14.3 A 43pJ/Cycle Non-Volatile Microcontroller with 4.7μs Shutdown/Wake-up Integrating 2.3-bit/Cell Resistive RAM and Resilience Techniques

Tony F. Wu[1], Binh Q. Le[1], Robert Radway[1], Andrew Bartolo[1], William Hwang[1], Seungbin Jeong[1], Haitong Li[1], Pulkit Tandon[1], Elisa Vianello[2], Pascal Vivet[2], Etienne Nowak[2], Mary K. Wootters[1], H.-S. Philip Wong[1], Mohamed M. Sabry Aly[3], Edith Beigne[2], Subhasish Mitra[1]

[1]Stanford University, Stanford, CA
[2]CEA-LETI-MINATEC, Grenoble, France
[3]Nanyang Technological University, Singapore, Singapore

Non-volatility is emerging as an essential on-chip memory characteristic across a wide range of application domains, from edge nodes for the Internet of Things (IoT) to large computing clusters. On-chip non-volatile memory (*NVM*) is critical for low-energy operation, real-time responses, privacy and security, operation in unpredictable environments, and fault-tolerance [1]. Existing on-chip NVMs (e.g., Flash, FRAM, EEPROM) suffer from high read/write energy/latency, density, and integration challenges [1]. For example, an ideal IoT edge system would employ *fine-grained temporal power gating* (i.e., shutdown) between active modes. However, existing on-chip Flash can have long latencies (> 23ms latency for erase followed by write), while inter-sample arrival times can be short (e.g., 2ms in [2]).

Our chip monolithically integrates two heterogeneous technologies: 18KB of on-chip resistive RAM (emerging on-chip NVM, technology details in Fig. 14.3.1) on top of commercial 130nm silicon CMOS (16b general-purpose microcontroller core with 8KB of SRAM). For various applications (in machine learning, control, and cryptography), we demonstrate active mode average energy of 43pJ/cycle (up to 5.7× lower vs. similar chips at similar speeds / technology nodes using on-chip Flash and FRAM), fine-grained temporal power gating (0.25μW during shutdown) with up to 8μs (average 4.7μs) transition from active to shutdown mode (up to 5,878× quicker vs. on-chip Flash), and 2-clock cycle (200ns) transition from shutdown to active mode. We also demonstrate a complete chip that stores multiple bits per on-chip RRAM cell (5 resistance values, i.e., 2.3b per cell) and processes stored information correctly (vs. previous demonstrations using standalone RRAM cells or few cells in standalone RRAM array). Such multi-bit storage improves the accuracy of neural network inference (2.3× for MNIST) on same hardware (vs. 1b per cell).

RRAM (like other emerging NVMs, such as phase change memory) exhibits write failures [1]. We overcome these challenges through the critical combination of two resilience techniques: 1) *dynamic address remapping*, which overcomes write failures during system operation with 0.5% active-mode energy increase and negligible execution time impact; 2) periodic *ENDUrance REsiliency using random Remapping (ENDURER* – Fig. 14.3.5*)* [3] – a new technique implemented here. This combination enables our chip to achieve a 10-year functional lifetime when running MNIST inference continuously.

To demonstrate fine-grained temporal power gating enabled by on-chip RRAM, our chip operates as follows (Fig. 14.3.1). During *active mode*, instructions are read from the on-chip 12KB instruction RRAM and executed by the microcontroller core (MSP430 instruction set). During this time, data is accessed from peripheral ports (e.g., off-chip sensors), on-chip 4KB data RRAM, or on-chip 8KB scratchpad SRAM (loop counters, temporary variables with repeated writes: memory-mapped using the compiler). After the data is processed, to transition to *shutdown mode*, results are written back to the 4KB on-chip data RRAM (consuming 168pJ over 5 clock cycles per 16b word, Fig. 14.3.2) and the hardware scheduler unit power-gates (i.e. turns off power) the core, memory controllers, and memory. Our chip performs this transition 5,878× quicker than those with on-chip Flash due to the low write latency of RRAM (500ns vs 23ms for Flash). The chip returns to active mode upon data arrival (e.g., from sensors).

