Spin-orbit torque MRAM for ultrafast cache and neuromorphic computing applications

Siddharth Rao
Compute and memory technology
Imec
Leuven, Belgium
Siddharth.Rao@imec.be

Kaiming Cai
Compute and memory technology
Imec
Leuven, Belgium
Kaiming.Cai@imec.be

Giacomo Talmelli
Unit process module
Imec
Leuven, Belgium
Giacomo.Talmelli@imec.be

Nathali Franchina-Vergel
Unit process Module
Imec
Leuven, Belgium
Nathali.Franchina@imec.be

Ward Janssens
Unit process Module
Imec
Leuven, Belgium
Ward.Janssens@imec.be

Hubert Hody
Compute and memory technology
Imec
Leuven, Belgium
Hubert.hody@imec.be

Farrukh Yasin
Compute and memory technology
Imec
Leuven, Belgium
Farrukh.Yasin@imec.be

Kurt Wostyn
Compute and memory technology
Imec
Leuven, Belgium
Kurt.Wostyn@imec.be

Sebastien Couet
Compute and memory technology
Imec
Leuven, Belgium
Sebastien.Couet@imec.be

***Abstract*— Spin-orbit torque (SOT) magnetic random-access memory (MRAM) is a 3-terminal non-volatile memory technology promising high speed up to multi-GHz, high endurance and non-volatility. Here we show how SOT-MRAM stack can be optimized to reach performance towards an embedded last level cache memory replacing SRAM. Moreover, we show how the stack and device geometry can be optimized to increase density and how the stack properties can be optimized to perform analog in-memory computing (AiMC) functions using high resistance devices.**

Keywords—MRAM, NVM, SOT, in-memory compute

I. Introduction

Magnetic random-access memories (MRAM) have emerged as an attractive non-volatile memory (NVM) technology owing to their high speed, tunable retention, small footprint and compatibility with CMOS core voltage at advanced nodes [1-3]. Spin-transfer torque (STT)-MRAM has been in mass production since 2019 at major foundries as an option to replace embedded-flash at advanced nodes [4,5]. There is currently a large industrial effort to extend the STT-MRAM application space to last level of cache (LLC) [6]. Indeed, such a memory would provide higher density than SRAM and minimize standby power leakage. However, the STT switching mechanism and device geometry suffers from other trade-offs for LLC. The switching dynamics is expected to limit switching speed down to a few ns. Moreover, since STT-MRAM is a 2-terminal device sharing read/write path, the higher switching current required will impact the lifetime to breakdown and limit endurance, which needs to be extremely high for cache applications. Hence it is not (yet) clear whether STT will be capable of delivering LLC-like performance.

Spin-orbit-torque (SOT)-MRAM is an alternative 3-terminal device geometry [7]. Switching of the free layer is induced by passing a charge current in an adjacent high spin-orbit coupling heavy metal layer, such as Ta, W, and Pt. As shown in Fig. 1, a top-pinned magnetic tunnel junction (MTJ) is deposited on top of the SOT track so as to ensure direct contact between the SOT heavy metal layer and the free layer of the MTJ. The MTJ stack is necessary to enable read-out via the tunnel magneto-resistance (TMR) effect. This effectively provides separate read/write path and promises much higher endurance since only a smaller current (compared to STT) has to go through the thin MgO dielectric for reading only. For perpendicular MTJs, ultrafast switching below 300 ps has been demonstrated [8]. These two advantages (‘unlimited’ endurance and GHz operation) makes it an attractive candidate for non-volatile cache applications.

SOT-MRAM technology provides more opportunities for performance engineering since both the SOT track and the free layer design can be adjusted, to reduce further the writing

