Multi-pillar SOT-MRAM for Accurate Analog in-Memory DNN Inference

J. Doevenspeck[1], K. Garello[1,2], S. Rao[1], F. Yasin[1], S. Couet[1], G. Jayakumar[1], A. Mallik[1]
S. Cosemans[1], P. Debacker[1], D. Verkest[1], R. Lauwereins[1,3], W. Dehaene[1,3] and G.S. Kar[1]
[1]imec, Leuven, Belgium, [2]SPINTEC, Grenoble, France, [3]KU Leuven ESAT, Leuven, Belgium, email: jonas.doevenspeck@imec.be

Abstract: Deep neural network (DNN) inference can be performed efficiently with analog in memory computing (AiMC). MRAM is an attractive solution to implement the DNN weights due to its non-volatility and scalability. However, accurate inference requires memories with multi-level conductance values. while MRAM is binary. In this work, we propose and demonstrate a multi-level SOT-MRAM device concept by placing multiple MTJ pillars between a single SOT track and common top electrode. Selective level programming is achieved by smartly using a pillar-position dependent VCMA-assist effect. Three DNN algorithm-driven technology requirements are derived: number of conductance levels, bit-error rate and conductance variation. This work demonstrates that multi-pillar SOT-MRAM meets all derived specifications, making it a promising candidate as memory device for accurate analog in-memory DNN inference.

Introduction: Analog in memory computing (AiMC) is a promising architectural solution to process deep neural networks (DNNs) (Fig.1). Very high efficiencies can be achieved by implementing the DNN weights with non-volatile memories with large resistance values and small variability [1]. Recently [2], SOT-MRAM was demonstrated to fulfil these requirements with adjustable resistances to target array size, tight distributions, and no compromise on writing, thanks to a separate write and read path (Fig.2). Moreover, SOT-MRAM combines excellent endurance, fast and low power switching and good scalability [3,4], making it an appealing candidate as AiMC weight memory.

In [2], an SOT-MRAM based differential conductance pair (Fig. 1c) was proposed to implement ternary DNN weights with two binary SOT-MRAM devices (anti-parallel AP and parallel P). However, despite advances in quantization-aware training, ternary weights are not sufficient to recover the floating-point (fp.) accuracy on popular datasets such as CIFAR-10. To verify the required number of weight levels, a ResNet-20 DNN was quantized to a low number of weight levels by using an advanced quantization-aware training methodology described in [5]. Input activations and output accumulations were quantized to DAC and ADC compatible levels (64 levels). As shown in Fig. 3, the fp. accuracy is retrieved for 9 weight levels. For binary memories, this can be achieved at circuit level by utilizing multiple array columns for a single weight, incurring a severe area and energy efficiency penalty.

We propose a new approach at device level to reach multi-level conductance SOT-MRAM without additional array and little peripheral overhead. The concept exploits multiple MTJs on a single SOT track and selective VCMA-base writing. Since the front-end of line (FEOL) dominates the total cell area [6], FEOL area is utilized without introducing any additional terminals, or further increasing the cell area (Fig. 4). In the following, we demonstrate single-cell 2- and 4-pillar SOT-MRAM devices with electrical control of the 3- and 5-conductance values, reaching 9 levels in a differential conductance pair for the 4-pillar device.

Experimental demonstration: Fig. 5 shows a TEM micrograph of a representative 4-pillar SOT-MRAM device integrated on 300 mm wafers using CMOS-compatible processes. The devices considered in this study have a MTJ pillar CD of 80nm and pitch of 280nm. The MTJs consists of a perpendicularly-magnetized (PMA) stack in a top-pinned configuration: SOT track/composite free layer MgO(RA2000Ω·μm^2)/CoFeB/SAF and is annealed at 300°C. As sketched in Fig. 6, the in-plane write current is applied across the SOT-track while the device resistance is read out between the SOT-track and common top electrode (TE).

We illustrate in Fig. 7 the device behaviour as a function of applied SOT current (I_{SOT}) pulses of 1ns duration. We observe five different conductance values *G*, as expected from the different MTJ combinations (4AP, 3AP/1P, 2AP,2P, 1AP/3P, 4P). The selectivity of different levels is attributed to the VCMA effect (Voltage Control of Magnetic Anisotropy, Fig.2b, [7,8]). Each MTJ experiences a different voltage drop across the SOT-track (Fig. 6), hence a different VCMA effect, resulting in the lowering of each MTJ switching barrier linearly with its position on the SOT track. As a result, each combination of MTJ states can be selectively obtained by applying the appropriate SOT switching current.

