16.3 A 40nm 60.64TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems

Muya Chang[1], Samuel D. Spetalnick[1], Brian Crafton[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

2022 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731679

Resistive RAM (RRAM) is an exciting technology that exhibits various new properties that have been long absent in traditional charge-based memories. RRAM features high-bit density, non-volatile storage, accurate compute in-memory (CIM), and both process and voltage compatibility. Each of these properties makes RRAM a compelling candidate for AI applications, particularly at the edge. To demonstrate the utility of these properties, we direct our effort to real-world event-driven and memory-constrained applications, such as recommendation systems and natural language processing (NLP). To enable these applications at the edge, higher memory capacity and bandwidth must be achieved despite irregular data access patterns that prevent effective caching and data reuse. Furthermore, we find that these applications are rarely (if ever) run continuously, but instead execution is triggered by events. The combination of these two challenges makes RRAM an ideal candidate given its high density and non-volatility enabling near-zero leakage power and complete power down. To address these challenges, this paper presents a 2.25MB RRAM based CIM accelerator with 765kB of SRAM and an embedded Cortex M3 processor for edge devices.

In Fig. 16.3.1, we show the application flow required for a heterogeneous AI application featuring a recommendation system as the key workload. Next, we detail the architecture and software kernel for executing the application. The Cortex M3 receives events in the form of notifications or speech and initiates inference on the RRAM processing elements. The neural network model is distributed across the RRAM (shallow layers) and the SRAM (last ~20% of the network). Then training is performed in the last layer(s) when feedback is received from the user. Training is only performed in SRAM using CMOS SIMD units to limit writes to the RRAM and thus provide both energy and performance advantages while preserving the limited endurance of the RRAM cells. If sufficient time has elapsed since the last event, RRAM is completely powered down to enable up to 89.21% power reduction. Lastly, we present the CIM core with ECC for both sparse length sum (SLS) operation and vector-matrix multiplication (VMM). Embedding tables and dense layers are mapped to the same array and partitioned to maximize throughput. To overcome device-level challenges such as variation and resistance drift, we implement a novel CIM ECC scheme to detect, localize, and correct soft errors that requires only 12.5% memory overhead.

Figure 16.3.2 illustrates a complete system which supports full software programmability with: (1) A centralized Cortex M3 microprocessor capable of running up to 200MHz with 128KB ROM for storing the application image and 512KB SRAM for the application to use. (2) 288 fully integrated RRAM modules via AHB-Lite with 8KB RRAM cells per module which supports 1-to-8b inputs/weights and 1-to-32b output over 1-to-8 clock cycles. The RRAM modules are selected based on a 9b mask and a 9b target index, the unselected ones can be completed turned off through power gating. (3) A fully integrated vector module via AHB-Lite which contains a 128KB SRAM inside to store intermediate results, and 8 sets of ALUs capable of various functions. (4) 32 GPIOs and Serial-Wire-Debug (SWD) port for external communication/debugging. For dataflow, the application image as well as the workload data are first stored in an external dataflash and transferred into the testchip via SWD when the system starts. Once the system starts, the Cortex M3 sends the first set of inputs to the selected RRAM modules, afterwards the intermediate data is transferred between the RRAM modules and the Vector module to maximize throughput.

Figure 16.3.3 shows the detail of the RRAM module. The inputs/outputs can be divided into different categories: (1) Power control. (2) Targeted address. (3) Read configurations. (4) Write configurations. (5) MAC configurations. (6) Fault-tolerance configurations. As briefly mentioned previously, for the unselected RRAM modules, the power gates can be turned off to minimize power consumption. For read operations, we can turn on 1-to-9 word lines simultaneously with the desired cycles with the tradeoff between throughput and the sensing accuracy. For write operations, besides the digitally adjustable WL/WR voltages which will be discussed later, similar to the read operations, we can control the pulse width by changing the targeted cycle. The MAC unit supports both signed/unsigned operations and is capable of handling 1-to-8b number formats.

To overcome inherent device variation in RRAM that yields sum-of-products errors in CIM, we incorporate a single error detection and single error correction (SECSED) ECC scheme in five steps: (1) Encoding is performed by appending a single parity bit to each weight. (2) Single errors are detected when the LSB of the ADC checksum does not match the parity bit. (3) If an error is detected, we read one word line at a time to avoid the sum-of-product error and correct the CIM error. As a result, we pay the penalty of serializing the read operation temporarily. (4) We localize the BL where the sum-of-product error has occurred by comparing the CIM result and the correct serial result. (5) Lastly, victim BLs are tracked with a status register and regularly refreshed to combat resistance drift and variation.

