Integrated HfO_2-RRAM to Achieve Highly Reliable, Greener, Faster, Cost-Effective, and Scaled Devices

ChiaHua Ho*, Shuo-Che Chang, Chao-Yi Huang, Yu-Cheng Chuang, Seow-Fong Lim, Ming-Huei Hsieh, Shu-Cheng Chang, and Hsiu-Han Liao
Winbond Electronics Corp., Taiwan, email: chho1@winbond.com

Abstract – For the first time, this work demonstrated a 90nm 512Kb SPI HfO_2-RRAM product vehicle successfully with reducing read / write power by 18X / 2X, boosting read / write speed by 5X / 10X, and scaling feature size by 2X, compared to presented 512Kb SPI EEPROM; while sustaining high reliability on million cycle endurance, even better post-cycle retention (85ºC retention 100years for post 100K cycles), and 150ºC high temperature operation, by optimized mismatching, read-integrity, relaxation, and noise as discussed in this work. Technology also offers alternative solution for greener, highly-reliable, and scaled NOR Flash applications. A new plasma dicing technology was implemented to further increase gross die per wafer.

I. Introduction

Being compatible with CMOS process, RRAM (Resistive-Random-Access-Memory) are attractive and much investigated for decays [1-8]. However, due to cell-mismatching, randomly soft error during cycling [1], degradation by cycling / retention / 260ºC re-flow [3-4], oxygen-vacancy relaxation and random noise [4], long algorism [5], and oxygen-ion diffusion-out by BEoL thermal cycles, and large feature size as well, standalone marketing segment positioning of RRAM is still not clear yet. This paper for the first time reports overall integrated solution to fix above issues and provides a highly reliable, greener, faster, cost-effective, and scaled non-volatile memory branch.

In this work, reproducibly high-yielded 512Kb SPI HfO_2-RRAM product vehicle was to target SPI-EEPROM's high-reliability with extra benefit: 10X write speed boosting, 18X read power consumption reducing, and 2X area scaling, as shown in Fig. 1-2. The read speed and write power are also improved by respectively 5X and 2X. Contrary to the most of emerging memories [9-10], HfO_2 based TMO stack (Transition-Metal-Oxide) of this work is inserted between M1 and M2, shown in Fig. 3(c), to sustain availability of circuit routing for boosting product speed. Such advantages in power, speed, and area are highly encouraging for IoT applications beyond SPI-EEPROM.

II. Manufacturing HfO_2-RRAM

Fig. 3 discloses the detail of top- and cross-sectional views of HfO_2-RRAM product vehicle. TMO structure in Fig. 3(c-d) with two extra masks of RRAM cell and top-via lands on Co-silicide based 90nm logic comparable island-shaped active-area, as shown in Fig. 3(e-f). TE and BE (top- and bottom-electrodes) are conventional metals of TiN and Ti/TiN. Both CMP defects and surface oxidation of BE are well controlled to offer atomically flat interface with further TMO film. Monoclinic preference HfO_2-TMO is deposited by conventional ALD tool. Direct stack reactive-ion-etching patterns RRAM cell by considering effects of dielectric plasma damage and sidewall oxygen distribution. Composite oxygen ion reservoir stack also dominates RRAM reliability for reducing high load resistance caused from improper TiO_x phase-transition [1]. The whole TMO structure experiences extra thermal cycles of M2 ~ M4 with 350ºC ~ 450ºC processing temperature for metal routing availability of product applications.

III. Mismatch Optimization and Yield

A. Forming Optimization

Forming soft-breakdown dominates Set/Reset variation and RRAM reliability. Phenomena as function of ion-vacancy bonding energy, external bias, compliance current, and pulse-width can be well examined by 3D Kinetic Monte-Carlo (KMC) simulation with factors of ion-vacancy generation-recombination rate at interface of reservoir and TMO, vacancy diffusion within TMO film, vacancy re-oxidation nearby filament, hopping transport between vacancies, heat and Poisson equations, as shown in Fig. 4(a). Fig. 4(b) reveals a simulated Forming transient state in which the most of oxygen ions accumulate at interface; while few of ions distribute nearby filament. Resistance distribution with various scenarios is shown in Fig. 4(c). Tailing behaviors occurs for Condition-1 (under-Forming) and Condition-3 (complimentary-switching by over-Forming). For reliability concern, optimization of Forming resistance should be higher than further stabilized LRS (Low-Resistance-State).

