TAU: A 28nm 120M Weights×TOPs/W/mm² PCM-Based MVM Accelerator featuring Concurrent Datastream and Dynamic-Steering CCO-ADCs

R. Zurla[†]*, A. Lico[‡‖]*, A. Antolini[‡]*, E. Calvetti[§], L. Roverselli[¶], L. Croce[§], R. Vignali[¶], F. Zavalloni[‡], L. Greco[‡], B. Bozkurt[¶], A. Cabrini[¶], E. Franchi Scarselli[‡], M. Pasotti[§]
[†] STMicroelectronics, Pavia, Italy; [‡] University of Bologna, Italy; [‖] Now with Chips-IT, Italy;
[§] STMicroelectronics, Agrate Brianza, Italy; [¶] University of Pavia, Italy. *Equal contribution.*

***Abstract*—This paper presents TAU, an Analog in-Memory Computing core (AiMC), that has been designed for Matrix-Vector-Multiplication hardware acceleration. TAU has been manufactured in a 28 nm FDSOI CMOS technology exploiting high-density phase-change memory cells to store 2M weights used for the implementation of deep neural networks with up to 8 (512×512)-layers. TAU core features a concurrent datastream (thanks to its τ-shaped digital architecture running with a 500MHz clock) together with 512 dynamically-steered CCO-based ADCs (running at a maximum frequency of 8 GHz) which have been optimized to achieve an highly-linear response as required by high-accuracy analog computation. These two key characteristics of TAU test-chip, together with an improved low-power readout scheme, allowed obtaining a FoM, calculated in terms of Weights×TOPs/W/mm^2, in the order of 120M with a normalized computation error of 0.86% which represents an important advancement with respect to the state-of-the-art.**

***Index Terms*—Analog In-Memory Computing (AiMC), Phase-Change Memory (PCM), concurrent datastream, dynamic-steering CCO.**

I. Introduction

In-memory computing (IMC) has emerged as an effective approach to address the limitations of Von Neumann architectures especially for data-intensive applications such as deep neural network (DNN) inference. Analog IMC (AiMC) systems perform matrix-vector multiplications (MVMs) directly within memory arrays by mapping weights to cell conductances and exploiting Ohm's and Kirchhoff's laws. This enables high parallelism and significantly reduces data transfer, improving both energy efficiency and throughput.

This paper presents the TAU test-chip, an AiMC core based on Ge-rich GST Phase-Change Memory (PCM) cells. TAU core has been designed and fabricated in a 28nm FDSOI CMOS STMicroelectronics technology introducing Dynamic-steering CCO-based ADCs (to overcome intrinsic limitation of CCOs at high currents which causes severe non-linearity adversely affecting MVM accuracy) significantly improving the computation energy efficiency by means of an enhanced low-voltage Bit-line Voltage Regulator (BLVR) which enables low power readout on the PCM memory array. Moreover,

This work is supported in part by NeuroSoC (G.A. 101070634) project under the European Union Horizon research and innovation program.

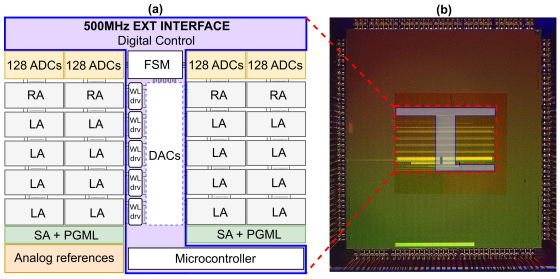

Fig. 1. (a) Architecture of the TAU AiMC core showing the internal organization of the memory array and of the analog/digital circuitry. The name of the AIMC core refers to the τ-shaped layout of digital circuitry. (b) Micrograph of TAU testchip.

TAU core architecture increases the system bandwidth through a dedicated (τ-shaped) digital interface which enables concurrent execution of inputs and outputs stream with MVM computation. As a result, TAU core achieves a storage-energy efficiency per area of 120M Weights×TOPs/W/mm^2, thus significantly advancing the state-of-the-art.

This paper is organized as follows: in Sec. II the architecture of the AiMC core is described along with the pipelined MVM computation; Sec. III presents the main features of the analog circuitry involved in MVM computations; Sec. IV demonstrates the experimental results and concludes the paper.