In Fig. 16.3.4, we illustrate physical design considerations, power plan, software programmability, and the testing procedure. Due to the considerable number of RRAM modules, proper physical design becomes critical in terms of power delivery, balanced delay, and routing congestion. To minimize the routing congestion and balance the delay between the Cortex M3 and all the RRAM modules, placing the microprocessor at the center of the chip turned out to be the best solution. To minimize the IR drop and get the best power delivery, multiple hierarchical power rings/stripes and layers as listed in the table are used. In addition, multiple voltage domains are separated for cleaner power delivery and measuring power consumption for each functional block. To provide a simple and complete platform, a complete application programming interface (API) is developed, as shown in the table. Furthermore, we develop an evaluation board which is capable of digitally adjusting reference voltages, power voltages via SPI and I2C, and is easily controllable through a UART interface. The system operation comprising: (1) Kickoff, (2) DFU mode, (3) testchip initialization, and (4) testchip application are shown in the figure.

In Fig. 16.3.5, we first demonstrate the impact of power gates. To make the system low power, a set of dedicated power gates are integrated inside each RRAM module and the measured power consumption with different number of RRAM modules turned on is shown in the figure. The average measured power consumption per RRAM module is 70μW. The power distribution with all RRAM modules turned on/off, as well as the area distribution are shown. The results show that the RRAM modules occupy the maximum amount of area, and are also responsible for 89.37% of total power when all of them are turned on. Next we show measured compute in-memory error and its impact on application performance. Each bin shows the percentage of actual ADC output codes obtained for the expected ADC output code. When the number of LRS cells is low (< 4), the result is always correct for the experiment's sample size (8192 total). When more LRS cells are read, errors occur with increasing frequency. However, we note that errors are always constrained to ±1 errors (i.e., |measured – expected ADC code| <= 1). This property has special implications for both error correction and detection. Like traditional single-error detection, a ±1 can be detected using a single parity bit that we demonstrate in Fig. 16.3.3. With this data, we simulate the Neural Collaborative Filtering (NCF) model on the standard MovieLens dataset [1]. We show both BER and the standard hit-rate (HR) metric using 2 write configurations and ECC.

Figure 16.3.6 shows the measured energy efficiency plot for binary IN/W along with the table showing energy efficiency under other different IN/W/Sparsity configurations. Under binary IN/W with no sparsity and 0.9V power supply, we measured 60.64TOPS/W @192MHz. The second figure shows the block diagram on the PCB board along with a table listing the component models being used. A comparison with state-of-the-art CIM architectures [2-7] illustrates competitive metrics, while addressing key technological challenges. The die-shot and the chip-characteristics are shown in Fig. 16.3.7.

Acknowledgement:
MC, SS, BC, and AR were supported by the Semiconductor Research Corporation under the Center for Brain-Inspired Computing (C-BRIC) under Grant 2777.005 and 2777.006, and the Applications and Systems-driven Center for Energy-Efficient Integrated Nano Technologies (ASCENT) under Grant 2776.037. The authors would also like to thank TSMC for technical discussions and chip fabrication support.

References:

[1] U. Gupta et al., "The Architectural Implications of Facebook's DNN-Based Personalized Recommendation," *IEEE HPCA*, pp. 488-501, 2020.
[2] W.-H. Chen et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with Sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, 2018.
[3] C.-X. 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.
[4] C.-X. 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.
[5] W. Wan et al., "A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," *ISSCC*, pp. 498-499, 2020.
[6] J. Wang et al., "A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration," *ISSCC*, pp. 224-225, 2019.
[7] J. -H. Yoon et al., "A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification," *ISSCC*, pp. 404-405, 2021.

