Fully Row/Column-Parallel In-memory Computing SRAM Macro employing Capacitor-based Mixed-signal Computation with 5-b Inputs

Jinseok Lee, Hossein Valavi, Yinqi Tang and Naveen Verma
Princeton University, Princeton, NJ, USA (jinseokl@princeton.edu)

Abstract

This paper presents an in-memory computing (IMC) macro in 28nm for fully row/column-parallel matrix-vector multiplication (MVM), exploiting precise capacitor-based analog computation to extend from binary input-vector elements to 5-b input-vector elements, for 16x increase in energy efficiency and 5x increase in throughput. The 1152(row)x256(col.) macro employs multi-level input drivers based on a digital-switch DAC implementation, which preserve compute accuracy well beyond the 8-b resolution of the output ADCs, and whose area is halved via a dynamic-range doubling (DRD) technique. The macro achieves the highest reported IMC energy efficiency of 5796 TOPS/W and compute density of 12 TOPS/mm^2 (both normalized to 1-b ops). CIFAR-10 image classification is demonstrated with accuracy of 91%, equal to the level of ideal SW implementation.

IMC Macro Architecture

Fig. 1 shows a block diagram of the macro, comprising: an array of 10T SRAM multiplying bit cells (M-BCs), with capacitors implemented as upper-layer metal-fringing structures; write/read periphery (WL/BL decoders/drivers); 5-b input-vector element drivers [Dynamic-Range Doubling (DRD) DACs]; and 8-b SAR ADCs for IMC readout. While previous IMC works explored multibit vector/matrix-elements [1,2], this work exploits precision capacitor-based computation [3] to enable the high dynamic range required. The macro implements an inner-product operation in each column, where all 1152 rows are driven at once with input-vector data on complementary IA_i/IAb_i lines. Each M-BC selects between IA_i/IAb_i to drive its local capacitor, thus performing multiplication with stored data and then accumulation across the column via charge redistribution of all capacitors coupled by the compute line (CL_j). Using capacitors enables high precision CL_j data, which is then digitized.

Previous capacitor-based IMC [4] employed binary input-vector/matrix elements and achieved extension to multibit via bit-parallel/bit-serial (BP/BS) operation. BP/BS stores matrix bits in parallel columns and provides input bits serially. The columns thus give binary-vector inner products, which are then digitized, bit-shifted (for proper binary weighting), and summed for extension to multibit. The drawback is that this requires B_X bit-serial (BS) IMC cycles for B_X-bit input data. Instead, in this work, the M-BCs, which can take analog inputs, are provided 5-b input data in one cycle. This is done using energy/area-efficient digital-switch-based DACs, which preserve the precision of capacitor-based compute, needed for the required dynamic range.

Input Drivers with Dynamic-Range Doubling (DRD) DACs

Fig. 2 shows the DRD approach, which improves DAC energy and area by halving the number of levels required to 16 for B_X=5 input data. While previous capacitor-based IMC involved setting IA_i/IAb_i both to GND during a reset phase, and then setting one to V_{DD} based on the input bit, during an evaluate phase. This causes charge redistribution on CL_j by driving a voltage change of V_{DD} across the M-BC capacitors in the column. On the other hand, DRD explicitly utilizes the reset phase to apply 16-level input data on IA_i or IAb_i if the data's MSB (taken as sign bit for

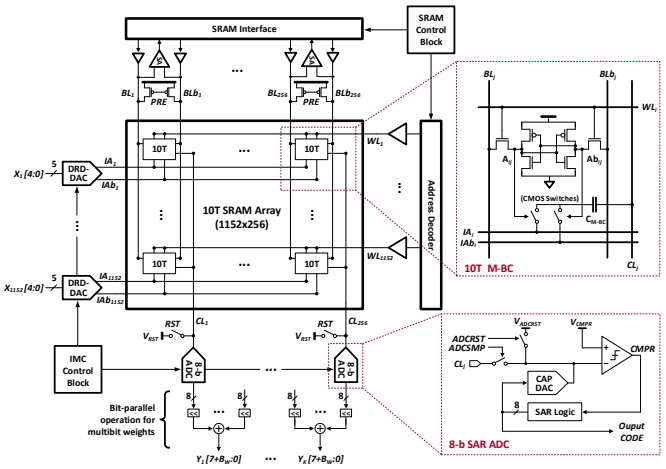


Fig. 1 IMC macro based on capacitor compute, to enable 5-b inputs.

