High temperature stability embedded ReRAM for 2x nm node and beyond

Conference Proceedings: 2022 IEEE International Memory Workshop (IMW)
Author: G. Molas; G. Piccolboni; A. Bricalli; A. Verdy; I. Naot; Y. Cohen; A. Regev; I. Naveh; D. Deleruyelle; Q. Rafhay; N. Castellani; L. Reganaz; A. Persico; R. Segaud; J. F. Nodin; V. Meli; S. Martin; F. Andrieu; L. Grenouillet
Publisher: IEEE
Date: 15-18 May 2022

https://ieeexplore.ieee.org/document/9779293

High temperature stability embedded ReRAM for 2x nm node and beyond

G. Molas[1], G. Piccolboni[1], A. Bricalli[1], A. Verdy[1], I. Naot[1], Y. Cohen[1], A. Regev[1], I. Naveh[1], D. Deleruyelle[2],Q. Rafhay[3], N. Castellani[4], L. Reganaz[4], A. Persico[4], R. Segaud[4], J. F. Nodin[4], V. Meli[4], S. Martin[4], F. Andrieu[4], L. Grenouillet[4]

[1]Weebit Nano Ltd (https://www.weebit-nano.com/), gabriel@weebit-nano.com, [2]INSA Lyon, INL, UMR5270, 69621 Villeurbanne, France, [3]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, Grenoble INO, IMEP-LAHC, 38000 Grenoble, France, [4]CEA, LETI, MINATEC Campus, GRENOBLE, France

***Abstract*—We report the performances and reliability of our ReRAM technology integrated in 28nm node. Low raw BER approaching 10^{-5} without ECC or redundancy is achieved. 10^6 cycles endurance without significant window degradation is shown. We report stable memory window after 15h bake at 210°C after 10kcycles, which is one of the best results reported so far to our knowledge. Technology passed basic (3x reflow) and extended (9 cycles) SMT tests with zero failures. Bitcell and memory stack engineering improved the window margin statistics. Optimized forming protocols are developed to increase memory yield over cycling. Program and verify algorithms allowed to insure no overlap between high and low resistive states on 1Mb arrays.**

Keywords—ReRAM, reliability, endurance, retention, SMT, program and verify

I. Introduction

As embedded memories are reaching the 28nm node, it is becoming more and more complex and expensive to scale standard charge-based memory concepts and integrate them with advanced CMOS nodes [1]. However, there is a high demand to scale memory nodes further for microcontrollers and AI circuit applications.

For this reason, emerging back-end memories are driving more and more interest from the industry, various demonstrations of NVM macro in 28nm node and beyond are reported in the literature [2].

Among the various existing emerging memory concepts, Resistive Random-Access-Memory (ReRAM) is a very promising technology for embedded Non-Volatile Memory thanks to faster speed, lower power consumption and lower additional mask count compared to Flash memory, and its ability to be integrated in the BEOL above the logic area. ReRAM technology recently underwent a significant speed up in maturity; Intel [3] and TSMC [4] reported macros in 1T1R configuration in 22nm node for embedded applications with comparable features (10 years retention at 85°C, 10kcycles endurance).

In order to address automotive market, strong requirements have to be met in terms of stability, implying high filament robustness at high temperatures. Panasonic reported equivalent 10 years retention at 85°C after 10kcycles on 8Mb array statistics [5] and Infineon recently demonstrated good behavior and memory window at 1ppm after 175°C temperature storage for 1000h [1].

In this paper, we present the reliability of our ReRAM technology, integrated in 28nm node. We discuss the memory potential and performances in terms of endurance, retention, immunity to disturb. Impact of program and verify algorithm on memory characteristics is also presented.

II. Technological Details

The ReRAM is integrated in the BEOL of 28nm CMOS process. The select element is a logic core (GO1) transistor. The memory is programmed at conditions overdriving nominal GO1 voltage, but GO1 transistor reliability is known to be not significantly impacted during OxRAM cycling [6]. 16kb and 1Mb test vehicles were used in this work in order to prove the intrinsic reliability of the technology.

