16.2 eDRAM-CIM: Compute-In-Memory Design with Reconfigurable Embedded-Dynamic-Memory Array Realizing Adaptive Data Converters and Charge-Domain Computing

Shanshan Xie[1], Can Ni[1], Aseem Sayal[1], Pulkit Jain[2], Fatih Hamzaoglu[2], Jaydeep P. Kulkarni[1]

[1]University of Texas, Austin, TX
[2]Intel, Hillsboro, OR

The unprecedented growth in deep neural networks (DNN) size has led to massive amounts of data movement from off-chip memory to on-chip processing cores in modern machine learning (ML) accelerators. Compute-in-memory (CIM) designs performing analog DNN computations within a memory array, along with peripheral mixed-signal circuits, are being explored to mitigate this memory-wall bottleneck: consisting of memory latency and energy overhead. Embedded-dynamic random-access memory (eDRAM) [1,2], which integrates the 1T1C (T=Transistor, C=Capacitor) DRAM bitcell monolithically along with high-performance logic transistors and interconnects, can enable custom CIM designs. It offers the densest embedded bitcell, a low pJ/bit access energy, a low soft error rate, high-endurance, high-performance, and high-bandwidth: all desired attributes for ML accelerators. In addition, the intrinsic charge sharing operation during a dynamic memory access can be used effectively to perform analog CIM computations: by reconfiguring existing eDRAM columns as charge domain circuits, thus, greatly minimizing peripheral circuit area and power overhead. Configuring a part of eDRAM as a CIM engine (for data conversion, DNN computations, and weight storage) and retaining the remaining part as a regular memory (for inputs, gradients during training, and non-CIM workload data) can help to meet the layer/kernel dependent variable storage needs during a DNN inference/training step. Thus, the high cost/bit of eDRAM can be amortized by repurposing part of existing large capacity, level-4 eDRAM caches [7] in high-end microprocessors, into large-scale CIM engines.

This work demonstrates a 65nm CIM prototype that repurposes 1T1C eDRAM columns as charge domain circuits to perform DNN computations (Fig. 16.2.1). The key attributes of which are, (1) support of in-eDRAM DNN analog computations: such as data conversion, dot-product, averaging, pooling, and rectified linear unit (ReLU) activation; (2) support for 8b input and 8b signed/unsigned weight multiply-accumulate-averaging (MAV) operations; (3) a modified WL controller to configure some of the 1T1C eDRAM columns as charge-sharing compute units in CIM mode; (4) performing dot products with non-destructive weight reads, thus avoiding weight duplication, extra control logic and not requiring a pre-initialized array; (5) an in-eDRAM adaptive dynamic-range successive-approximation (SAR) analog-to-digital converter (ADC) using narrow range of dot-product distribution to minimize the ADC latency/energy; and (6) quantify eDRAM-CIM benefits in an advanced eDRAM technology node.

The eDRAM-CIM based DNN computations and dataflow are shown in Fig. 16.2.2. An 8b digital input, X_{IN}, and its 1's complement are converted to differential analog voltages, V_a and V_{a_bar}, which are centered around $½V_{DD}$, in two conversion steps (4b/conversion). Two 1T1C eDRAM columns are used as digital to analog converters (DACs) for the conversion: DAC_{POS} for V_a and DAC_{NEG} for V_{a_bar}. Next, a MAV operation is performed by first reading all 8b weights from the eDRAM array. The most significant bit W_7, representing the sign bit, is used to select between the differential voltage V_a or V_{a_bar}. The dot products are performed using 2:1 multiplexers and sampled on the binary scaled eDRAM capacitors depending on the weight bit position. Using this method, the read/write/refresh operations for the weight array are unaffected. Averaging, ReLU, and average/max pooling computations are performed with comparators and additional 1T1C eDRAM bitcells, as charge sharing steps, to generate the MAV output (V_{MAV}). Figure 16.2.3 shows the captured oscilloscope waveforms from the 65nm prototype test-chip (Fig. 16.2.7) demonstrating the eDRAM-CIM functionality and the dataflow.

