MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology

O. Golonzka, J. -G. Alzate, U. Arslan, M. Bohr, P. Bai, J. Brockman, B. Buford, C. Connor, N. Das, B. Doyle, T. Ghani, F. Hamzaoglu, P. Heil, P. Hentges, R. Jahan, D. Kencke, B. Lin, M. Lu, M. Mainuddin, M. Meterelliyoz, P. Nguyen, D. Nikonov, K. O'brien, J. ODonnell, K. Oguz, D. Ouellette, J. Park, J. Pellegren, C. Puls, P. Quintero, T. Rahman, A. Romang, M. Sekhar, A. Selarka, M. Seth, A. J. Smith, A. K. Smith, L. Wei, C. Wiegand, Z. Zhang and K. Fischer

Intel Corporation, Santa Clara, CA, USA, email: oleg.golonzka@intel.com

Abstract— This paper presents key features of MRAM-based non-volatile memory embedded into Intel 22FFL technology. 22FFL is a high performance, ultra low power FinFET technology for mobile and RF applications with extensive high voltage and analog support, and a high level of design flexibility at low cost[1]. Embedded NVM technology presented here achieves 200°C 10-year retention capability combined with $>10^6$ cycle endurance and high die yield. Technology data retention, endurance and yield capabilities are demonstrated on 7.2Mbit arrays. We describe device-level MTJ characteristics, key integration features, cell characteristics, array operation specifics, as well as key yield milestones.

I. Introduction

Embedded Non-Volatile Memory (e-NVM) technology has generated interest as a potential solution for several key market segments. Internet of Things applications, field-programmable arrays and chipsets with on-chip boot data are among potential customers for e-NVM. Existing solutions are provided by external flash and embedded flash memories. These technologies suffer from latency delay in the case of external flash and high manufacturing costs for embedded flash. Several new technologies are emerging as competitive replacements for flash memory. Among them, Magnetoresistive Random-Access Memory (MRAM) offers low manufacturing cost as well as exceedingly competitive data retention and switching endurance capabilities. In addition to embedded nonvolatile applications, Magnetic Tunnel Junction (MTJ)-based memory has the potential to be competitive as higher level SRAM or e-DRAM replacement, as well as a basic building block for future logic devices[2]. What makes MTJ-based technology unique is the large range of switching energy vs. retention vs. endurance tunability, allowing for large application flexibility. These features of MRAM have recently attracted investments from major semiconductor companies[3,4,5].

II. Device Characteristics

This work uses dual-MgO MTJs with a CoFeB-based free layer (Fig. 1). Typical Resistance-vs-Voltage (R-V) characteristics are shown on Fig. 2. Tunneling Magnetoresistance Ratio (TMR) is >180% with Resistance-Area product (RA) of 9 $\Omega\mu m^2$. Target device size is between 60nm and 80nm. Coercivity of the device can be easily tuned to achieve the desired retention or switching characteristics. Fig. 3 shows write-error-rate curves for a typical device. Write-error-rate curves were collected with 20ns, 80ns, 500ns and 1µs pulses and show relatively flat Vc values for pulses down to 80ns and transition to increasingly precessional switching at 20ns pulses.

III. Test Vehicle

Fig. 4 shows a cross-sectional TEM of MTJ array embedded between Metal 2 and Metal 4 of the 22FFL logic process. Fig. 5 shows the layout of the 1T-1R MRAM cell used in this work. The MTJ device is centered on a Metal 2 pad placed on a 216nm x 225nm pitch grid (cell area of $0.0486\mu m^2$). We employ single-ended sensing and 128b-Triple error correction. We use an adaptive Write-Verify-Write (WVW) scheme with a sequence of write pulses of increasing pulse lengths and amplitudes. The WVW scheme can be tuned to trade off write energy vs retention vs endurance depending on the intended application. Read sensing is done with a short (<10ns) low amplitude pulse and data readout is accomplished by detecting the differential current signal between the MTJ and a thin film precision resistor tuned to provide optimal read margin.

IV. Endurance and Retention

Fig. 6 shows wafer-level bit error rates observed on full 7.2Mbit arrays subjected to 10^6 switching cycles with two switching protocols optimized for high retention or low switching energy, correspondingly. The data demonstrate cycling induced fallout <1E-6 rate for the high retention switching protocol and <5E-7 for the low energy switching protocol. Fig. 7 shows the projected wafer-median array-level 10-year retention capability for stacks with three different free layer thicknesses. The data demonstrate 155°C, 175°C and 200°C retention capability depending on the free layer thickness choice. Retention projections were made using raw bit failure rates observed at multiple bake temperatures and times, followed by extrapolation of those data to extract the temperature at which the fail rate would fall below 1E-5 after 10 years.