B. Read Integrity

Higher read bias enhances sensing margin to compensate tailing induced margin loss but its read-disturb weakness would challenge product qualification, especially High-Temperature Operating Lifetime (HTOL) at 150ºC environment. Fig. 5 shows DC current-voltage extractions of die-to-die (hundreds) and cycle-to-cycle (within 10K cycles) variations. Unexpectedly wide Set transient behavior is found and independent of Reset ability even with optimized Forming condition as described above, suspected due to the grown mini-filament during beginning of Set process. However, a narrow and common operation window in Fig. 5 can be designed to sustain cell-to-cell and cycle-to-cycle stability and successfully passed 150ºC HTOL for 1000hrs, as shown in Fig. 11. Such read scheme can also support 10years read-disturb immunity at 90ºC [1].

C. Relaxation and Random-Telegraph-Noise (RTN)

Both effects of short-time RTN (< seconds, due to trap assist tunneling and prefers low-temperature) and long-time filament relaxation (< hours, due to ion-vacancy diffusion and prefers high-temperature) on High-Resistance-State (HRS) consume RRAM sensing margin even by proper program / read schemes. Combination can be illustrated in Fig. 6 with tightened initial tailing to 0.1uA (tailing > several million Ohms). Such behavior is found to depend on intrinsic TMO stack process.

D. Yield

Fig. 7 shows CP yield (combine 90ºC, -25ºC, data retention loss, and few cycles pre-checking) of small die product vehicle with respect to various process splits for improving TMO stack healthy and tailing from external resistance loading. Insert of Fig. 7 is die mapping of one of better CP yielded RRAM splits (CP 96%). All device / process knobs described above were integrated to achieve yielded and reproducible RRAM technology.

IV. Reliability and Power Consumption

A. Endurance

Fig. 8(a) depicts sensed currents of LRS and HRS of 256 RRAM cells with external YMUX circuits within 200K cycling. Extraction per every cycle avoids misjudgment from cycle randomly soft-error [1]. Fig. 8(b) shows the endurance extreme test of one cell up to billion cycles. Those imply HfO_2-RRAM intrinsic endurance characteristics. Furthermore, fully functional whole chip cycling (512Kb; total 880 units) close to million endurance is obtained in Fig. 9 at both 90ºC / 25ºC temperatures and also with / without 260ºC IR re-flow thermal budget. Testing is still ongoing. There is no cycle induced write time degradation, as shown in Fig. 10, compared to 90nm low density SPI NOR Flash, indicating the maturity of designed RRAM algorism.

B. Extremely High Temperature Operating

Fig. 11 illustrates designed testing flow for 150ºC operating and results of 512Kb SPI HfO_2-RRAM product vehicle of this work. Fully functional EFR test (Early Failure Rate) at 150ºC for 168hrs continuing read is achieved after 260ºC IR re-flow 3 times and post 150ºC 20K cycling, contributed from proper read scheme for read-integrity. Those units are also fully functional for further dynamic HTOL at 150ºC for 1000hrs, HTSL (High-Temperature Storage Lifetime) at 150ºC, and LTDR (Low-Temperature Data Retention) at 25ºC.

C. Overall Benchmark vs SPI-NOR and SPI-EEPROM

Overall benchmark of this work 90nm 512Kb SPI HfO_2-RRAM and 90nm 2Mb SPI NOR Flash is shown in Fig. 12. Except read / write comparable speed, all reliability and power indexes of this work exhibit much advantage than NOR Flash. Regarding to power consumption, thanks for low read / write voltages of RRAM with write algorism assistance, power loss in pumping circuit can be minimized. High RRAM's activation energy of 1.54eV ~ 1.67eV helps its equivalent data retention with orders of magnitude higher than NOR Flash.

