution of channel inverted charge results in a localized current density which must be taken into account in predicting V_T characteristics during P/E and data loss depending on cell size [13].

To compensate for short-channel degradation of V_T and to exploit 3D stacking allowing virtually unlimited cost/density scaling, vertical charge-trap architectures have been proposed. The most interesting concept is the bit-cost scalable (BiCS) Flash architecture, featuring vertical pillars as series-connected channels and nitride trapping layers [16]. 3D cell structures with better short-channel performance than planar Flash have also been demonstrated [17]. The development of these novel concepts requires detailed 3D simulation tools with sound physical models for electrostatics, tunneling and trapping/detrapping processes in high-k and trapping layers.

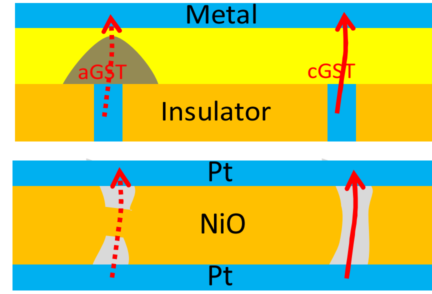


Fig. 6. Sketch for the working principle of resistive memory devices, namely PCM (top) and RRAM (bottom). The high and low resistance states are depicted on the left and right, respectively.

IV. Resistive Memories

Alternative memory concepts that may allow a better scalability rely on the change of resistance due to phase transition or filament formation/rupture in specific active materials. The working principle of these resistive memory devices is schematically shown in Fig. 6: The PCM (top schematic) consists of an active thin film of chalcogenide material, typically $Ge_2Sb_2Te_5$ (or, briefly, GST), with top and bottom electrodes. The chalcogenide material can be switched from a low-resistance crystalline phase to a high-resistance

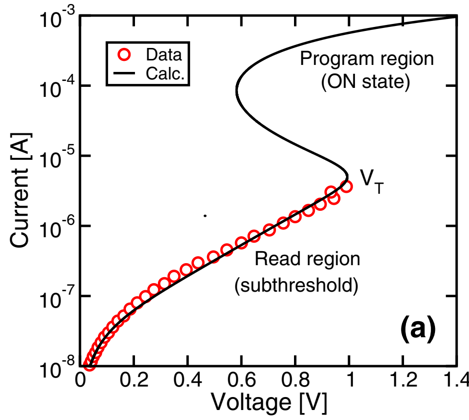


Fig. 7. Measured and calculated I-V characteristics for a PCM cell in the reset state. Calculations by the thermally-assisted hopping model [33] match the subthreshold data and reproduces NDR at the basis of threshold switching (from [31]).

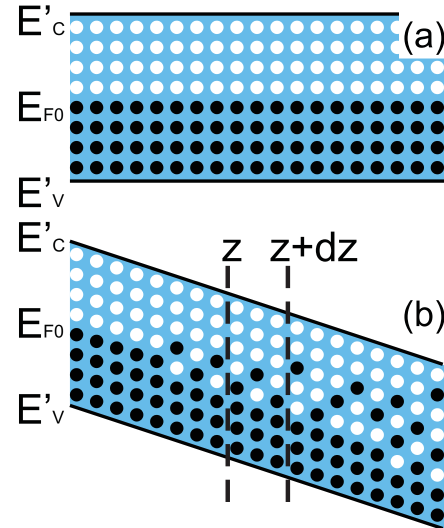


Fig. 8. Sketch for the energy-gain threshold switching model, showing equilibrium (a) and high-field conditions (b) (from [35]).

amorphous phase. The latter is formed on top of the bottom electrode (the 'heater'), which is usually very narrow for better confinement of the current density and temperature during the programming operation. The high-resistivity 'plug' on top of the heater results in a change of resistance by two-three orders of magnitude as compared to the crystalline state [18].

RRAM relies on resistive-switching insulator materials. An electrically-operated dielectric breakdown results in the formation of a conductive filament (CF), which can then be erased and reformed by subsequent electrical operations. The resistance-switching material is typically a binary metal oxide, such as NiO [19], Cu_2O [20], TiO_2 [21, 22], HfO_2 [23] and several others. The RRAM-like operation of Ag-doped chalcogenide materials [24, 25] and organic materials [26] has been demonstrated as well. The scaling advantages of these resistive concepts are due to (i) the extremely compact cell layout, requiring only two terminals instead of the usual three for the MOS transistor, (ii) the inherent scalability of the storage concept, based on nm-sized filaments or amorphous regions, free of any electrostatic interference or few-electron phenomena, like those affecting Flash, and (iii) the capability for on-chip 3D stacking of several active layers [27]. These properties allow for high-density cross-bar structures with $4F^2$ size of the physical cell, for each memory layer. In addition, programming resistive memories is extremely fast, due to the thermally-activated, localized character of the CF switching or phase change.