We run 5 applications representing machine learning (logistic regression, support vector machine, convolutional neural network), control (Kalman filter) and cryptography (SHA256 hash) to demonstrate the effectiveness of our chip (Fig. 14.3.2). To put our results into perspective, we select a similar clock rate for our chip (10MHz, vs. industry chips with existing on-chip NVM such as FRAM and Flash) that is sufficient for fine-grained temporal power-gating, while avoiding excessive energy consumption. The active mode power of our chip varies between 407μW to 477μW (average active mode energy: 43pJ/cycle). We achieve average 4.7μs/1.6nJ transition from active to shutdown mode and a 200ns/152pJ transition from shutdown to active mode (Fig. 14.3.2). Although the industry chips might be engineered to include additional margins, the overall benefits demonstrated by our chip are expected to stay significant even after margins are taken into consideration.

We store multiple resistance levels (up to 5 in our chip) inside on-chip RRAM cells (e.g., neural network model weights, only read during inference) by special algorithms that change wordline voltage (V_{WL}) and bitline voltage (V_{BL}) in addition to modifying the pulse width (Fig. 14.3.3) and allocating larger resistance windows for levels with higher resistance values. With greater effective memory capacity (2.3b vs. 1b per RRAM cell) on the same hardware, higher-precision weights (e.g., 4b vs 8b) or larger neural network models (e.g., 6,490 vs. 9,402 weights) can be used (Fig. 14.3.3). Despite errors (cells with resistance values outside its intended resistance window) in 5 levels-per-cell storage, we achieve a 2.3× improvement in inference accuracy (i.e., 2.3× decrease in inference error) for neural networks (on the MNIST dataset, Fig. 14.3.3) when the weights are encoded as follows: two 5-level cells for magnitude and one 2-level cell for sign bit.

RRAM is subject to temporary write failures (*TWFs*) and permanent write failures (*PWFs*, resulting in limited *endurance*: maximum number of successful writes to a cell) [4] that degrade application accuracy over time (Fig. 14.3.4). Cell-level parameter adjustment to improve write failures is not sufficient [4]. To address TWFs, we employ a write-verify scheme with retries [4]. If a write to an RRAM address is unsuccessful after 4 retries, we map that address (during runtime) to another location in a separate backup RRAM array using *dynamic address remapping* (Figs. 14.3.1, 14.3.4). Our chip contains a backup RRAM array (256 16b words) for every 4KB of RRAM; 128 words of that backup array are used for this mapping. The mapping information is stored in a 128-entry volatile look-up table (*volatile LUT*, implemented using flip-flops, Fig. 14.3.1). During transition from active to shutdown mode, the contents of each volatile LUT are stored in the remaining 128 words of the corresponding backup array (*non-volatile LUT*). A write failure to a non-volatile LUT entry results in that entry marked invalid (majority vote over 5 RRAM bits decides entry validity). When the chip boots, the contents of the volatile LUTs are loaded from the corresponding non-volatile LUT. We use dynamic address remapping for our data RRAM, incurring 0.5% energy and negligible (0.005%) execution time costs; our data RRAM tolerates TWFs and PWFs in 17.3% and 2% of words, respectively (Fig. 14.3.4). We use stronger programming conditions (higher voltage, more retries) to mitigate TWFs and insert dummy instructions to avoid PWFs in instruction memory (as writes occur only during programming).

Despite limited write endurance of the 4KB data RRAM, we achieve 10-year lifetime using ENDURER (Fig. 14.3.5, software on FPGA + our chip) combined with dynamic address remapping, when running our neural network application (MNIST dataset) continuously (Fig. 14.3.6). We accelerate our tests to account for 10 years of running an application by first obtaining a sequence of all writes to RRAM (which account for 258 out of 617,669 total memory operations for a single inference) for the application. Then, we repeatedly perform the sequence of writes, through the ENDURER module on the FPGA, on the RRAM (skipping any read operations, writes to non-RRAM, and computation to save time). In our implementation of ENDURER, remapping is performed every 30 minutes and we use an SRAM buffer of 8 16b words.