II. AiMC Architecture and MVM Computation

A. Core Architecture

TAU core is a PCM-based MVM computation hardware accelerator. The architecture, shown in Fig. 1(a), integrates a computational core with dedicated digital and analog circuits adopting the same inputs and weights encodings (as well as drift and temperature compensation techniques) presented in [1]. More specifically, TAU exploits a tile-based organization composed of multiple Local Arrays (LAs), arranged as a 4×4 grid of 256k-cell arrays, together with Reserved Arrays (RAs) used for reference cells, trimming data, and storage of parameters. The system supports MVM operations up to 512×512, with signed 8-bit input activations and up to 11-bit signed outputs, while weights are stored as analog conductance

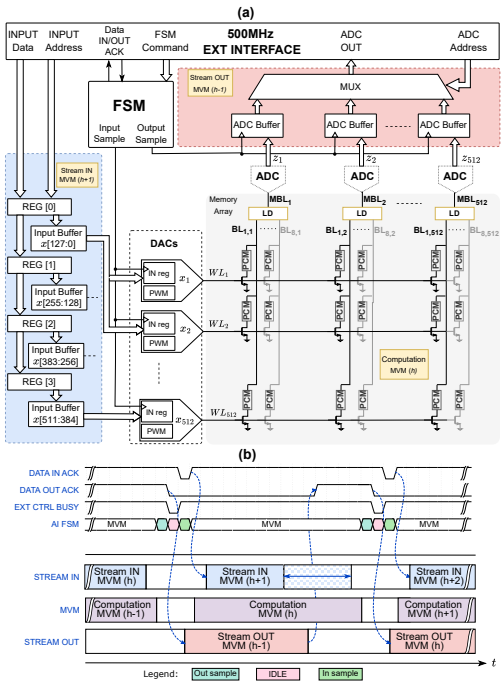

Fig. 2. (a) Block diagram of TAU architecture handling Input/Output datastream and MVM computation. (b) Timing diagram of the concurrent datastream which allows simultaneous streaming of both inputs and outputs during MVM computation.

levels in PCM devices exploiting an analog programming procedure [2].

The core includes a dedicated 32-bit microcontroller (μC), which operates using on-chip SRAM-based instruction and data memories (16 KB Instruction RAM and 8 KB Data RAM), and acts as the master controller of the core, in charge of handling weight programming, memory verification and drift compensation. The 500MHz τ-shaped digital domain is responsible for runtime operations such as input and output streaming as well as execution control.

B. AiMC MVM Computation

The MVM computation is executed in the analog domain. As shown in Fig. 2(a), the ADCs are connected to the main bit-line (MBL) nodes of the memory array. Each MBL is then connected, via the Layer Decoder (LD), to one of 8 selectable bit-lines (BLs) corresponding to the 8 memory layers, i.e., one of the 8 512×512 weight matrices stored in the LAs. The DACs convert input data (x) into pulse-width-modulated (PWM) signals applied to the word-lines, while the ADCs convert the integrated BL current, $\mathrm{I_{BL}}$, into an 11-bit digital output, denoted as $z[0{:}10]$.

C. Concurrent Datastream

In order to guarantee a continuous-like high-throughput computation (and, hence, improving effective TOPs of the system) a concurrent datastream has been introduced (Fig. 2(a)). The proposed datastream is based on a high-speed digital interface featuring decoupled input and output double-buffered channels. This enables, during the h-th computation, a temporal overlap between consecutive MVMs, thus allowing input loading of the next $(h{+}1)$-th MVM (operated by the Stream IN block) to occur during the h-th computation, while simultaneously providing the results of the previous $(h{-}1)$-th MVM through the Stream OUT interface as it is shwon by the timing diagram in Fig. 2(b). The Stream IN block is composed of a 4-stage pipeline register stack ($\mathrm{REG}[3...0]$), each one controlling a 128, 8-bits input buffer connected to the WL DACs while the Stream OUT block consists in a interleaved multiplexing of the BL ADCs.

