A 278-514M Event/s ADC-Less Stochastic Compute-In-Memory Convolution Accelerator for Event Camera

Jiyue Yang, Alexander Graening, Wojciech Romaszkan, Vinod K. Jacob, Puneet Gupta, Sudhakar Pamarti
University of California, Los Angeles, CA, USA. Email: jyang669@ucla.edu.

Abstract

We present a Compute-In-Memory (CIM) convolution accelerator for object tracking applications using event camera, which is a new imaging technology that significantly improves latency and dynamic range over conventional cameras. Previous works proposed an efficient ADC-less Stochastic CIM for deep learning but need to store 2^N stochastic bits for an n-bit number. We propose to store binary numbers in memory and convert them to stochastic bits by in-situ Stochastic Number Generators on the fly, which reduce the storage requirement by >10x. The CIM macro embeds 32 tiny MAC units per weight and uses an early termination technique to skip unnecessary computation of zeros. The accelerator achieves energy efficiency of 485 TOPS/W and throughput of 278-514 Mevent/s. The proposed SCIM macro can also be used to accelerate convolution in most deep learning applications.

Keywords: Compute-In-Memory Accelerator, Event Camera.

Introduction

Compute-In-Memory (CIM) is a promising solution to overcome the memory bottleneck of computing. Previous analog-based CIMs, however, sacrifice both energy efficiency and area density due to costly ADCs. Stochastic Compute-In-Memory (SCIM) was proposed to embed Stochastic Computing (SC)'s 1-b digital logic in memory to achieve higher energy efficiency and macro density [1], Fig.1. The SCIM macro needs to store 2^N stochastic bits for N bit number, which limits the on-chip storage capacity. Event cameras and other deep learning applications demand extremely high processing throughput and energy efficiency in convolution [2]. We propose an ADC-less low-power and high-throughput Stochastic CIM accelerator for convolution-based object tracking algorithm on event cameras, Fig.1. The accelerator supports both event polarity(2b) and non-event binary(6b) inputs. Two SCIM macros perform convolution by 186K SC MACs and near-memory fixed-point processing circuits supports cross-macro accumulation and max pooling.

In-Situ Stochastic Number Generator: Our proposed in-situ SNG can accurately convert binary to stochastic bit in memory sequentially and reduces required storage by >10x than [3], Fig.2. SRAM cells store 6b binary weights. The magnitude bits, W_0–W_4, are multiplexed by random streams, RN_0–RN_4, with binary weighted means, ($1/2^1$–$1/2^5$), using a wired-OR to generate the desired stochastic bit stream Wsc. The sign bit, W_5, is used to control the demultiplexer from Wsc to Wsc+/Wsc- such that P(Wsc+)-P(Wsc-) represents a signed value. The maximal length LFSR pseudo-random sequence with an extra zero state has an autocorrelation of 0. The sequence is stored in cyclic shift registers achieving 5x smaller number of flip flops. The LFSR states are uniformly distributed and make the in-situ SNG accurate. The random binary weighted streams, RN_0–RN_4, in each SNG are generated using a different set of five bits from the shift registers e.g., L_0-L_4 for SNG1, L_1-L_5 for SNG2 etc. Since stochastic streams in different dot products do not interact, random numbers are shared by SNGs in the same row without correlation.

Compact In-Memory MAC Array: A massive number (32x) of MACs are embedded for every stored weight to increase parallelism with only 30% extra area. Conventional CIM solutions cannot embed a large MAC array in memory due to the large area of ADCs, Fig.3. SCIM only requires a 1-bit sense amp and the area overhead is very small. The macro supports both 2b event and 6b binary inputs. In-situ SNG converts weight to stochastic bits and shares with 32 MAC units, which use 3 NMOS transistors to perform two AND-multiplication between Wsc+/Wsc- and input IN. MAC units only account for 30% of the area. The 81 rows form an 81-long dot product at each compute line(CLp/ CLn), which performs a 1-bit wired-OR operation. The sense amplifier converts 1-bit OR accumulation output(High/Low). Counters accumulate CLp and CLn over 64 clock cycles (for 6-bit binary output) and compute the difference to generate binary MAC output.

