29.5 A 73.53TOPS/W 14.74TOPS Heterogeneous RRAM In-Memory and SRAM Near-Memory SoC for Hybrid Frame and Event-Based Target Tracking

Muya Chang*[1], Ashwin Sanjay Lele*[1], Samuel D. Spetalnick[1], Brian Crafton[1], Shota Konno[1], Zishen Wan[1], Ashwin Bhat[1], Win-San Khwa[2], Yu-Der Chih[3], Meng-Fan Chang[2], Arijit Raychowdhury[1]

[1]Georgia Institute of Technology, Atlanta, GA
[2]TSMC Corporate Research, Hsinchu, Taiwan
[3]TSMC Design Technology, Hsinchu, Taiwan
*Equally Credited Authors (ECA)

2023 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067544

Vision-based high-speed target-identification and tracking is a critical application in unmanned aerial vehicles (UAV) with wide military and commercial usage. Traditional frame cameras processed through convolutional neural networks (CNN) exhibit high target-identification accuracy but with low throughput (hence low tracking speed) and high power. On the other hand, event cameras or dynamic vision sensors (DVS) generate a stream of binary asynchronous events corresponding to the changing intensity of the pixels capturing high-speed temporal information, characteristic of high-speed tracking. Such event streams with high spatial sparsity processed with bio-mimetic spiking neural networks (SNN) provide low power consumption and high throughput. However, the accuracy of object detection using such event cameras and SNNs is limited. Thus, a frame pipeline with a CNN and an event pipeline with a SNN (Fig. 29.5.1) possess complementary strengths in capturing and processing the spatial and temporal details, respectively. Hence, a hybrid network that fuses frame data processed using a CNN pipeline with event data processed through an SNN pipeline provides a platform for high-speed, high-accuracy and low-power target-identification and tracking. To address this need, we present a fully-programmable heterogeneous ARM Cortex-based SoC with an in-memory low-power RRAM-based CNN and a near-memory high-speed SRAM-based SNN in a hybrid architecture with applications in high-speed target identification and tracking.

Figure 29.5.1 motivates the system architecture illustrating predation in animals that fuses visual and vestibular data with multiple accurate and parallel operations such as self-motion cancellation and prey detection. This motivates the design of an electronic vision system, where CNNs and SNNs provide complementary speed vs. accuracy advantages. Our approach envisions event-camera and SNN-based high-speed target tracking that sacrifices accuracy for speed and power while a periodically triggered, slower but reliable CNN-driven compute restores high target-identification accuracy. The spatio-temporal continuity of the noisy SNN outputs is compared against previous CNN outputs to determine whether the SNN output is usable for actuating the tracker. Design constraints include (1) an SNN to operate at a high speed to match the bandwidth of the DVS (>10M events/sec) for real-time operation, while (2) compute-hungry CNN cores need to save battery power until triggered for validation and to provide accurate detection. This unique set of requirements for the hybrid compute requires modality-matched domain-specific compute fabrics. Our SoC integrates (1) an SRAM-based highly localized compute near memory (CNM) based SNN for real-time operation and low data movement and (2) an RRAM-based non-volatile compute-in-memory (CIM) fabric for sporadically triggered CNNs to achieve energy savings through power-gating. High density and integration with the CMOS process make RRAM a viable non-volatile memory candidate. We use (1) two-level power gating for up to 95% energy savings on energy-intensive CNN; (2) a triple error correction (TEC) scheme for tolerating high variations in RRAM CIM operations to restrict the application-accuracy degradation to <1%; and (3) an SRAM-based parallel-operating processing elements (PE) for handling up to 11.1Mevents/s for real-time operation. An embedded processor accepts external sensor data and concurrently allocates it to the hybrid compute fabric.

Figure 29.5.2 shows the overall architecture of the system. An ARM Cortex M3 with dedicated SRAM and ROM acts as the central processor accepting frame, event and inertial data. The Cortex-M3 interfaces with the RRAM-CIM and SRAM-CNM modules for sensor data allocation and coordinates the SNN and CNN interplay. The RRAM-CIM unit contains 32 parallel-accessible PEs containing one RRAM module each. Each module contains 5 RRAM macros with independent read-write circuits, MAC computation, and error correction units. Two SNN modules containing 12 PEs store a total of 54kB of SRAM each in smaller banks processing the FIFO buffered event stream. The SNN carries out self-motion cancellation by fusing the event stream, depth information and inertial data to filter out events corresponding to the self-motion of the tracker and retaining only the events corresponding to the external movement. The weights in this filter are tuned using experimental insights from the compound eyes of houseflies and it is implemented as a 4-layer network. Layers 2 and 3 are partitioned into 64×64 patches, each processed on 1 PE using integrate and fire (IF) operation as shown. Layer 4 is implemented on the processor for calculating the spike histograms. Our SNN exceeds the bandwidth of a typical DVS allowing the output spikes to be processed at any target rate. We calculate SNN outputs every 10ms to generate 100 outputs per second.