III. Results and Discussion

A. Memory window and endurance

High endurance (above 100k programming operations) is required to replace current embedded non-volatile memories [1]. Endurance failure is due to a collapse of the resistance window, resistance can drop at an intermediate resistance between HRS and LRS [7] or be stuck at the LRS due to dielectric breakdown [8]. Endurance optimization requires adjustment of the programming conditions [9, 10], optimization of the programming energy [11], and tradeoff with the window margin [12]. Endurance failure follows a log-normal law with the number of cycles [8] and has to be addressed and engineered on memory arrays to reduce the degradation acceleration factor.

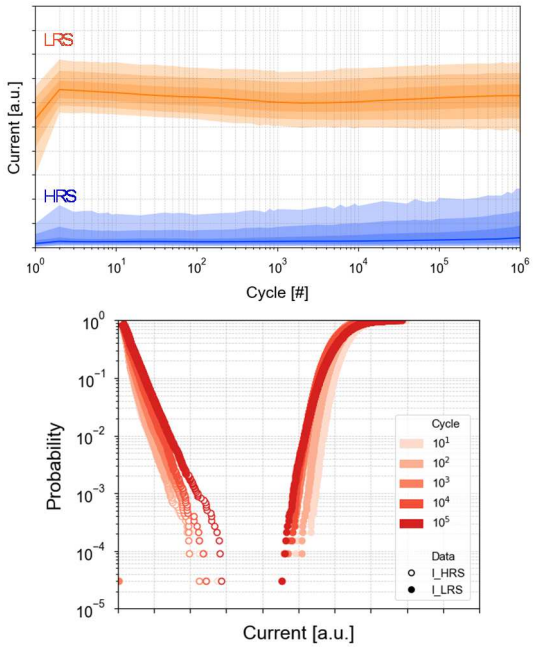

Fig.1. (a) Cell current over cycling for high (orange curve) and low (blue curve) resistive states. Line shows the median value while 1σ, 2σ, and 3σ contours are showed in light colours. (b) LRS and HRS current distribution measured during endurance, up to 10^5 cycles. No program and verify algorithm was used.

Our ReRAM can endure more than 4-σ distributions at 10^5 cycles with sufficient margin and no memory window degradation (fig.1). Again, in this case, no program & verify algorithm was used, showing the intrinsic quality of the technology and memory stack.

B. Retention and stability

In order to target automotive applications, very stable retention along with no fail has to be demonstrated at high temperatures. Retention failures are usually caused by filament dissolution resulting from motion and recombination of oxygen vacancies. Both LRS and HRS can drift over time. It was reported that no clear correlation exists between resistance in LRS after programming and resistance after bake [13]. In other words, bits that will fail during high temperature bake can hardly be screened in test by measuring the resistance after programing.

Information was retained after accelerated thermal stress and endurance cycling (10^4 cycles), with slightly small current drift after 15h at 210°C (Fig.2.a), showing solid cell reliability demonstration. One should note that most of the drift occurs during a short period, which is generally explained by the cell relaxation due to filament stabilization. Analyzing LRS retention BER which increases like a power-law over time [14], and based on Arrhenius extrapolations, data retention predicts 10 years operation at 220°C (Fig.2.b). Using lower programming current (low power operation) leads to 10 years retention at 130°C. Also note that the results presented here are raw data without write termination techniques, while improved reliability is expected when adapted programming protocols or verification schemes are used [13, 15].