Benchmarking to SPI-EEPROM (Fig. 1-2), all read / write power consumption and speed of HfO_2-RRAM exhibit great benefits; while sustaining its high reliability characteristics. Upon those results, HfO_2-RRAM is highly encouraging for EEPROM's reliability applications for IoT era.

V. Scaling Energy and Size

A. Energy Scaling

Set / Reset pulse widths of above achievements are 100nsec. For further write energy 5X reduction, Set / Reset pulse widths scaling to 10nsec, as shown in Fig. 13(d), achieves comparable reliability, as shown in Fig. 13(a-b, e), indicating ion-vacancy dynamics can narrow down to such short period.

B. Size Scaling

Due to filament localization characteristics [6], RRAM can be successfully scaled with technology node. Fig. 14 depicts a consistent reliability performance of 30% scaled TMO dimension for 55nm node RRAM. Million cycle performance without major degradation is observed for with and without 260ºC IR re-flow 3 times thermal stress. Achievement is benefit for new RRAM products with higher-density or even smaller die size.

In addition to TMO and logic device scaling, an innovated plasma dicing was developed to shrink scribe-line width from 65μm to 15μm successfully for gross die per wafer (GDW) further boosting, as shown in Fig. 15. Conventional Bosch type Si substrate etching by volume cycles of etching / deposition / clean accomplish the dicing without chipping damage.

VI. Summary

We have for the first time reported million-cycle highly reliable, 150ºC high-temperature operative, IoT power greener, cost-effective, and reproducible yielded SPI interface HfO_2-RRAM product vehicle with overall integrated solution by process / device / design. Technology is highly encouraging for IoT era SPI-EEPROM and also offers alternative potential application of IoT era NOR Flash.

Acknowledgment

The authors gratefully acknowledge the contributions of Xu Xu and M. P. Anantram / University of Washington to the RRAM 3D KMC simulation.

References

[1] ChiaHua Ho, T. Y. Shen, P. Y. Hsu, S. C. Chang, S. Y. Wen, M. H Lin, P. K. Wang, S. C. Liao, C. S. Chou, K. M. Peng, C. M. Wu, W. H. Chang, Y. H. Chen, F. Chen, L. W. Lin, T. H. Tsai, S. F. Lim, C. J. Yang, M. H. Shieh, H. H. Liao, C. H. Lin, P. L. Pai, T. Y. Chan, and Y. C. Chiao, IEEE VLSI symposia on Technology, T3P4 (2016).

[2] C.Y. Chen, A. Fantini, L. Goux, R. Degraeve, S. Clima, A. Redolfi, G. Groeseneken, and M. Jurczak, IEEE IEDM, S7-6 (2015).

[3] Z.-Q. Wang, S. Ambrogio, S. Balatti, S. Sills, A. Calderoni, N. Ramaswamy and D. Ielmini, IEEE IEDM, S7-6 (2015).

[4] J.-K. Lee, H. Y. Jeong, I.-T. Cho, J. Y. Lee, S.-Y. Choi, H.-I. Kwon, and J.-H. Lee, IEEE Electron. Device. Lett. 31, 603 (2010).

[5] Y. Meng, X. Y. Xue, Y. L. Song, J. G. Yang, B. A. Chen, Y. Y. Lin, Q. T. Zou, R. Huang, and J. G. Wu, IEEE VLSI Symposia on Technology, T22P3 (2014).

[6] ChiaHua Ho, Cho-Lun Hsu, Chun-Chi Chen, Jan-Tsai Liu, Cheng-San Wu, Chien-Chao Huang, Chenming Hu, and Fu-Liang Yang, IEEE IEDM, S19-1 (2010).

[7] ChiaHua Ho, E. K. Lai, M. D. Lee, C. L. Pan, Y. D. Yao, K. Y. Hsieh, Rich Liu, and C. Y. Lu, IEEE VLSI Symposia on Technology, 12B-2 (2007).

[8] ChiaHua Ho and Fu-Liang Yang, "Overview of Metal-Oxide Resistive Memory" in "Nonvolatile Memories: Materials, Devices and Applications", American Scientific Publishers, ed T.Y Tseng and S.M Sze (2012).