The CNN computation is supported by five parallel-operating RRAM macros located within each of the 32 RRAM modules (Fig. 29.5.3). Both modules and macros can be hierarchically isolated from the power source to take advantage of non-volatility. One macro within each module can be used as both an error-correcting and a regular CIM macro depending upon the error compensation requirement. The outputs of the CIM are provided to MAC units. Buffers at input and output store the weights and activations during the operation. Each macro contains 8kB RRAM devices in 2-bitline, 1-sourceline architecture, and 16 lanes of 6b SAR ADCs. A multi-WL driver implements a tradeoff between overall throughput and accuracy. In the case of ADC precision overflow, the system can continue with the noisy, high-throughput data or revert the computation to reliable but lower throughput. Fig. 29.5.3 also summarizes the key architectural features.

Figure 29.5.4 characterizes the two-level power gating that ensures only the fraction of active macros dissipate power in the sequential processing of CNN. Power gating switches occupy a total 0.58% of the macro area and the 0.7% of the module area. The system communicates to the external sensors using a USB-connected programmable Python interface. The operation involves sending in the vision and inertial sensor data to the processor on the test chip through an off-chip MCU. The outputs of the SNN, CNN and fused target position can be received for verification and actuation of the tracking UAV. The measured passive power on every module and macro on the chip shows a tight distribution with a standard deviation of 36.8μW and 12μW for the macro and module, respectively. Two levels of power gating allow the macros and module to be sequentially activated with a linear dependence of power consumption on the required parallelism. The typical operation involves a continuously operating SNN, while the CNN modules are power-gated. This is compared to a case where 25% of the RRAM-CIM (CNN) modules are operational. Results indicate a 78% power reduction owing to power gating. The peak on-power consumed by the chip is dominated by the CNN modules, making the power-gating important. The CNN-only operation achieves 75.02TOPS/W while the overall chip achieves 73.53TOPS/W at 90 MHz.

Figure 29.5.5 shows the intended application with a tracking UAV. The operating region for SNN shows maximum throughput of 11.1Mevents/s. The on-chip area is dominated by the RRAM-CIM cores. The functional operation of the system illustrates the high throughput SNN with noisy outputs at each inference. A low spatio-temporal continuity within the SNN outputs and the most-recent CNN output indicates a high suspicion of incorrect detection. This triggers the CNN to provide reliable detection. The object detector CNN provides its output, which is used as the baseline in the next spatio-temporal continuity calculation. Until triggered, all RRAM-CIM modules remain fully power-gated consuming only 0.43mW. The modules are turned on sequentially to execute the layers of a compact version of the MobileNet-YOLO object detector with 8b quantization for weights and activations. The models fit on the on-chip memory utilizing 94% of the RRAM bits. Power-gating offers 91.8% power savings with a sporadically triggered CNN. A detection is considered accurate if the intersection over union for the bounding box exceeds 0.5. The CNN is trained on a 2000 image public dataset of UAV images. Variation-induced noise in the RRAM-CIM shows a measured bit error rate (BER) within each macro that causes 40.6% accuracy degradation without ECC. This is minimized to <1% with double error correction (DEC). A further increase in BER induced by endurance issues can also be compensated using TEC. The fused system achieves small accuracy degradation compared to CNN-only with 10× higher throughput and lower average power consumption for two CNN inferences/second. The high throughput of the SNN and power gating and ECC of the RRAM-CIM makes the system suitable for target tracking. Fig. 29.5.6 compares this work with prior art. Fig. 29.5.7 shows the die micrograph and performance specifications.

Acknowledgement:
MC, AL, SS, BC, AB, ZW, AR were supported by the Semiconductor Research Corporation under the C-BRIC, ASCENT and EXPERT grants. The authors would also like to thank TSMC for technical discussions and chip fabrication support.

