2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) | 978-1-6654-9772-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/VLSITechnologyandCir46769.2022.9830153

A 22nm 128-kb MRAM Row/Column-Parallel In-Memory Computing Macro with Memory-Resistance Boosting and Multi-Column ADC Readout

Peter Deaville, Bonan Zhang, and Naveen Verma
Princeton University, Princeton, NJ, USA (deaville@princeton.edu)

Abstract

This paper presents a 128-kb in-memory computing (IMC) macro for fully row/column-parallel matrix-vector multiplication (MVM), implemented using a foundry MRAM in 22nm FD-SOI. Previous IMC in eNVM relied on RRAM with significantly higher resistance and resistance-state contrast than typical in foundry processes [1-3] or where parallelism was substantially reduced [4]. MRAM addresses distinct application requirements (e.g., temperature, radiation). This work advances previous MRAM IMC by improving area-normalized EDP by 60× over [5] and by employing a standard high-density bit cell without additional devices, as in [6]. This is achieved via a readout architecture that performs column-resistance boosting, with integrated auto-zeroing, and conductance-to-current sampling, to simultaneously feed four IMC columns to a single ADC for conversion to 6-b outputs (highest ADC precision among eNVM IMC designs).

IMC Macro Architecture and Overview

Fig. 1 shows the IMC macro, which performs single-shot 128×128 MVM, on 1-b input-vector elements and 4-b matrix elements (higher precision is supported as described later). It comprises a 256(row)x512(col.) MRAM array of 1T1R cells and periphery for single-bit and MVM readout, bit-line/source-line (BL/SL) driving, word-line (WL) driving, and row decoding. Standard foundry MRAM is used, with only BL/SL metal strapping added to reduce parasitic resistances.

For signed binarized multiplication between input-vector and matrix elements, pairs of bit cells in a column store complementary data and are driven by complementary WL data, implementing an XNOR operation. Accumulation is achieved as the sum of parallel conductances across column bit cells between BL-SL. Multi-bit input-vectors are supported by bit-serial processing of parallel input-vector elements over multiple cycles. Multi-bit matrix elements are supported by mapping bits to parallel columns, where four columns feed one readout channel, which performs binary weighting, summation, and 6-b digitization. This eliminates area-consuming column-multiplexing circuitry and maximizes column parallelism.

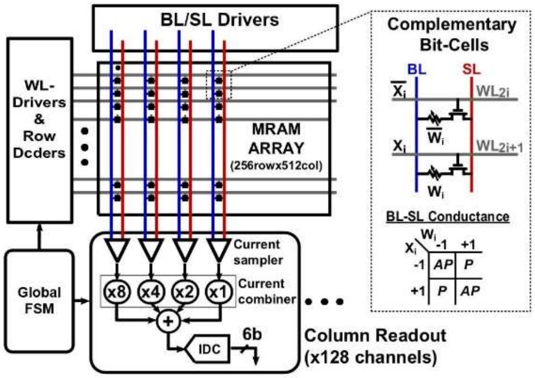


Fig. 1. Demonstrated MRAM IMC macro.

Readout Channel Details

The readout channel consists of: (1) four current-sampling blocks, which convert column conductances to sampled currents; (2) a current-combiner block, which binary-weights and sums the four sampled currents, to feed a single digitizer; and (3) a 6-b current-to-digital converter (IDC), based on a SAR architecture. The 128 readout channels share a single global FSM, which controls auto-zeroing phases in the current-sampling blocks and bit-cycling phases in the SAR IDCs, and which selects between MVM and memory-data readout modes.

Fig. 2 shows details of the current-sampling block during conductance-to-current conversion. To achieve high sensitivity, it first employs auto-zeroing (AZ), where a fixed input reference conductance G_{AZ} is generated in-situ by the MRAM array, and an MVM output conductance G_{MVM} is then processed as a relative change in conductance. AZ is described later. Conversion from G_{MVM} to current involves four stages: (1) cascode; (2) amp; (3) feedback; and (4) sampler.

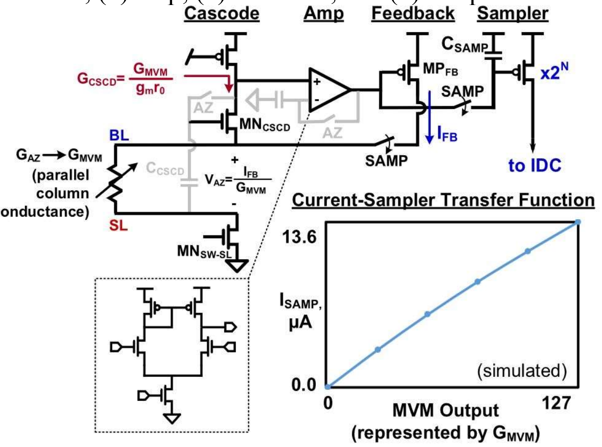

Cascode
Amp
Feedback
Sampler
$G_{CSCD}=\frac{G_{MVM}}{g_m r_0}$
AZ
MN_{CSCD}
AZ
+
-
MP_{FB}
C_{SAMP}
x2^N
SAMP
I_{FB}
SAMP
to IDC
BL
$G_{AZ}\rightarrow G_{MVM}$ (parallel column conductance)
C_{CSCD}
+
$V_{AZ}=\frac{I_{FB}}{G_{MVM}}$
-
SL
$MN_{SW\text{-}SL}$
Current-Sampler Transfer Function
13.6
I_{SAMP}, μA
0.0
0
127
(simulated)
MVM Output (represented by G_{MVM})

Fig. 2. Current-sampling block for converting column BL-SL conductance to a sampled current.

