Unified Characterization Platform for Emerging NVM Technology: Neural Network Application Benchmarking Using off-the-shelf NVM Chips

Supriya Chakraborty, Abhishek Gupta, and Manan Suri
Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India
Email: manansuri@ee.iitd.ac.in

Abstract—In this paper, we present a unified FPGA based electrical test-bench for characterizing different emerging Non-Volatile Memory (NVM) chips. In particular, we present detailed electrical characterization and benchmarking of multiple commercially available, off-the-shelf, NVM chips viz.: MRAM, FeRAM, CBRAM, and ReRAM. We investigate important NVM parameters such as: (i) current consumption patterns, (ii) endurance, and (iii) error characterization. The proposed FPGA based testbench is then utilized for a Proof-of-Concept (PoC) Neural Network (NN) image classification application. Four emerging NVM chips are benchmarked against standard SRAM and Flash technology for the AI application as active weight memory during inference mode.

I. INTRODUCTION

The increasing trend of memory content in System-on-Chip (SoC) designs demand embedded or off-the-shelf Non-Volatile Memory (NVM) with low power consumption, high speed operations, and high endurance [1], [2]. Flash EEPROM is the current state-of-art NVM technology used for commercial and industrial SoCs. However, Flash memories suffer from limitations such as physical scaling, erase-before-write operation, limited endurance, cell to cell interference, high power consumption, low programming speed, complex controller structures. The limitations of the Flash memories are overcome by exploring emerging NVM technologies such as magnetoresistive random access memory (MRAM), resistive RAM (ReRAM), ferroelectric RAM (FeRAM), conductive bridge RAM (CBRAM) [3]–[6].

Most of the research work on emerging NVM technologies in literature are based on model simulations of device and circuit, at architectural/system level [7]–[10], single standalone device or an array of devices [11], [12]. Researchers are focused on improving the emerging NVM technologies at different levels like materials [13], stack engineering [14], and circuit-level strategies [15]. Moreover, testing of matured NVM technology requires efficient characterization setup. Single standalone setup for characterizing multiple NVM technologies is rare to find. Commercially available packaged memory chip testing setups are complex and dedicated to a particular memory technology [16]–[18]. The data-sheet specifications of the commercially available NVM chips provide typical and maximum values for write current and endurance of the NVM chips. However, variations may occur when the chips are used for real-time applications. The major factors include i) incoming data to be written in the memory location and ii) aging effect. Detailed characterization of the NVM chips is required for system-level integration in a variety of applications. In this paper, we present a unified test platform for characterizing multiple commercially available off-the-shelf NVM technologies. Our study helps in exploring electrical and endurance properties of fabricated NVM chip and exploits them by designing hardware/software techniques for performance enhancement. The proposed hardware setup can be used generically for characterizing off-the-shelf emerging NVM chips (SPI or parallel interface). Moreover, the setup presents an efficient way to indirectly extract certain analog characteristics from the packaged chip without having access to a dedicated analog interface. The contributions of this paper are

1) Electrical characterization of different commercially available emerging NVM technologies specifically (1) toggle MRAM [19], (2) FeRAM [20], (3) CBRAM [21], and (4) ReRAM [22] from different vendors.
2) Error characterization of NVM chips.
3) Implementation of the proposed FPGA based setup for basic neural network (NN) application case study comprising of a hybrid CMOS-NVM pipeline.

The rest of the paper is organized as follows: Section II explains the experiments performed on the emerging NVM technologies. Results and discussions are presented in Section III. A case study of NVM technologies for basic NN application using the proposed characterization platform is presented in Section IV. Section V concludes the paper.

©2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
II. EXPERIMENT PERFORMED

A. Characterization Platform

The characterization platform (shown in Fig. 1) comprises of FPGA evaluation board, custom designed daughter board to interface the NVM ICs, Keithley 4200 SCS characterization system for current measurement and a host computer. Two soft-core MicroBlaze processors are implemented on the FPGA to control digital interfaces between the host computer, NVM chips and other peripheral ICs. Both the soft-core processors with dedicated hardware are used to run parallel tasks. One of the processors (Processor 1) performs continuous read/write/erase operations from/to the NVM chips. Another processor (Processor 0) reads the write current measured using the current sensor IC. The operating clock frequency of the NVM chips (SPI based) is set to 1.5625 MHz for all the experiments. Write current is measured using two different methodologies: (i) using on-board I²C interface based current sensor [23] to measure the write current of toggle MRAM. (ii) using Keithley 4200 SCS instrument to measure write current of FeRAM, CBRAM, and ReRAM. One terminal of the current characterization instrument is connected to supply voltage from the FPGA board while the other terminal of the instrument is connected to the VDD supply pin of the NVM chips. The host machine collects the data using virtual serial port connection. A detailed description of similar kind of experimental setup and procedure for current measurement and endurance characterization is explained in [24], [25].

B. Current Characterization

Write current variations in NVM technologies due to input data pattern is analyzed by performing extensive data write operations with fixed number of bits toggled per byte. Initially, all the bits in a byte are set to either ‘0s’ or ‘1s’ followed by toggling specific bit(s) in a byte. Write current is measured for a particular page (64 bytes). The write operations are performed for multiple times (500 cycles) to measure the average write current for a particular type of bits toggled. The same experiment is repeated for different data patterns varying the number of bits toggled (from 1 to 8) in a byte. Write current variation due to the aging of the memory devices is performed by writing random data at a particular location for multiple (more than 50k) cycles.