V. PCM MODELING

PCM cells are programmed applying electrical pulses resulting in a high local temperature by Joule heating in the active chalcogenide region. To form the high-resistance amorphous phase, the pulse voltage is high enough to melt a portion of the active material, and the pulse is switched off fast enough to quench the molten region in a solid amorphous state. For re-crystallization, a lower voltage is applied so that Joule heating accelerates nucleation and growth restoring the low-resistance crystalline phase. The two states are called reset (amorphous phase) and set (crystalline state) respectively.

The Joule-heating-based programming of the phase-change can be modeled by self-consistent drift-diffusion coupled with heat transport equations. After calibration against I-V and set/reset characteristics for several cell geometries [28], the model can be applied to explore the optimization/scaling strategies [29] and for an investigation of the thermal interference in the PCM array [30]. Although plausible for the crystalline set state, the drift-diffusion approach is not adequate for describing the amorphous chalcogenide conduction. The I-V characteristic for the amorphous reset state is shown in Fig. 7: at sufficiently low values, the current increases exponentially up to a threshold voltage V_T where a negative differential resistance (NDR) region takes place. In the high-current ON-state, the I-V curve displays a positive slope. Subthreshold conduction and the switching threshold / dynamics are essential for an accurate simulation of PCM read and program, particularly for the reset transition [31].

For subthreshold conditions, carrier transport occurs at localized states in the mobility gap [32]. Thermally-assisted hopping can account for the temperature dependence and the exponential voltage dependence of the subthreshold current and for the presence of the characteristic threshold switching effect [33, 34]. The latter can be explained as a hopping-conduction instability taking place at sufficiently large voltage, as a result of the energy gain of trapped carriers [35].

The threshold switching mechanisms is schematically shown in Fig. 8: at low field (a), electrons contribute to the trap-limited current occupying localized states below the equilibrium Fermi level E_F. The energy gain due to the electric field is effectively balanced by energy relaxation processes in the trapped state. At sufficiently large fields (b), energy relaxation cannot compensate energy gain, as electrons do not spend enough time in the same localized state. This results in a 'heating' mechanism enhancing conductivity, since the trap emission rate is exponentially increasing with the carrier energy [35]. The mobility enhancement allows for a decrease of the electric field in most of the amorphous chalcogenide

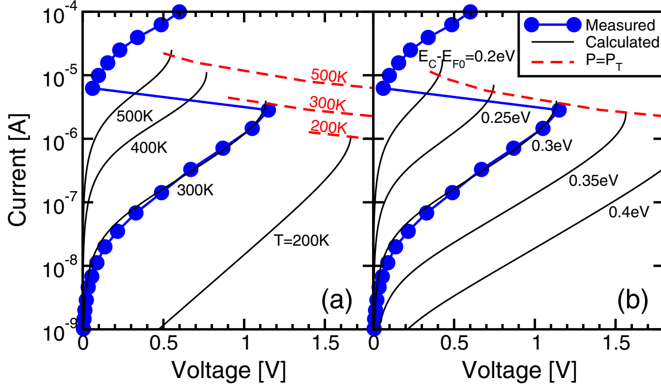


Fig. 9. Calculated I-V curves of a PCM cell for increasing temperature (left) and for increasing mobility gap of the active material. Data from a for GST at 25°C are shown for reference (from [35]).

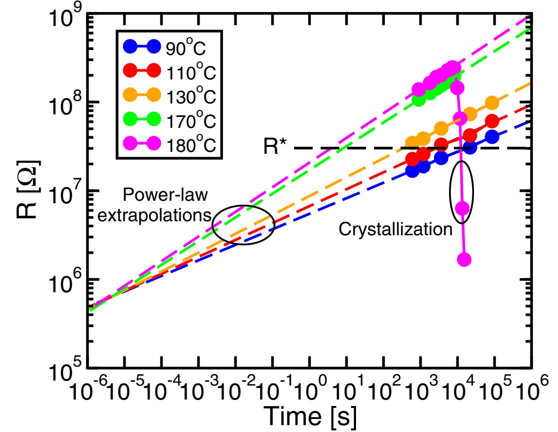


Fig. 10. Measured resistance as a function of time for increasing annealing temperature. R was measured at 25°C (from [42]).