For the eDRAM based DAC design, the lower 4 bits of an 8b digital input, which are stored in different portion of the array, are first loaded into a 1T1C DAC column in a thermometer scaled fashion (Fig. 16.2.3). Conversion is performed by activating all WLs of this column simultaneously. Since the bitcells in this column are initialized just before charge sharing, the bitcell charge leakage would be minimal. A similar step is performed for the higher 4 bits with a 16C sampling capacitor to reflect its scaling factor. A programmable gain amplifier is used to compensate for any potential gain error. Finally, C_{x1} and C_{x16}, where the two-step converted analog voltages are stored, are added using a charge-sharing step to realize differential analog voltages V_a and V_{a_bar}. The measured DAC dynamic range matches the simulations for both X_{IN} (V_a) and its 1's complement (V_{a_bar}) and the measured dynamic non-linearity (DNL) is within 1 LSB at lower input values.

In the final step, a SAR ADC is realized by comparing V_{MAV} with a reference voltage (V_{REF}), which is successively refined in each SAR step by initializing the DAC column bitcells appropriately before a charge sharing operation (Fig. 16.2.4). To mitigate the ADC energy and area overheads, a 1T1C eDRAM-based adaptive dynamic-range ADC is devised, by leveraging the narrow V_{MAV} distribution. By trimming infrequent appearing V_{MAV} values and skipping the subsequent SAR steps, ADC energy is reduced by 1.14× at 60% clipping threshold. The 1T1C bitcells in the V_{REF} generation DAC column can be set to be always ON/OFF while being initialized, such that V_{REF} can be bounded between adjustable low and high voltage bounds (V_{LB} and V_{HB}) depending on the V_{MAV} distribution for a specific convolution layer. For illustration, in Fig. 16.2.4, ADC dynamic range is set between $¼V_{DD}$ and $¾V_{DD}$. This configuration saturates the quantization of very low or very high V_{MAV} at 64 or 191 (out of 256) codeword values, respectively, without undergoing subsequent SAR cycles. The V_{MAV} values, within the bounded range, are quantized using the remaining non-clipping 128 levels. In addition, a 2b/cycle conversion technique is implemented using $3V_{REF}$ comparisons/cycle; thereby, shortening the ADC conversion latency to minimize the effect of V_{MAV} capacitor leakage during SAR cycles. This improves the adaptive ADC throughput by 1.14× while incurring a 3.79% (1.21%) drop in Top-1 (Top-5) CIFAR-10 classification accuracy using 8b integer operands.

Figure 16.2.5 shows the simulated and measured Top-1 and Top-5 CIFAR-10 classification accuracy as a function of the ADC clipping threshold using a neural network having 4 convolution, 2 pooling, and 2 fully connected layers. The ADC clipping decision incurs extra cycles for comparing V_{MAV} with V_{LB} and/or V_{HB} thresholds, and incurs additional energy for low clipping thresholds. As clipping threshold is increased to an optimal point (0.6) the ADC energy drops by 1.14×.

Figure 16.2.6 compares the eDRAM-CIM design with prior multi-bit CIM designs using the CIFAR-10 dataset [3-6]. The test-chip measurement setup, die-micrograph, macro area, and energy breakdown are shown in Fig. 16.2.7. In the scalability analysis, the presented eDRAM-CIM approach when adopted to an advanced eDRAM technology node [2,7], shows promising energy efficiency and throughput metrics; suggesting its potential for deployment in large-scale energy-efficient CIM designs to mitigate the memory bottleneck challenges.

Acknowledgement:
The authors would like to thank Clifford Ong for technical discussions and Intel for funding support.

References:
[1] G. Fredeman et al., "A 14nm 1.1Mb Embedded DRAM Macro with 1ns Access," *IEEE JSSC*, vol. 51, no. 1, pp. 230–239, Jan. 2015.
[2] F. Hamzaoglu et al., "A 1Gb 2GHz 128Gb/s Bandwidth Embedded DRAM in 22nm Tri-Gate CMOS Technology," *IEEE JSSC*, vol. 50, no. 1, pp. 150–157, Jan. 2014.
[3] J. Su et al., "A 28nm 64Kb Inference-Training Two-Way Transpose Multibit 6T SRAM Compute-in-Memory Macro for AI Edge Chips," *ISSCC*, pp. 240–242, 2020.
[4] X. Si et al., "A 28nm 64Kb 6T SRAM Computing-in-Memory Macro with 8b MAC Operation for AI Edge Chips," *ISSCC*, pp. 246–248, 2020.
[5] C. Xue et al., "A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices," *ISSCC*, pp. 244–246, 2020.
[6] S. K. Gonugondla et al., "A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training," *ISSCC*, pp. 490-492, 2018.
[7] N. Kurd et al., "Haswell: A Family of IA 22nm Processors," *IEEE JSSC*, vol. 50, no. 1, pp. 49-58, Jan. 2015.
[8] A. Biswas et al., "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks," *IEEE JSSC*, vol. 54, no. 1, pp. 217-230, Jan. 2019.

