A 1-to-4b 16.8-POPS/W 473-TOPS/mm^2 6T-based In-Memory Computing SRAM in 22nm FD-SOI with Multi-Bit Analog Batch-Normalization

Adrian Kneip, Martin Lefebvre, Julien Verecken and David Bol
ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
{adrian.kneip, martin.lefebvre, julien.verecken, david.bol}@uclouvain.be

***Abstract*—Computing in-memory (CIM) is rapidly becoming an enticing solution to accelerate convolutional neural networks (CNNs) at the edge. Yet, low-precision current-based CIM-SRAMs face severe SNR degradation due to numerous analog non-idealities and high quantization noise when performing analog-to-digital conversion prior to digital batch-normalization (DBN). In this paper, we propose a dual-supply 1-to-4b CIM-SRAM macro in 22nm FD-SOI using 6T foundry bitcells, co-designed with a CIM-aware CNN training framework to overcome these challenges. The macro includes a multi-bit analog BN (ABN) unit combined with self-calibrating dual-phase sense-amplifiers (SCDP-SAs). Measurement results show peak 1b-normalized power and area efficiencies of 16.8POPS/W and 473TOPS/mm^2 at 0.4/0.8V supply and 100MHz, surpassing existing low-precision designs.**

***Index Terms*—Compute in-memory (CIM), SRAM, Analog batch-normalization (ABN), ADC, Convolutional neural networks (CNNs), 22nm FD-SOI.**

I. Introduction

NOWADAYS, edge devices face ever-growing processing and communication challenges to execute data-driven AI algorithms, such as convolutional neural networks (CNNs) [1]. The compute in-memory (CIM) paradigm has recently proven to be a valuable solution by performing highly efficient analog dot-product (DP) operations inside SRAM arrays [2], [3]. However, CIM-SRAMs still face several design challenges, as depicted in Fig. 1(a). First, CIM-SRAMs based on high-density 6T bitcells from the foundry suffer from exacerbated variability in deeply scaled bulk CMOS technology, leading to accuracy loss [4]. Novel architectures based on charge [3] or digital [5] computing help to reduce this loss, yet at the expense of area and/or energy efficiency. Alternatively, the FD-SOI technology allows further downscaling at constant accuracy. Second, most conventional CIM-SRAMs focus on DP acceleration and perform remaining CNN operations in the digital domain, after ADC quantization of the DP results [6], [7]. In the case of (digital) batch-normalization (DBN), ADCs with less than 8b precision do not provide enough BN input data diversity and hinder the learning of DBN parameters, resulting in up to 20% accuracy loss on MNIST as shown in Fig. 1(a). The accuracy drop can be reduced by pre-training the DBN weights with full-precision (32b)

This work is supported by the Fonds de la Recherche Scientifique - FNRS, Belgium, under CDR Grant J.0014.20 and A. Kneip's Research Fellowship.

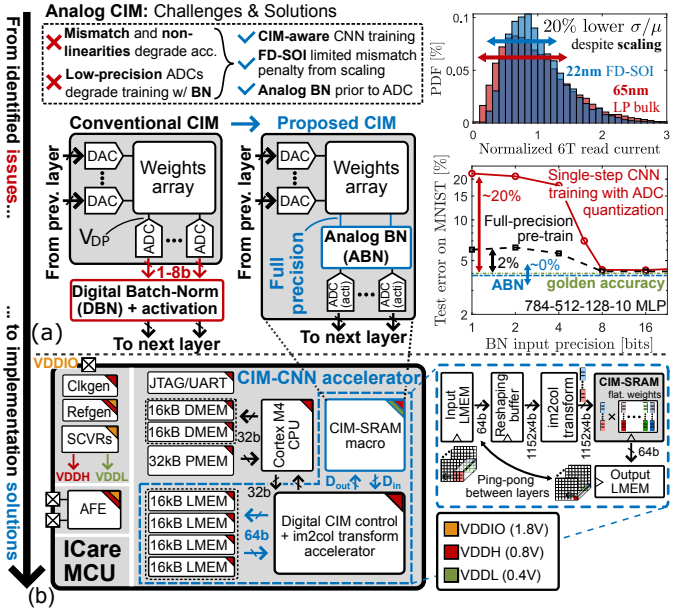


Fig. 1. (a) Design challenges in analog CIM-SRAMs and possible solutions. (b) Architecture of the ICare MCU, embedding a custom CIM-SRAM macro surrounded by a massively parallel digital controller, with im2col acceleration capability and tightly-coupled local memories (LMEMs).