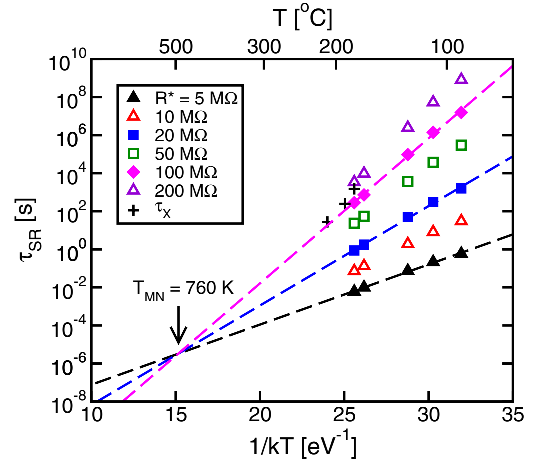


Fig. 11. Arrhenius plot of relaxation times τ_{SR} extracted from Fig. 10 and crystallization time τ_x (from [42]).

layer, for increasing current, which is at the basis of the NDR above the switching point in Fig. 7.

A quasi-analytical modeling of threshold switching was shown to account for the thickness dependence of the threshold voltage V_T and of the threshold current I_T [35]. The carrier heating model for threshold switching also allows for a prediction of the subthreshold current and of switching points as a function of experimental and material parameters. Fig. 9 shows the calculated I-V curves for variable temperature T (left) and for variable mobility gap of the chalcogenide material (right). Data for GST at 25°C are also shown for reference. V_T decreases and I_T increases for increasing T, in qualitative agreement with literature data [36]. This is because switching is triggered by a critical carrier energy which corresponds to a power density P_T''' given by:

$$P_T'''=\frac{\gamma_T N_T (kT)^2}{\tau_{rel}(E_C^{'}-E_{F0})} \quad (2)$$

where $\gamma_T \approx 1$ is a constant, N_T is the density of localized states, k is the Boltzmann constant, τ_{rel} is a time constant for energy relaxation and E_C' is the conduction-band mobility edge [35]. The crossing between I-V curves and Eq. (2) yields the switching points in the figure. For increasing mobility gap, I_T decreases and V_T increases, due to the strong suppression of current. These results are consistent with reported V_T as a function of the mobility gap for Ge-Sb-Te compounds [37].

A. PCM reliability modeling

The inherent metastability of the amorphous phase makes the PCM cell prone to structural relaxation (SR) and phase crystallization. During SR, the disordered structure rearranges toward more stable configurations [38–40]. Structural defects annihilates and the mobility gap increases, resulting in a regular increase of resistance R [41]. Fig. 10 shows measured R as a function of time for variable annealing temperature: R drifts faster for increasing T, revealing the thermal activation of the SR process [42]. Crystallization (see data at 180°C in the figure) is instead a proper phase transition from the high-R amorphous phase to the low-band-gap, doped-semiconductor crystalline phase, resulting in a dramatic R drop [43, 44].

Modeling PCM reliability requires adequate laws to predict the temperature acceleration of SR and crystallization. SR can be described by the standard Arrhenius law:

$$\tau_{SR}=\tau_0\, e^{\frac{E_A}{kT}} \quad (3)$$

where τ_{SR} is the time to reach a resistance R* (see Fig. 10) and E_A is the activation energy. Eq. (3) can similarly account for the crystallization time τ_x, with an activation energy E_x of about 2.6 eV and a pre-exponential time $\tau_{x0} \approx 10^{-23}$ s [42].