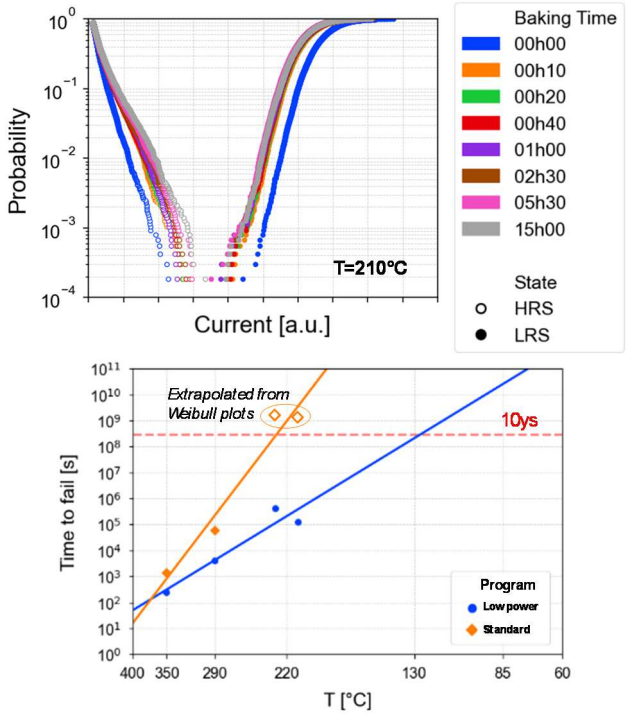

Fig.2. (a) LRS and HRS current distributions during bake at 210°C, up to 15h, for memory cells after 10kcycles. No program and verify algorithm was used. (b) Arrhenius graphs for two programming modes and corresponding programming currents.

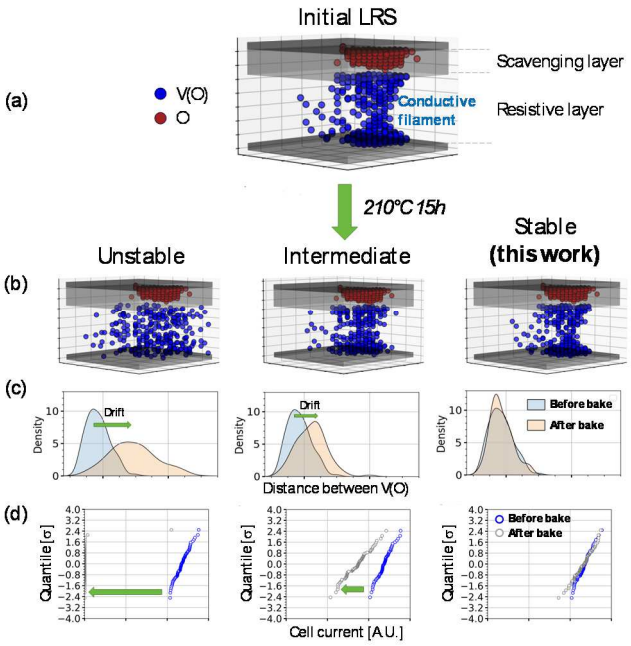

Fig.3. Kinetic Monte Carlo simulations of filament dissolution at high temperature. The motion of oxygen vacancies and oxygen ions is simulated over time and cell current is calculated. (a) Initial filament corresponding to Low Resistive State. (b) Simulated filament morphology after 15h at 210°C. Various set of microstructure features of the memory stack are considered, leading to unstable (left) to stable (right, this work) filament. (c) Simulated distribution of the distance between the adjacent oxygen vacancies constituting the conductive filament. More dissolution of the filament generates longer distance between the sites, leading to resistance drift after bake. In the stable system, oxygen vacancies profile is kepts unaltered. (d) Simulated cell current distributions before and after bake.

Retention enhancement is strictly correlated to the microstructural properties of the memory stack. Using Kinetic Monte Carlo calculations, we started from an initial conductive filament corresponding to the Low Resistive State and ran simulations of the filament dissolution over time at high temperature, taking into account the migration of oxygen vacancies in the resistive layer and potential recombination with Oxygen atoms. Playing with the material characteristics, it is possible to mitigate the filament spread and reduce the resistance drift (Fig.3). Thus, memory stack engineering enables strong improvement of retention behavior.