Early Termination: Event camera's images or inputs of CNNs' hidden layers may have very high sparsity (>90%). An Early Termination (ET) technique for SC is proposed to skip computing zeros. As SC computes sequentially, the output counter increases proportionally with clock cycles and the variation gradually reduces, Fig.4. We propose a 2-step fine/coarse-grained ET to turn off inactive counters. The fine-grained ET compares the counter output of each sense amp with a programmable threshold and clock-gates the counters by "mac_disable" signal, Fig.4. In coarse-grained ET, if all the counter units in the macro meet the criteria of fine-grained ET, entire chip's operation can be terminated. The chip will move on to compute the next inputs to save energy and time. The threshold of the ET can be set to trade off tracking accuracy and energy consumption. Our evaluation shows >1.9x improvement of energy efficiency by using early termination.

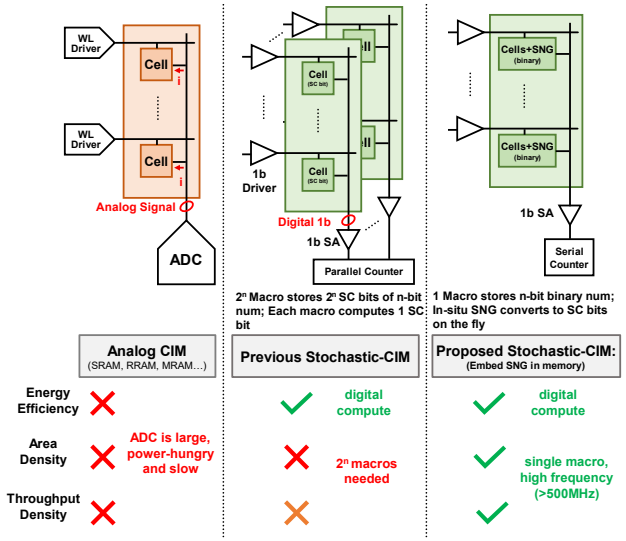
Measurement Results

Our prototype is fabricated in 12nm with a 0.5mm^2 core area, Fig.6. The system operates at 600~850MHz under 0.64V~0.85V supply, Fig.5(mid). The error of computation at the macro level is characterized and shown in Fig.5(left). The error RMS between measured SC dot-products and FP ground truth is 0.74. Object tracking is demonstrated using a data set of high-speed flying objects captured by DVS event camera. Event inputs (2b) are convolved with 32 Gabor filters (6b), followed by a max pooling. The accuracy is measured using the Higher Order Tracking Accuracy metric, characterized with different ET thresholds, Fig.5. Higher ET thresholds terminate more computation, but sacrifice tracking accuracy. When ET is enabled, the system achieves a max throughput of 514M event/s. The energy efficiency is 158 TOPS/W for system and 495 TOPS/W for macro. Energy efficiency at different sparsity levels is shown in Fig.5(right). In non-event 6bx6b processing mode, energy efficiency is 46TOPS/W for system and 86TOPS/W for macro. Table 1 compares us with other works.

Conclusion: The object tracking system achieves 60x higher throughput than the previous event camera accelerator [3]. In conventional non-event mode, the macro's energy efficiency is 2~3x higher than other CIM solutions[4][5], and similar to the previous SCIM work[1]. The throughput density is >10x higher than other works due to the compact MAC array.

References

[1] J. Yang et *al.,* ASSCC, 2022. [2] D. Falanga *et al*., IEEE RA-L, 2019. [3] M. Chang *et al*., ISSCC, 2023. [4] B. Yan *et al.,* ISSCC, 2022. [5] J. Yue *et al.,* ISSCC, 2021. [6] J. Yue *et al.,* ISSCC, 2023.