III. Dynamic-Steering CCO-Based ADC

The architecture of the proposed linearity-enhanced CCO-based ADC is shown in Fig. 3. The current of the selected BL is split in two branches. A fraction $(1-\beta)I_{\mathrm{BL}}$ (with $\beta = 0.8$) is drawn by the High-Accuracy Bit-Line Voltage Regulator (HA-BLVR), which bias $\mathrm{BL_k}$ (k being the selected layer) at a constant voltage $\mathrm{V_R}$ independent from the $\mathrm{I_{BL}}$ amplitude. The remaining portion, $\beta\mathrm{I_{BL}}$, is drawn by a programmable Current Dump Path (CDP) which is in charge of dynamically steer the CCO current in order to enhance the ADC linearity. This is obtained by means of a Current Window Decoder (CWD) which identifies the operating region and generates the CDP control signals (S1, S2) producing the configuration bits $\mathrm{m}[0{:}2]$ that route the CCO output $\mathrm{z}[0]$ to the appropriate stage of the Asynchronous Ripple Counter (ARC). This way, the ADC operates in three distinct current windows, whose boundaries are set at one-half and one-quarter of the maximum bit-line current $\mathrm{I_{BL_{MAX}}}$, as illustrated by the waveforms in Fig. 4. Thanks to this approach, it is possible to limit the CCO oscillation-frequency (by reducing its effective input current) thus ensuring that the circuit operates in its linear (i.e. low-current) frequency–current region.

A. High-Accuracy Bit-Line Voltage Regulator

The HA-BLVR (Fig. 5(a)) forces $\mathrm{V_{MBL}} = \mathrm{V_{MBL'}}$ with a current–voltage mirror (M1-M2-M3-M4) and a differential pair (DP) which equalizes the drain voltages of the all the transistors ($\mathrm{V_A} = \mathrm{V_B}$) thus preventing channel-length modulation, responsible of mirroring errors, and, hence, significantly improving the accuracy of the MBL/MBL' voltage equalization. The mirroring factor of M1–M4 and M2–M3 is set to 1/2 to limit power consumption. A Layer Decoder Replica (LDR), matched and scaled to the actual LD (see Fig. 3), compensates for the voltage drop occurring during computations, ensuring that BL voltage of the selected k-th layer, $\mathrm{BL_k}$, matches $\mathrm{V_{BL'}}$. When L = 0, the op-amp buffer (OA) sets $\mathrm{V_{BL'}} = \mathrm{V_{TR}}$. To compensate for process and mismatch variations, $\mathrm{V_{TR}}$ is trimmed so that the regulated reference $\mathrm{V_{BL_k}}$ matches $\mathrm{V_R}$. Trimming is performed through the Offset Trim Network (OTN), which uses a resistive ladder. The HA-BLVR, enabled by a start-up circuit, regulates the BL voltage within a nominal 10÷30 mV range, while the

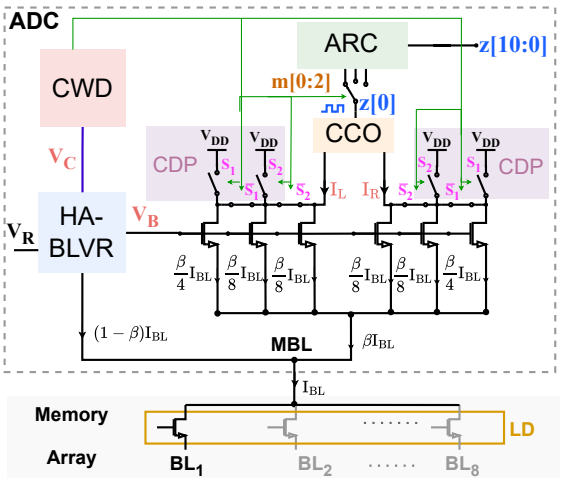

Fig. 3. Architecture of the proposed linearity-enhanced CCO-based ADC featuring dynamic current steering (CWD and CDP blocks) and its interface with the memory array. CCO currents I_L and I_R are a fraction of I_{BL} and are dynamically generated as explained in Fig. 4 depending on CWD block.