Fig. 11 shows the Arrhenius plot of τ_{SR} and τ_x: All data obey the Arrhenius law, crossing at T_{MN} = 760 K and $\tau_0 \approx 4\times10^{-6}$ s. This is evidence for the Meyer-Neldel rule [42]:

$$\tau_0=\tau_{00}\, e^{-\frac{E_A}{kT_{MN}}} \quad (4)$$

which relates the pre-exponential time to E_A by a compensation law, where higher activation energies are favored by a decreased τ_0. The presence of a common relationship satisfying

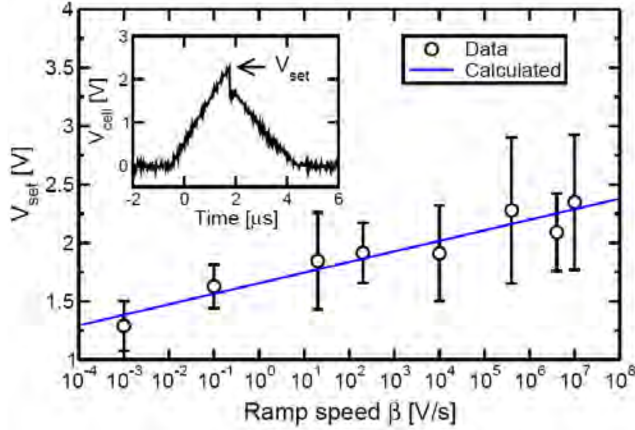


Fig. 12. Measured and calculated V_{set} as a function of sweep rate $\beta = dV/ds$. The inset shows an oscilloscope trace of cell voltage during the set triangular pulse (from [50]).

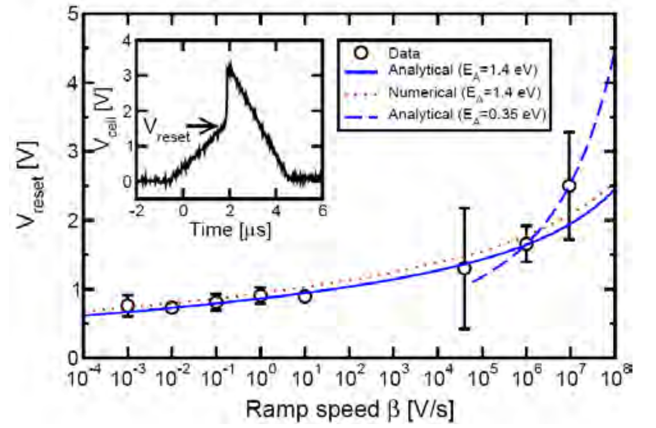


Fig. 13. Measured and calculated V_{reset} as a function of sweep rate $\beta = dV/ds$. The inset shows an oscilloscope trace of cell voltage during the reset triangular pulse (from [50]).

Eq. (4) demonstrates that SR and crystallization, although being different processes, share the same multi-phonon excitation mechanism for overcoming their respective energy barriers. This allows for accurate kinetic models for reliability prediction and temperature-accelerated data analysis [45].

VI. RRAM MODELING

Resistance switching effects in metal oxides are known since the mid 1960s [46], however this concept has been recently revived as a potential NVM approach. Nickel oxide (NiO) has shown the most promising properties including stable and repeatable switching voltages, sufficient endurance and unipolar switching [27, 47]. Bipolar switching displayed by other materials may instead prevent a straightforward implementation within a diode-rectified cross-bar architecture.

Contrary to phase-change mechanisms in PCMs, modeling of RRAM operation is still in its infant stage. This is because of the localized nature of resistance switching in RRAM devices, which makes it difficult to investigate the switching phenomena on a larger scale by thermal or optical methods. Switching in RRAM may be compared to the time-dependent dielectric breakdown in SiO_2 and high-k dielectrics. Similar to the breakdown effect, the low-to-high resistance (set) transition consists in the formation of a microscopic CF through the high-resistance dielectric layer. CF formation in NiO has been shown to proceed via a preliminary threshold switching process [48]. The large current density, Joule power dissipation and temperature increase in the threshold switching spot is responsible for the structural/chemical transformation at the basis of the CF formation, through electromigration effects and non-equilibrium oxygen diffusion. In fact, the metallic nature of set resistance strongly suggests that the CF consists of a Ni-rich region, similar to the reported over-stoichiometric Si in the damaged spot in post-breakdown MOS transistors [49]. The intimate link between set and threshold switching allows to estimate the set voltage V_{set} based on analytical threshold switching models [50]. Fig. 12 shows the measured V_{set} as a function of the voltage sweep rate $\beta = dV/ds$ in triangular pulses. The inset shows a typical oscilloscope trace for the cell voltage during the pulse, marking the voltage drop at the set transition. Assuming an exponential increase of the switching probability with voltage, the V_{set} can be obtained as:

$$V_{set} = V_0 \log \frac{\beta \tau_0}{V_0}, \tag{5}$$

where V_0 and τ_0 are constants. Calculations by Eq. (5) are also shown in the figure. The increase of V_{set} with sweep time is due to the 1/f-noise-driven switching mechanism [31]: for increasing β, a smaller region of the noise power spectrum contributes to fluctuations, thus a larger voltage must be reached for threshold switching [31].

