CMOS-embedded STT-MRAM Arrays in 2x nm Nodes for GP-MCU applications

D. Shum, Sr. Member, IEEE, D. Houssameddine, S.T.Woo, Y.S.You, J. Wong, K.W. Wong, C.C.Wang, K.H. Lee, K. Yamane, V.B. Naik, C.S. Seet, T. Tahmasebi, C. Hai, H. Yang, N. Thiyagarajah, R. Chao, J.W. Ting, N.L. Chung, T. Ling, T.H. Chan, S.Y. Siah and R. Nair

GLOBALFOUNDRIES Singapore Pte, Ltd., Singapore, 738406.

S. Deshpande, R. Whig, K. Nagel, S. Aggarwal, M. DeHerrera, J. Janesky, M. Lin, H.-J. Chia, M. Hossain, H. Lu, S. Ikegawa, F.B. Mancoff, G. Shimon, J.M. Slaughter, J.J. Sun, M. Tran, S.M. Alam, T. Andre

Everspin Technologies, Inc., Chandler, Arizona 85224 (USA)

VLSI Symposia on Technology and Circuits
Kyoto, Japan
June 2017

CMOS-embedded STT-MRAM Arrays in 2x nm Nodes for GP-MCU applications

D. Shum, Sr. Member, IEEE, D. Houssameddine, S.T.Woo, Y.S.You, J. Wong, K.W. Wong, C.C.Wang, K.H. Lee, K. Yamane, V.B. Naik, C.S. Seet, T. Tahmasebi, C. Hai, H. Yang, N. Thiyagarajah, R. Chao, J.W. Ting, N.L. Chung, T. Ling, T.H. Chan, S.Y. Siah and R. Nair

GLOBALFOUNDRIES Singapore Pte, Ltd., Singapore, 738406

S. Deshpande, R. Whig, K. Nagel, S. Aggarwal, M. DeHerrera, J. Janesky, M. Lin, H.-J. Chia, M. Hossain, H. Lu, S. Ikegawa, F.B. Mancoff, G. Shimon, J.M. Slaughter, J.J. Sun, M. Tran, S.M. Alam, T. Andre

Everspin Technologies, Inc., Chandler, Arizona 85224 (USA)

ABSTRACT — Perpendicular Spin-Transfer Torque (STT) MRAM is a promising technology in terms of read/write speed, low power consumption and non-volatility, but there has not been a demonstration of high density manufacturability at small geometries. In this paper we present an unprecedented demonstration of a robust STT-MRAM technology designed in a 2x nm CMOS-embedded 40 Mb array. Key features are full array functionality with low BER (bit error rate), process uniformity and reliability, 10 years data retention at 125C with extended endurance to ~ 107 cycles. All achieved with standard BEOL process temperatures. Data retention post 260ºC solder reflow temperature cycle is demonstrated.

INTRODUCTION

The growth of semiconductors in the consumer, industrial and automotive sectors has increased the demand of embedded non-volatile memories (eNVM) for general purpose (GP) general purpose microcontrollers. But the scalability trend of eFlash is limited by high voltage operation and tunnel oxide thickness. This limitation reduces the competitiveness of embedded flash (eFlash) in terms of cost and compact product design [1]. The advancement of CMOS technology posts additional challenges in integrating eFlash with features like HKMG, FDSOI, FINFET etc. MRAM, which can be embedded in CMOS BEOL with less process complexity, offers advantages in shorter learning cycles and better CMOS matching allowing design library re-usability. STT-MRAM with pMTJ devices extends MRAM technology to densities beyond those achieved with eFlash [2], enabling potential shrink beyond the 2x nm node thus making STT-MRAM an attractive candidate for Flash replacement.

Recent developments have improved our understanding of pMTJ bits and their magnetic properties, but reliable high memory density arrays embedded on 300mm CMOS Logic with standard BEOL processes have yet to be reported [3]. Manufacturing issues such as process repeatability, yield stability and factory cross-contamination control need to be addressed before embedded MRAM (eMRAM) can become a commercial success. It is the intention of this paper is to address these concerns and demonstrate a functionally competitive, logic-compatible embedded process in a 40 Mb array designed at 2x nm ground rules to debug any process complexities that arise from the add-on MRAM module. Array performance and key parameters to meet standard CMOS BEOL will be presented.