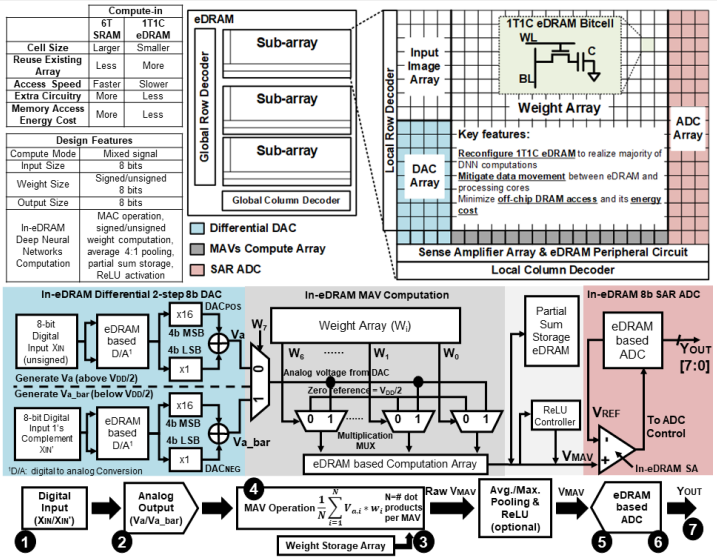


Figure 16.2.1: Big picture: compute-in-SRAM and compute-in-eDRAM comparison, design highlights, high-level array structure, and step-by-step eDRAM-CIM dataflow.

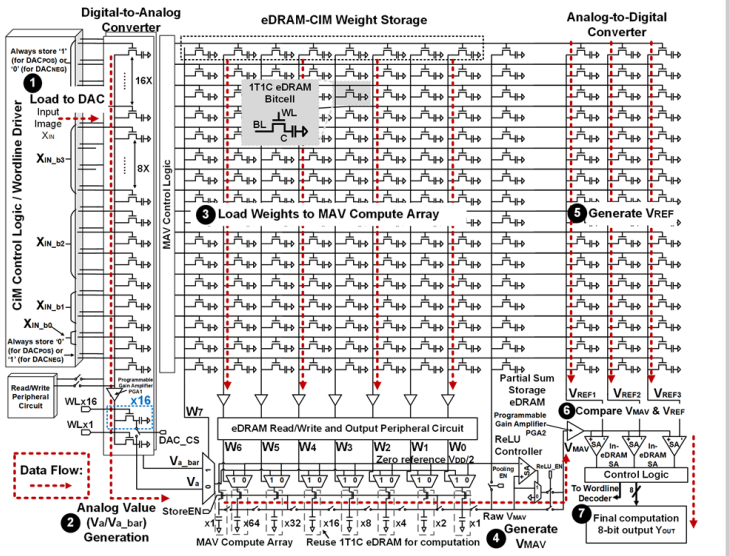


Figure 16.2.2: Overall circuit schematic of eDRAM-CIM performing majority of DNN computations: digital-to-analog conversion (DAC), multiplication, averaging, pooling, ReLU, and analog-to-digital conversion (ADC).

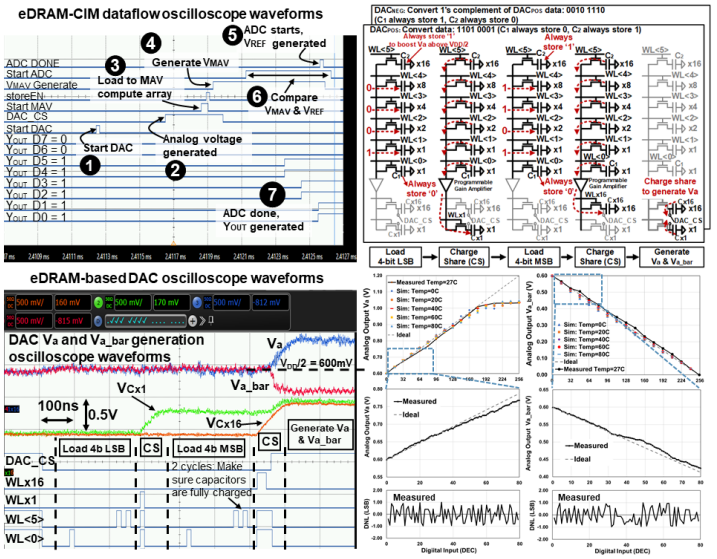


Figure 16.2.3: eDRAM-CIM dataflow demonstration with oscilloscope waveforms, in-eDRAM DAC circuit schematics, functional oscilloscope waveforms, simulated and measured DAC characteristics including DNL.