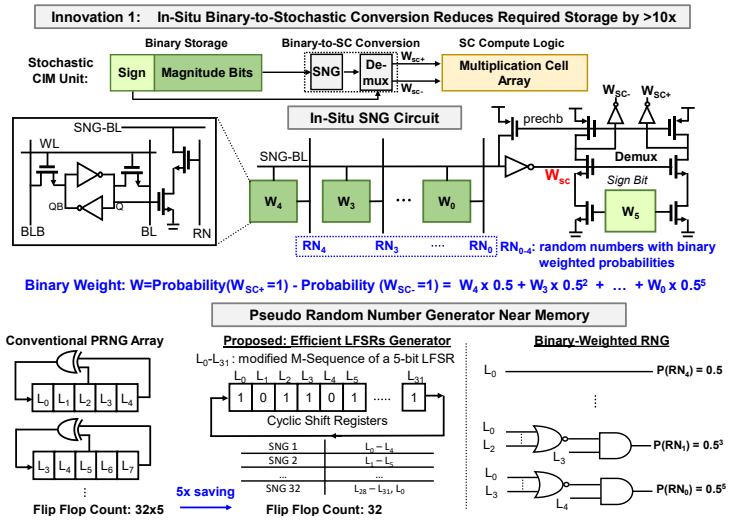


Fig.2 In-Situ SNG allows binary weights and reduce required storage.

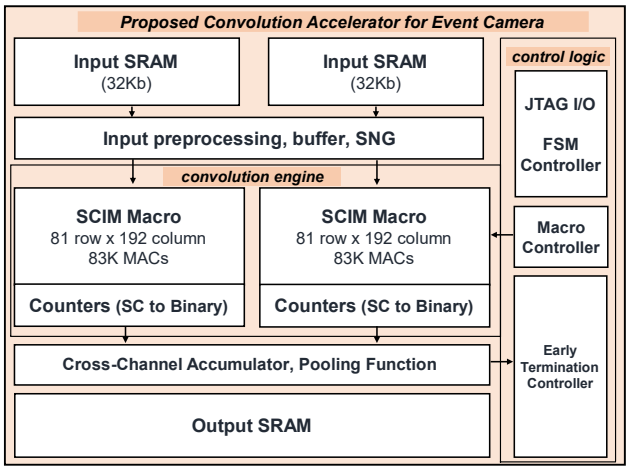


Fig.1 Comparison with other works and proposed architecture.

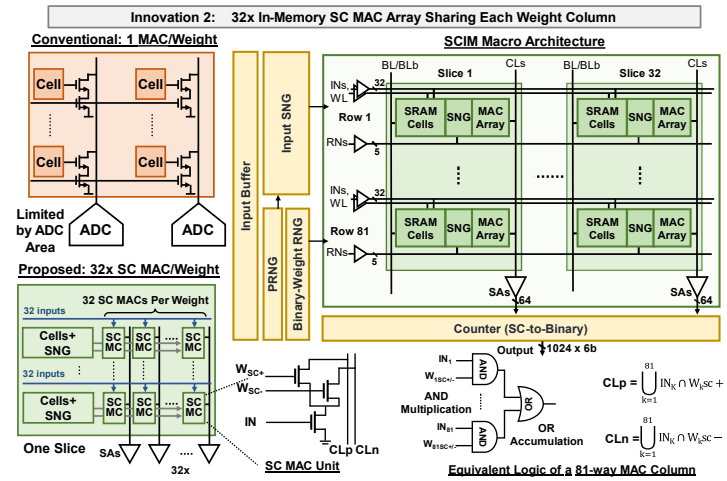


Fig.3 In-memory 32x SC MAC array shares each weight.