illustration) is 1 (positive) or 0 (negative), respectively (note, generally the MSB need not be a sign bit, but can be formatted as such by applying a fixed offset to the IMC inputs and outputs). Then, during the evaluate phase, the voltages of IA_i and IAb_i are swapped, causing CL_j charge redistribution by driving a voltage change of up to $2\times V_{DD}$ across the M-BC capacitors in a column. In terms of dynamic range, this has two advantages: (1) it enables 5-b input activations with just half of the required DAC levels; (2) it increases the CL_j voltage swing to overcome attenuation from CL_j parasitic capacitance. In terms of energy and throughput, this also has two advantages: (1) M-BC capacitors are driven by lower voltages, both due to the range of DAC levels and due to reduction in the maximum DAC level enabled by DRD, which together result in ~3.3× energy reduction; (2) BS cycles are eliminated, increasing energy efficiency and throughput further by 5×.

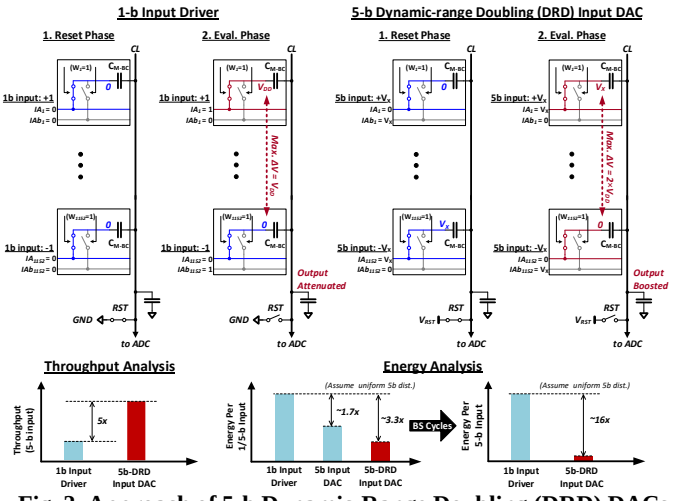
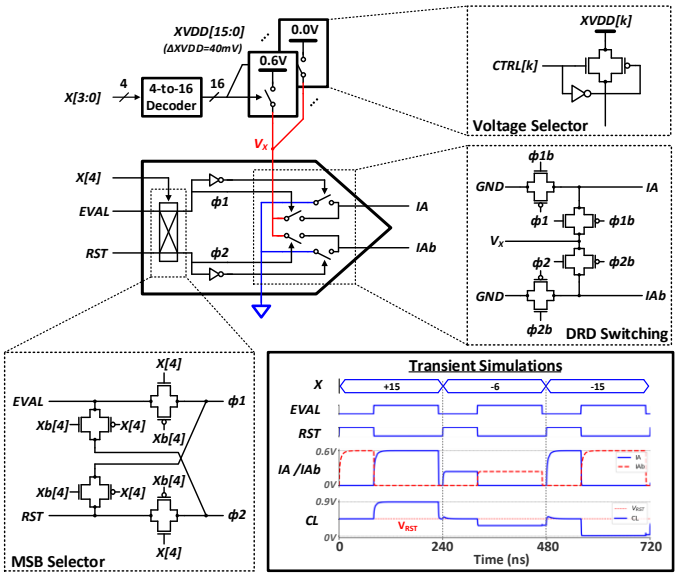


Fig. 2. Approach of 5-b Dynamic-Range Doubling (DRD) DACs.