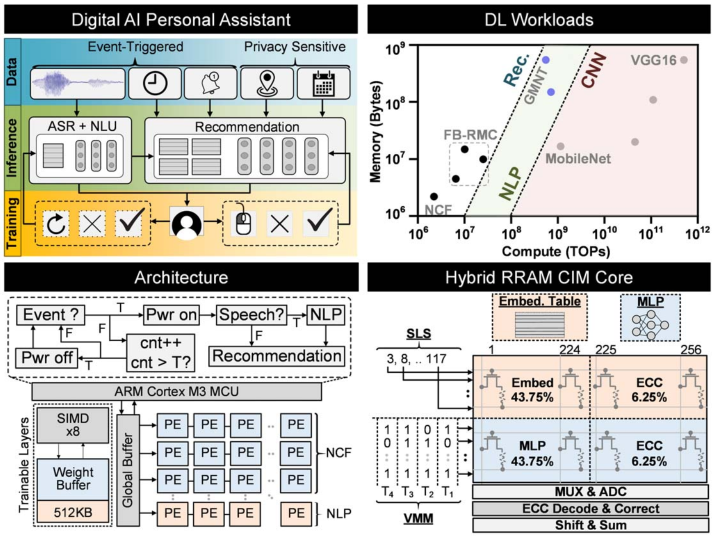


Figure 16.3.1: Motivation and overall architecture of the proposed programmable hybrid digital/CIM RRAM system.

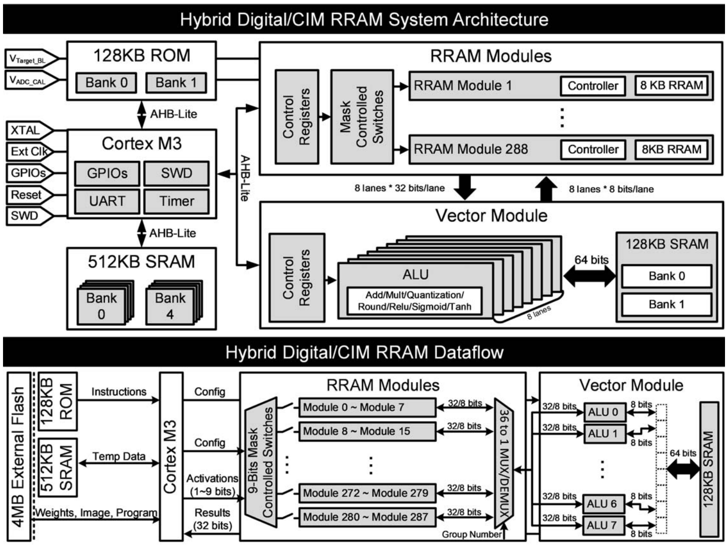


Figure 16.3.2: Hybrid digital/CIM RRAM system architecture and dataflow.

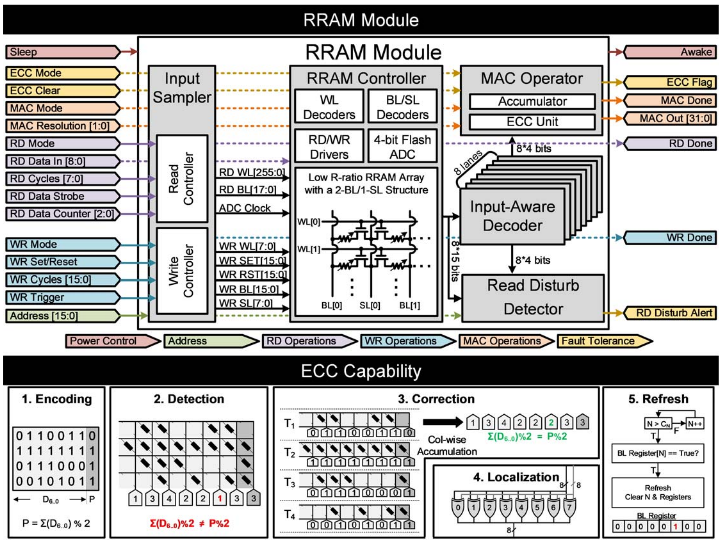
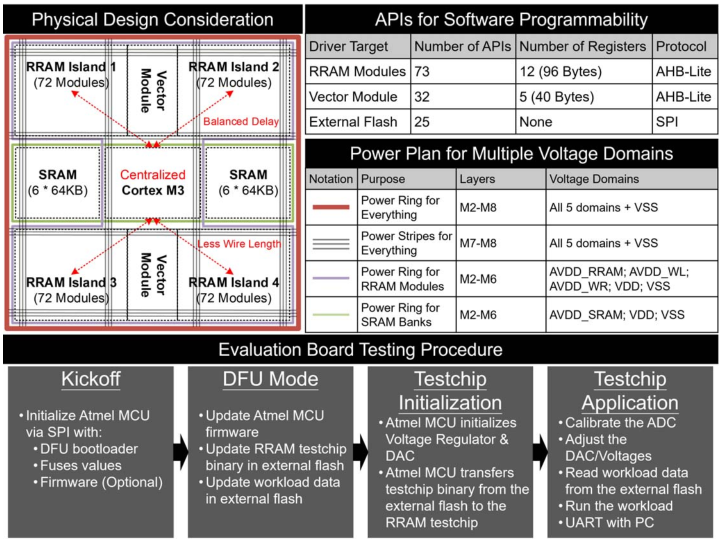


Figure 16.3.3: Proposed RRAM module with ECC capability.