[9] K. Ohmori, A. Shinoda, K. Kawai, Z. Wei, T. Mikawa, and R. Hasunuma, IEEE VLSI Symposia on Technology, T7-1 (2017).

[10] J.R. Jameson, P. Blanchard, C. Cheng, J. Dinh, A. Gallo, V. Gopalakrishnan, C. Gopalan, B. Guichet, S. Hsu, D. Kamalanathan, D. Kim, F. Koushan, M. Kwan, K. Law, D. Lewis, Y. Ma, V. McCaffrey, S. Park, S. Puthenthermadam, E. Runnion, J. Sanchez, J. Shields, K. Tsai, A. Tysdal, D. Wang, R. Williams, M.N. Kozicki, J. Wang, V. Gopinath, S. Hollmer, M. Van Buskirk, IEEE IEDM S30-1 (2013).

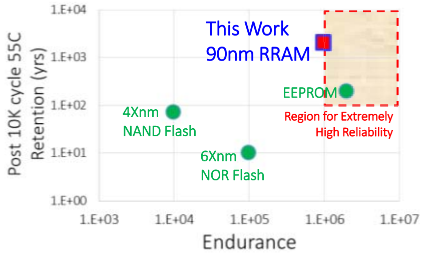

Fig. 1. High reliability performance of 90nm SPI HfO_2-RRAM product vehicle of this work.

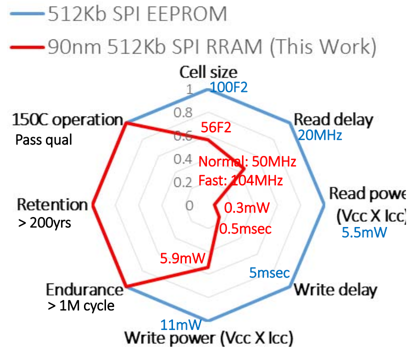

Fig. 2. Compared to SPI-EEPROM, SPI HfO_2-RRAM of this work achieves reduced read / write power consumption, boosted read / write speed, and scaled feature size; while sustaining million cycle reliability, high post-cycle retention, and high-temperature operation ability.