On-chip RRAM NVM enables significantly lower energy during active mode (vs. existing on-chip NVM such as Flash and FRAM), fine-grained temporal power gating, and multiple bits per RRAM cell. Correct computation using multi-bit RRAM cells in a complete chip successfully improves neural network inference accuracy. Effective resilience techniques enable chips with on-chip RRAM to achieve 10-year lifetime (for neural network inference applications) despite write failures in the underlying RRAM. Our results can be further enhanced through domain-specific accelerators, bit-cost scalable 3D Vertical RRAM [5], and monolithic 3D integration of multiple RRAM layers [5]. The presented techniques (fine-grained temporal power gating, resilience) may be used for other emerging on-chip NVM (e.g., phase change) technologies as well.

Acknowledgements:
Work supported in part by DARPA, NSF/NRI/GRC E2CDA, and the Stanford SystemX Alliance.

References:
[1] A. Chen, "A Review of Emerging Non-Volatile Memory (NVM) Technologies and Applications," *Solid-State Electronics*, vol. 25, pp. 25-38, 2016.
[2] R. Braojos, et al., "Nano-Engineered Architectures for Ultra-Low Power Wireless Body Sensor Nodes," *CODES+ISSS*, 2016.
[3] M. M. S. Aly, et al., "The N3XT Approach to Energy-Efficient Abundant-Data Computing," *Proc. IEEE*, 2019.
[4] A. Grossi, et al. "Fundamental Variability Limits of Filament-based RRAM," *IEDM*, pp. 4.7.1-4.7.4, 2016.
[5] H.-S. P. Wong, et al., "Memory Leads Way to Better Computing," *Nat. Nanotech.*, vol. 10, pp 191-194, 2015.

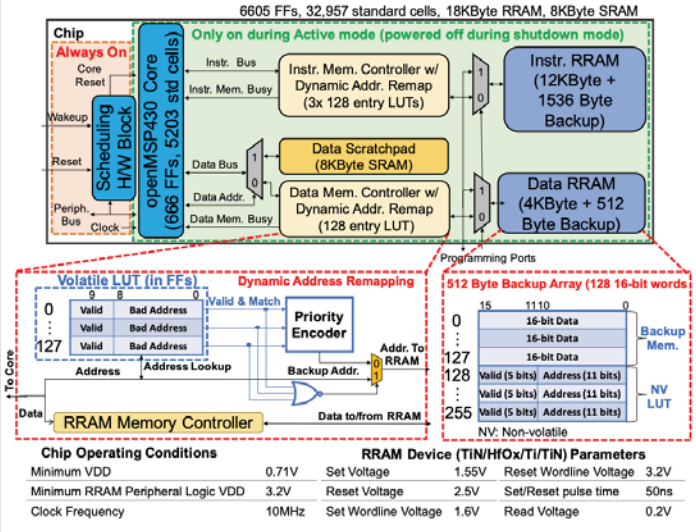


Figure 14.3.1: Block diagram of our chip.

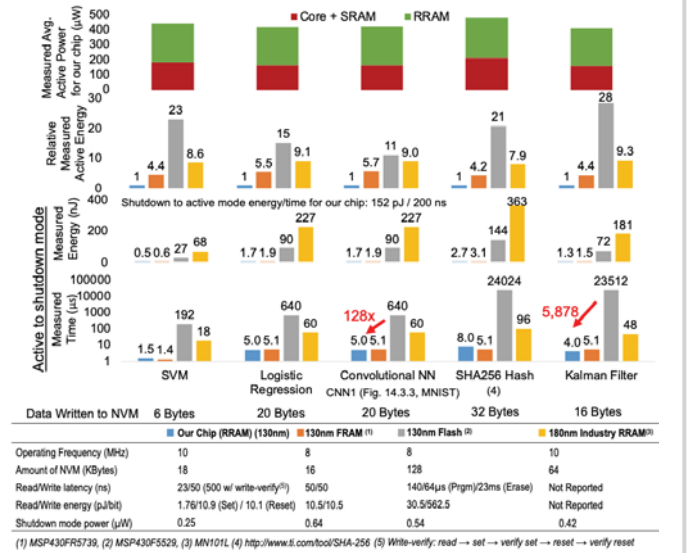


Figure 14.3.2: Benefits of using our chip with on-chip RRAM.