We further confirm the role of VCMA using two- and four-MTJs on one SOT-track device configuration (Fig. 8). To get better insights on the selectivity of each conductance level, we quantified the switching probability distributions (P_{SW}) as a function of I_{SOT} by performing bit-error rate (BER) measurements over 1500 events per I_{SOT} step. This is illustrated for the AP to P transition in Fig. 9 for 2- and 4- pillar devices where good separation of each level is achieved. We report in Fig. 10 the median of the required write current I_{SOT} of each level as function of the MTJ position, measured over 80 devices for both 2- and 4-pillar devices. For each conductance level, the required write current linearly scales with the location of the MTJ on the SOT-track. This confirms that the switching selectivity originates from a position-dependent VCMA effect, as the drop of voltage also linearly varies with the MTJ position.

We characterized more extensively the reliability of the switching selectivity for the 4-pillar case (to achieve 9 levels to obtain the fp. accuracy, Fig. 3) by repeating the BER experiment from Fig. 9 on 80 devices. The maximum P_{SW} for all five conductance levels from all measured devices are summarized in Fig. 11. The two extreme states (4AP and 4P) are easily reached with ~100% P_{SW}. For the intermediate states (3AP/1P, 2AP,2P, 1AP/ 3P), P_{SW} ranges between 50% and 80% due to the overlap between the different distributions (Fig. 9). The mean P_{SW} across all five conductance levels across all devices is 74%.

Top-down algorithm study: In Fig. 13, the tolerance of DNNs against these write errors is explored by training and testing a ResNet-20 (Fig. 12) on CIFAR-10 with different BERs. To maintain a test accuracy above 92%, a BER of 1e-3 is required. In Fig. 14, we show the number of write attempts, determined by the device switching probability P_{SW} and required BER. For the measured SOT-MRAM devices with a mean P_{SW} of 74%, 5 write attempts are needed to reach the BER of 1e-3.

The second metric for the developed devices is conductance variation, resulting in DNN weight noise. Therefore, in Fig. 15, we explore the tolerance of DNNs against weight noise by training and testing a ResNet-20 (quantized to 9 weight levels) with weight noise. To maintain a test accuracy of 92%, a weight noise of 10% can be tolerated. Experimental σ_w is obtained from the five measured conductance distributions, as plotted in Fig. 16. In Fig. 17, we report the resulting weight noise for all 9 weight levels. We find that it ranges from 4.8% to 8.2%, which is lower than the targeted 10%. As summarized in Table 1, the specifications from an algorithm point of view are all well satisfied with the proposed multi-pillar SOT-MRAM devices.

Conclusions: Using the voltage-drop VCMA-assist, multi-pillar SOT-MRAM with high selectivity is demonstrated for the first time to implement multi-level DNN weights for AiMC. Devices with 4 pillars on one SOT-track are used to implement DNN weights with 9 levels and as a result retrieve the fp. accuracy of 92% on CIFAR-10 with a ResNet-20. The required BER of 1e-3 is achieved with the developed devices with five write attempts. Finally, the measured weight noise is within the algorithm-driven specification of 10%. Noteworthy, each of these metrics (selectivity, density, efficiency) will be significantly improved by on-going material and integration efforts, opening appealing perspectives for using SOT-MRAM in future AiMC hardware.

Acknowledgments: This research is conducted within the imec IIAP entitled "Machine Learning".

References: [1] S. Cosemans et al., IEDM, pp. 22.2.1-22.2.4 (2019)
[2] J. Doevenspeck et al., VLSI, (2020)
[3] K. Garello et al., VLSI, p. 81 (2018)
[4] K. Garello et al., VLSI, JFS4-5 (2019)
[5] B. Verhoef et al., arXiv:1912.09356 (2019)
[6] M. Gupta et al., IEDM (2020)
[7] H. Yoda IEDM (2016)
[8] Y.C. Wu et al., VLSI (2020)
[9] K. He et al., CVPR, p. 770 (2015)

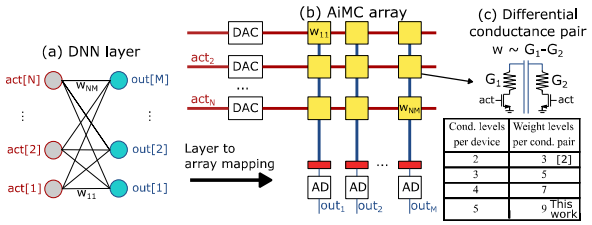


Fig.1: ACIM concept: DNN layer to analog matrix-vector-multiply (MVM) mapping. The matrix-vector product can be calculated in the analog domain using Ohm's and Kirchhoff's law. Each weight is encoded by a differential conductance pair.