V. YIELD

A. Shorting across the MgO barrier

One of the key issues in developing a high yielding MRAM process is overcoming shorting across the extremely thin MgO barrier layer. Several mechanisms contribute to this defect mode: shorting around the perimeter of the MTJ device due to remaining conducting metal residue, direct shorting between the free and reference layers due to non-uniformity of the MgO barrier layer, and shorting due to MTJ top contact wrap around. Fig. 8 shows the 12-month time trend of the "Shorting across the MgO barrier" defect indicator. With meticulous optimization of the MTJ stack, MTJ etch and integration we were able to achieve <1E-6 wafer-level bit error rate due to this defect mode, which is well below the ECC budget.

B. Thermal stability of the MTJ stack

Integrating MTJ-based memory into CMOS process brings an additional challenge of preserving magnetic properties of the stack through the thermal cycle of the subsequent back-end-of-line (BEOL) processing. Typical cumulative thermal cycle of modern BEOL will add close to one hour of exposure at temperatures >400°C. An implementable stack needs to show little-to-no degradation in TMR, coercivity, and switching voltage distributions when subjected to such exposure. Fig. 9 shows blanket-level TMR data of the stack developed for this process as a function of anneal temperature, demonstrating best-in-class capability to withstand 440°C one hour exposure.

C. Synthetic antiferromagnet locking

Device parametrics play a key role in achieving high yields. Fig. 10 shows a comparison of full M-H sweeps for two stacks: a typical stack (Stack A) and an advanced stack (Stack B). Stack A exhibits 3 stable states at zero external field: parallel and antiparallel states of the free and reference layers under antiparallel configuration of the Synthetic Antiferromagnet (SAF) as well as the state originating from the undesired parallel configuration of the bottom and top parts of the SAF. A small percentage of Stack A devices are manufactured with the SAF locked in the parallel state, leading to a strong stray field on the free layer, ultimately resulting in the free layer magnetization being locked parallel to the reference layer magnetization. This defect can be overcome by subjecting the devices to the appropriate external magnetic field at the end of manufacturing process, see Fig. 11. The downside of Stack A is that its functionality will be compromised if the device is maliciously subjected to a very large "positive" magnetic field. This vulnerability is completely eliminated in case of Stack B due to substantially stronger coupling in the SAF.

D. Existence of states with intermediate resistances

Another key consideration affecting MRAM yield is the switching time of MTJ devices. Fig. 12 shows two examples of time-resolved traces for AP to P switching under applied low-voltage 1µs pulses, showing typical switching behavior (left) and rare behavior in which the device exhibits a stalled state having an intermediate resistance value, in this case persisting for several hundred nanoseconds (right). We will further refer to these states as "intermediate" states. A bit can remain in the intermediate state well after the end of the write pulse and ultimately complete switching into the P state or return back to the AP state. With an adaptive WVW switching scheme, returning back to the original AP state after falsely passing a verify read directly contributes to the write error rate. We attribute the physical origin of the intermediate states to the formation of a metastable, multi-domain magnetic configuration of the free layer. As expected, experimentally observed occurrence rate of the intermediate states drops dramatically with reducing the size of the device (Fig. 13). For a given device size, among other effects, formation of the intermediate states is driven by the inherent non-uniformity of the stray dipole field resulting from the incomplete cancellation of fields produced by the top and the bottom halves of the SAF structure. Fig. 14 shows simulations of the stray dipole field experienced by the free layer of an 80nm device of Stack A and Stack B discussed in the previous paragraph. Simulations were normalized to produce zero average stray field integrated across the full device and show that Stack A exhibits substantially larger non-uniformity in the radial distribution of the stray field with device edges favoring P configuration. This non-uniformity ultimately results in early nucleation of AP to P switching on some devices and formation of a metastable domain structure with P configuration around the device edges and AP configuration in the device center. Fig. 13 shows experimental results comparing the intermediate states metric for Stack A vs Stack B at the same device CD and matched average stray field. As expected, due to the improved within-bit stray field uniformity Stack B shows a significant reduction in the intermediate states.

VI. SUMMARY

An industry leading combination of high-retention, high-endurance MRAM-based e-NVM and high performance, ultra low power FinFET CMOS logic technology has been developed. e-NVM technology uses a 216nm x 225nm 1T-1R memory cell and demonstrates 200°C 10-year retention capability and $>10^6$ write endurance. Technology retention, endurance and high yield capabilities were shown on 7.2Mbit arrays and full 300mm wafers.

REFERENCES

[1] B. Sell et al., "22FFL: A High Performance and Ultra Low Power FinFET Technology for Mobile and RF Applications", IEDM, 2017.