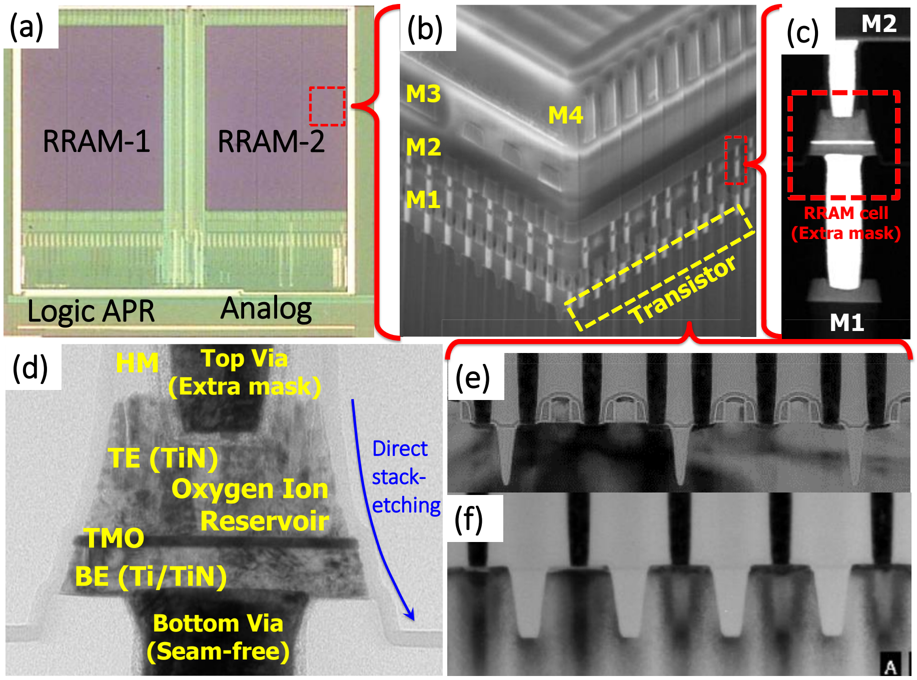

Fig. 3. (a) Top-view of 512Kb SPI HfO_2-RRAM product vehicle with 1.2V core and 3.3V IO. (b) Cross-sectional view of whole RRAM structure. (c) TMO stack locates between M1 and M2. Two extra masks are needed to pattern RRAM cell and top-via. (d) Detail TMO stack, including BE, TE, TMO, and oxygen reservoir stack. BE lands on seam-free bottom-via Tungsten-plug. Composite oxygen ion reservoir film stack for reducing high load resistance caused from improper TiOx phase-transition. Direct stack etching performs the whole RRAM cell patterning. (e-f) 90nm Co-silicide based driving transistor of RRAM cells.

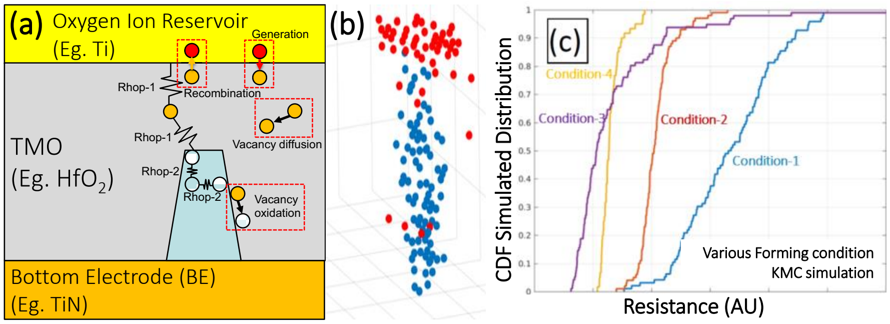

Fig. 4. (a) 3D Kinetic Monte-Carlo (KMC) simulation with various factors and as function of bonding-energy, external bias, compliance current, and pulse-width. (b) Simulated Forming transient state. (c) Resistance distribution with various scenarios.

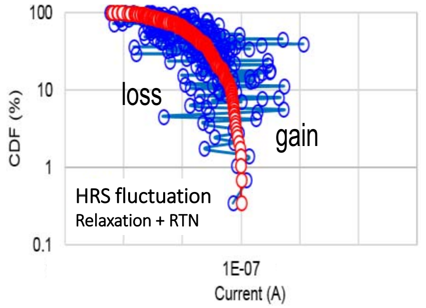

Fig. 6. Combination of filament relaxation and RTN effects on HRS with tightened initial tailing to 0.1uA (tailing > several million Ohms).

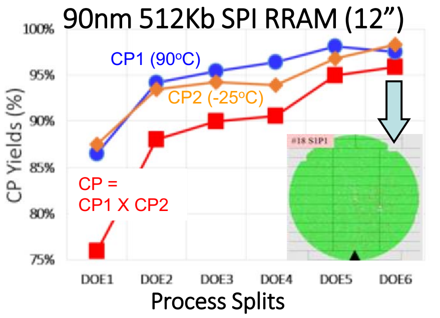

Fig. 7. CP yield of product vehicle with respect to various process splits for improving TMO stack and tailing from external resistance loading.

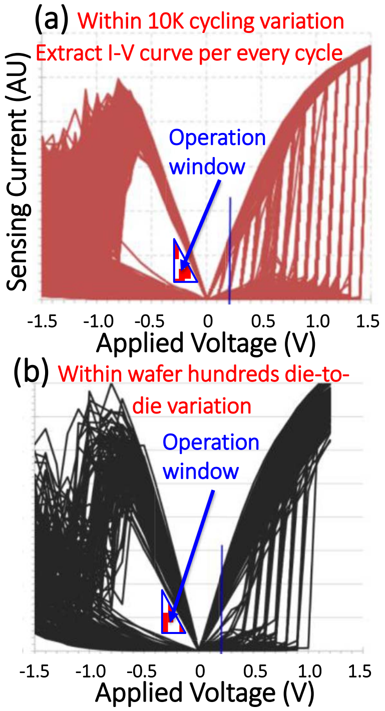

Fig. 5. DC extractions of die-to-die and cycle-to-cycle variations. A common operation window can be designed to fix those variations.