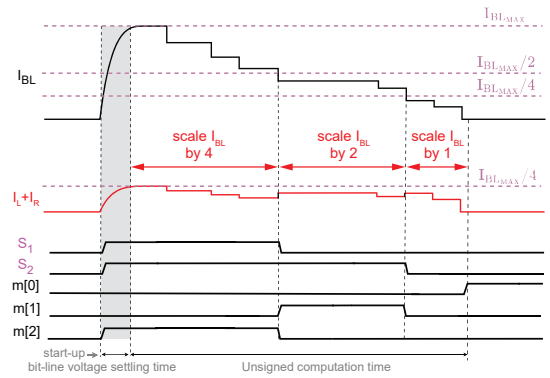

Fig. 4. Example of I_{BL} time evolution showing the operation of the CCO dynamic current steering. CWD control-signals S_1 and S_2 flip as I_{BL} crosses the defined thresholds determining the I_{BL} scaling factors which generates the effective CCO current $I_L + I_R$ (red curve).

OTN offers ± 10 mV offset calibration through 16 steps. This feature allows boosting computation energy efficiency since the largest part of the power consumption comes from the inherent biasing of the cells.

The Linearity-Enhancing Network (LEN), activated by setting L = 1, introduces a resistor which adjusts V_{MBL} and $V_{MBL'}$ so that, for high values of $\mathrm{I_{BL}}$ the CCO oscillation frequency increases thus compensating the residual (since not totally compensated by means of the dynamic steering of the CCO) non-linearity of the ADC as demonstrated in [3].

B. Dynamically-Steered CCO

The CCO (Fig. 5(c)) alternatively sinks a portion of the BL current ($\mathrm{I_L}$ or $\mathrm{I_R}$) to generate the oscillating signal $\mathrm{z}[0]$, whose frequency is proportional to the current. Oscillation is established by two identical capacitors, $\mathrm{C_L}$ and $\mathrm{C_R}$, which are alternately charged until their voltages ($\mathrm{V_L}$ or $\mathrm{V_R}$) reach the switching threshold of the corresponding inverters ($\mathrm{A_L}$ or $\mathrm{A_R}$). Each transition toggles the internal latch signals A and B, effectively digitizing the transfer of a fixed charge packet during each half-cycle. The CWD (Fig. 5(b)) consists of two current comparators each one driven by a scaled copy of the BL current, $\gamma \mathrm{I_{BL}}$, which are compared against two reference levels, $\mathrm{I_{REF}}$ and $\mathrm{I_{REF}}/2$. The parameters γ and $\mathrm{I_{REF}}$ are selected so that the resulting operating-window boundaries match those shown in Fig. 4. Based on these comparisons, the CWD identifies the operating region and generates the control signals $\mathrm{S_1}$ and $\mathrm{S_2}$ used to steer the CDPs (Fig. 3). The signals $\mathrm{S_1}$ and $\mathrm{S_2}$ are sampled by registers clocked by $\mathrm{z}[0]$, ensuring accurate synchronization between current detection and CCO activity. The selection bits $\mathrm{m}[0{:}2]$ configure the ARC (Fig. 5(d)) by routing $\mathrm{z}[0]$ to the corresponding counting stage.

TABLE I
COMPARISON WITH STATE-OF-THE-ART.

	[4] (2026)	[5] (2026)	[6] (2022)	[1] (2025)	[7] (2026)	**This work (2026)**
CMOS Technology	28nm	22nm	14nm	28nm	18A	**28nm**
Memory Technology	SRAM	ReRAM	PCM	PCM	Digital	**PCM**
Weight Unit Cell	6T	1T1R	8T4R	2T2R	12T	**2T2R**
I/O Precision	8/22 bit	8 bit	8/8 bit	8/11 bit	8 bit	**8/11 bit**
Weights Precision	8 bit	8 bit	Analog	Analog	8 bit	**Analog**
MVM Error [%]	NA	NA	1.94	1.7	NA	**0.86**
Throughput [TOPs]	0.125	NA	1	1.7	21.4	**1.7**
Energy Eff. [TOPs/W]	40.74	170.2**	10.5	24*	147	**80***
Core Area [mm^2]	0.072	38.55	0.64	1.7	0.086	**1.34**
# Weights [NW]	32K	12M	65.5K	2M	8K	**2M**
FoM***	17.6M	53M	1.1M	28M	14M	**120M**

* Weights and inputs equal to the 20% of their corresponding maximum.
** Weights and inputs equal to the 10% of their corresponding maximum.
*** [Weights×TOPs/W/mm^2].