Physical Design Consideration
RRAM Island 1 (72 Modules)
RRAM Island 2 (72 Modules)
RRAM Island 3 (72 Modules)
RRAM Island 4 (72 Modules)
Vector Module
SRAM (6 * 64KB)
Centralized Cortex M3
Balanced Delay
Less Wire Length

APIs for Software Programmability

Driver Target	Number of APIs	Number of Registers	Protocol
RRAM Modules	73	12 (96 Bytes)	AHB-Lite
Vector Module	32	5 (40 Bytes)	AHB-Lite
External Flash	25	None	SPI

Power Plan for Multiple Voltage Domains

Notation	Purpose	Layers	Voltage Domains
🟥	Power Ring for Everything	M2-M8	All 5 domains + VSS
≡	Power Stripes for Everything	M7-M8	All 5 domains + VSS
🟪	Power Ring for RRAM Modules	M2-M6	AVDD_RRAM; AVDD_WL; AVDD_WR; VDD; VSS
🟩	Power Ring for SRAM Banks	M2-M6	AVDD_SRAM; VDD; VSS

Evaluation Board Testing Procedure

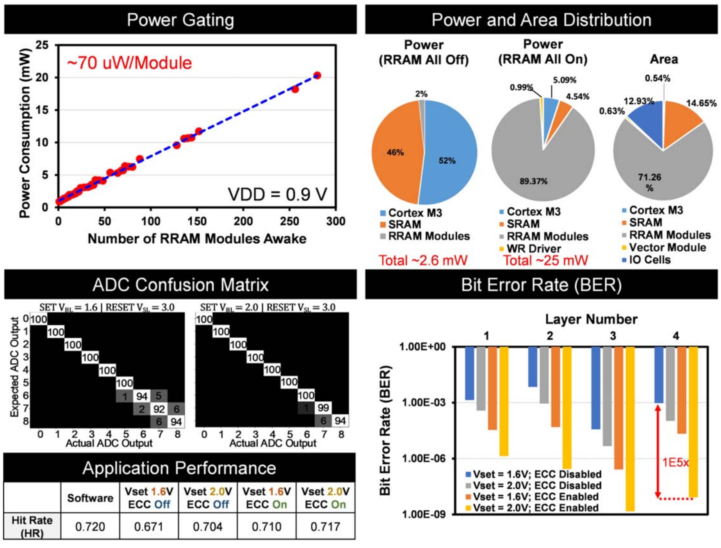
Kickoff
- Initialize Atmel MCU via SPI with:
  - DFU bootloader
  - Fuses values
  - Firmware (Optional)

DFU Mode
- Update Atmel MCU firmware
- Update RRAM testchip binary in external flash
- Update workload data in external flash

Testchip Initialization
- Atmel MCU initializes Voltage Regulator & DAC
- Atmel MCU transfers testchip binary from the external flash to the RRAM testchip

Testchip Application
- Calibrate the ADC
- Adjust the DAC/Voltages
- Read workload data from the external flash
- Run the workload
- UART with PC

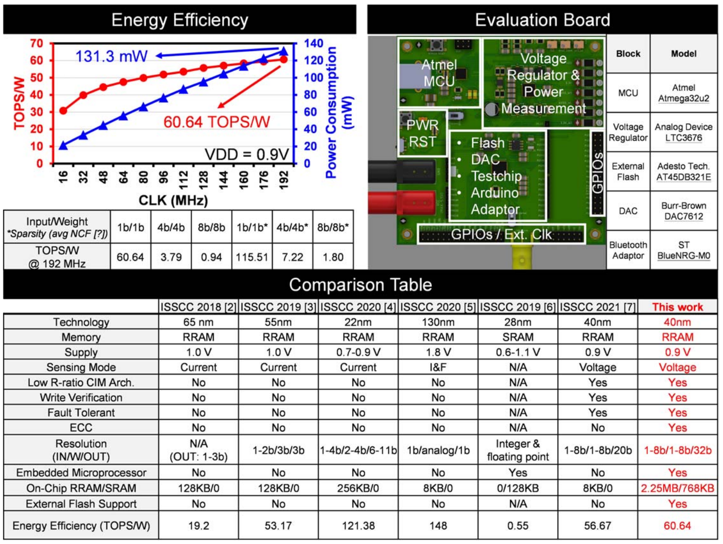
Figure 16.3.4: Software programmability, physical design considerations and evaluation board testing procedure.