Fig. 3 shows details of the DRD-DAC. Rather than an analog implementation, the DAC employs CMOS transmission gates (TGs) to select and drive 16 explicit voltage levels XVDD[15:0], thereby providing fast settling and maintaining the linearity of the provided levels. While it may seem that 16

separate supplies introduce additional macro complexity, this is not the case. In practice, IA_i/IAb_i drivers dominate power consumption in fully row-parallel IMC, and driver supply tracks are allocated all above the array to provide current. In this case, the tracks are simply allocated across the 16 supplies, whose total current is in fact lowered thanks to the reduced voltage levels. Post-layout simulations (bottom right) illustrate operation, showing fast DAC settling and full-swing range on CL_j, despite parasitics.



nFig. 3. Circuit and simulation waveforms of the DRD-DACs.

Dynamic-range Analysis

An important consideration is the dynamic range required with 5-b inputs, both for computation on CL_j (i.e., accumulation across 1152 M-BCs) and then for the column ADC. In terms of CL_j computation, which involves only M-BC capacitors, both previous analysis [3] and analysis of the current process technology show that 10's of thousands of compute levels can be supported before SNR is limited by matching of the metal-fringing capacitors. In terms of the 8-b ADC, Fig. 4 shows that 5-b activation DACs retain roughly the same signal-to-quantization-noise ratio (SQNR, referenced to floating-point baseline) as BP/BS operation, across different weight precisions and column dimensionalities, both of which, together with the ADC, set quantization effects. This is because, during each BS cycle, BP/BS benefits from bit growth via ADC output-code bit shifting, but also incurs loss of an 8-b residue of the CL_j analog signal, introducing rounding.

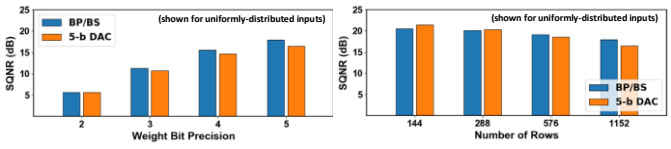


Fig. 4. Analysis of SQNR resulting from 5-b inputs with 8-b ADC.

Measurements and Demonstrations

Fig. 5 shows a die photo and basic measurements of the 28nm prototype. The M-BC size (0.936 μm^2) is roughly 2× larger than the layout of a standard 6T SRAM cell using technology logic rules. At the left, the column transfer function shows low variability (error bars show standard deviation across all 256 columns), low noise (mean of 0.98 LSB_{RMS}), and high linearity (INL < 1LSB), due to precision capacitor-based IMC. At the bottom right, the detailed energy breakdown shows both that the dominant input-driver energy is reduced specifically by DRD, and the overall energy is further reduced with 5-b inputs, thanks to the range of DAC voltage levels and elimination of BS cycles.

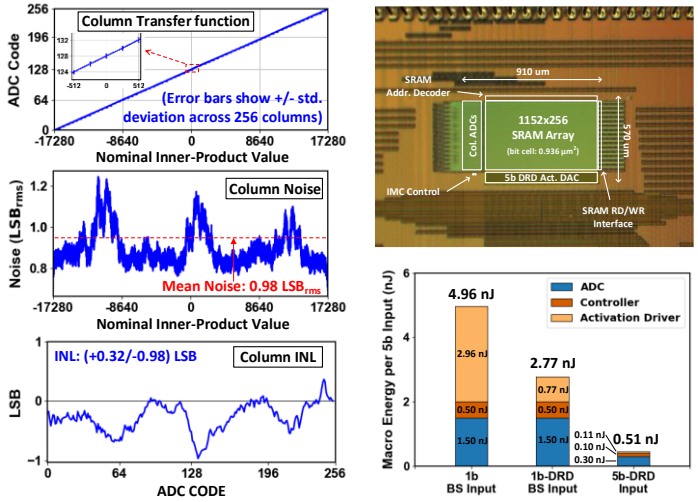
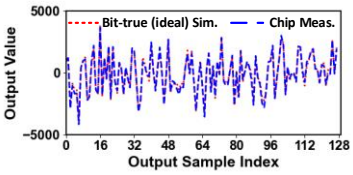


Fig. 5. Protype and measurements of basic operation.