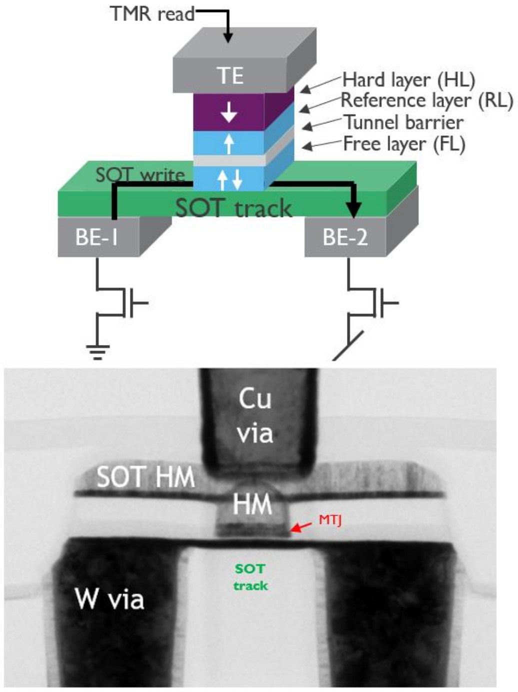

Fig. 1. Schematic view and TEM micrograph of a SOT-MRAM integrated device.

current [7]. The 3-terminal geometry however implies an intrinsically less favorable bit density compared to 2-terminal STT-MRAM devices [9]. Hence, methods for creating a SOT-multi-pillar (MP) structure where several MTJs share a common SOT track are currently being explored [10].

Finally, the write-independent tunability of the MTJ resistance has also led research into its use in Analog in-memory compute (AiMC) architectures, where individual read currents of several MTJs are summed in the SOT track channel [11].

Here, we review the development of a common stack, capable of serving both high performance cache and AiMC and with 400 °C backend-of-line (BEOL) thermal budget compatibility. We show how device performance can be optimized to satisfy both high performance cache and AiMC applications.

II. Stack development

The SOT track and MTJ stack largely define the device performance in terms of switching current, read window and retention. LLC and AiMC applications share several common interests such as large read window (linked to the junction TMR) and low write current. One of the major differences between the two lies in the junction resistance, which needs to be low for enabling fast read in LLC [10] and very high (MΩ) for analog summation of sense currents [11]. This figure of merit can be translated into CD-independent resistance- area (RA) product target for the film stack. A summary of key requirements is listed in Table I. Included only for TMR in this table but also important for other parameters are: tight distribution control is key to both LLC and AiMC, where the accuracy will also be linked to variability [11]. Overall, stack development for switching performance can be combined while a specific development is needed to adjust the junction resistance to fit AiMC application.

TABLE I. Key SOT device targets for LLC and AiMC

Target	*Units*	*LLC*	*AiMC*
TMR	%	150	150
TMR	σ/μ		19
RA	$\Omega.\mu m^2$	5-10	5000
I_{write}	μA	100	100
t_{write}	ns	0.3-1	1
Δ	k_BT	60	75

In principle, this can be readily done by adjusting the deposition time of the MgO tunnel barrier. However, since the interfacial quality of the MgO/CoFeB layers is critical to ensure a good perpendicular magnetic anisotropy (PMA) of the CoFeB layer, it is important to ensure that the PMA is unaffected by the RA increase. Indeed, a degradation of the interfacial PMA will reduce the range of CoFeB thicknesses that can be used, thereby directly impacting the data retention and TMR metrics at device level.

Multilayer MTJ film stacks are deposited by PVD on 300mm wafers and subsequently annealed to either 300 °C for the W-based systems or 400 °C for the Pt-based systems. We evaluate first a relatively conventional stack based on a β-W SOT track and CoFeB free layer. Fig.2 shows the remanent (zero-field) magnetization of the free layer as a function of CoFeB thickness for different junction RA. For the W/CoFeB system, there is a continuous decrease of the PMA window as you increase the RA, ultimately leading to a too narrow window for proper device functionality as you reach RA above 1000 $\Omega{\cdot}\mu m^2$.

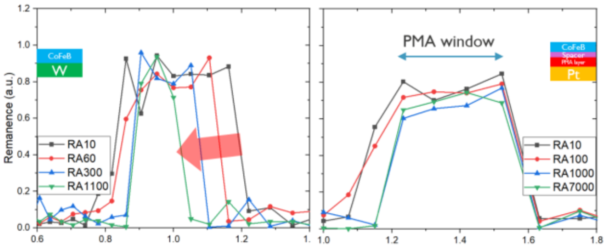

Fig. 2: PMA window for different CoFeB thickness and junction RA for (left) W/CoFeB and (right) Pt/hybrid free layer.