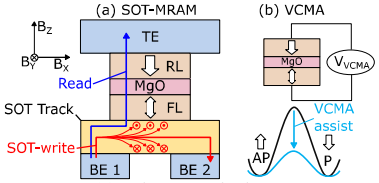


Fig. 2: (a) Three-terminal SOT-MRAM device with decoupled read and write path. The SOT-track converts an in-plane current into a spin-current perpendicular to the free layer (FL). (b) Voltage Control of Magnetic Anisotropy (VCMA) effect

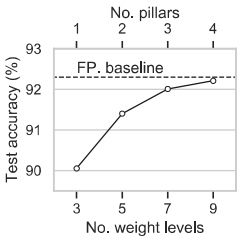


Fig. 3: No. weight levels vs. test accuracy on CIFAR-10 with a quantized ResNet-20. No. of input act. levels = 64, no. of output acc. levels = 64.

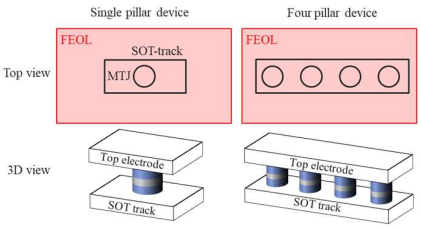


Fig. 4: 2D top view and 3D sketch of a standard and multi-pillar SOT-MRAM device.

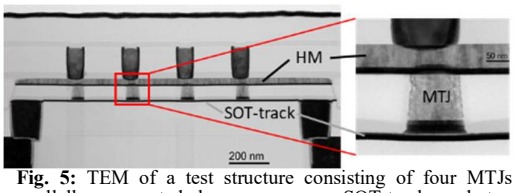


Fig. 5: TEM of a test structure consisting of four MTJs parallelly connected by a common SOT-track and top electrode. MTJs are a perpendicularly-magnetized (PMA) stack in a top-pinned configuration: SOT track/composite free layer/MgO(RA2000Ω·μm^2)/CoFeB/SAF annealed at 300°C.

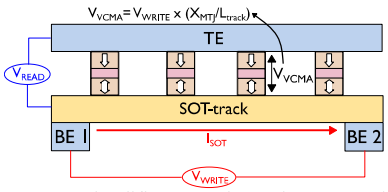


Fig. 6: Simplified experimental setup and illustration of pillar dependent VCMA assist determined by pillar position X_{MTJ} on the track.

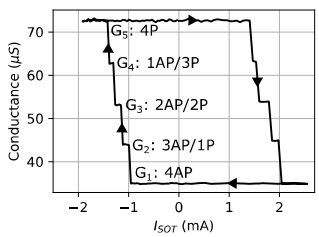


Fig. 7: Conductance switching loop as function of applied SOT current pulses with pulse width τ_p = 1ns. All five conductance levels can be observed.

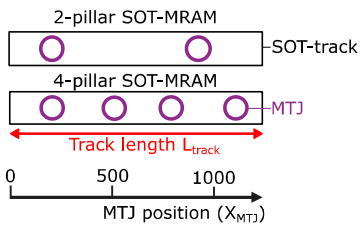


Fig. 8: Physical location X_{MTJ} of MTJ pillars on a SOT-track with two and four pillars. SOT track length L_{track} = 1.18μm.

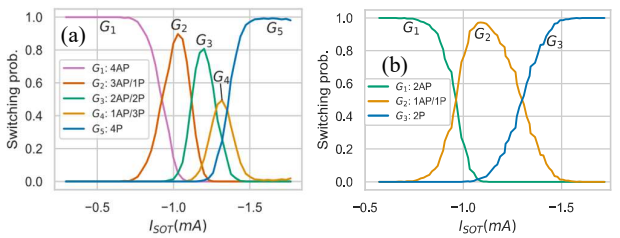


Fig 9: Switching probabilities from BER experiment starting from all devices in the AP state for (a) 4 MTJs on 1 SOT track and (b) 2 MTJs on 1 SOT track. Each switching prob. is obtained from 1500 cycles and pulse width of τ_p = 1ns. In-plane field B_X = -9mT.