The low-to-high resistance (reset) transition in NiO has been instead explained by a thermally-activated reaction / diffusion process resulting in a CF disruption, restoring the reset state [51, 52]. Thermal activation results in a remarkable voltage acceleration. Fig. 13 shows measured and calculated reset voltage V_{reset} as a function of β during triangular reset pulses. The inset shows measured cell voltage with a sharp increase at reset. Calculated results were obtained assuming an Arrhenius-accelerated reset process where the CF diameter ϕ shrinks according to:

$$v_G = -\frac{d\phi}{2dt} = v_{G0} e^{-\frac{E_{reset}}{kT}}, \tag{6}$$

where v_G is the velocity of the CF edge, v_{G0} is a pre-exponential constant and E_{reset} is the activation energy for the reset process. The edge drift in Eq. (6) may relate to atomic diffusion (e.g. O vacancies or Ni from the CF, or O back to the CF) and/or chemical reaction, e.g. oxidation of metallic Ni. NiO is in fact stable below about 2270°C, according to the Ellingham diagram for oxidation free enthalpy [53].

Data in the figure below $\beta = 10^6$ Vs^{-1}, analyzed with a thermal model for the voltage-dependent T profile in the CF, yield $E_{reset} = 1.4$ eV. This parameter has a key importance for NiO-based RRAM modeling, since it controls the voltage acceleration of reset in Fig. 13, which allows to extrapolate the fast-pulse reset voltage from DC data. In this regard, note that, for efficient

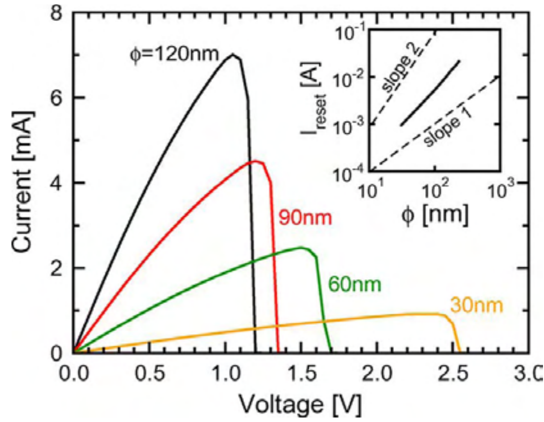


Fig. 14. Calculated I-V curves for set-state RRAM cells for variable CF size. I_{reset} scales with CF size, as reported in the inset (from [52]).

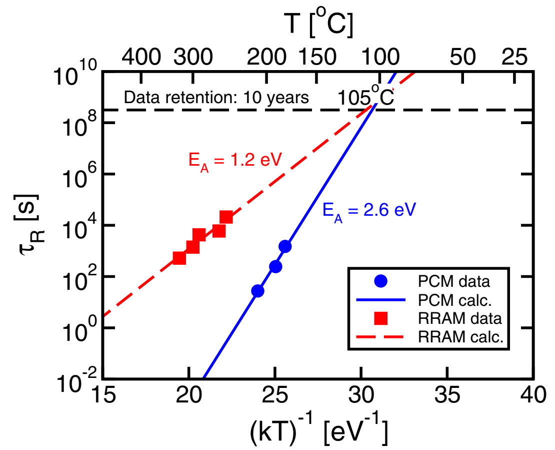


Fig. 15. Arrhenius plot of retention times for intrinsic RRAM and PCM cells. Despite the lower activation energy, RRAM appears compliant with NVM applications.

pulsed reset in the 10-100 ns time range, a V_{reset} of about 3 V should be expected from the figure, as compared to the sub-1 V reset typically reported in the low-frequency range [50].