STT-MRAM INTEGRATION

Figure 1 shows the ST-MRAM cell scaling path on a normalized scale and compares well with eFlash ≤ 40nm [4]. The pMTJ layers are integrated between two Cu levels M5-M6 as shown in the Figure 2 flow chart. A TEM cross

section of the 40 Mb array is shown in Figure 3. Reliable interconnect integration requires smooth surface interfaces between MTJ films, the bottom electrode (BE) and the top electrode (TE) [5]. The MTJ is placed directly above the BE to minimize cell pitch. Figure 3a shows the MTJ tilted-angle SEM image after TE patterning. The insert image Figure 3b displays a flat and uniform morphology of the MgO tunnel barrier. The liner encapsulation is in-situ after MTJ etch to prevent oxidation of the MTJ sidewalls prior to vacuum break. The post-MRAM processes use standard 300 mm production tools with proper MRAM contamination control protocols as reported elsewhere [6].

ARRAY PERFORMANCE & DISCUSSIONS

MTJ stack and integration have been optimized for 400ºC, 60 minute post MTJ-patterning thermal budget as shown Figure 4. Figure 5 demonstrates stable magnetic properties on patterned MTJ bits across MTJ diameters down to 55nm. Figure 6 shows film optimization for two stacks where A is intended for high endurance and B for high data retention. Stack B was optimized for higher interfacial PMA (perpendicular magnetic anisotropy) resulting in higher Hk value.

Array data were taken from a 1Mb sub-array at wafer level. Figure 7 shows data retention measurements on both stacks after submitting the wafers to three consecutive 260ºC reflow anneals [7]. Stack B provides a significant improvement over stack A with most die having less than 10ppm fails after reflow with no ECC, which is sufficient to guarantee data integrity through solder reflow after ECC. Both stacks show a large programming window with zero fails for 50ns write pulses, as presented in Figure 8. The operating window shrinks on stack B because of increased switching voltage to achieve high data retention. Figure 9 shows a typical Rmin distribution and virtually no degradation in read margin post 107 write cycles, unlike Flash where the read window degrades after prolonged W/E cycling. Measurements were performed under bi-polar bias conditions to 107 cycles for each bit and confined to 1kb because of test time limitation. Work is on-going for a fully functional 40 Mb array read shmoo with frequency range up to 80MHz. Preliminary results show a sweet spot observed at ~ 20ns at nominal Vdd.

CONCLUSIONS

Perpendicular STT-MRAM is a promising candidate for eFlash replacement in MCU and IoT applications. It offers fast write, high endurance, high retention, 20ns read access at nominal Vdd operating voltage and IP re-usability.

References

[1] Y.S. Shin et al., VLSI Symp. Tech. Dig.,2005, pp. 156-159J; R. Strenz, IEDM 2011, pp. 211-214.

[2] J.M. Slaughter et al., IEDM 2012, pp. 673-676; IEDM 2016, pp. 568-571.

[3] A. Driskill-Smith at al., IMW 2011, pp.50-52. Y. J. Song et. Al., IEDM 2016, pp. 663-666.

[4] D. Shum et al., IMW 2015, pp.165-168; L.Q. Luo et al., IMW 2016, pp. 149-152.

[5] H. Honjo et al., VLSI Symp. Tech. Dig., 2015, pp. 160-161.

[6] S. Aggarwal, SEMICON WEST 2015, ”Solving Manufacturing Challenges: ST-MRAM on advanced nodes“.

[7] M.C. Shih et al., VLSI Symp. Tech. Dig.,2016, pp. 144-145; J. A. Rodriguez et al., IMW 2016, pp. 157-160.

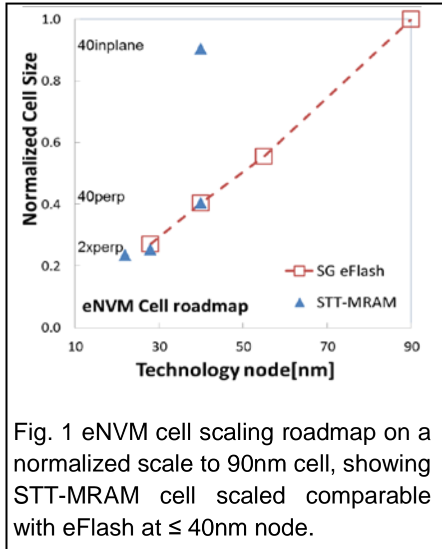

Fig. 1 eNVM cell scaling roadmap on a normalized scale to 90nm cell, showing STT-MRAM cell scaled comparable with eFlash at ≤ 40nm node.