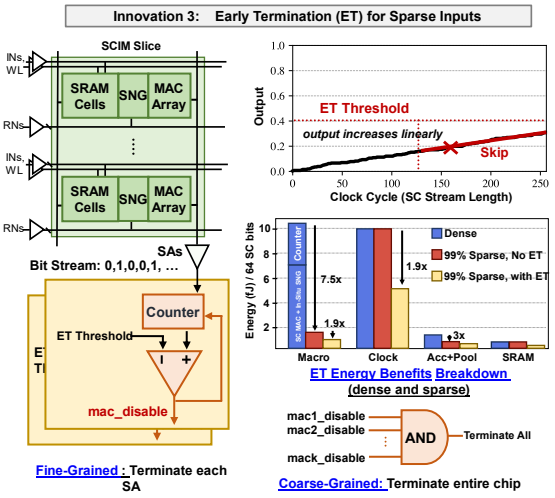
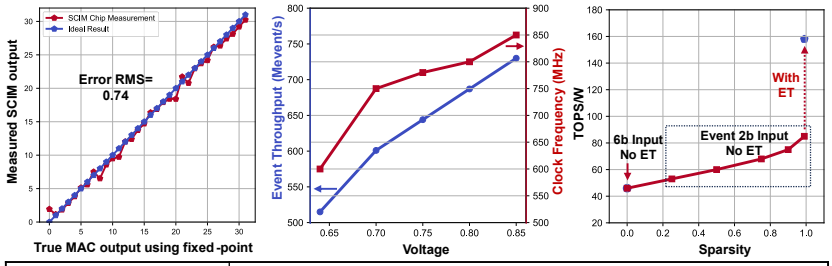


Fig.4 Early Termination skips unnecessary computes.



Dataset	DVS Event Camera, Input format: 2b, Pixel Array Size: 640x480 **Preprocessing:** Accumulated over time and down-sampled to 250 frames					
Tracking Algorithm	32 Gabor 3-D Convolutional Filters (9x9x2) + Max Pooling (column-wise)					
	Object Tracking Accuracy Metric (HOTA Score)				**Hardware Performance**	
	HOTA-0	Detection	Association	Localization	Throughput (Meveni/s)	Energy Efficiency (TOPS/W)
Software Fixed-Pt Simulation	94	36.9	39.2	50	Not Applied	Not Applied
SCIM Chip, No Early Termination	93.3	36.8	37.1	44.8	278 (1.0x)	System: 85, Macro: 267
SCIM Chip, ET Threshold: 0.125	92.5	36.4	36.4	44.5	425 (1.53x)	System: 131, Macro: 411
SCIM Chip, ET Threshold: 0.25	91.1	35.5	37.4	50	461 (1.66x)	System: 142, Macro: 445
SCIM Chip, ET Threshold: 0.375	93.3	37.2	22.4	43.9	514 (1.85x)	System: 158, Macro: 458
SCIM Chip, ET Threshold: 0.5	72	35.6	18.7	36.8	556 (2.0x)	System: 174, Macro: 545

Fig.5 Performance characterization and object tracking demonstration result.

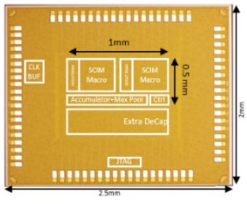


Fig.6 Die photo.

Table.1 Comparison with other works.

		Yan ISSCC22	Yue, ISSCC21	Yang, ASSCC22	Chang, ISSCC23	This Work	
Application		ML	CNN	CNN	Event Tracking	Event Tracking	
Technology		28nm SRAM Digital CIM	65nm SRAM Analog CIM	65nmSRAM Stochastic CIM	28nm RRAM CIM + NM	12nm, SRAM SCIM	
Area (mm²)		0.03	8.3	9.4	20.3	0.5 (Core)	
Voltage (V)		0.45-1.1	1	0.75	0.9	0.64	
Clock (MHz)		333	100	10	100	**600**	
Sparsity Support		**No**	**1.4x Boost**	**No**	**No**	**1.9x Boost**	
Sparsity Level		80%-99% *1				**Event 99%**	**Frame 0%**
Energy Efficiency (TOPS/W)	**Op** *2	8b	4b	8b	4b	6b x 2b	6b x 6b
	Macro	**27.4**	**24**	**20**		**267 ~ 495(ET)**	**86**
	System		**9**	**7.96**	**73.5**	**85 ~158(ET)**	**46**
Throughput		**5.4 GOPS**	**2 TOPS**	**0.06 TOPS**	**11.1M event/s 14.7 TOPS**	**278~514M event/s**	**3.4 TOPS**
Throughput Density		**0.18 TOPS/mm²**	**0.24 TOPS/mm²**	**0.0064 TOPS/mm²**	**3.7M event/mm² 0.86 TOPS/mm²**	**556~1112M event/mm²**	**6.8 TOPS/mm²**