The saturation-biased cascode device (MN_{CSCD}) has the property that the resistance (conductance) at its source is amplified (attenuated) when seen at its drain, by a factor equal to the MOSFET intrinsic gain ($g_m r_0 \approx 20$). Thus, amplification is achieved without separate current-consuming branch, reducing the current required to generate sensing voltages from small (large) parallel resistances (conductances). Following this, the differential-to-single-ended amp provides additional gain (~10). This enhances feedback by driving a device (MP_{FB}), whose current I_{FB} through G_{MVM} works to restore the AZ biasing condition at the cascode's source node: $V_{AZ}=I_{FB}/G_{MVM}$. Thus, I_{FB} in proportion to G_{MVM} is generated, and sampled via a capacitor C_{SAMP} at the gate of a binary-weighted current-sampling device, which then feeds the IDC.

Fig. 3 shows details of the AZ operation, which generates a fixed reference conductance G_{AZ} to establish bias conditions for the current-sampling block. G_{AZ} generation exploits the

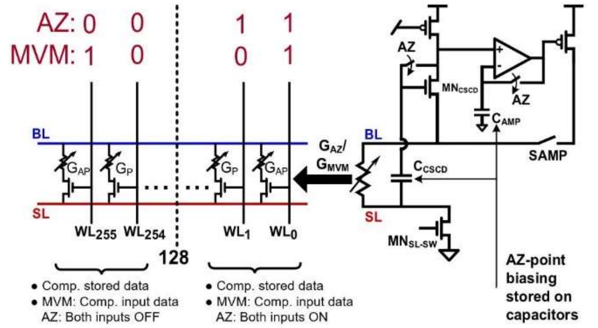


Fig. 3. Current-sampling block auto-zeroing (AZ) operation.

complementary data storage in column bit cells. Specifically, during MVM compute, complementary WL driving and bit-cell data storage over 256 MRAM rows means that G_{MVM} ranges from $128 \times G_P$ to $128 \times G_{AP}$ (where G_{AP}/G_P are the MRAM anti-/parallel-state conductances). However, during AZ, WL drivers perform non-complementary driving over the first 128 rows, yielding a fixed conductance of $64 \times G_{AP} + 64 \times G_P$, at the midpoint of the G_{MVM} range. The resulting G_{AZ} is used to self-bias stages of the current-sampling block, enhancing channel sensitivity and mitigating the effects of channel-to-channel offsets. As shown, the cascode stage employs negative-feedback biasing on a gate capacitor C_{CSCD} tied to the SL. In addition to input-offset cancellation, this helps cancel effects of SL voltage variation from data-dependent I_{FB} through the SL-switch ($MN_{SL\text{-}SW}$) resistance (prominent when $MN_{SL\text{-}SW}$ is small to minimize area). The amp stage employs negative-feedback biasing on a capacitor C_{AMP}.

The current combiner takes the binary-weighted current outputs from four current-sampling blocks and provides these to the $V_{G\text{-}DAC}$ node of the SAR IDC. The sampled currents are provided through cascode PMOS's to shield against parasitic $V_{G\text{-}DAC}$ to C_{SAMP} coupling (from $C_{D\text{-}G}$). The SAR IDC employs a conductance-feedback DAC, implemented by binary weighted NMOS's with supply-tunable drivers, which provide conductance trimming for IDC gain control. The DAC's code-dependent conductance sets the settling time constant and thus the max readout frequency, as shown in the bottom waveforms.

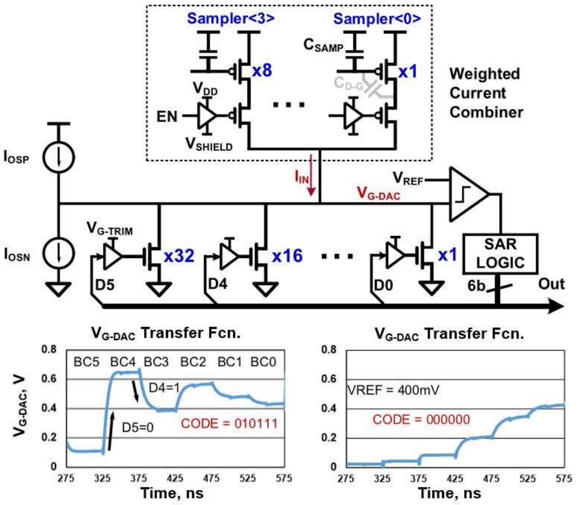


Fig. 4 shows details of the current-combiner and IDC blocks.