ADCs, then fine-tuning with the actual low-precision (1-8b) ADC quantization. Still, a 2% accuracy loss remains below the 6b ADC precision compared to golden accuracy, obtained for digital hardware with 32b BN inputs. This puts significant pressure on the ADC, requiring a 6-to-8b design and making it a bottleneck with respect to throughput and energy efficiency [6]. Instead, performing BN in the analog domain prior to ADC quantization allows to bypass the precision bottleneck, thereby recovering golden accuracy with compact 1-to-4b ADCs. Yet, analog BN (ABN) has only been proposed for binary CIM-SRAMs to this day [3].

In this work, we propose a mixed-signal CIM-CNN accelerator based on a 1152×512 dual-supply CIM-SRAM macro in 22nm FD-SOI using high-density 6T-SRAM bitcells and featuring a custom multi-bit ABN unit. The accelerator is embedded within the ICare microcontroller unit (MCU) shown in Fig. 1(b) and features a digital controller with a 1-to-4b configurable datapath and four 16kB local SRAM memories (LMEMs). The controller uses highly parallel pipelined oper-

ESSCIRC 2022- IEEE 48th European Solid State Circuits Conference (ESSCIRC) | 978-1-6654-8494-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ESSCIRC55480.2022.9911348

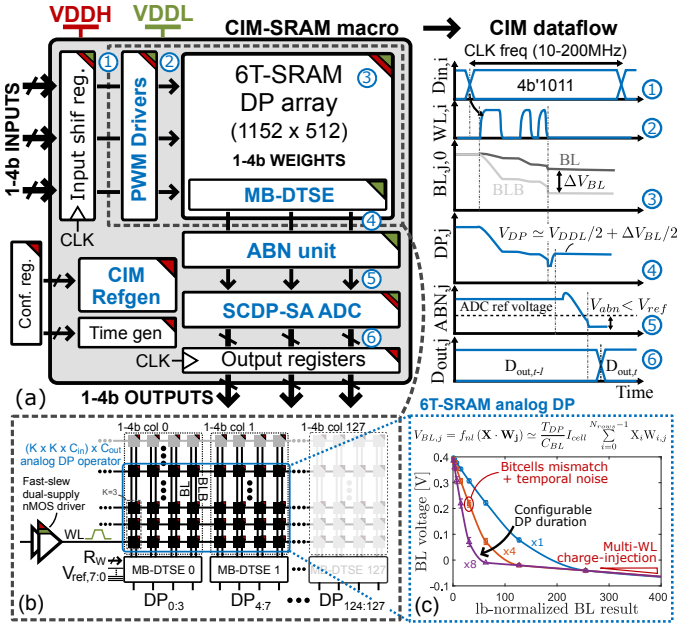


Fig. 2. (a) Top-level architecture of the CIM-SRAM macro, with qualitative depiction of a CIM operation. (b) Zoom on the analog 6T current-based DP array, relying upon differential BL discharge and DTSE conversion. (c) One-sided post-layout transfer function of the DP operation. Non-linearity and variability are included in a custom CIM-aware training framework.