Application Performance

	Software	Vset 1.6V ECC Off	Vset 2.0V ECC Off	Vset 1.6V ECC On	Vset 2.0V ECC On
Hit Rate (HR)	0.720	0.671	0.704	0.710	0.717

Figure 16.3.5: Measured result of power gating, power and area distribution, ADC confusion matrix, bit error rate (BER) and application performance.



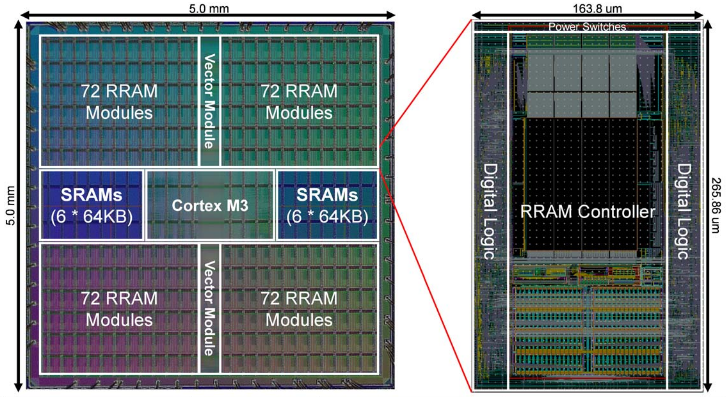
Input/Weight *Sparsity (avg NCF [?])	1b/1b	4b/4b	8b/8b	1b/1b*	4b/4b*	8b/8b*
TOPS/W @ 192 MHz	60.64	3.79	0.94	115.51	7.22	1.80

Block	Model
MCU	Atmel Atmega32u2
Voltage Regulator	Analog Device LTC3676
External Flash	Adesto Tech. AT45DB321E
DAC	Burr-Brown DAC7612
Bluetooth Adaptor	ST BlueNRG-M0

Comparison Table

	ISSCC 2018 [2]	ISSCC 2019 [3]	ISSCC 2020 [4]	ISSCC 2020 [5]	ISSCC 2019 [6]	ISSCC 2021 [7]	This work
Technology	65 nm	55nm	22nm	130nm	28nm	40nm	40nm
Memory	RRAM	RRAM	RRAM	RRAM	SRAM	RRAM	RRAM
Supply	1.0 V	1.0 V	0.7-0.9 V	1.8 V	0.6-1.1 V	0.9 V	0.9 V
Sensing Mode	Current	Current	Current	I&F	N/A	Voltage	Voltage
Low R-ratio CIM Arch.	No	No	No	No	N/A	Yes	Yes
Write Verification	No	No	No	No	N/A	Yes	Yes
Fault Tolerant	No	No	No	No	N/A	Yes	Yes
ECC	No	No	No	No	N/A	No	Yes
Resolution (IN/W/OUT)	N/A (OUT: 1-3b)	1-2b/3b/3b	1-4b/2-4b/6-11b	1b/analog/1b	Integer & floating point	1-8b/1-8b/20b	1-8b/1-8b/32b
Embedded Microprocessor	No	No	No	No	Yes	No	Yes
On-Chip RRAM/SRAM	128KB/0	128KB/0	256KB/0	8KB/0	0/128KB	8KB/0	2.25MB/768KB
External Flash Support	No	No	No	No	N/A	No	Yes
Energy Efficiency (TOPS/W)	19.2	53.17	121.38	148	0.55	56.67	60.64

Figure 16.3.6: Measurement results of energy efficiency, evaluation board diagrams and comparison table.



Technology	TSMC 40nm ULP	**General Purpose IOs**	32 (16*2)
Chip Size	5 mm x 5 mm	**Debug Interface**	JTAG / Serial Wire(SW)
Package	QFN7x7-60	**Voltage Domains**	6 + VSS
Embedded Microprocessor	Cortex M3	**Low Power Design Technique**	Clock gating / Power gating
Number of RRAM Module	288	**Clock Source / Max. Clock Rate**	Crystal or External / 200 MHz
On-Chip RRAM / SRAM	2.25 MB / 768 KB	**Core / IO Supply Voltage**	0.9 V / 3.3 V

Figure 16.3.7: Micrograph of the test-chip and summary of performance.