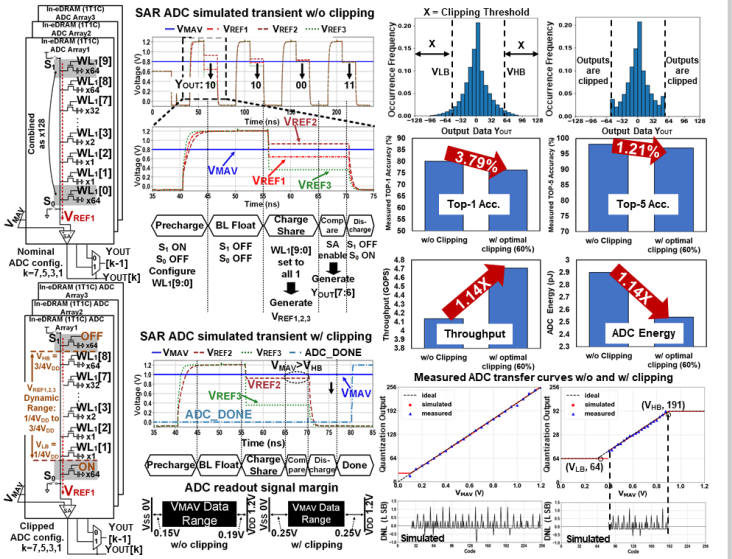
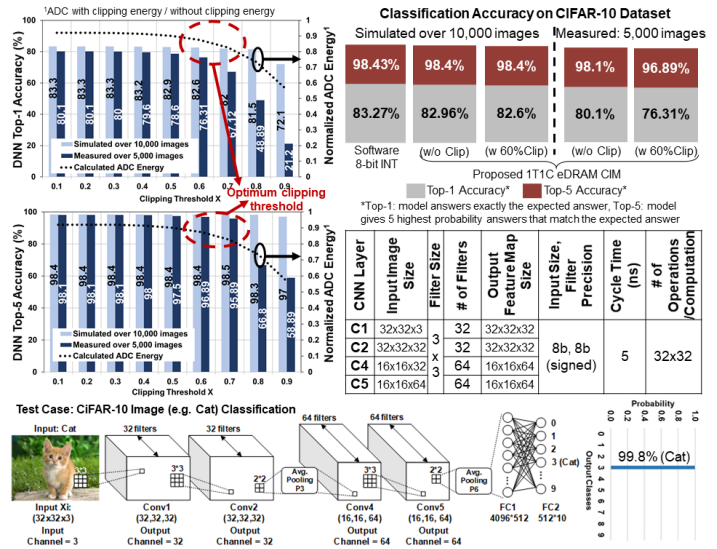


Figure 16.2.4: In-eDRAM SAR ADC with V_{REF} generation, ADC operation waveforms, ADC characteristics and DNL, V_{MAV} distribution with and without clipping and its effect on ADC accuracy, throughput and energy.



CNN Layer	Input Image Size	Filter Size	# of Filters	Output Feature Map Size	Input Size, Filter Precision	Cycle Time (ns)	# of Operations /Computation
C1	32x32x3	3 x 3	32	32x32x32	8b, 8b (signed)	5	32x32
C2	32x32x32		32	32x32x32			
C4	16x16x32		64	16x16x64			
C5	16x16x64		64	16x16x64			

Figure 16.2.5: Simulated and measured CIFAR-10 dataset Top-1 and Top-5 classification accuracy variation with the in-eDRAM ADC clipping threshold along with neural network details.