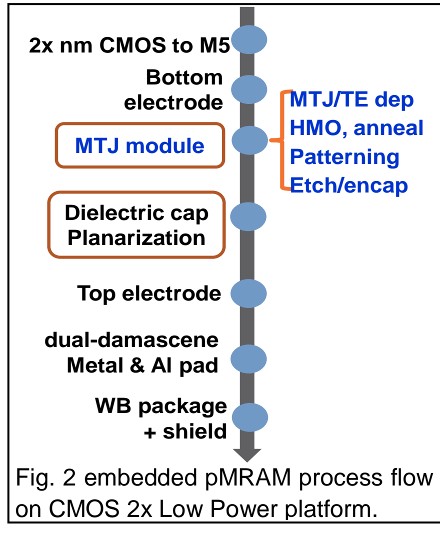

Fig. 2 embedded pMRAM process flow on CMOS 2x Low Power platform.

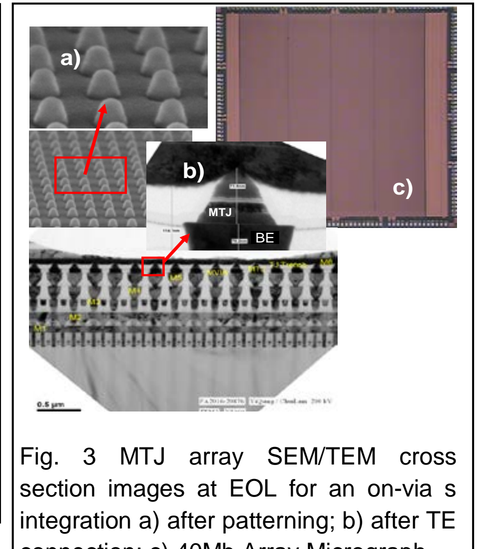

Fig. 3 MTJ array SEM/TEM cross section images at EOL for an on-via s integration a) after patterning; b) after TE connection; c) 40Mb Array Micrograph.

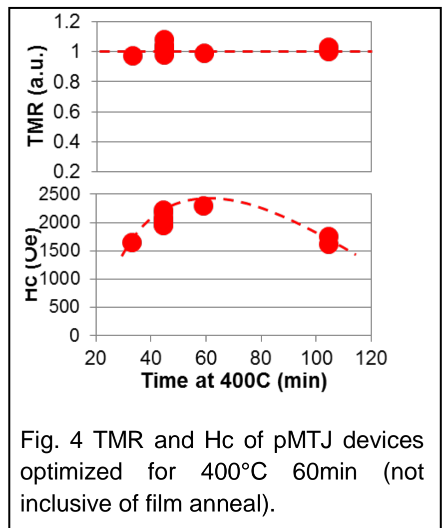

Fig. 4 TMR and Hc of pMTJ devices optimized for 400°C 60min (not inclusive of film anneal).

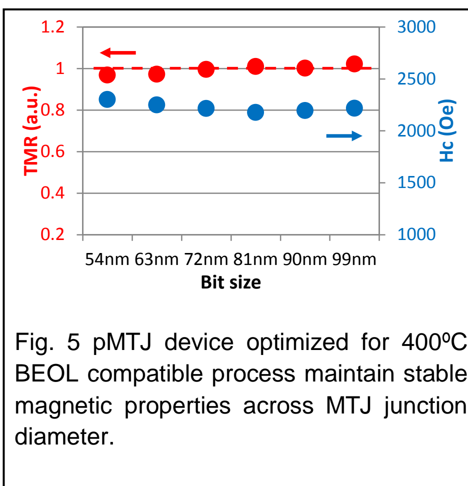

Fig. 5 pMTJ device optimized for 400ºC BEOL compatible process maintain stable magnetic properties across MTJ junction diameter.