[2] D. E. Nikonov et al., "Proposal of a Spin Torque Majority Gate Logic", IEEE Electron. Device Lett., v. 32, n. 8, pp. 1128-30, 2011.

[3] M-C. Shih, et al., "Reliability study of perpendicular STT-MRAM as emerging embedded memory qualified for reflow soldering at 260 C", VLSI Symposium, 2016.

[4] Y. J. Song, et al., "Highly Functional and Reliable 8Mb STT-MRAM Embedded in 28nm Logic", IEDM, 2016.

[5] D. Shum, et al., "CMOS-embedded STT-MRAM Arrays in 2x nm Nodes for GP-MCU applications", VLSI Symposium, 2017.

Fig. 1. Simplified stack schematic.

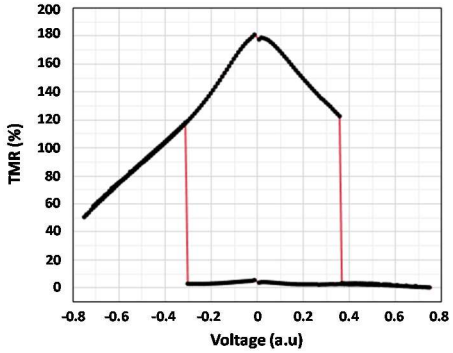

Fig. 2. Typical R-V curve. TMR is 180%.

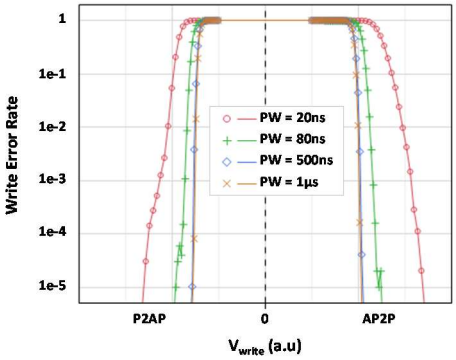

Fig. 3. Write error rates collected with 20ns, 80ns, 500ns and 1µs pulses. Significant increase in Vc between 80ns and 20ns data demonstrates transition into recessional switching regime.

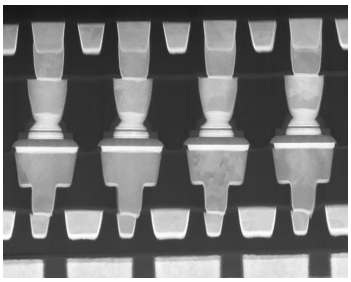

Fig. 4. Cross-sectional TEM of an MTJ array embedded between Metal 2 and Metal 4 of the 22FFL logic process.

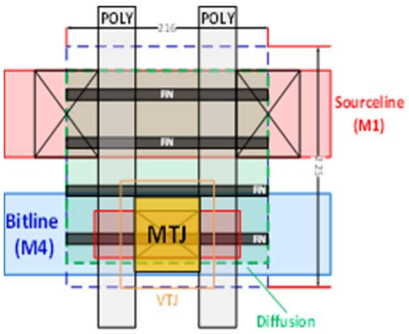

Fig. 5. Layout of the 1T-1R MRAM cell used in this work.

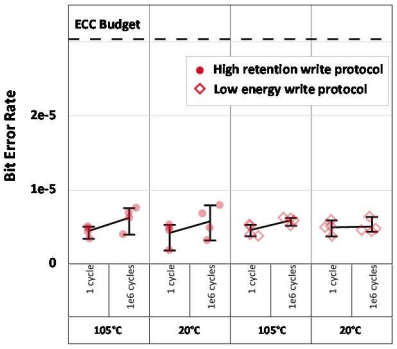

Fig. 6. Median wafer-level bit error rates before and after 10^6 write cycles at 105°C and 20°C with high retention and low energy switching protocols.

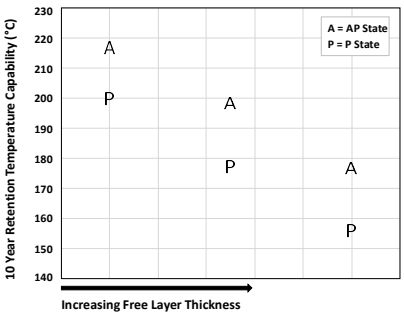

Fig. 7. Projected array level retention capability for stacks with different free layer thicknesses.

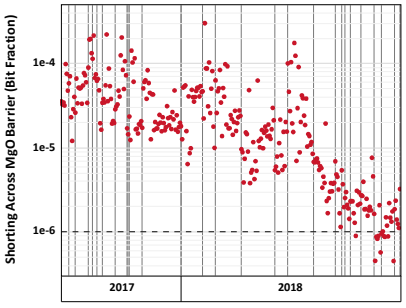

Fig. 8. 12-month time trend for the "Shorting across MgO Barrier" defect.