Fig. 6 shows demonstration of a 14-layer CNN using 5-b weights/activations for CIFAR-10 image classification (which is significantly more challenging than MNIST), with both overall accuracy and intermediate pre-activation computation values shown, closely matching ideal SW-modeled (quantized) computation. Below is a comparison table to previous work. With 5/1-b native input/weight precision, this work achieves the highest energy efficiency (especially normalized for technology) of 1159 TOPS/W, corresponding to 5796 TOPS/W normalized to 1/1-b inputs/weights. High area-normalized throughput of 12 TOPS/mm^2 (normalized to 1/1-b) is achieved, with precision compute enabling scaling to large macro dimensions and fully row/column-parallel operation.

Demonstrated CNN for CIFAR-10 Classification

Computed Pre-activations



Demonstration Summary

Bit Precision	5b / 5b / 12b (Act. / Weight / Out)
Energy/classification	1.73 μJ / classification
Est. Throughput	650 fps
Network Architecture	L1 : 64 CONV L2 : 64 CONV L3 : 64 CONV L4 : 64 CONV L5 : POOL L6 : 128 CONV L7 : 128 CONV L8 : POOL L9 : 128 CONV L10: 128 CONV L11: POOL L12: 1024 FC L13: 1024 FC L14: 10 FC
Accuracy	91.1% (vs. 91.3% ideal SW)

Summary and Comparison Table

	Valavi, JSSC 19	Biswas, JSSC 19	Yang, ISSCC 19	Dong, ISSCC 20	Su, ISSCC 20	Si, ISSCC 20	This Work
Technology	65nm	65nm	28nm	7nm	28nm	28nm	28nm
Area (mm^2)	12.6	0.063	0.216	0.053	-	-	0.510
Supply Voltage (V)	0.94/0.68/1.2	1.2/0.8/1.0	0.6-0.9	0.8	0.85-1	0.7-0.9	0.9/0.6
Array Size	295 KB	2 KB	19.5 KB	4 KB	8 KB	8 KB	36 KB
Native Bit Precision (act. / weight / output)	1b/1b/1b	7b/1b/7b	8b/1b/8b	4b/4b/4b	2b/4b/10b (extendable)	4b/4b/12b (extendable)	5b/1b/8b (extendable)
Throughput (GOPS) normalize to 1-b ops.	18876	8	-	5958	-	-	6144
Comp. Density (GOPS/mm^2) normalize to 1-b ops.	1498	127	-	112423	-	-	12047
Energy Efficiency (TOPS/W) normalize to 1-b ops.	866	51.3	119.7	5616	486	1064	5796
Neural Network Demo (accuracy)	CIFAR-10 (83.3%)	MNIST (98.3%)	-	MNIST (~98.5%)	CIFAR-10 (91.94%)	CIFAR-10 (92.02%)	CIFAR-10 (91.1%)

Fig. 6: Chip CNN demonstration and comparison summary.

References

[1] Y.-C. Chiu et al., *IEEE JSSC*, vol. 55, no. 10, pp. 2790–2801, Oct. 2020.

[2] X. Si et al., *IEEE ISSCC*, pp. 246–248, 2020.

[3] H. Valavi et al., *IEEE JSSC*, vol. 54, no. 6, pp. 1789-1799, Jun. 2019.

[4] H. Jia et al., *IEEE JSSC*, vol. 55, no. 9, pp. 2609—2621, Sep. 2020.