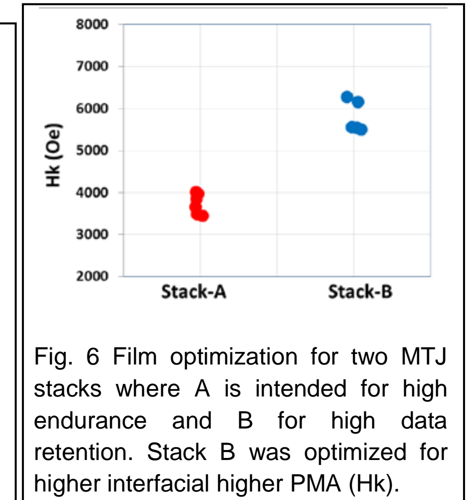

Fig. 6 Film optimization for two MTJ stacks where A is intended for high endurance and B for high data retention. Stack B was optimized for higher interfacial higher PMA (Hk).

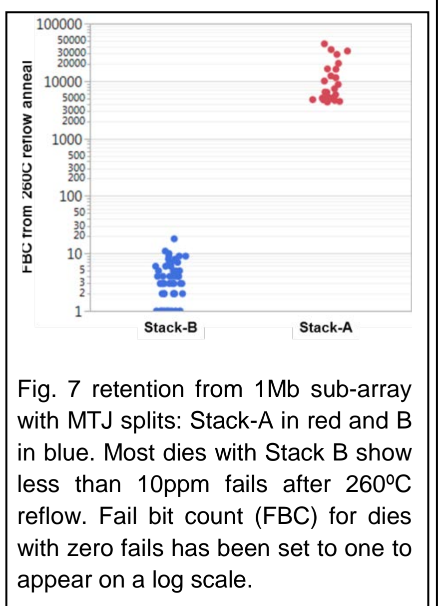

Fig. 7 retention from 1Mb sub-array with MTJ splits: Stack-A in red and B in blue. Most dies with Stack B show less than 10ppm fails after 260ºC reflow. Fail bit count (FBC) for dies with zero fails has been set to one to appear on a log scale.

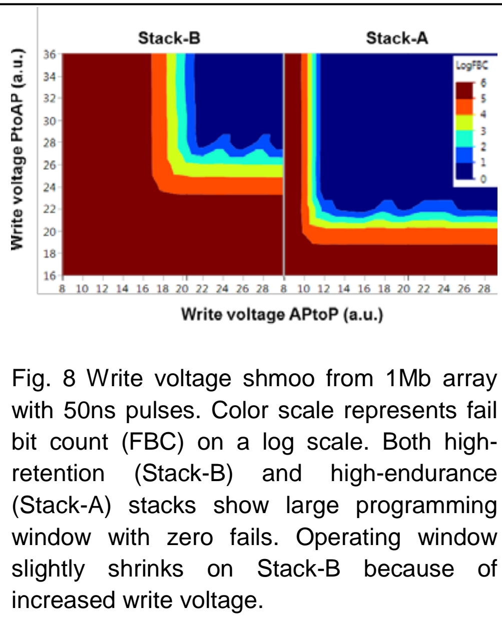

Fig. 8 Write voltage shmoo from 1Mb array with 50ns pulses. Color scale represents fail bit count (FBC) on a log scale. Both high-retention (Stack-B) and high-endurance (Stack-A) stacks show large programming window with zero fails. Operating window slightly shrinks on Stack-B because of increased write voltage.

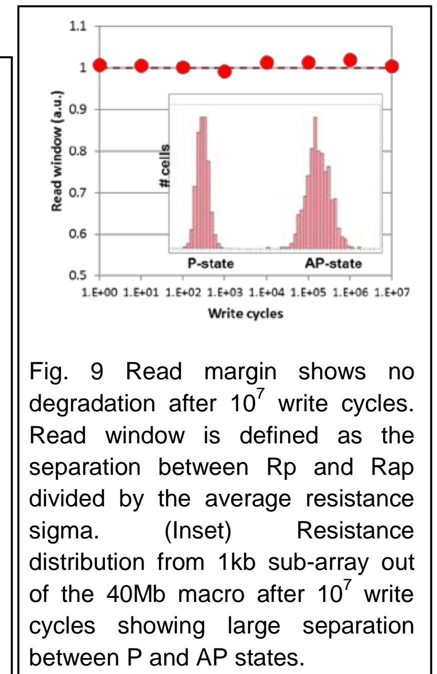

Fig. 9 Read margin shows no degradation after 10^7 write cycles. Read window is defined as the separation between Rp and Rap divided by the average resistance sigma. (Inset) Resistance distribution from 1kb sub-array out of the 40Mb macro after 10^7 write cycles showing large separation between P and AP states.