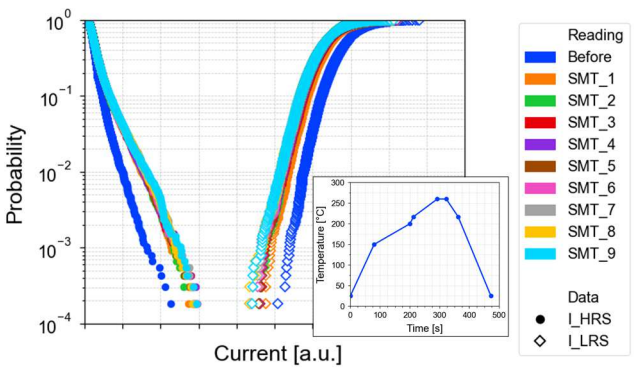

Fig.4. LRS and HRS current distributions before and after 1 to 9 reflow cycles (260°C SMT temperature profile is showed in the inset).

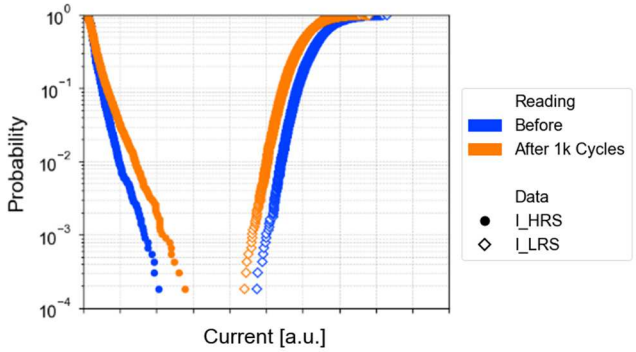

Fig.5. LRS and HRS resistance distributions before and after 1k thermal shock cycles at -65°C and 200°C.

In order to be solder reflow compliant, the technology has to sustain the Pb-free solder reflow profile, as described on JEDEC standards (IPC/JEDEC J-STD-020D.1), with a 260°C temperature peak for at least three reflow cycles (one for each of the board's side and another cycle in case rework is required). Our technology passed basic (three cycles of the temperature profile emulating solder reflow) and extended reflow (nine repeated cycles) with zero failures (Fig.4).

For packaging constraints, we evaluated immunity to extreme Thermal shock using the harshest temperature changes conditions, applying successive thermal stress cycles from -65°C to 200°C with 60s at the target temperatures based on JEDEC documentation (JESD22A104F). 1000 thermal cycles were achieved without a single measured failure (Fig.5).

C. *Disturb*

When integrated in a memory array, the memory cells undersee program disturb each time a neighbour cell sharing the same BL or WL is programmed. However, ReRAM technology is inherently robust to program disturb due to the select transistor in series with the memory that cuts the cell current. Fig.6 shows that LRS and HRS resistance distributions are unaltered after 10^5 disturb cycles.

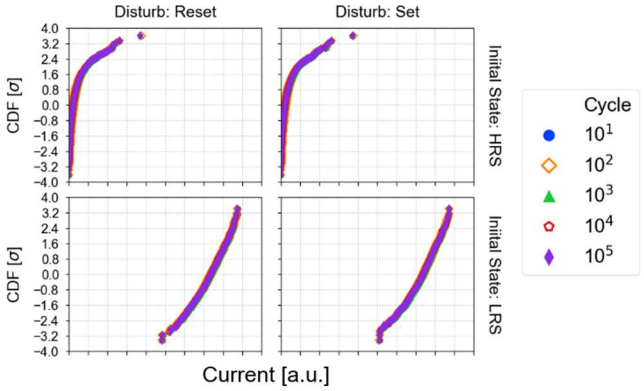

Fig.6. HRS (top) and LRS (bottom) current distributions after various numbers of disturb cycles. No resistance change is measured.

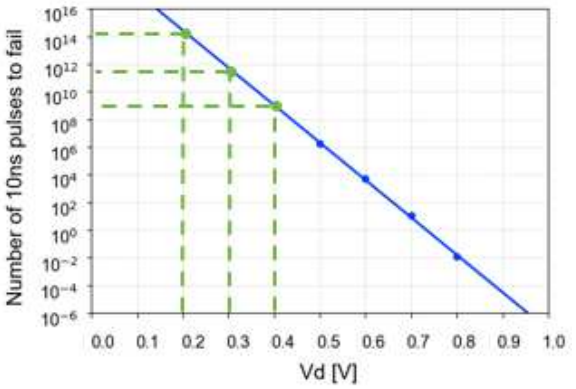

Fig.7. Number of 10ns reading pulses leading to failure as function of reading voltage Vd. Memory cells were cycles 10k times before disturb.