	This work	ISSCC'20 [3]	ISSCC'20 [4]	ISSCC'20 [5]	ISSCC'18 [6]
Technology	65nm	28nm	28nm	22nm	65nm
Memory Cell Structure	1T1C eDRAM	6T SRAM	6T + Local Computing SRAM	1T1R SLC ReRAM	6T SRAM
Array Size	16Kb	64Kb	64Kb	2Mb	128Kb
Input Precision (bit)	8	8	8	4	8
Weight Precision (bit)	8	8	8	4	8
Supply Voltage (V)	1~1.2	0.85~1.0	0.7~0.9	0.8	1
Dataset	CIFAR-10	CIFAR-10			
Model	CNN: 4 CONV + 2 Pooling + 2 FC	CNN: ResNet-20	CNN: ResNet-20	N/A	SVM
Measured Accuracy	80.1% (Top-1), 98.1 % (Top-5)	[5]91.91%	[5]92.02%	N/A	[5]83.27%
Throughput (GOPS)	[1,3]4.71	N/A	N/A	N/A	4
Average Energy Efficiency (TOPS/W)	[1]4.76	7.3 [2](1.35)	14.08 [2](2.61)	28.93 [2](3.31)	3.125
GOPS/mm^2	8.26	N/A	N/A	N/A	2.78
[4]FoM	304.6	86.4	167	53	201.6

[1]measured at 1.1V
[2]Scaled to 65nm, assume energy ∝ (Tech.)^2 [8]
[3]Limited by clocking infrastructure, chip size, technology and bit cell area
[4]FoM = input precision x weight precision x energy efficiency (scaled to 65nm)
[5]Top-1 or Top-5 is not mentioned

Figure 16.2.6: Comparison with prior works supporting multi-bit input and weight integer operands and using CIFAR-10 dataset. FoM is calculated based on input precision, weight precision and energy efficiency.

		Test-chip Summary	Estimated
Technology		65nm CMOS	22nm tri-gate CMOS
Supply Voltage (V)		1~1.2	1.05
Total eDRAM Size		16Kb	[2]38.4MB
Bit-cell Size (um^2)		22.08	0.029
Unit eDRAM Capacitor Size (fF)		13	13
eDRAM Capacitor Type		Metal-oxide-metal (MOM)	Capacitor-over-bitline (COB)
Input Clock Frequency		[4]50 MHz (measured) 200MHz (simulation)	1.6GHz
[3]eDRAM-CIM Macro Area (mm^2)		0.57	[2]17
GOPS/mm^2		8.26	1305
MAV Operation Latency (ns)		10 (2 cycles)	1.25 (2 Cycles)
Input, Weight, Output Precision		8b, 8b, 8b	
Dataset		CIFAR-10	
Software Classification Accuracy (INT8) (Top-1, Top5)		83.27%, 98.43%	
Measured Classification Accuracy for 5000 Images (Top-1, Top-5)	w/o ADC Clipping	80.1%, 98.1%	
	with 60% ADC Clipping	76.31%, 96.89%	
[1]Throughput @ 1.1V	w/o ADC Clipping	4.037 GOPS	19.5 TOPS
	with 60% ADC Clipping	4.71 GOPS	22.2 TOPS
[5]Energy Efficiency (TOPS/W) @ 1.1V		4.76	11.12
Energy/MAV (fJ/MAV) with 60% ADC Clipping @ 1.1V		30.6	[5]0.21

[1]Assume 1MAV = 2 operations
[2]Assume 30% utilization of 128MB (77mm^2) [6] eDRAM memory for CIM usage
[3]Main design area (not including others structure, e.g. scanchain, test structure)
[4]Limited by jumper wire bandwidth
[5]Average Energy Efficiency and excluded scan-chains power
[6]Increased parallelism in large-capacity eDRAM array

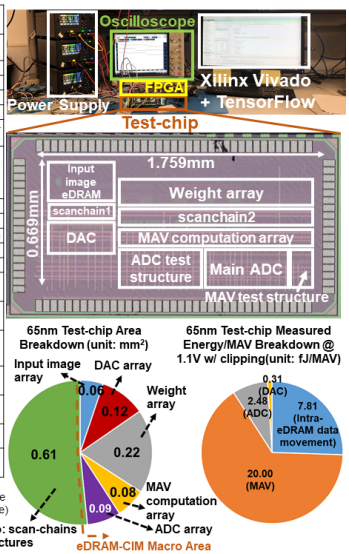


Figure 16.2.7: Test-chip measurements summary, characterization setup, die micrograph with functional blocks highlighted, area and energy/MAV breakdown and scaling analysis to 22nm tri-gate eDRAM technology.

978-1-7281-9549-0/21/$31.00 ©2021 IEEE