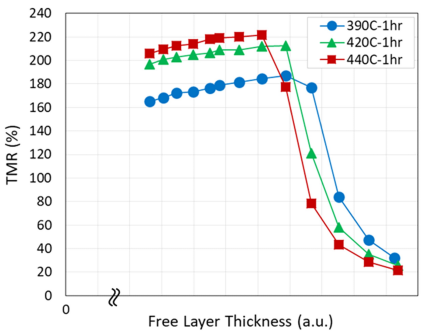

Fig. 9. TMR vs free layer thickness for a blanket film MTJ stack, with RA= 9 $\Omega\mu m^2$ annealed at different temperatures.

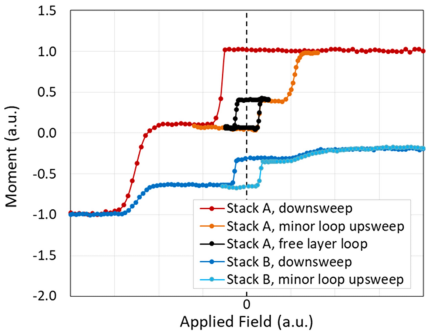

Fig. 10. M-H curves of patterned MTJ arrays. Stack A exhibits three states under no external magnetic field, while Stack B eliminates the undesired parallel SAF configuration.

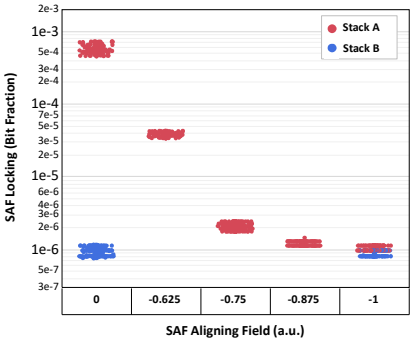

Fig. 11. Fraction of bits showing SAF locking vs the magnitude of SAF-aligning magnetic field. Stack B shows immunity to SAF locking behavior.

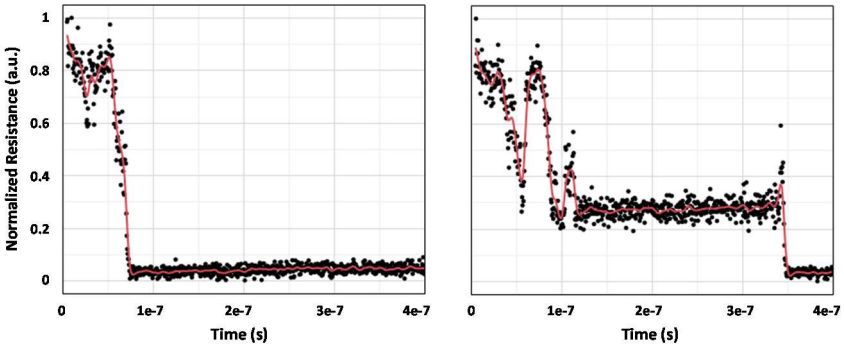

Fig. 12. Examples of time resolved resistance traces collected during AP to P device switching with low-voltage 1µs pulses. Typical device switching (left) vs device switching which exhibits a 250ns-long metastable state having an intermediate resistance value (right).

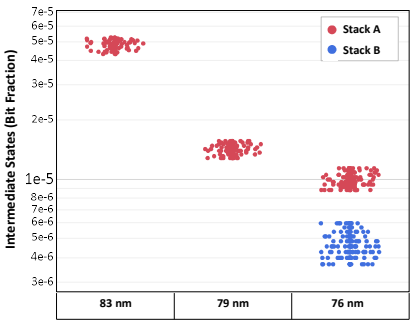

Fig. 13. Fraction of bits showing intermediate states vs device size. For a given device size Stack B shows lower intermediate state counts due to improved uniformity of the uncompensated stray field profile arising from the SAF.

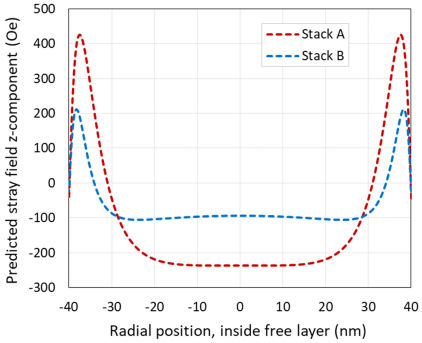

Fig. 14. Simulations of the stray field profile experienced by the free layer of an 80nm device with Stack A and Stack B.