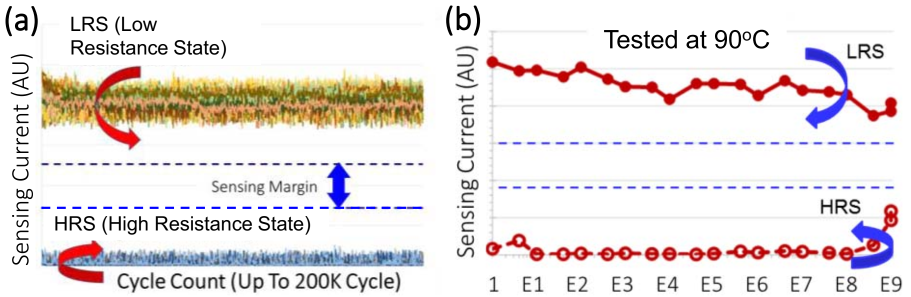

Fig. 8. (a) Sensed LRS / HRS currents of 256 RRAM cells with external YMUX circuits within 200K cycle. Extraction per every cycle avoids misjudgment from cycle randomly soft error. (b) Endurance extreme test of one RRAM cell up to billion cycles. Obvious HRS degradation can be observed after 100 million cycles.

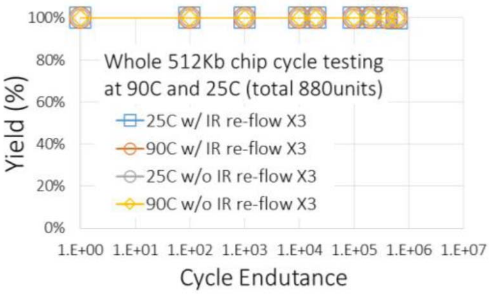

Fig. 9. Fully functional whole chip close to million cycle at both 90ºC / 25ºC and with / without 260ºC IR re-flow thermal budget. Testing is still ongoing.

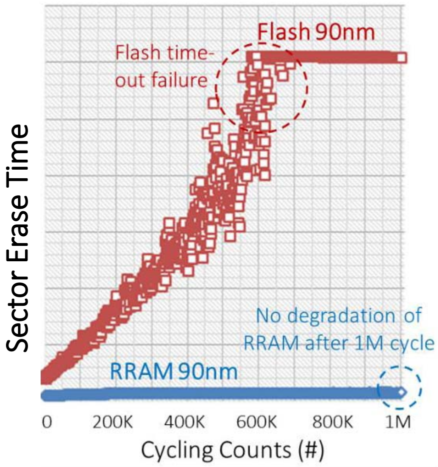

Fig. 10. HfO_2-RRAM exhibits negligible write time degradation during million cycles, compared to NOR Flash.