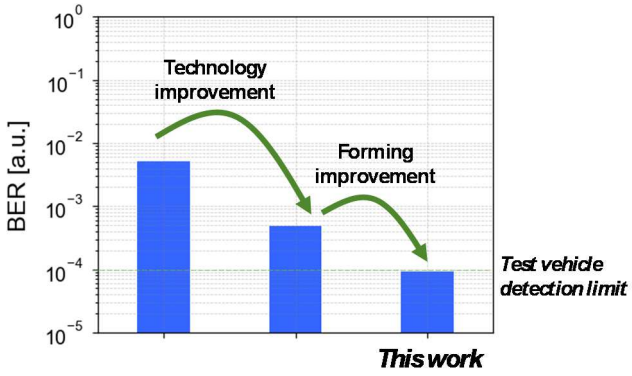

Fig.8. BER optimization after 10^4 cycles thanks to technology and forming improvements.

Due to time voltage dilemma, repeated reading cycles in ReRAM may end up with device switching [16]. Thus, high non-linearity between applied voltage and switching time is required to combine fast programming and robustness against read disturb. Fig.7 plots the number of reading cycles leading to cell switching as function of the reading voltage. Very high 10^{14} read cycles can be achieved before disturbance for a reading voltage of 200mV (Vwl=0V), without degradation of the target speed.

D. *Forming optimization*

Forming operation is a very important step that impacts the reliability of the memory [17, 18]. Indeed, the right amount of energy has to be provided in order to create a robust conductive filament, without degrading the dielectric, what affects endurance. We developed and compared various forming protocols and compared the BER during endurance. Optimized forming protocols allow to reduce the BER after 10^4 cycles below the detection limit of the test vehicle (Fig.8).

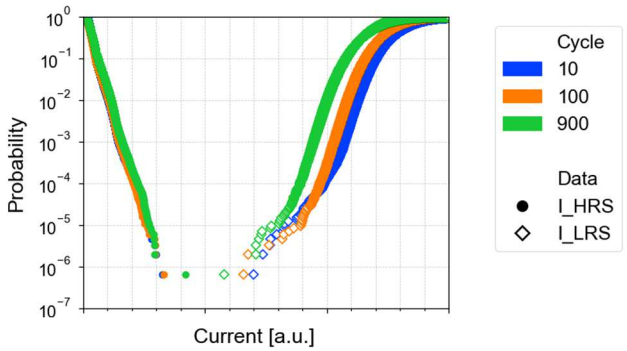

Fig.9. LRS and HRS resistance distributions with Program & Verify scheme after various cycles.

E. *Program and verify*

As previous sections were focusing on raw data showing the intrinsic performances of the technology, we present here the improvements that can be reached using program and verify (P&V) program schemes. In this case, each SET and RESET cycle is followed by a reading step in order to compare the cell resistance to a certain threshold, and to chose whether an additional pulse is needed or not. Various P&V schemes exist in the litterature, repeating the programming operation, or adapting the voltage, time or current (in the case of SET) on each subsequent cycle [19]. P&V has to handle with various issues related to current instabilities and reported in the litterature as stochastics nature of filament formation, resistance relaxation [20], Random Telegraph Noise (RTN) [21] and Ionic Telegraph Noise [22] or fluctuation induced false reading [23].

We used optimized P&V algorithms on 1Mb array to avoid the overlap between LRS and HRS resistance distributions. Both SET and RESET reprogram strategies had to be separately optimized to improve the reliabilbity and reduce the Bit Error Rate. For each cycle, most of the cells can be SET and RESET with nominal conditions. However, a certain population (tens of cells) require reprogram to reach the resistance threshold. Actually, it appears that any cell can require this additional operation at a certain cycle of its lifetime, evidencing a stochastic behavior rather than an intrinsic signature of erratic behavior. Figure 9 shows that LRS and HRS are clearly separated with no fails on a 1Mb array.