In order to reach 400 °C thermal budget compatibility, another system was developed based on a Pt track and a hybrid free layer (HFL) concept [7] that enables higher magnetic volume and hence higher retention at scaled dimensions. The hybrid free layer still contains CoFeB for the TMR effect but rely on other materials to increase the PMA. Fig.2 shows a much more stable PMA window in the tested RA range. From the W/CoFeB behavior, we clearly observe that interfacial PMA provided by the MgO/CoFeB decreases with increasing RA, thus reducing the PMA window for the CoFeB. There is a small decrease for the Pt/HFL system, but since an intrinsic PMA layer is present in the free layer, it turns out to be less sensitive to the quality of the MgO/CoFeB interface. Hence, Pt track and hybrid-free layer design were selected to enable LLC and AiMC applications at relevant BEOL thermal budget processing.

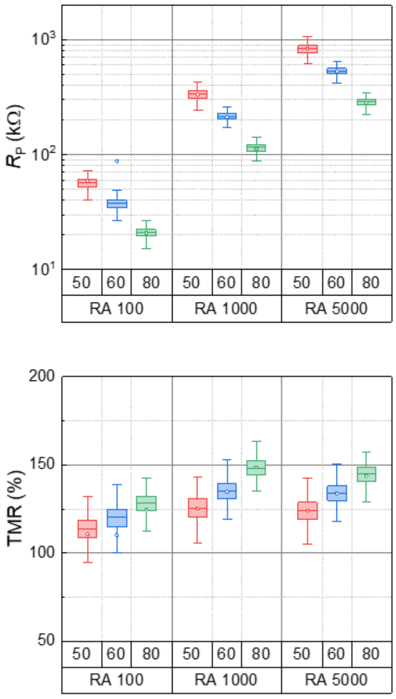

Fig. 3. R_p and TMR of Pt/HFL MTJs for different RA and CD.

III. Device performance

A. Static performance

As mentioned in the previous section, RA scaling is one of the main significant differences between LLC and AiMC requirements. Fig. 3 shows the basic MTJ properties - parallel resistance (R_p) and TMR – as a function of RA and device diameter (CD) for the Pt/HFL system. As can be seen, scaling RA from 100 to 5000 $\Omega\cdot\mu m^2$ enables MΩ resistances at 50 nm CD. Thanks to the higher resistance, the TMR also increases and reaches close to 150 % at 80 nm CD. Further optimization of the MTJ etch step is expected to enable similar TMR at smaller CD in the future.

B. Switching performance

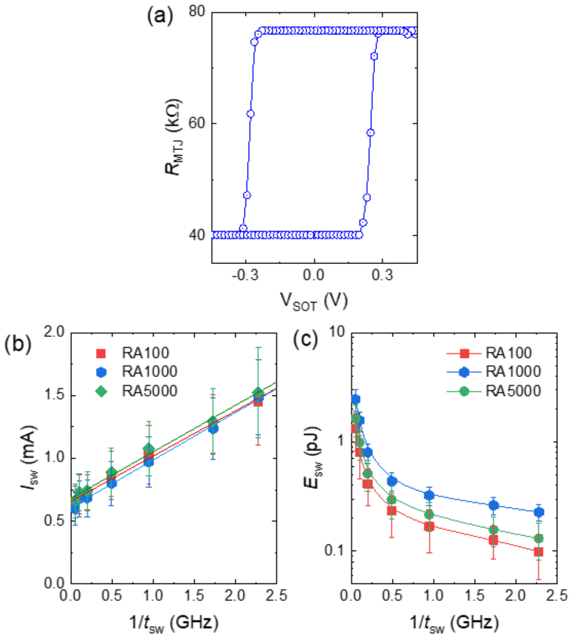

Fig. 4. SOT switching of Pt/HFL device. (a) R-V curve of a 60-nm device with RA100 under t_{PW} = 0.3 ns and $|B_x|$ = 35mT. (b) and (c) I_{sw} and E_{sw} as a function of $1/t_{sw}$ for devices with different RA.