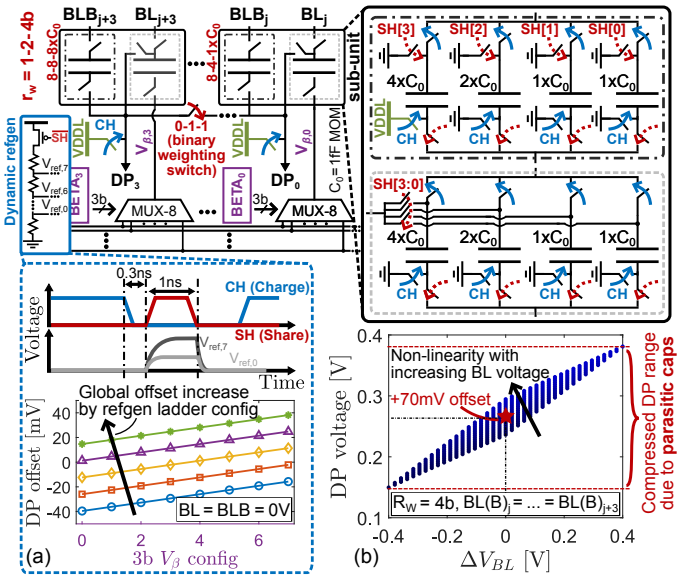


Fig. 3. Architecture of the MB-DTSE unit, consisting of four adjacent DTSE sub-units with tunable total integration capacitance. The chosen weight's precision defines the total capacitance value to perform inter-column binary weighting during the share (SH) phase. (a) MB-DTSE supports DP voltage offsetting with 3b local V_β and global ladder configurability. (b) Post-layout transfer characteristic of the MB-DTSE unit, suffering from non-linearity and dynamic range reduction due to parasitic coupling.

ation to minimize data transfer latency and handles image-to-column (im2col) transforms. Altogether, the CIM-CNN accelerator achieves a peak throughput of 7.4TOPS at 100MHz and a total energy efficiency of 2.8POPS/W, including the controller and LMEM access energy. Considered alone, the CIM-SRAM macro reaches peak 1b-normalized area and energy efficiencies of 473TOPS/mm^2 and 16.8POPS/W, respectively. Finally, a custom CIM-aware CNN training framework is developed to account for analog non-idealities.

The rest of the paper focuses on the CIM-SRAM macro and is organized as follows. Section II presents the overall CIM-SRAM architecture, highlighting its dataflow. Then, Section III focuses on the co-design of the ABN and ADC units within the macro, while Section IV presents measurement results.

II. CIM Macro Architecture

The CIM-SRAM macro architecture is depicted in Fig. 2(a) with its different power domains. CNN weights are flattened on a row-wise basis to perform column-parallel analog DP operations. For convolutional layers (CONV-2D), the macro supports a kernel size of 3×3 and enables data reuse thanks to an input shift register, limiting the amount of LMEM transfers. Moreover, using 0.8V V_{DDH} and 0.4V V_{DDL} voltages respectively for the control signals and the highly parallel datapath reduces the DP energy without significant speed degradation.

As the analog DP operation starts, 4b digital inputs are sequentially applied to the wordlines (WLs) with binary pulse-width-modulated (PWM) encoding. Depending on the weights stored in each column, the V_{DDL}-precharged bitline (BL) then discharges over time using pull-down currents, obtaining the transfer function shown in Fig. 2(c) by time integration. The configurable DP duration T_{DP} not only allows to compensate for PVT variations but also to tune the BL dynamic range to fit the expected distribution of BL results, shaped at train time. Therefore, we add T_{DP} as a custom learning parameter during CNN training to maximize information on the BL. Nevertheless, allowing a full BL swing makes bitcells subject to data corruption events as the BL goes close to 0V [4]. To avoid such events, we underdrive WLs to V_{DDL} with bitcells supplied at V_{DDH}. While this increases the sensitivity to mismatch, Fig. 2(c) shows that the relative 1-σ deviation of the BL voltage stays below an acceptable 10% for the longest T_{DP} configuration in the considered technology.

Once the analog DP finishes, the resulting voltage difference $\Delta V_{BL} = V_{BL} - V_{BLB}$ undergoes a multi-bit differential-to-single-ended (MB-DTSE) conversion before entering the ABN stage and being digitized by the ADC. The MB-DTSE unit shown in Fig. 3(a) combines ideas from [8] and [9] to simultaneously perform DTSE conversion and binary capacitive weighting between blocks of four adjacent columns. It operates in two phases. In the *charge phase*, which takes place during the analog DP operation, DTSE sub-units integrate their connected BL and BLB voltages on 8fF MoM capacitors. In the ensuing *share phase*, the outputs of the sub-units are shorted to perform binary-weighted charge-sharing, obtaining the single-ended DP result. The weights resolution r_w controls the sharing between DP outputs as well as the capacitance in each sub-unit such that the ideal DP voltage yields

$$V_{DP} = \frac{V_{DDL}}{2} + \sum_{i=0}^{r_w-1} \frac{2^i}{2^{r_w}-1} \times \frac{\Delta V_{BL,i}}{2}. \quad (1)$$

In practice, parasitic capacitances compress the effective DP

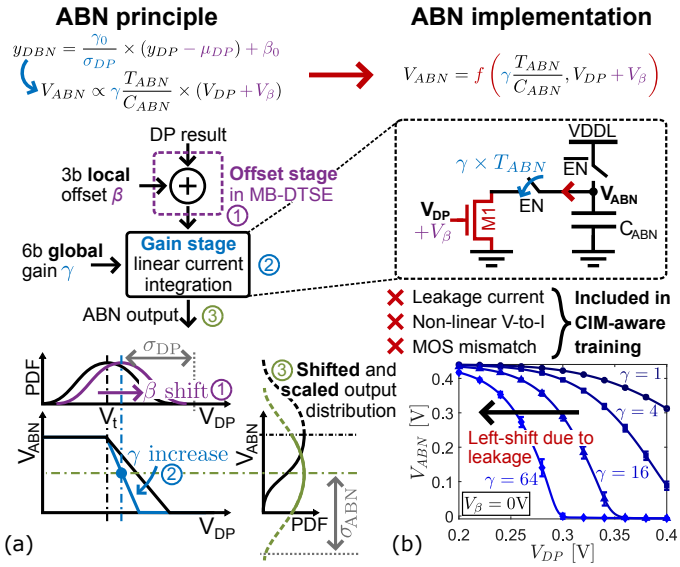


Fig. 4. Architecture of the in-memory ABN unit, with (a) a qualitative depiction of its working principle, with consecutive 3b local shift and 6b global scaling stages, (b) the implementation of the current integration stage, controlled by the shifted DP voltage. Its post-layout transfer function on the bottom-right highlights the main non-idealities to be considered at train time.

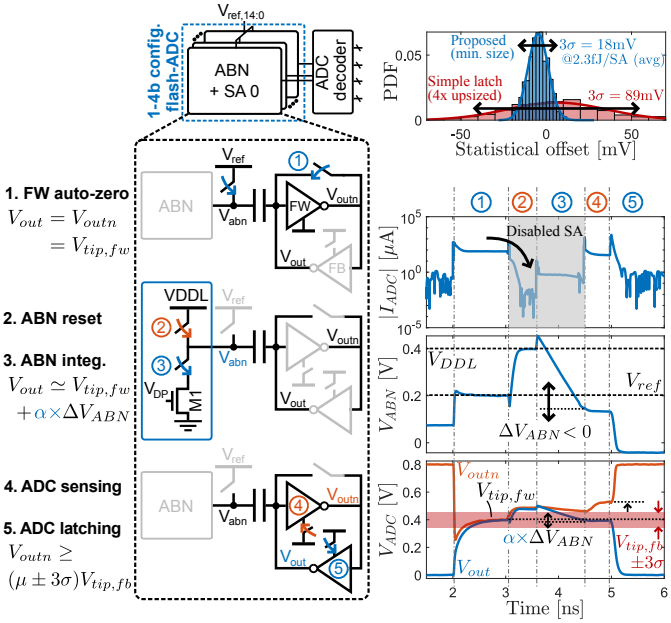


Fig. 5. The 1-to-4b flash ADC, co-integrated with the ABN unit. Self-calibrating dual phase sense-amplifiers (SCDS-SAs) achieve a 18mV 3-σ input offset with a sub-ns 2.3fJ/SA decision. Disabling SAs during ABN operations avoids high short-circuit current while keeping the AZ information.

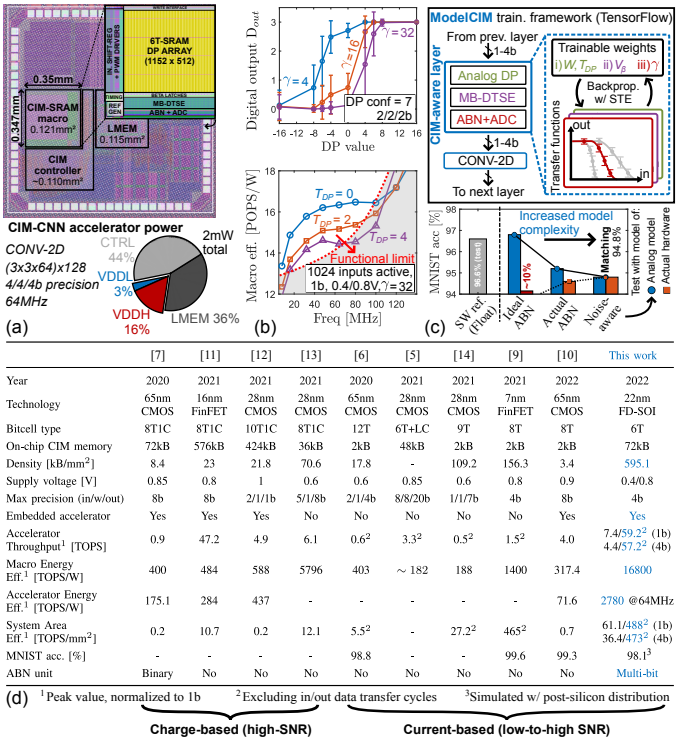
dynamic range and shift the result based on the DP node voltage during the charge phase. Interestingly, we can take advantage of this shift to align the limited DP range with the desired ABN input voltage range by precharging the output node to V_{DDL}. Still, parasitics also introduce additional dependence on V_{BL} in Eq. (1), leading to the non-linearity observed in Fig. 3(c). As the integration capacitance cannot be further increased due to the pitch-matching constraint, we accept this non-linearity and model it in training, thereby limiting the accuracy degradation on MNIST below 1%.

Furthermore, the MB-DTSE unit implements a configurable DP offset, as shown in Fig. 3(b). We control the offset value by applying a configurable V_β voltage on the bottom plate of the BL integration capacitance during the *share* phase. A 3b pitch-matched latch stores the offset configuration per column and selects one of eight reference voltages V_{ref} generated by a resistive ladder as V_β. The ladder is only enabled for the duration of the *share* phase, which avoids significant power overhead during the rest of CIM operations. Moreover, a five-level configuration provides PVT compensation as well as the ability to offset the DP globally.

III. Multi-Bit ABN-ADC Co-Design

The ABN unit intends to shift and scale the distribution of DP voltages to match the behavior of DBN. As qualitatively depicted in Fig. 4(a), we apply these operations in two steps. First, we use the configurable offset of the MB-DTSE to apply a 3b V_β shift per output channel. Then, this shifted DP voltage is used to control the discharge rate of a current integration unit. There, we apply the gain factor γ by globally tuning the integration time on 6b, changing the slope of the ABN response. While ideally linear, this integration stage suffers from MOS transistor non-idealities, altering its transfer function as seen in Fig. 4(b). Namely, M_1 subthreshold leakage limits the maximum ABN gain by transforming the linear scaling effect into a shift of the ABN transfer function towards lower V_{DP} values at high gain γ. Besides, such architecture also faces intrinsic non-linearity and column-to-column mismatch. These effects are critical and have to be considered during the off-line CNN training, as shown later in Section IV.

A dense 1-to-4b flash ADC based on custom self-calibrating dual-phase sense-amplifiers (SCDP-SAs) is connected to the ABN output, requiring four ABN+SA units per column. The 4b output precision is enabled by combining the results of four adjacent columns within the ADC decoder. Similar to the MB-DTSE unit, a resistive ladder dynamically generates the required SA decision voltages V_{ref} for the target output precision. A 2fF auto-zero (AZ) MoM capacitance C_{az} connects the ABN node to the SCDP-SA forward (FW) and feedback (FB) inverters, which are respectively driven by separate control signals. The SCDP-SA combines the advantages of the conventional latch and single-ended AZ topologies, achieving a sub-ns 2.3fJ/SA decision with minimum-sized devices and a 3-σ input-referred offset deviation of 18mV, which is 3.5× lower than a conventional latch built with 4× upsized transistors. The SCDP-SA operates in five phases, as described in Fig. 5. First, the FB inverter is disabled prior to ABN operations to perform FW offset compensation, setting V_{out} to the FW inverter's tipping point. During phases 2 and 3, the ABN operation takes place and both inverters are disabled. In this way, V_{out} follows changes in V_{ABN} through C_{az} coupling, with a reduction factor α related to parasitic charge-sharing. Moreover, disabling inverters during ABN avoids to suffer from significant short-circuit current. After ABN operations, the FW inverter is first enabled to sense the $\alpha \times \Delta V_{ABN}$



	[7]	[11]	[12]	[13]	[6]	[5]	[14]	[9]	[10]	This work
Year	2020	2021	2021	2021	2020	2021	2021	2021	2022	2022
Technology	65nm CMOS	16nm FinFET	28nm CMOS	28nm CMOS	65nm CMOS	28nm CMOS	28nm CMOS	7nm FinFET	65nm CMOS	22nm FD-SOI
Bitcell type	8T1C	8T1C	10T1C	8T1C	12T	6T+LC	9T	8T	8T	6T
On-chip CIM memory	72kB	576kB	424kB	36kB	2kB	48kB	2kB	2kB	2kB	72kB
Density [kB/mm²]	8.4	23	21.8	70.6	17.8	-	109.2	156.3	3.4	595.1
Supply voltage [V]	0.85	0.8	1	0.6	0.6	0.85	0.6	0.8	0.9	0.4/0.8
Max precision (in/w/out)	8b	8b	2/1/1b	5/1/8b	2/1/4b	8/8/20b	1/1/7b	4b	8b	4b
Embedded accelerator	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes
Accelerator Throughput[1] [TOPS]	0.9	47.2	4.9	6.1	0.6[2]	3.3[2]	0.5[2]	1.5[2]	4.0	7.4/59.2[2] (1b) 4.4/57.2[2] (4b)
Macro Energy Eff.[1] [TOPS/W]	400	484	588	5796	403	~ 182	188	1400	317.4	16800
Accelerator Energy Eff.[1] [TOPS/W]	175.1	284	437	-	-	-	-	-	71.6	2780 @64MHz
System Area Eff.[1] [TOPS/mm²]	0.2	10.7	0.2	12.1	5.5[2]	-	27.2[2]	465[2]	0.7	61.1/488[2] (1b) 36.4/473[2] (4b)
MNIST acc. [%]	-	-	-	-	98.8	-	-	99.6	99.3	98.1[3]
ABN unit	Binary	No	No	No	No	No	No	No	No	Multi-bit

(d) [1]Peak value, normalized to 1b [2]Excluding in/out data transfer cycles [3]Simulated w/ post-silicon distribution

Charge-based (high-SNR) Current-based (low-to-high SNR)

Fig. 6. (a) Chip microphotograph, highlighting the CIM-CNN accelerator and showcasing the layout of the CIM-SRAM macro. (b) Measured transfer function and energy efficiency under different conditions, at 0.4/0.8V. (c) CIM-aware training framework, with simulated accuracy improvements after each modelling step. (d) Comparison to the state of the art.

change, such that its output voltage V_{outn} resolves in the correct direction. If the sensed $\alpha \times \Delta V_{ABN}$ is large enough, the change in V_{outn} will overcome the FB inverter's offset deadband and ensure correct decision in the latching phase. As only a few mV is required on $\alpha \times \Delta V_{ABN}$ to overcome the FB deadband, the SCDP-SA offset is much lower than that of a conventional latch.

IV. Measurement Results

The proposed CIM-SRAM macro has been fabricated in the ICare 22nm FD-SOI MCU and occupies a total area of 0.121mm^2 with a 67% array efficiency, as shown in Fig. 6(a). Thanks to its 6T current-based DP architecture and ultra-low voltage datapath, the CIM-SRAM macro achieves 1b-normalized peak energy and area efficiencies of 16.8POPS/W and 473TOPS/mm^2 at 0.4/0.8V and 100MHz with all inputs activated. Importantly, system-level efficiency remains limited by the in/out data transfer through the controller and LMEM accesses, calling for a future holistic optimization of the full CIM-CNN accelerator. In order to limit the accuracy loss due to analog non-idealities, we developed the *ModelCIM* CIM-aware CNN training framework from Fig. 6(c), which relies on statistical hardware models and look-up tables.