IV. CONCLUSIONS AND PERSPECTIVES

We have developed a reliable ReRAM technology integrated in 28nm, showing intrinsic reliability of more than 10^5 cycles endurance, SMT compliance and stable resistance for 15h at 210°C. Both technology enhancements and forming protocol optimization allowed to optimize the device performances. Then, using program and verify algorithm, clear window margin was achieved with no fails on 1Mb array.

Besides the integration of ReRAM technology on 28nm test vehicles, silicon demo wafers integrating Weebit's embedded ReRAM module were processed, what represents a key step in allowing potential customers to test and validate the technology, by testing the ReRAM parameters on packaged chips within specific applications.

Next step is then to scale the ReRAM technology further down to 22nm, designing a full IP memory module that integrates a multi-megabit ReRAM block on FDSOI process.

Overall objective is to target MCU for embedded applications, including automotive markets, but also open to AI opportunities [24, 25] with state of art memory nodes.

REFERENCES

[1] R. Strenz, Proc. 2020 IEEE International Memory Workshop (IMW), pp.1-4.

[2] G. Molas et al., *Appl. Sci.* 2021, *11*, 11254.

[3] O. Golonzka et al., Proc. 2019 Symposium on VLSI Technology, pp. T230-T231.

[4] C. -C. Chou et al., Proc. 2020 Symposium on VLSI Circuits, pp.1-2.

[5] S. Ito et al., Proc. 2018 IEEE International Memory Workshop (IMW), pp.1-4.

[6] L. Grenouillet et al., Proc. 2021 IEEE International Memory Workshop (IMW), pp.1-4.

[7] S. Balatti et al., IEEE Transactions on Electron Devices, vol. 62, no. 10, pp. 3365-3372, Oct. 2015.

[8] D. Alfaro Robayo et al., IEEE Transactions on Electron Devices, vol. 66, no. 8, pp. 3318-3325, Aug. 2019.

[9] J. Sandrini et al., 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 30.5.1-30.5.4.

[10] B. Govoreanu et al., 2013 5th IEEE International Memory Workshop, 2013, pp. 48-51.

[11] G. Sassine et al., 2018 IEEE International Memory Workshop (IMW), 2018, pp. 1-4.

[12] C. Nail et al., 2016 IEEE International Electron Devices Meeting (IEDM), 2016, pp. 4.5.1-4.5.4.

[13] M. Ueki et al., in IEEE Transactions on Electron Devices, vol. 64, no. 2, pp. 419-426, Feb. 2017.

[14] Y. -F. Chang et al., 2020 IEEE International Reliability Physics Symposium (IRPS), 2020, pp. 1-4.

[15] C. -C. Chou et al., Proc. 2018 IEEE International Solid-State Circuits Conference (ISSCC), pp. 478-480,

[16] J. Guy et al., Proc. 2014 IEEE International Electron Devices Meeting (IEDM)Tech. Dig., pp-152-155.

[17] J. Guy et al., Proc. 2017 IEEE International Reliability Physics Symposium (IRPS), pp.2.1-2.5.

[18] T. Ninomiya et al., Proc. 2012 Symposium on VLSI Technology, pp. 73-74.

[19] C. -F. Lee et al., Proc. 2017 IEEE Asian Solid-State Circuits Conference (A-SSCC), pp. 9-12.

[20] E. Esmanhotto et al., Proc. 2020 IEEE International Electron Devices Meeting (IEDM), 2020, pp. 36.5.1-36.5.4.

[21] S. Ambrogio et al., in IEEE Transactions on Electron Devices, vol. 62, no. 11, pp. 3805-3811, Nov. 2015.

[22] S. Brivio et al., Scientific Reports, (2019) 9:6310, pp.1-9.

[23] D. M. Nminibapiel et al., IEEE Electron Device Letters, vol. 38, no. 6, pp. 736-739, June 2017.

[24] I. Muñoz-Martin et al., Proc. 2020 Symposium on VLSI Technology, pp.1-2.

[25] S. K. Kingra et al., Appl. Phys. Lett. 120, 034102 (2022).