Fig. 4a shows a typical current-induced switching loop in a 60 nm Pt/HFL device at write pulse-widths of 300 ps. The critical switching current (I_{sw}) of the Pt/HFL system shows a similar linear scaling on pulse width (t_{sw}) as reported in the W/CoFeB system [12], which can be expressed as:

$$I_{sw}(t_{sw})=I_0+Q/t_{sw} \tag{1}$$

, where I_0 is the intrinsic critical switching and Q factor is an effective charge parameter to describe the number of electrons to be injected into the SOT system for magnetization switching. Faster switching speeds could be obtained by applying higher currents. While there are some differences in switching energy for different RA, they follow a similar trend up to 2.5 GHz. At frequencies <= 1 GHz, thermal fluctuations play a stronger role in the reversal process. Owing to the current injection geometry, increased thermal effects at lower frequencies leads to higher switching energy requirements. As the frequency goes beyond 1 GHz, thermal effects are minimized and the device switches more reliably under the influence of solely the SOT torques., thereby leading to lower switching energies.. This confirms write operations can be performed at high speed for LLC. The large operating frequency range will allow to adjust the SOT-based AiMC core to specific requirements.

The presence of an intrinsic PMA layer in the free layer enables more flexibility to adjust the retention and switching current characteristics. In Fig. 5a, the device-level retention is estimated from switching field distributions fitted to a macrospin model of magnetization reversal [13]. To ensure accuracy, the switching fields for each device are measured from 3500 switching events. Different free layer stack designs enables us to cover a wide Δ range from 60 k_BT (>10 years at RT) to 140 k_BT (>20 years at RT). Fig. 5b shows the corresponding switching current performance for Stack flavours 1-3. One can see that there is no strict relationship between Δ and I_{sw}. We suspect this to be due to a more complex switching dynamics induced by SOT. Stack 1 proposes an already high $\Delta > 130$ k_BT while keeping the lowest switching current. While switching currents are presently some distance away from the target requirement, we expect performance improvement can be provided by further engineering of the SOT track and free layer design which is an active field of research both in academic and industrial research community.

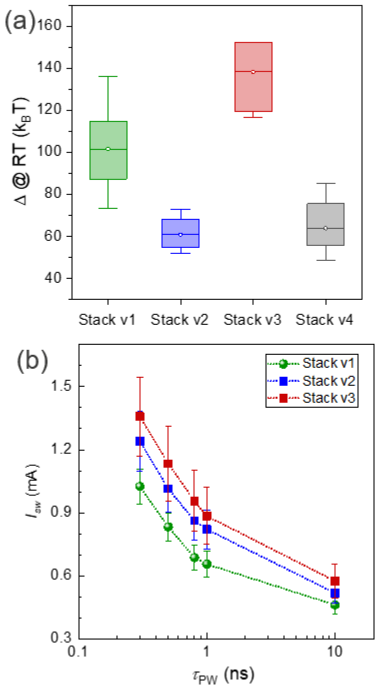

Fig. 5. a) Free layer retention for different hybrid free layer design and (b) corresponding switching performance for stack 1-3.

C. Single- and mutli- pillar devices

In order to enhance the density of SOT-MRAM devices, the concept of multi-pillar devices for AiMC has been proposed with multiple MTJ pillars on a shared SOT track [10, 11]. This configuration enables to store for instance 4-bits on one SOT track which would provide enough accuracy for weight storage. One important device parameter for neural network accuracy is to achieve sufficiently low resistance variation device to device, in particular in view of the low TMR window provided by the MTJ.

Fig. 6 show the distributions of device properties in W/CoFeB and Pt/HFL devices. The W/CoFeB devices show a large device-to-device variability for multi-pillar devices at different locations. Moreover, the low B_c (< 50 mT) for CD = 60 nm is insufficient for further device scaling. The Pt/HFL devices offer better uniformity and higher retention. We believe this less uniform distribution for the W/CoFeB system is at least partly linked to the sole reliance of the system to sustain PMA to the CoFeB/MgO interface which can be impacted by the etch process. In particular, the smaller 60nm device show even stronger variation while the Pt/HFL devices keep very good distributions.

Fig. 7 shows the typical SOT switching curves of 80-nm multi-pillar devices for a pulse width of 100 ns and $|B_x|$ = 33 mT. Thanks to the good uniformity of Pt/HFL devices (shown in Fig. 6), the switching loops of four MTJs on the shared SOT track are almost overlapped, which indicates similar critical switching currents. However, in the W/CoFeB system, switching loops from four bits show some drifts with certain differences in switching currents, which may result in some penalties in write-error rate (WER) performance and device reliability. The excellent device variations of the Pt/HFL system offer more flexibility and facility for large-scale integration process and circuit design in industrial production and application.