IV. Experimental Results

TAU core was implemented in a 28-nm FD-SOI CMOS technology with a power supply V_{DD} of 0.9 V (see micrograph in Fig. 1). Left plot in Fig. 6 shows the measured ADC output counts ($z[0{:}10]$) as a function of the normalized I_{BL} current. The diagram provides the measured count ($z[0{:}10]$) obtained when the ADC operates in four different conditions: no compensation (red circles), with the LEN enabled (green diamonds), with CWD enabled (yellow triangles), and with both compensation techniques active (light-blue squares). The results show that the simultaneous use of both linearity-enahancement approaches significantly improves the ADC current–count characteristic (and, hence, the accuracy of the MVM computations). The measured MVM results (collected with both linearity-enhancement techniques active) as a function of ideal MVM values are shown in the middle-plot of Fig. 6. The corresponding normalized MVM computation error is finally provided in the right diagram of Fig. 6 showing a measured sigma of the normalized error distribution approximately equal to $\sigma = 0.86$. Table I provides a comparison with state-of-the-art IMC cores showing a significantly improved FoM, calculated in terms of Weights×TOPs/W/mm^2, which turned out to be about 120M.

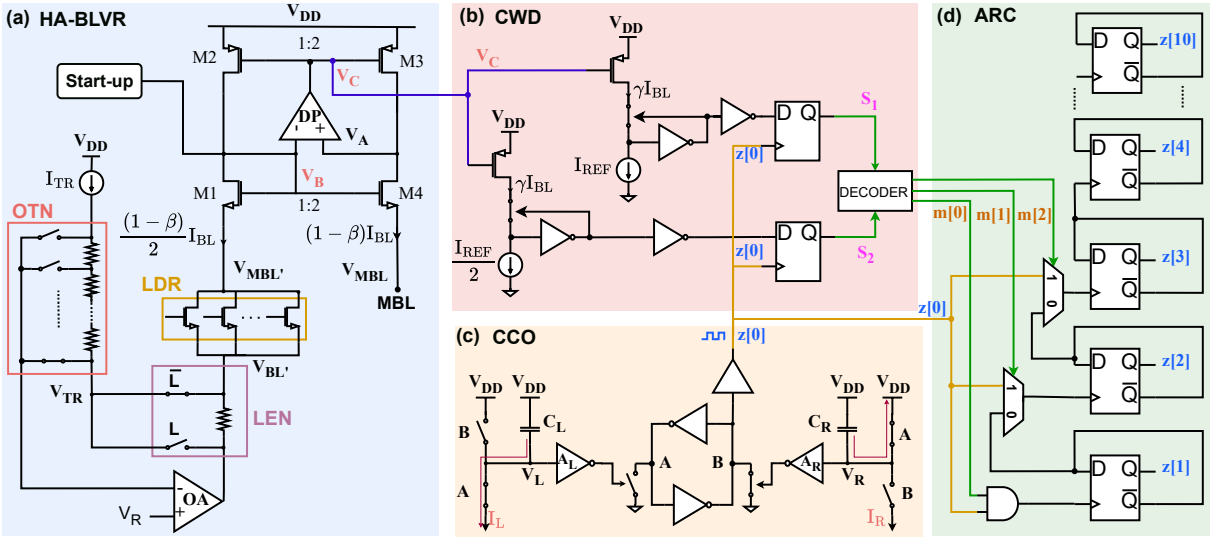

Fig. 5. Schematic of the (a) High-Accuracy Bit-Line Voltage Regulator (HA-BLVR); (b) Current Window Decoder (CWD); (c) Dynamically steered CCO (current $I_L + I_R$ is generated according to signals S_1 and S_2 generated by CWD block that are shown in Fig. 3 and Fig. 4); (d) Asynchronous Ripple Counter (ARC) which stores and provides, at the and of integration, the digital output.