From this analysis, one main drawback of RRAM appears to be the large reset current I_{reset}. In fact, for $V_{reset} \approx 3$ V and a set resistance in the 0.3 kΩ range, one gets a rough estimate of 10 mA for I_{reset}. This large I_{reset} is not compatible with the small chip size and the large throughput required in high-density NVM devices. Two paths are available for I_{reset} scaling, namely (i) geometry scaling by careful control of the CF size during set [54], and (ii) materials engineering. In the first approach, CF downscaling should result in an increase of R_{set}, hence a decrease of I_{reset}. $\propto \phi^2$. Fig. 14 shows the calculated I-V curves for variable ϕ, from 30 to 120 nm [52]. Calculations were obtained by a detailed finite-element-method (FEM) tool enabling self-consistent electrical/thermal transport. As expected, I_{reset} decreases for decreasing ϕ (see the inset). Note that V_{reset} increases for decreasing ϕ, resulting in a weaker scaling dependence of $I_{reset} \propto \phi^{1.5}$. This is due to the increasing additional contribution of heat loss through the surrounding NiO for small CFs, degrading thermal confinement [52].

A. RRAM reliability modeling

The activation energy E_{reset} is also of extreme value for predicting and extrapolating the thermal stability of the CF, which dictates the intrinsic data retention of the RRAM device. Fig. 15 shows the Arrhenius plot of measured retention time τ_{reset}, obtained from annealing experiments of the thermally-driven reset. The activation energy is 1.2 eV, in agreement with the 1.4 eV extracted from voltage-driven reset. A temperature of roughly 105°C is found at 10 years, which complies with the NVM reliability requirements. Data and extrapolations for GST-based PCM devices are shown for comparison: the larger activation energy of 2.6 eV results in a better tradeoff between fast crystallization during programming and long retention time (10 years at about 105°C). The larger energy barrier for data loss in GST is due to the large number of atomic transitions required for the first crystalline nucleation in the amorphous region [45], while CF rupture may proceed by single O vacancies independently diffusing out of the CF.

VII. Conclusions

Flash downscaling beyond 20 nm will pose severe challenges, while other memory concepts (e.g. TANOS, BiCS Flash, PCM and RRAM) may offer better scaling opportunities. In such a scenario, physics-based modeling is essential for providing comparative projections of NVM scaling. The review discusses the main modeling approaches, focusing on the challenges related to the broad spectrum of new materials and the complex blend of thermal, chemical and electrical processes involved in the memory programming and reliability.

Acknowledgment

The author acknowledges useful discussions with C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita, A. Pirovano, R. Bez, D. Kau and G. Spadini.

References

[1] S. Lai, “Non-volatile memory technologies: The quest for ever lower cost,” IEDM Tech. Dig. 1 (2008) 11–16

[2] C.-G. Hwang, “New paradigms in the silicon industry,” IEDM Tech. Dig. 19-26 (2006).

[3] A. Gehring and S. Selberherr, "Modeling of tunneling current and gate dielectric reliability for nonvolatile memory devices," IEEE Trans. Device and Materials Reliability 4, 306-319 (2004).

[4] D. Ielmini, A. S. Spinelli, A. L. Lacaita and A. Modelli, “A new two-trap tunneling model for the anomalous SILC in flash memories,” Microelectron. Eng. 59, 189-195 (2001).

[5] D. Ielmini, A. S. Spinelli, A. L. Lacaita and A. Modelli, “A statistical model for SILC in Flash memories,” IEEE Trans. Electron Devices 49, 1955-1961 (2002).

[6] N. Mielke, “NVM Reliability,” IRPS Tutorial (2008).

[7] C. Monzio Compagnoni, et al., "Statistical Model for Random Telegraph Noise in Flash Memories," IEEE Trans. Electron Devices 55, 388-395 (2008)

[8] A. Ghetti, et al., "Scaling trends for random telegraph noise in deca-nanometer Flash memories," IEDM Tech. Dig. 835-838 (2008).

[9] C. Monzio Compagnoni, et al., “Ultimate Accuracy for the NAND Flash Program Algorithm Due to the Electron Injection Statistics," IEEE Trans. Electron Devices 55, 2695-2702 (2008).

[10] J.-D. Lee, S.-H. Hur and J.-D. Choi, “Effects of floating-gate interference on NAND flash memory cell operation,” IEEE Eletron Device Lett. 23, 264-266 (2002).

[11] D. Ielmini, A. S. Spinelli, A. L. Lacaita and M. J. van Duuren, “Defect generation statistics in thin gate oxides,” IEEE Trans. Electron Devices 51, 1288-1295 (2004).