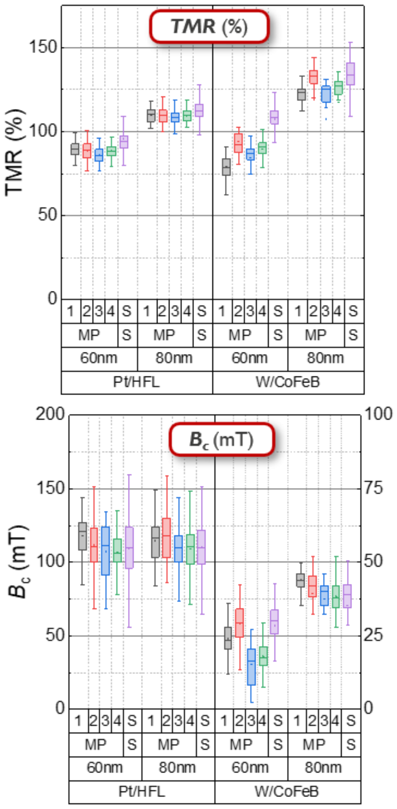

Fig. 6. TMR and B_c for different locations of single- and multi- pillar devices of Pt/HFL and W/CoFeB structures.

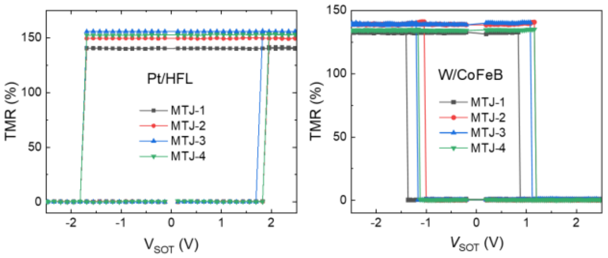

Fig. 7. SOT switching curves of Pt/HFL and W/CoFeB multi-pillar devices for t_{PW} = 100 ns and $|B_x|$ = 33 mT.

Conclusions

We have explored the device level performance of a new hybrid free layer concept developed for high performance SOT-MRAM applications. The design relying on the introduction of an intrinsic PMA layer within the free layer which is in direct contact to a Pt-based SOT track. This design enables higher PMA and retention, and is less sensitive to the PMA originating from the CoFeB/MgO interface. The MgO barrier thickness can be adjusted up to MΩ device-level resistance with almost no modification of the free layer PMA. The intrinsic PMA layer also allows easy adjustment of the retention and write current for different applications. Finally, we demonstrate 0.2pJ/bit switching up to 2GHz. While there remains several challenges for the industrialization of SOT-MRAM such as further reducing the switching current and bitcell size, the presented device sign shows that high performance applications such as last level cache or AiMC can be achieved.

Acknowledgment

This work was supported by IMEC's industrial affiliation program on MRAM device. We also acknowledge the support from the ECSEL Joint Undertaking Program (grant No. 876925- project ANDANTE).

References

[1] G. Jan et al., *IEEE* Symp. VLSI Tech., pp. 65 – 66 (2018)
[2] J.J. Kan et al., *IEEE* IEDM, pp. 27.4.1 – 27.4.4 (2016)
[3] G. Hu et al., *IEEE IEDM*, pp 19.38 – 19.41 (2019)
[4] V. B. Naik et al., *IEEE* IEDM, pp. 2.3.1 – 2.3.4 (2019)
[5] K. Lee et al., *IEEE* IEDM, pp. 2.2.1 – 2.2.4 (2019)
[6] C. Safranski et al. *IEEE* Symp. VLSI Tech., pp. 288-289 (2022)
[7] S. Couet et al., *IEEE* Symp. VLSI Tech., pp. 1-2, (2021)
[8] K. Garello et al., *IEEE* Symp. VLSI Tech., pp. 194-195 (2019)
[9] M. Gupta et al., *IEEE IEDM*, pp.24.5.1-24.5.4 (2020)
[10] K. Cai et al., *IEEE* Symp. VLSI Tech., pp. 375-376 (2022)
[11] J. Doevenspeck et al., *IEEE* Symp. VLSI Tech., pp. 1-2 (2021)
[12] K. Garello et al., *Appl. Phys. Lett.* 105, 212402 (2014)
[13] L. Thomas et al., *J. Appl. Phys*. 115, 172615 (2014)