References:
[1] C. Xue et al., “A 1Mb Multibit ReRAM Computing-in-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN-Based AI Edge Processors,” *ISSCC*, pp. 388-389, 2019.
[2] 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. 244-245, 2020.
[3] M. Chang et al., “A 40nm 60.64 TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25 MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems,” *ISSCC*, pp. 270-271, 2022.
[4] J. Park et al., “A 65nm 236.5nJ/Classification Neuromorphic Processor with 7.5% Energy Overhead On-Chip Learning Using Direct Spike-Only Feedback,” *ISSCC*, pp. 140-141, 2019.
[5] F. Buhler et al., “A 3.43 TOPS/W 48.9 pJ/Pixel 50.1 nJ/Classification 512 Analog Neuron Sparse Coding Neural Network with On-Chip Learning and Classification in 40nm CMOS,” *IEEE Symp. VLSI Circuits*, pp. C30-C31, 2017.
[6] S. Kim et al., “Neuro-CIM: A 310.4 TOPS/W Neuromorphic Computing-in-Memory Processor with Low WL/BL activity and Digital-Analog Mixed-mode Neuron Firing,” *IEEE Symp. VLSI Tech. and Circuits*, pp. 38-39, 2022.

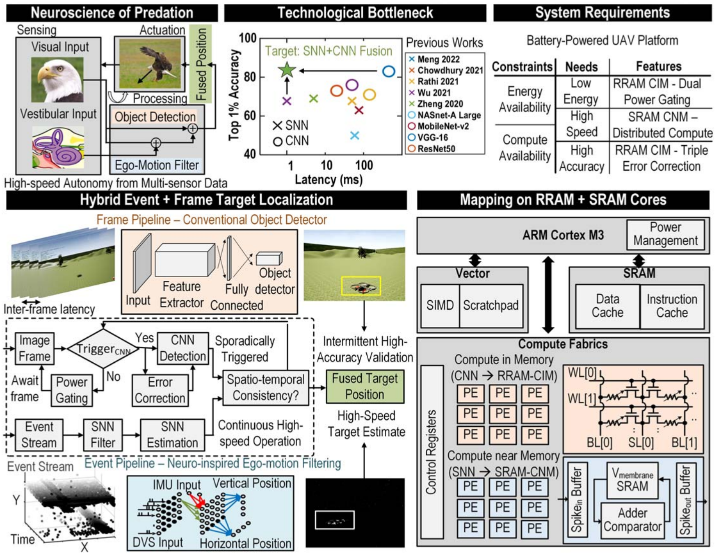


Figure 29.5.1: Satisfying requirements of hybrid network with an SNN mapped on SRAM-CNM and a CNN mapped on RRAM-CIM.

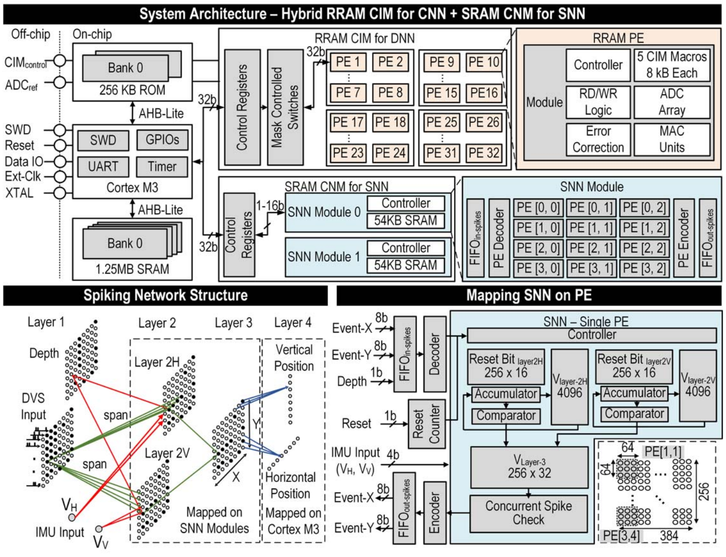
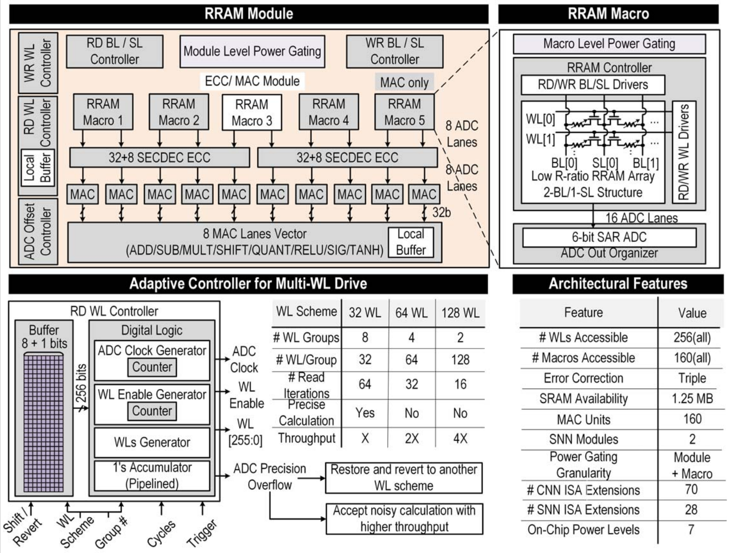