[12] N. Mielke, et al., “Flash EEPROM threshold instabilities due to charge trapping during program/erase cycling,” IEEE Trans. Device Materials Reliab. 4, 335-344 (2004).

[13] D. Ielmini, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita and C. Gerardi, “A new channel percolation model for Vt shift in discrete-trap memories,” Proc. IRPS, 515-521 (2004).

[14] G. Molas, et al., “Degradation of floating-gate memory reliability by few electron phenomena," IEEE Trans. Electron Devices 53, 2610 (2006)

[15] C. Lee, et al., “A novel SONOS structure of SiO2/SiN/Al2O3 with TaN metal gate for multi-giga bit flash memories,” IEDM Tech. Dig. 613-616 (2003).

[16] H. Tanaka, et al., "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory," IEEE Symp. VLSI Tech. 14-15 (2007).

[17] D. Kwak, et al., "Integration Technology of 30nm Generation Multi-Level NAND Flash for 64Gb NAND Flash Memory," IEEE Symp. VLSI Tech. 12-13 (2007).

[18] D. Ielmini, A. L. Lacaita, A. Pirovano, F. Pellizzer and R. Bez, “Analysis of phase distribution in phase-change non volatile memories,” IEEE Electron Device Lett. 25, 507-509 (2004).

[19] U. Russo, D. Ielmini, C. Cagli and A. L. Lacaita, “Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices,” IEEE Trans. Electron Devices 56, 186-192 (2009).

[20] T.-N. Fang, et al., “Erase mechanism for copper oxide reistive switching memory cells with nickel electrode,” IEDM Tech. Dig. 789-792 (2006).

[21] B. J. Choi, et al., “Resistive switching mechanism of TiO_2 thin films grown by atomic layer deposition,” J. Appl. Phys. 98, 033715 (2005).

[22] D. B. Strukov, G. S. Snider, D. R. Stewart and R. S. Williams, “The missing memristor found,” Nature 453, 80–83 (2008).

[23] Y.-M. Kim and J.-S. Lee, “Reproducible resistance switching characteristics of hafnium oxide-based nonvolatile memory devices,” J. Appl. Phys. 104, 114115 (2008).

[24] M. N. Kozicki, M. Park and M. Mitkova, “Nanoscale memory elements based on solid-state electrolytes,” IEEE Trans. Nanotechnology 4, 331-338 (2006).

[25] U. Russo, D. Kalamanathan, D. Ielmini, A. L. Lacaita and M. Kozicki, “Study of multilevel programming in programmable metallization cell (PMC) memory,” IEEE Trans. Electron Devices 56, 1040-1047 (2009).

[26] R. Müller, et al., "Bipolar Resistive Electrical Switching of CuTCNQ Memories Incorporating a Dedicated Switching Layer", IEEE Electron Device Letters 30 (2009) 620-622

[27] I. G. Baek, et al., “Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND storage application,” IEDM Tech. Dig. 769-772 (2005).

[28] A. Redaelli, D. Ielmini, U. Russo and A. L. Lacaita “Modeling and simulation of conduction characteristics and programming operation in nanoscaled phase-change memory cells,” J. Computational and Theoretical Nanoscience 5, 1183-1191 (2008).

[29] U. Russo, D. Ielmini, A. Redaelli and A. L. Lacaita, “Modeling of programming and read performance in phase-change memories - Part I: Cell optimization and scaling,” IEEE Trans. Electron Devices 55, 506-514 (2008).

[30] U. Russo, D. Ielmini, A. Redaelli and A. L. Lacaita, “Modeling of programming and read performance in phase-change memories – Part II: program disturb and mixed scaling approach,” IEEE Trans. Electron Devices 55, 515-522 (2008).

[31] S. Lavizzari, D. Ielmini, D. Sharma and A. L. Lacaita, “Transient effects of delay, switching and recovery in phase change memory (PCM) devices,” IEDM Tech. Dig., 215-218 (2008)

[32] D. Ielmini and Y. Zhang, “Evidence for trap-limited transport in the subthreshold conduction regime of chalcogenide glasses,” Appl. Phys. Lett. 90, 192102 (2007).

[33] D. Ielmini and Y. Zhang, “Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices,” J. Appl. Phys. 102, 054517 (2007).