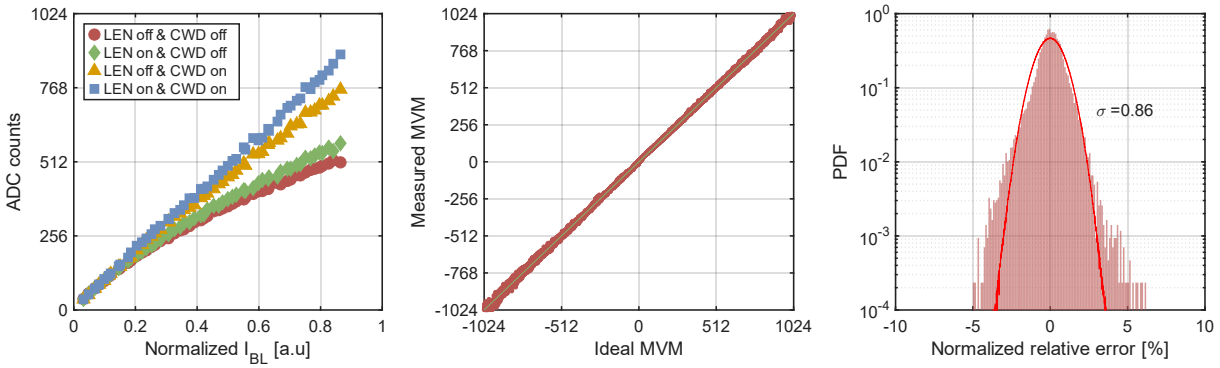

Fig. 6. Measured ADC output $z[0{:}10]$ as a function of normalized I_{BL} collected in four different operating conditions: without any compensation; with LEN enabled ($L = 1$, Fig. 5(a)), with CWD enabled, and with both compensation techniques enabled (left plot). Measured MVM results (middle plot) and its corresponding normalized computation error (right plot).

V. Conclusions

In this paper an AiMC core, named TAU, has been presented. The TAU core has been designed for MVM hardware acceleration. The core features a concurrent datastream (thanks to the τ-shaped 500MHz digital architecture) together with a dynamic-steering CCO-based ADC. These two key characteristics of TAU test-chip allowed obtaining a FoM, calculated in terms of Weights×TOPs/W/mm^2, in the order of 120M and a normalized computation error of 0.86%.

References

[1] M. Pasotti *et al.*, "28 m weights × tops /W/mm^2 pcm-based analog in-memory computing core with 8 512 × 512-weight layers in 28 nm fd-soi cmos," in *2025 IEEE ESSERC*, 2025, pp. 129–132.

[2] A. Antolini *et al.*, "High-precision close-to-analog programming of pcm cells as devices for aimc edge-ai," in *2025 IEEE ESSERC*, 2025, pp. 461–464.

[3] A. Lico *et al.*, "Linearization technique for cco-based adcs for aimc mvm architectures using resistive memory devices," *IEEE Trans. on Circuits and Systems II: Express Briefs*, pp. 1–1, 2026.

[4] L. Feng *et al.*, "A 28nm 106.85tops/w and 77.68tflops/w cim macro with stage-wise-enabled lossless compressors based on sign-bit-embedded transition-counting-lines for edge-ai devices," in *2026 IEEE Int. Solid-State Circuits Conference (ISSCC)*, 2026, pp. 518–519.

[5] H.-H. Hsu *et al.*, "A 22nm 96mb 50.6-to-90.2tflops/w non-linear mlc reram cim macro with high-retention for mamba/transformer/cnn," in *2026 IEEE Int. Solid-State Circuits Conference (ISSCC)*, vol. 69, 2026, pp. 516–518.

[6] R. Khaddam-Aljameh *et al.*, "Hermes-core—a 1.59-tops/mm2 pcm on 14-nm cmos in-memory compute core using 300-ps/lsb linearized cco-based adcs," *IEEE J. of Solid-State Circuits*, vol. 57, no. 4, pp. 1027–1038, 2022.

[7] A. Agarwal *et al.*, "30.9 a 147tops/w, 250tops/mm2, fully synthesizable, digital compute-in-memory accelerator supporting int8×int8 with zero-point quantization in intel 18a technology," in *2026 IEEE Int. Solid-State Circuits Conference (ISSCC)*, vol. 69, 2026, pp. 528–530.