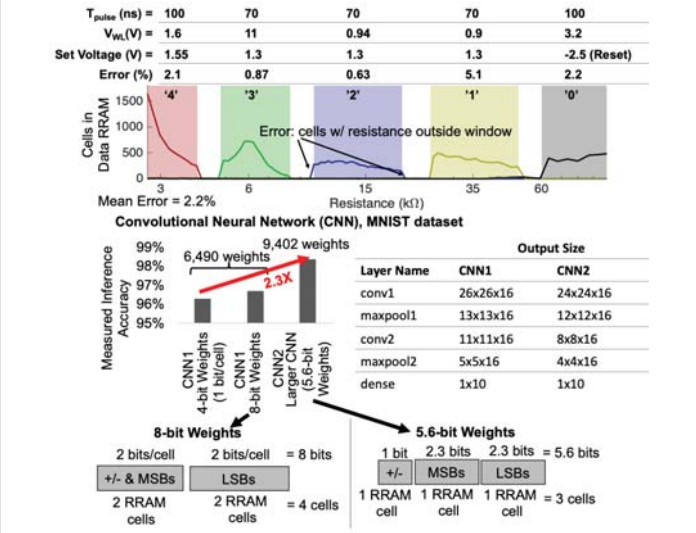


Figure 14.3.3: Using 2.3 bits per RRAM cell for convolutional neural network applications.

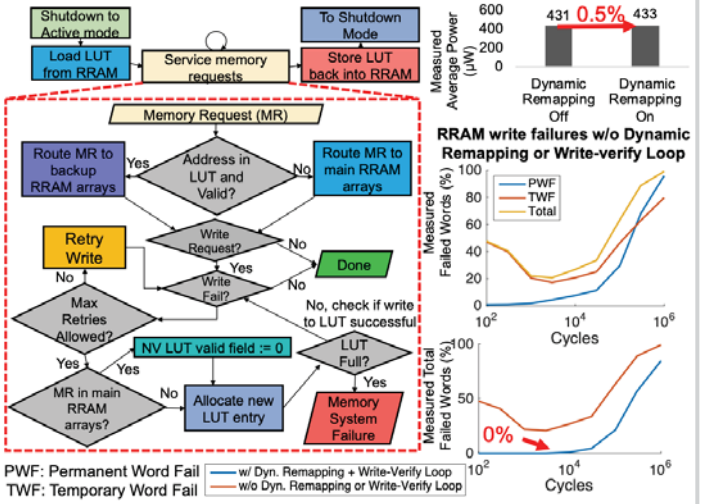


Figure 14.3.4: Dynamic address remapping with write-verify loop.

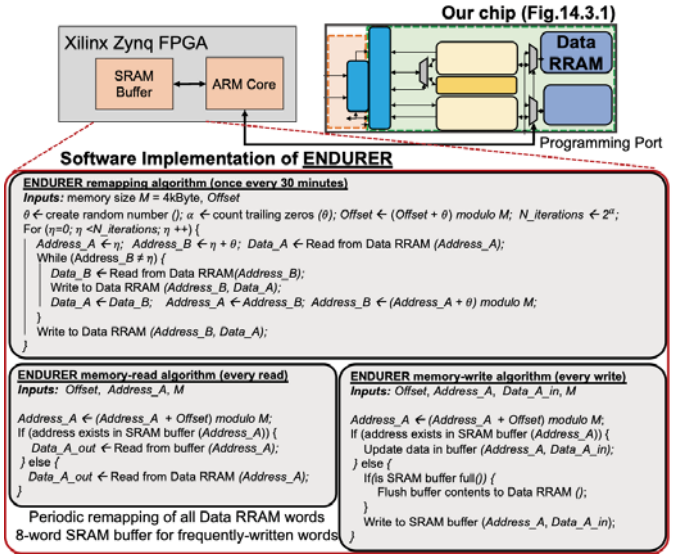


Figure 14.3.5: ENDURER test setup and remapping, read, and write algorithms.