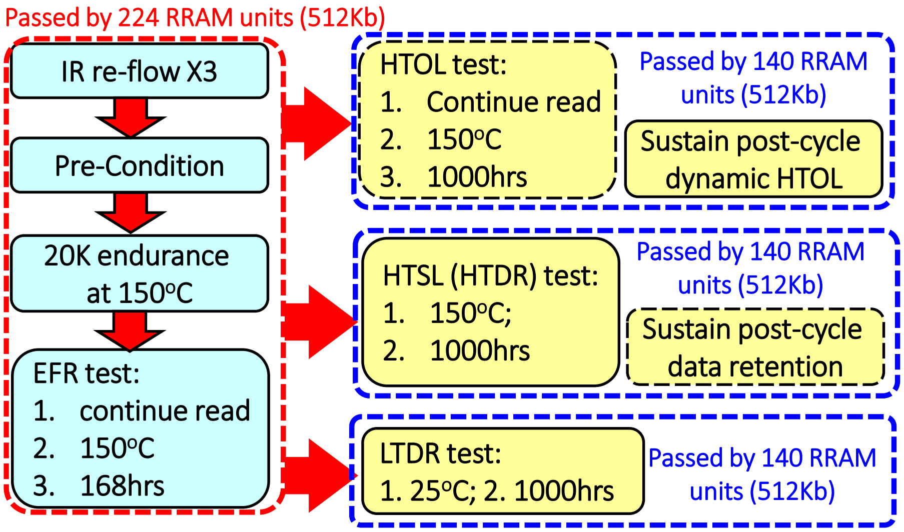

Fig. 11. Designed testing flow for 150ºC operating of HfO_2-RRAM product vehicle. Achieved fully functional EFR at 150ºC for 168hrs continuing read after 260ºC IR re-flow 3 times and post 150ºC 20K cycling, contributed from proper read scheme for read-integrity. Units are also fully functional for further dynamic HTOL at 150ºC for 1000hrs, HTSL at 150ºC, and LTDR at 25ºC.

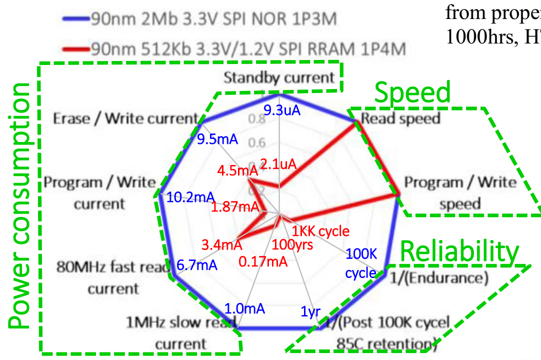

Fig. 12. Overall benchmark of this work 90nm 512Kb SPI HfO_2-RRAM and 90nm 2Mb SPI NOR Flash. Except read / write comparable speed, all reliability and power indexes of this work exhibit much advantage than NOR Flash. This offers alternative potential application of IoT era NOR Flash.

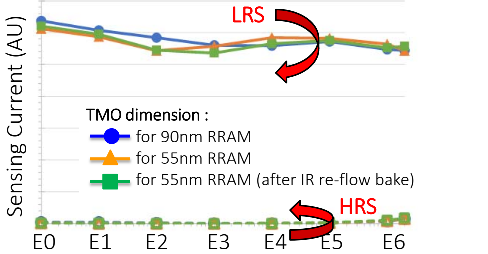

Fig. 14. Consistent million-cycle endurance of 30% scaled TMO dimension for 55nm node RRAM application. Same as 90nm node, endurance is independent of IR re-flow thermal cycle.

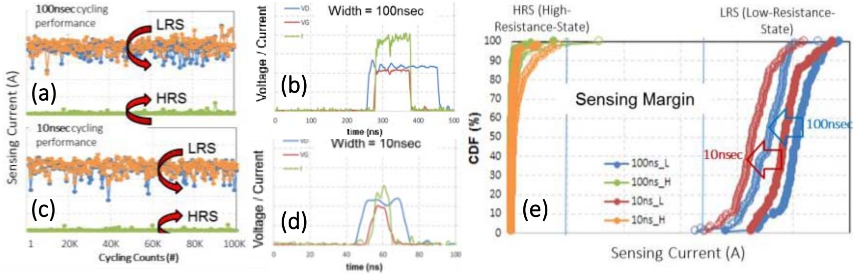

Fig. 13. For further write energy 5X reduction, pulse widths scaling to 10nsec achieves comparable reliability, indicating ion-vacancy dynamics can narrow down to such short period.

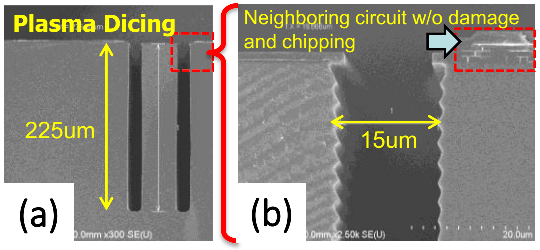

Fig. 15. Developed Bosch type plasma dicing to shrink scribe-line width from 65µm to 15µm for further increasing gross die per wafer. No chipping damage to neighboring circuit is