Chip Results and NN Demo

Fig. 5 (left) shows chip-measured IMC column transfer functions in the form of heatmaps. Data is shown for both 128 complementary bit cells (full 256-row parallelism) and 64 complementary bit cells (128-row parallelism). A fully-mapped 6-layer conv. neural network (CNN) is demonstrated for CIFAR-10/100 image classification. Fig. 5 (right) shows ideal- and chip-measured average transfer functions for tiled MVM computations at each layer. At lower-right is ideal and chip-measured pre-activation data, showing raw MVM results.

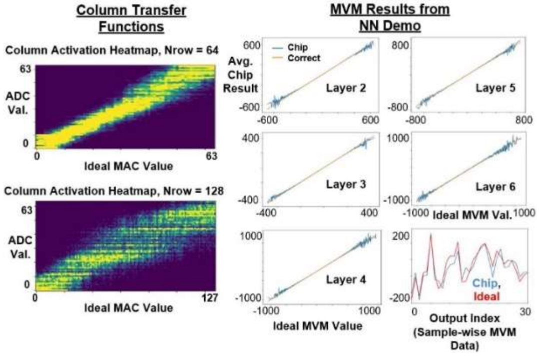
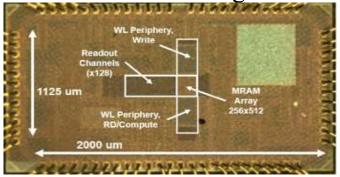
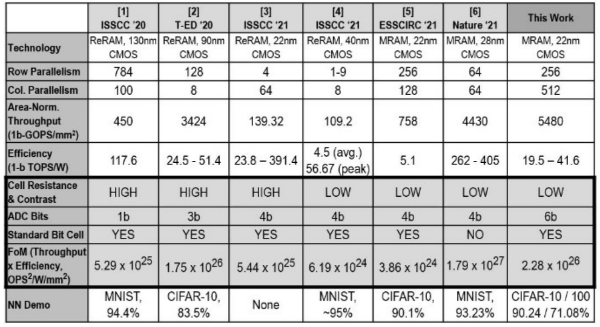


Fig. 5. Heatmaps of the measured column transfer function (left), and MVM data from CNN CIFAR-10 demo (right).

Fig. 6 summarizes the results and comparison with previous work. The presented macro achieves high energy efficiency (41.6 1b-TOPS/W), area-norm. throughput (5480 1b-GOPS/mm^2), and output ADC resolution, without modifying the bit-cell circuit (as in [6]) or the bit-cell resistance levels (as in [1-3]), relative to typical foundry technologies (energy range corresponds to configurable biasing points of current-sampling block). High application performance is verified for CIFAR-10/100 classification, mapping all conv. layers to the chip, with accuracy of 90.24%/71.08% matching ideal SW computation.





	[1] ISSCC '20	[2] T-ED '20	[3] ISSCC '21	[4] ISSCC '21	[5] ESSCIRC '21	[6] Nature '21	This Work
Technology	ReRAM, 130nm CMOS	ReRAM, 90nm CMOS	ReRAM, 22nm CMOS	ReRAM, 40nm CMOS	MRAM, 22nm CMOS	MRAM, 28nm CMOS	MRAM, 22nm CMOS
Row Parallelism	784	128	4	1-9	256	64	256
Col. Parallelism	100	8	64	8	128	64	512
Area-Norm. Throughput (1b-GOPS/mm^2)	450	3424	139.32	109.2	758	4430	5480
Efficiency (1-b TOPS/W)	117.6	24.5 - 51.4	23.8 – 391.4	4.5 (avg.) 56.67 (peak)	5.1	262 - 405	19.5 – 41.6
Cell Resistance & Contrast	HIGH	HIGH	HIGH	LOW	LOW	LOW	LOW
ADC Bits	1b	3b	4b	4b	4b	4b	6b
Standard Bit Cell	YES	YES	YES	YES	YES	NO	YES
FoM (Throughput x Efficiency, OPS^2/W/mm^2)	5.29×10^{25}	1.75×10^{26}	5.44×10^{25}	6.19×10^{24}	3.86×10^{24}	1.79×10^{27}	2.28×10^{26}
NN Demo	MNIST, 94.4%	CIFAR-10, 83.5%	None	MNIST, ~95%	CIFAR-10, 90.1%	MNIST, 93.23%	CIFAR-10 / 100 90.24 / 71.08%

Fig. 6. Die photo and comparison table, with neural-network demonstration summary.

Acknowledgements: Research sponsored by AFRL and DARPA, agreement no. FA8650-18-2-7866.

References: [1] Q. Liu et al., ISSCC, 2020 [2] S. Yin, et al., IEEE T-ED, Oct. 2020 [3] C.-X. Xue et al., ISSCC, 2021 [4] J.-H. Yoon, et al., ISSCC, 2021 [5] P. Deaville, et al., ESSCIRC, 2021 [6] S. Jung et al., Nature, 2022