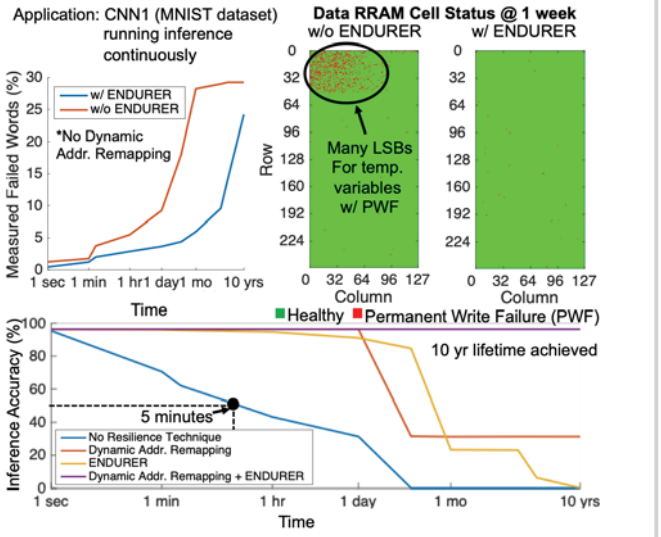
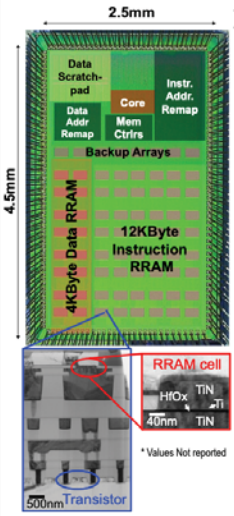


Figure 14.3.6: ENDURER and dynamic address remapping improves chip lifetime to 10 years.



* Values Not reported

	This work	Liu, et al. [6]	Su, et al. [7]	Chen, et al. [8]
Year	2019	2016	2017	2018
Supply Voltage (V)	0.71 - 1.2	0.8	0.8	1
Technology node (nm)	130	65	150	65
Clock Frequency (MHz)	10	100	20	64
Memory Technology	RRAM	RRAM	RRAM	RRAM
Amount of NVM(1) (KBytes)	18	12.2	1.3	128
# of bits/cell demonstrated	2.3	1	1	1
Type	16-bit microcontroller	8-bit NV(2) Processor	8-bit NV Proc. + Accelerator	In-memory compute macro
Applications (energy, pJ/cycle) / (Active to shutdown mode time) / (Active to shutdown mode energy):				
CNN (5x5 images, 4 class)	Dataset Not Avail.	No	110/0.1ms/0.5 µJ	Dataset Not Avail.
CNN (MNIST, 28x28 images)	42/5 µs/1.68 nJ	No	No	Yes*
SVM	44/1.5 µs/0.5 nJ	No	No	No
Linear Regression	42/5 µs/1.68 nJ	No	No	No
Kalman Filter	41/4 µs/1.34 nJ	No	No	No
SHA256 Hash	48/8 µs/2.69 nJ	No	No	No
Counter	24.2/0.5 µs/0.3 nJ	33/4µs - 1.02 ms/400 nJ	No	No
Active Power @ 25C (mW)	0.47 @ 0.71V	3.3	22	Not Reported
Fine-grained Temporal Power-gating demonstrated	Yes	Yes	Yes	N/A
Shutdown to Active mode time	200 ns	130 ns	50 ns	N/A
Shutdown to Active mode energy	152 pJ	450 pJ	510 pJ	N/A
RRAM Read/Write Latency (ns)	23/50	Not Reported	Not Reported	5/(Not Reported)
RRAM Read/Write energy (pJ/bit)	1.76/10.9 (Set) / 10.1 (Reset)	Not Reported /99	Not Reported /46.1	Not Reported
Resilience addressed by	Dynamic Addr. Remapping & ENDURER	None	None	None
Lifetime (years)	10	Not Reported	Not Reported	Not Reported

(1) NVM: Non-volatile memory
(2) NV: Non-volatile

[6] Y. Liu et al., ISSCC, 2016. [7] F. Su et al., VLSI Circuits, 2017. [8] W. Chen et al., ISSCC, 2018.

Figure 14.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE