2023 IEEE Custom Integrated Circuits Conference (CICC) | 979-8-3503-9948-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/CICC57935.2023.10121312

A 138-TOPS/W Delta-Sigma Modulator-Based Variable-Resolution Activation In-Memory Computing Macro

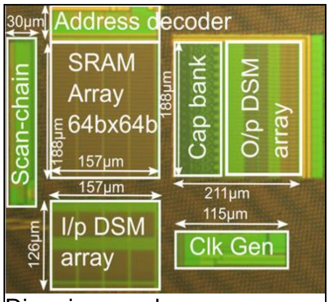
Vasundhara Damodaran, Ziyu Liu, Jae-sun Seo and Arindam Sanyal

Arizona State University, Tempe, AZ 85287, USA

In-memory computing (IMC) is a widely used technique that performs low-precision computations inside memory elements to break the von-Neumann bottleneck in conventional AI/ML hardware. Recently SRAM based IMC [1-4] has gained significant attention due to its high energy efficiency and easy integration with CMOS ICs. A fundamental limitation of SRAM based IMC is nonlinearity in the multiply-and-accumulate (MAC) operation. For large values of MAC result, the proportional large discharge current through SRAM bitline pushes the access transistors into linear region and makes the discharge current, and hence the MAC result, a nonlinear function of bitline voltage [1-3]. Recent works have tried to address this fundamental limitation by applying pulsed input activations [3] and adding capacitors to SRAM bitcell [4-5] for charge-domain computation which has much lower sensitivity to bitline voltage, and hence, higher linearity than current-domain computation in traditional SRAM bitcell. While pulsed input makes each SRAM bitcell linear, accumulation of partial products is still performed in current-domain and the overall MAC result is still nonlinear [3]. The capacitive SRAM in [4-5] improves linearity over current-domain accumulation by making the MAC result independent of the discharge current. However, linearity of MAC is still limited since the analog input is sampled on the capacitor in the bitcell through an NMOS switch. The analog input activation modulates threshold voltage (*Vth*) of the NMOS switch making the voltage sampled on the capacitor nonlinear. *Vth* drop in the NMOS capacitor also limits the maximum input swing that can be handled by the SRAM bitcell and restricts the supply voltage to relatively high values.

This work addresses the fundamental non-linearity in SRAM bitcell through two key techniques: 1) using delta-sigma modulators (DSM) to convert analog input activations into a binary pulse train; 2) using a 9T1C SRAM bitcell to perform computations in charge-domain. Compared to SAR or flash ADC, resolution of DSM can be re-configured easily without requiring changes in hardware. For the same oversampling ratio (OSR), quantization noise-shaping in DSM allows higher resolution of input activation compared to the counter-based technique in [3] which averages quantization error. The proposed IMC macro is shown in Fig. 1 and consists of a 64x64 9T1C bitcell array with 4-bit signed weights, 64 input 1-bit DSMs, switched-capacitor circuits for MAC and 16 output 1-bit DSMs for macro readout. Resolution of the input and output activations can be dynamically re-configured by changing OSR of the input and output DSMs. The output activation is reconstructed by digital decimation of the DSM output. Fig. 2 shows a single 9T1C SRAM bitcell. The pulsed-input is applied through the RWL and charges the capacitor *Csram* inside the bitcell. Since the RWL is driven by a binary pulse (0/1) instead of an analog input, the voltage sampled on *Csram* is restricted to two discrete voltage levels thus making dot-product in each SRAM bitcell perfectly linear. Use of CMOS switch instead of NMOS switch for sampling on the capacitor removes *Vth* drop in the sampled voltage and allows the SRAM bitcell to operate from very low supply voltages that improves power efficiency. The 1-bit switched-capacitor DSM used in this work is shown in Fig. 2. The DSM ensures that the temporal average of the output closely matches the temporal average of its input due to noise-shaping. A dynamic amplifier with gain *G* is used as loop filter which can have low gain since the comparator after loop filter performs 1-bit quantization based on only the sign of its input.

Fig. 3 shows a 4-bit slice of the IMC macro with the switched-capacitor circuits used for computation and the associated timing diagram. The SRAM bitcell capacitors, RBL lines, compute and compensation capacitors are discharged during ø1. The compensation capacitors ensure that all the RBL lines see the same capacitive load during computations. The MAC values are computed during ø2 and charged-shared with binary weighted compute capacitor bank as shown in Fig. 3. The compute capacitors are disconnected from the SRAM array and compensation capacitors during ø3. The compute capacitors are charge-shared with additional balancing capacitors in ø3 to ensure correct binary weighted MAC result with sign bit operation. The sign bit (*Vim*) and remaining 3-bit MAC result (*Vip*) is applied differentially to the output DSM to provide binary pulsed readout of the macro outputs. The macro takes *n*-cycles for computation where *n* corresponds to the OSR. The DSM output is digitally decimated using FIR low-pass filters. Fig. 4 shows the frequency response of the decimation filters for OSR=2 and 8 as well as the amplifier and comparator circuits. The amplifier and comparator are fully dynamic which improves power efficiency of the DSM. For realizing a full neural network with multiple macros mapped to each layer, decimation is not needed for DSMs in intermediate layers and only the final DSM output needs the decimation filter.



Die micrograph.

The test-chip is fabricated in 65nm process and occupies an area of 0.1mm^2 with the 64x64 SRAM array occupying an area of 0.03mm^2. The test-chip operates from a supply voltage of 0.5V-1.2V for the SRAM array and 1.2V for the DSM. The operating speed of the entire macro at 0.5V is 325kHz which is limited by buffers driving the sampling capacitor in the output DSM. The DSMs and clock generator consume 1.8μW from 1.2V supply while the SRAM array consumes 0.6μW-3.6μW from 0.5V-1.2V supply. Offset in the input and output DSMs are calibrated once in the foreground before characterization of the complete macro. Fig. 4 shows the measured average differential RBL voltage (*Vip-Vim*) versus MAC value for a SRAM bitcell, with the RBL voltage obtained by spatial averaging of bitcells in a column and the output DSM bitstreams 1M times. The maximum differential RBL swing is limited by charge sharing with balancing and compensation capacitors and linearity of the buffer driving the output DSM. Nonlinearity in the RBL transfer curve is due to static random mismatch between individual bitcells and capacitors, as well as charge-injection error in the switched-capacitor circuits. The maximum RMSE varies between 2.5mV at 0.5V SRAM power supply to 0.69mV at 1.2V SRAM power supply which corresponds to >6-bit linearity that is calculated by the ratio of maximum RBL swing to worst-case RMSE. The proposed macro is used to benchmark performance on MNIST dataset using a 5-layer neural network with two 2-D convolutional layers followed by three fully connected layers. Each convolution layer is followed by maxpool layer and ReLU activation. The first two fully connected layers use ReLU activation and the last layer uses softmax activation. To characterize the test-chip, analog inputs are directly applied to the test-chip using data acquisition system from National Instruments, and the chip outputs are captured using logic analyzer and digitally decimated off-chip. Fig. 5 shows the power and area breakdown, and accuracy and power efficiency (TOPS/W) as a function of SRAM supply voltage and OSR. The accuracy changes from 97.69% at 0.5V and OSR of 2 to 98.62% at 1.2V and OSR of 8 compared to software baseline accuracy of 98.68%. The power efficiency varies from 138.6TOPS/W at 0.5V to 61.6TOPS/W at 1.2V at OSR of 12. The normalized power efficiency computed for 1-bit input resolution and 1-bit weight resolution varies from 908.8 TOPS/W at 0.5V and OSR of 2 to 286.4 TOPS/W at 1.2V and OSR of 8. Fig. 6 summarizes this work and compares with state-of-the-art reported on MNIST dataset. The proposed test-chip has higher macro linearity than relevant prior works due to the adoption of capacitive SRAM and DSM for computation using binary pulses. The proposed test-chip has comparable accuracy on MNIST dataset and the best normalized power efficiency as state-of-the-art macros using similar CMOS process.

Acknowledgment: This work is partially supported by NSF grant CCF-1948331.

References:

[1] C. Yu *et al.*, CICC, April 2020. [2] X. Si *et al.*, JSSC Jan 2020. [3] Q. Dong *et al.*, ISSCC, Feb. 2020. [4] B. Zhang *et al.*, CICC, April 2022. [5] Z. Jiang *et al.*, JSSC, July 2020. [6] S. Jain *et al.*, JSSC, Oct. 2021.

979-8-3503-9948-6/23/$31.00 ©2023 IEEE

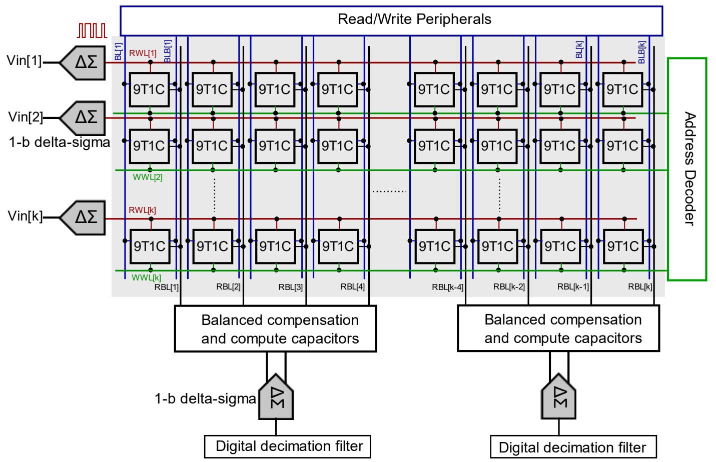
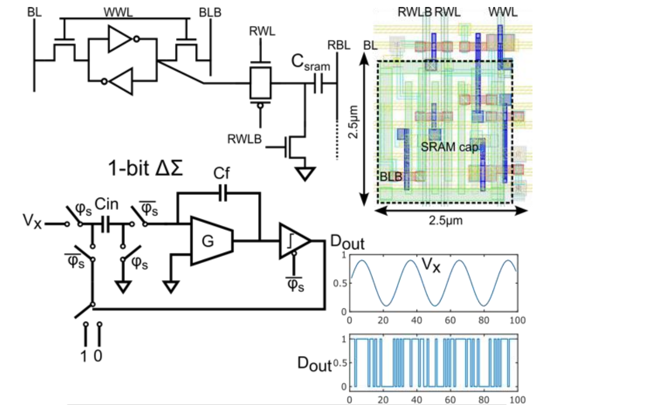
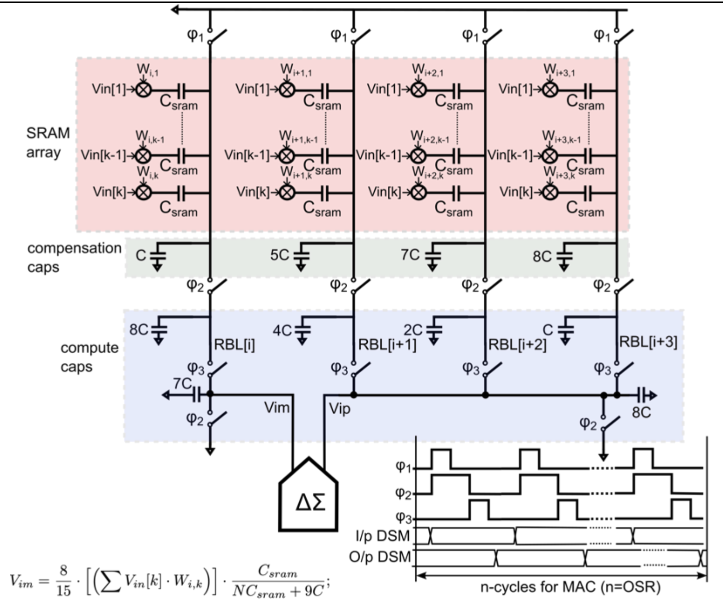


Fig. 1. Proposed DSM-based pulsed activation IMC macro.



$$D_{out} = \left[z^{-1}V_x + \underbrace{\left(1+2/G-(1+1/G)z^{-1}\right)Q}_{\text{high-pass shaped quantization error}}\right] \cdot \frac{1}{1+2/G-z^{-1}/G}$$

Fig. 2. 9T1C SRAM bitcell and 1-b DSM circuits.



$$V_{im} = \frac{8}{15} \cdot \left[\left(\sum V_{in}[k] \cdot W_{i,k}\right)\right] \cdot \frac{C_{sram}}{NC_{sram}+9C};$$

$$V_{ip} = \left[\frac{4}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+1,k}\right) + \frac{2}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+2,k}\right) + \frac{1}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+3,k}\right)\right] \cdot \frac{C_{sram}}{NC_{sram}+9C}$$

Fig. 3. Switched-capacitor circuit showing 4-bit macro and timing diagrams.

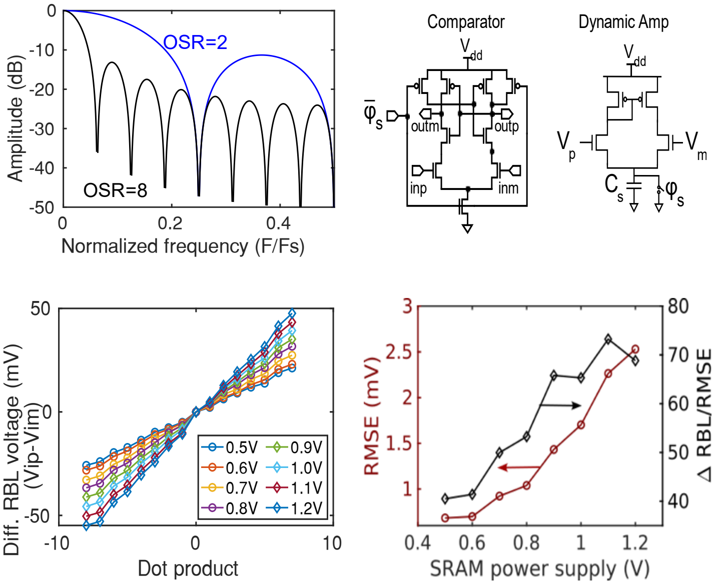


Fig. 4. DSM decimation filtering and key circuits, measured mean differential RBL voltage and RMSE as a function of MAC for 1 SRAM bit-cell for different SRAM supply

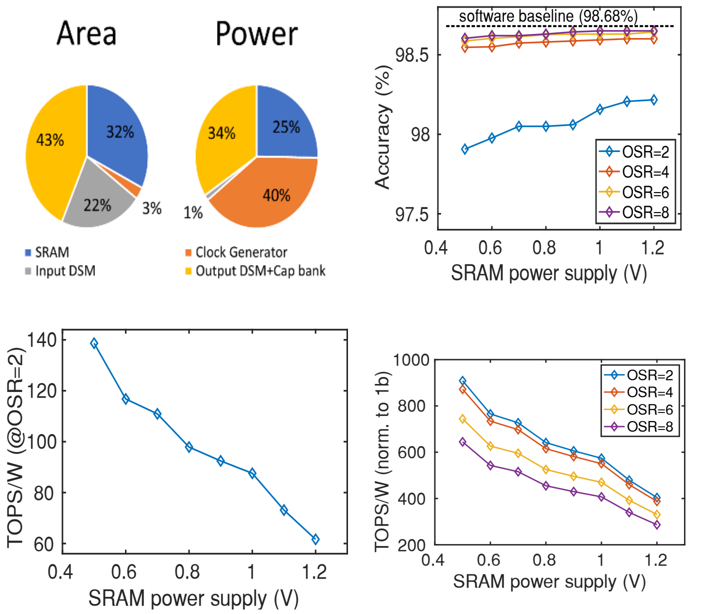


Fig. 5. Area, power pie-chart; and model accuracy and macro power efficiency as functions of SRAM power supply and OSR.

	This work	JSSC'20 [2]	CICC'20 [1]	JSSC'20 [5]	JSSC'21 [6]
Process	**65nm**	55nm	65nm	65nm	40nm
Supply	**0.5V-1.2V(SRAM)/ 1.2V(DSM)**	0.45V(SRAM)/0.8V	0.8V	0.8V(SRAM)/1V/0.6 V(ADC)[a]	1.1
SRAM type	**9T1C**	Twin 8T	8T	8T1C	8T
Bitcell size	**2.5μm x 2.5μm**	0.5μm x 1.7μm	1.8μm x 1.8μm	-	-
MAC type	**Charge-domain**	Current-domain	Current-domain	Charge-domain	Current-domain
MAC Linearity[b]	**40.5@0.5V/68.8 @1.2V**	39.3@1b, 12.6@2b	-	44	29.1
Read-out circuit	**DSM**	C2PU[c]	SA+ADC	Flash ADC	SAR ADC
Macro size	**64x64**	64x60	128x128	256x64	512x64
Activation resolution	**1.6b(@OSR=2) 4.6b(@OSR=8)**	1b	5b	1b	2b
Weight resolution	**4b**	3b	1b	1b	2b
Acc. (%) (MNIST)	**97.69(@0.5V, OSR=2)/ 98.62(@1.2V, OSR=8)**	98.2	96.2	98.3	98.2
Power Eff. (TOPS/W)[d]	**138.6@0.5V/61.6 @1.2V (OSR=2)**	144	31.6[f]	671.5	41
Norm. Power Eff. (TOPS/W)[e]	**908.8@0.5V/ 644.6@1.2V (OSR=2)**	432	158[f]	671.5	164

[a]only array (0.8V driver, 0.6V ADC);[b]max RBL swing/RMSE; [c]two's complement mapping and processing unit; [d]1 MAC=2 OP; [e]normalized energy efficiency = energy efficiency x input resolution x weight resolution; [f]only SRAM array

Fig. 6. Performance summary and comparison with state-of-the-art in similar CMOS technology.