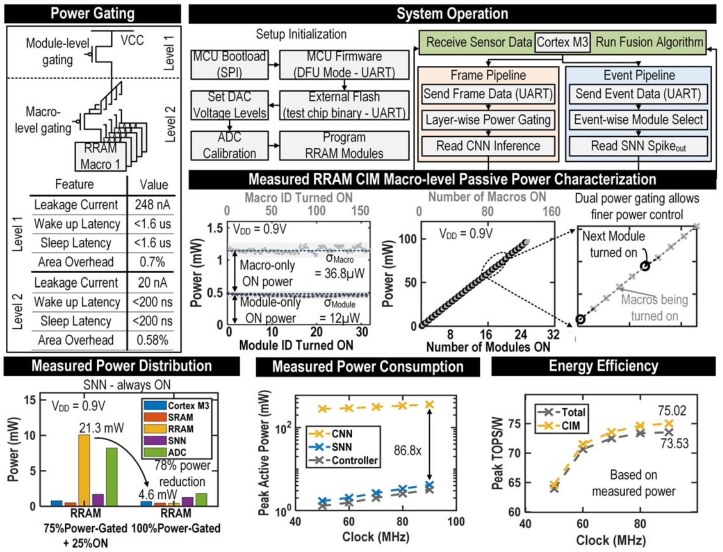
Figure 29.5.2: Overall SoC architecture with heterogeneous compute-cores; mapping SNN filter on distributed PEs.



WL Scheme	32 WL	64 WL	128 WL
# WL Groups	8	4	2
# WL/Group	32	64	128
# Read Iterations	64	32	16
Precise Calculation	Yes	No	No
Throughput	X	2X	4X

Feature	Value
# WLs Accessible	256(all)
# Macros Accessible	160(all)
Error Correction	Triple
SRAM Availability	1.25 MB
MAC Units	160
SNN Modules	2
Power Gating Granularity	Module + Macro
# CNN ISA Extensions	70
# SNN ISA Extensions	28
On-Chip Power Levels	7

Figure 29.5.3: Architecture of RRAM modules; WL controller for driving multiple WLs; architectural features.



	Feature	Value
Level 1	Leakage Current	248 nA
	Wake up Latency	<1.6 us
	Sleep Latency	<1.6 us
	Area Overhead	0.7%
Level 2	Leakage Current	20 nA
	Wake up Latency	<200 ns
	Sleep Latency	<200 ns
	Area Overhead	0.58%

Figure 29.5.4: Characterization of two-level power gating; operating setup; measured power and efficiency.