[34] D. Ielmini, “Phase change memory device modeling,” in Phase Change Materials – Science and Applications, S. Raoux and M. Wuttig Eds., 299-330 (2009). ISBN 978-0-387-84873-0.

[35] D. Ielmini, “Threshold switching mechanism by high-field energy gain in the hopping transport of chalcogenide glasses,” Phys. Rev. B 78, 035308 (2008).

[36] P. J. Walsh, R. Vogel and E. J. Evans, “Conduction and Electrical Switching in Amorphous Chalcogenide Semiconductor Films,” Phys. Rev. 178, 1274 (1969).

[37] X. S. Miao, et al., “Temperature Dependence of Phase-Change Random Access Memory Cell”, Jpn. J. Appl. Phys. 45, 3955 (2007).

[38] A. Pirovano, et al., “Low-Field Amorphous State Resistance and Threshold Voltage Drift in Chalcogenide Materials,” IEEE Trans. Elecctron Devices 51, 714-719 (2004).

[39] D. Ielmini, A. L. Lacaita and D. Mantegazza, “Recovery and drift dynamics of resistance and threshold voltages in phase change memories,” IEEE Trans. Electron Devices 54, 308-315 (2007).

[40] I. V. Karpov, et al. “Fundamental drift of parameters in chalcogenide phase change memory,” J. Appl. Phys. 102, 124503 (2007).

[41] D. Ielmini, D. Sharma, S. Lavizzari and A. L. Lacaita, “Reliability impact of chalcogenide-structure relaxation in phase change memory (PCM) cells – Part I: Experimental study,” IEEE Trans. Electron Devices 56, 1070-1077 (2009).

[42] D. Ielmini and M. Boniardi, “Common signature of many-body thermal excitation in structural relaxation and crystallization of chalcogenide glasses,” Appl. Phys. Lett. 94, 091906 (2009).

[43] A. Redaelli, D. Ielmini, U. Russo and A. L. Lacaita, “Intrinsic data retention in nanoscaled phase-change memories - Part II: Statistical analysis and prediction of failure time,” IEEE Trans. Electron Devices 53, 3040-3046 (2006).

[44] U. Russo, D. Ielmini and A. L. Lacaita, “Analytical modeling of chalcogenide crystallization for PCM data-retention extrapolation,” IEEE Trans. Electron Devices 54, 2769-2777 (2007).

[45] D. Ielmini, M. Boniardi, A. L. Lacaita, A. Redaelli and A. Pirovano, “Unified mechanisms for structure relaxation and crystallization in phase-change memory,” Microelectron. Eng. 86, 1942-1945 (2009).

[46] K. L. Chopra, “Avalanche-induced negative resistance in thin oxide films,” J. Appl. Phys. 36, 184 (1965).

[47] I. G. Baek, et al., “Highly scalable non-volatile resistive memory using Simple binary oxide driven by asymmetric unipolar voltage pulse,” IEDM Tech. Dig. 587-590 (2004).

[48] D.Ielmini, C. Cagli and F.Nardi, “Resistance transition in metal oxides induced by electronic threshold switching,” Appl. Phys. Lett. 94, 063511 (2009).

[49] X. Li, C. H. Tung, K. L. Pey and V. L. Lo, “The chemistry of gate dielectric breakdown,” IEDM Tech. Dig. 779-782 (2008).

[50] C. Cagli, F. Nardi and D. Ielmini, ”Modeling of set/reset operations in NiO-based resistive-switching memory (RRAM) devices,” IEEE Trans. Electron Devices 56 (2009).

[51] U. Russo, D. Ielmini, C. Cagli and A. L. Lacaita, “Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices,” IEEE Trans. Electron Devices 56, 186-192 (2009)

[52] U. Russo, D. Ielmini, C. Cagli and A. L. Lacaita, “Self-accelerated thermal dissolution model for reset programming in NiO-based resistive switching memory (RRAM) devices,” IEEE Trans. Electron Devices 56, 193-200 (2009).

[53] C. E. Housecroft and E. C. Constable, “Chemistry: an introduction to organic, inorganic, and physical chemistry, “ Pearson Education (2006).

[54] K. Kinoshita, et al., “Reduction in the reset current in a resistive random access memory consisting of NiO_x brought about by reducing a parasitic capacitance,” Appl. Phys. Lett. 93, 033506 (2008).