Compared to the state of the art in Fig. 6(d), this work achieves the densest design (relative to technology) on top of the highest 1b-normalized energy and area efficiency to date. Furthermore, to the best of our knowledge, this is the first design including an in-memory multi-bit ABN unit. However, this architecture targets low-precision computing and can therefore not properly map more complex tasks with limited memory footprint (e.g., ImageNet), contrary to [10], [11]. Moreover, capacitance-based binary weighted precision and column-parallel flash ADCs are hardly scalable above the 4b precision. These points highlight the well-known SNR-efficiency trade-off in CIM designs, the present work focusing on improving the efficiency boundary at low SNR.

V. Conclusion

In this work, we proposed a novel 72kB CIM-SRAM macro in 22nm FD-SOI with 6T foundry current-based dot-product (DP) operators and a dual-supply strategy enabling ultra-low power datapath. The macro features a multi-bit analog batch-normalization (ABN) unit to overcome the ADC quantization bottleneck in low-precision CIM-based CNNs. Dense 4b flash ADCs based on self-calibrating dual-phase sense-amplifiers (SCDP-SAs) showcase high resilience to input-referred offset with only 2.3fJ per SA decision. Measurement results showcase 1b-normalized peak macro area and energy efficiencies of 473TOPS/mm^2 and 16.8POPS/W respectively, surpassing previous low-precision CIM-SRAM designs to date.

References

[1] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE ISSCC*, 2014, pp. 10–14.

[2] J. Zhang *et al.*, "A Machine-Learning Classifier Implemented in a Standard 6T SRAM Array," in *IEEE VLSI*, 2016, pp. 1–2.

[3] H. Valavi *et al.*, "A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute," *IEEE JSSC*, vol. 54, no. 6, pp. 1789–1799, 2019.

[4] A. Kneip *et al.*, "Impact of Analog Non-Idealities on the Design Space of 6T-SRAM Current-Domain Dot-Product Operators for In-Memory Computing," *IEEE TCAS-I*, vol. 68, no. 5, pp. 1931–1944, 2021.

[5] J.-W. Su *et al.*, "A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips," in *IEEE ISSCC*, 2021, pp. 250–251.

[6] S. Yin *et al.*, "XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks," *IEEE JSSC*, vol. 55, no. 6, pp. 1733–1743, 2020.

[7] H. Jia *et al.*, "A Programmable Heterogeneous Microprocessor Based on Bit-Scalable In-Memory Computing," *IEEE JSSC*, vol. 55, no. 9, pp. 2609–2621, 2020.

[8] M. Lefebvre *et al.*, "A 0.2-to-3.6TOPS/W Programmable Convolutional Imager SoC with In-Sensor Current-Domain Ternary-Weighted MAC Operations for Feature Extraction and Region-of-Interest Detection," in *IEEE ISSCC*, 2021, pp. 118–120.

[9] M. Sinangil *et al.*, "A 7-nm Compute-in-Memory SRAM Macro Supporting Multi-Bit Input, Weight and Output and Achieving 351 TOPS/W and 372.4 GOPS," *IEEE JSSC*, vol. 56, no. 1, pp. 188–198, 2021.

[10] J. Yue *et al.*, "STICKER-IM: A 65 nm Computing-in-Memory NN Processor Using Block-Wise Sparsity Optimization and Inter/Intra-Macro Data Reuse," *IEEE JSSC (Early Access)*, pp. 1–14, 2022.

[11] J. Lee *et al.*, "A Programmable Neural-Network Inference Accelerator Based on Scalable In-Memory Computing," in *IEEE ISSCC*, 2021, pp. 236–238.

[12] S. Yin *et al.*, "PIMCA: A 3.4-Mb Programmable In-Memory Computing Accelerator in 28nm for On-Chip DNN Inference ," in *IEEE VLSI*, 2021, pp. 1–2.

[13] J. Lee *et al.*, "Fully Row/Column-Parallel In-memory Computing SRAM Macro employing Capacitor-based Mixed-signal Computation with 5-b Inputs," in *IEEE VLSI*, 2021, pp. 1–2.

[14] V. Rajanna *et al.*, "SRAM with In-Memory Inference and 90Activity Reduction for Always-On Sensing with 109 TOPS/mm2 and 749-1,459 TOPS/W in 28nm," in *IEEE ESSCIRC*, 2021, pp. 127–130.