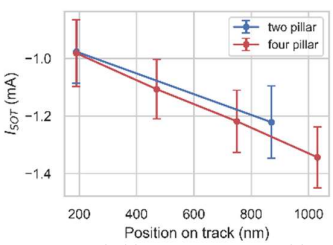


Fig. 10: Switching current vs. position on the track for multiple two-pillar and four-pillar SOT-MRAM devices.

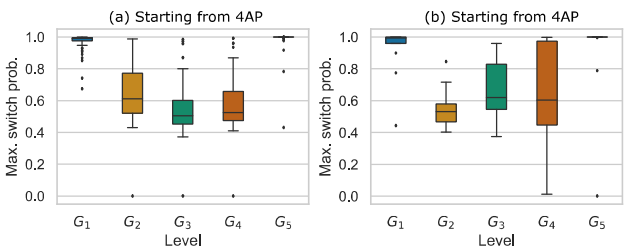


Fig. 11: Boxplot of maximum switching probability over 80 devices across a 300mm wafer. (a) For all devices starting in the AP state and (b) for all devices starting in the P state.

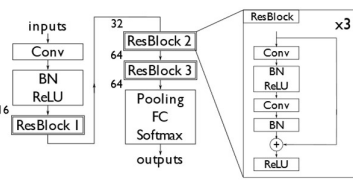


Fig. 12: Quantized ResNet-20 [9] trained and tested on the CIFAR-10 dataset. The quantization-aware training methodology from [5] was used and includes gradual quantization, learned scale quantization and network distillation.

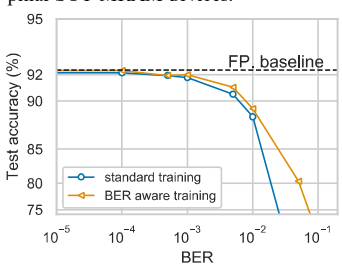


Fig. 13: Test accuracy on CIFAR-10 for different amounts of BER applied on 9 weight levels. The baseline accuracy of 92% is retained below 1e-3 BER.

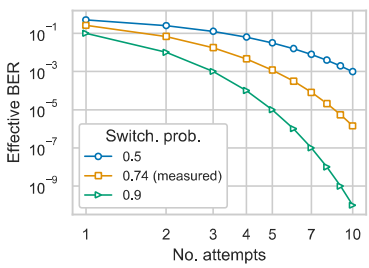


Fig. 14: Effective BER vs. number of write attempts. Measured switching probability of 74% is obtained as the mean of the switching probabilities measured on all devices.

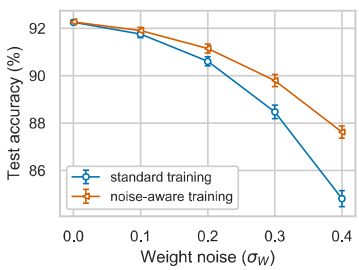


Fig. 15: Test accuracy on CIFAR-10 for different amounts of weight noise σ_W applied on 9 weight levels. Tested for 100 epochs.

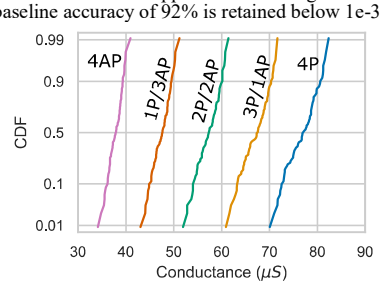
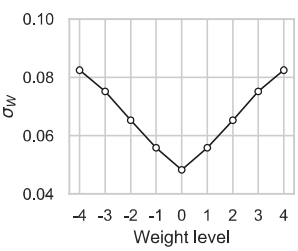


Fig. 16: Measured conductance distributions from 80 multi-pillar SOT-MRAM devices switched to the five different MTJ state combinations.

Weight	G1	G2
-4	4AP	4P
-3	4AP	1AP/3P
-2	4AP	2AP/2P
-1	4AP	3AP/1P
0	4AP	4AP
1	3AP/1P	4AP
2	2AP/2P	4AP
3	1AP/3P	4AP
4	4P	4AP

Fig. 17: Weight noise σ_W obtained for each weight level implemented by two measured conductance distributions G1 and G2 shown in Fig. 16.



Specification for 92%	Target	Achieved
No. Weight levels	9	9 with 4 pillars
BER	1e-3	1e-3 after 5 attempts
Weight noise	10%	4.8% - 8.2%

Table I: Target specifications to achieve the baseline accuracy (92%) on CIFAR-10 with a quantized ResNet-20 and achieved values for the presented multi-pillar SOT-MRAM devices in this work.