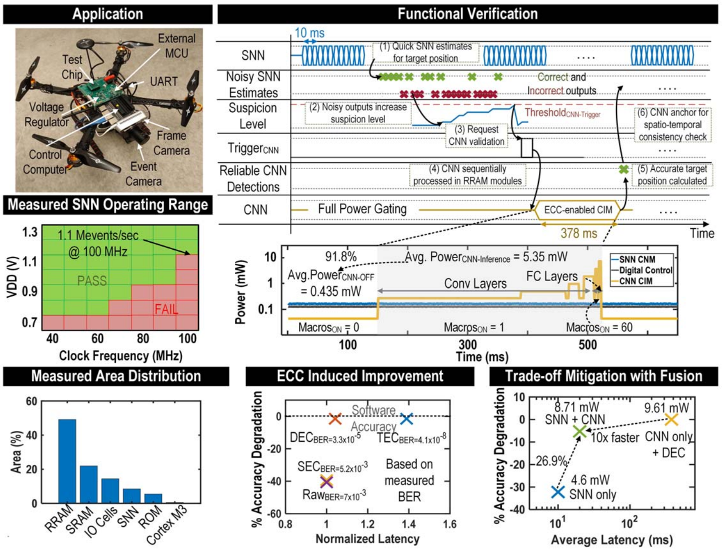


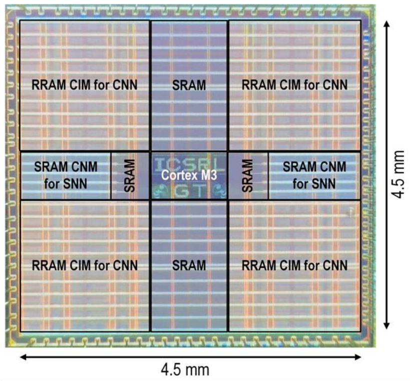
Figure 29.5.5: Application-level speed, power and accuracy with high throughput SNN, power-gated CNN with TEC.

Comparison Table

		ISSCC19 [1]	ISSCC20 [2]	ISSCC22 [3]	ISSCC19 [4]	VLSI17 [5]	VLSI22 [6]	This Work
System Features	Application	CNN CIM	CIM CNN	CIM CNN	MNIST Classify	MNIST Classify	CIFAR-10 Classify	**Event + Frame Target Identification and Tracking**
	Tech	55 nm	22 nm	40 nm	65 nm	40 nm	28 nm	**40 nm**
	On-chip SRAM	-	-	768 kB	4.8 kB	16 kB	32 kB	**1.25 MB**
	On-chip RRAM	128 kB	256 kB	2.25 MB	-	-	-	**1.25 MB**
	Nominal Supply	1	0.7-0.9	0.9	0.8	0.9	1.1	**0.9 V**
	Max Clock (MHz)	85**	NR*	200	20	250	200	**100**
	Embedded Proc	-	-	Yes	-	-	-	**Yes**
	Area (mm^2)	7.5	6	25	10.08	1.31	2.9	**20.25**
RRAM CIM Features	# Parallel Reads	NR*	NR*	20736				**81920**
	Sensing Mode	CSA	CSA	VSA				**CSA**
	Power Gating	-	-	Single				**Dual**
	ADC Resolution	3b	6-11b	4b				**6b**
	ECC	-	-	SEC				**TEC**
	BER with ECC	-	-	$1.34x10^{-6}$ (9 WL)				**10^{-12} (1 WL) $4.1x10^{-8}$(16 WL)**
SNN Features	Neuron				410 IF Digital	512 LIF Analog	1-8b MS	**1.96K IF Digital**
	Communication				Point-to-point	Bus-ring	Point-to-point	**Point-to-point**
System Performance	Efficiency (TOPS/W)	53.17	121.38	60.64	3.42	3.43	62.1	**73.53 (peak)**
	Avg Power (mW)	NR*	NR*	CIM_{Off}=2.6 CIM_{ON}=23.6	23.6	87	105.4-241.4	**SNN_{ONLY} = 4.6 $CIM_{25\%ON}$= 21.3 $SNN+CIM_{inference}$=8.71**
	Max Throughput	NR*	NR*	CNN Only 8.29 TOPS	78.4 Mpix/s	1778 Mpix/sec	NR*	**SNN: 11.1 Mevents/s CNN: 14.74 TOPS**

-: None **: Estimated NR*: Not reported

Figure 29.5.6: Previous work on vision applications where an RRAM-CIM approach was applied in a CNN and SNN architecture.



System Performance	
Peak TOPS	14.74
Peak TOPS/W	73.53
SNN Throughput	11.1 Mevents/ sec
System Throughput	100 outputs/sec
SNN + CIM_{Off}	4.6 mW
SNN + 25%CIM_{ON}	21.3 mW
BER w/o ECC	$7x10^{-3}$
BER with ECC	$4.1x10^{-8}$

System Features			
Technology	40 nm ULP	Microprocessor	Cortex M3
Chip Size	4.5 mm x 4.5 mm	Number of IO	62
Package	QFN 64	Communication	UART
On-chip RRAM	1.25 MB	Voltage Levels	7
On-chip SRAM	1.25 MB	IO Supply	3.3 V
Max Clock	100 MHz	Core Nominal Supply	0.9 V

Figure 29.5.7: Die micrograph and test chip characteristics.