C. Endurance Characterization

Error characterization is analyzed by performing extensive data write operations at a particular location (address) in the chip. For each experiment, random data write operations are performed in a page for 200k cycles. A read operation is performed between two consecutive write cycles. Error is estimated by calculating the number of bits found incorrect between the data that is “to be programmed” and the data that is “actually programmed” at any specific byte address. We characterize and term the nature of the error based on the number of bits found incorrect in a byte. For example, in a byte, if 3-bits are incorrect we termed it as a 3-bit error.
Fig. 4. Nature of error distribution of (a) a page (64 bytes), and (b) a byte program for 200k cycles in CBRAM.

Fig. 5. Comparison of decrease in nature of bit errors by implementing algorithms (soft technique) for error reduction.

data pattern in FeRAM (Fig. 2(b)). The current consumption for writing any data pattern is constant and does not vary with number of bits toggled. However, observed magnitude of the page write current consumption is low thus can be used for low power applications. Page write current consumption for ReRAM is random in nature for writing different bits toggled as shown in Fig. 2(d).

B. Current variation with aging

Page write current variation due to aging effect in CBRAM NVM chips is shown in Fig. 3(a). It is observed that current consumption increases with aging. One of the probable reasons that describe the phenomenon of increase in page write current consumption over cycling is the write-verify-write (WvW) scheme for write operation. In WvW scheme step voltages with increase in pulse width and voltage amplitude are applied to increase the success rate of write operation [26]. WvW scheme also increases the write latency. Significant results have been observed for CBRAM technology. In order to observe the WvW technique, write latency is also measured (as shown in inset Fig. 3(b)). The increasing trend of write latency values over cycling supports the effect of aging. However, it should be noted here that no significant change in current consumption over cycling is observed for the other NVM technologies (Toggle MRAM, FeRAM, and ReRAM).

C. Endurance Characterization

The data-sheet specified endurance of CBRAM chip is 100k write cycles while the minimum write endurance for other NVM technologies used for the study is $1.2 \times 10^6$ cycles. We focused our study on analyzing the nature and distribution of bit errors that occur in NVM technologies due to over-stressing a particular location. We select CBRAM chips for error characterization at page and byte level granularity for our study. Fig. 4 (a) and (b), shows the distribution of nature of errors occurred in a page and a byte level granularity respectively for random data write operations performed over 200k cycles (2X data-sheet specifications). It can be observed that the distribution of error in a page is random. The total error count for a particular byte in a page varies randomly. However, the total count for 2-bit error in a page is more compared to other types of bit errors. Moreover, higher bit errors (3-bit, 4-bit, etc.) counts are few. The probable reason for this specific nature of bit error is due to the implementation of specific bit error ECC within the chip. We have implemented Flip-N-Write (FNW) algorithm to analyze the effect of soft techniques in reducing the nature of error. It is observed from Fig. 5 that the implementation of FNW decreases the 2-bit error by $\sim 6.54X$ and overall total error count by $\sim 5.4X$. Thus, based on the nature of error and the implementation of soft techniques, the endurance of the NVM chips can be increased significantly [27]. This analysis helps in designing the soft/hard techniques like ECC within/next to the controller to enhance the endurance of the emerging NVM chips.

IV. CASE STUDY: FPGA BASED NN APPLICATION

A. Methodology

In this section, we benchmarked the emerging NVM chips for NN application using the proposed FPGA based test platform. We focused on characterizing the weights’ write latency and current consumption of the NVM chips for AI applications. The NN model used for the study consists of three layers: (i) input layer, (ii) hidden layer, and (iii) output layer (Fig. 6(a)). An RGB image ($16 \times 16 \times 3$ byte) is taken as an input. Therefore, the input layer consists of 768 neurons, where each neuron corresponds to a particular byte of the input image. The hidden layer is composed of 7 neurons feeding to a single output neuron. The neuron output from the hidden layer to the output layer is described by:

$$y = g \sum_{j=1}^{n} a_j \times w_j + b$$

where $n$ is the number of nodes in the previous layer, $a_j$ is the activation output represented as an 8-bit unsigned integer, $w_j$ and $b$ is the weight, and bias respectively and are represented by 8-bit signed integer. The activation function ($g(x)$) is a step function described by:

$$g(x) = \begin{cases} 0, & x < 0 \\ 255, & x \geq 0 \end{cases}$$
Fig. 6. (a) NN Model implemented on FPGA platform for characterizing different NVM technologies. (b) Software GUI for uploading weights and image for AI application.

Fig. 7. Block diagram of FPGA based setup for NN application using emerging NVM.

| Memory Type | Latency (a.u.) | Current (a.u.) |
|-------------|----------------|----------------|
|             | Weights Write  | NN Application | Average Erase | Average Byte Write |
| Toggle MRAM | $5.2 \times 10^{-5}$ | 0.05           | 0             | 1 |
| FeRAM       | 0.02           | 0.82           | 0             | 0.02 |
| CBRAM       | 0.07           | 0.82           | 0             | 0.03 |
| ReRAM       | 1              | 0.99           | 0             | 0.04 |
| Flash       | 0.26 a         | 1 b            | 1             | 1 |
| SRAM        | $5.2 \times 10^{-5}$ | 0.05 | 0             | 1 |

a Latency value includes read latency from emerging NVM chips to BRAM and NN computational latency.
b Latency value includes summation of erase and write operation together.

The network is trained using differential evolution (DE) algorithm [28] for the dataset used in [29]. The algorithm is suitable for network having integer weights, biases and non-linear activation function [30]. The integer weights and biases allow easier hardware implementation in FPGA.

B. Hardware Implementation

The hardware architecture, as shown in Fig. 7, is implemented on CMOD FPGA [31] evaluation board. The main difference between the FPGA based NN hardware architecture and the experimental setup used for chip characterization is the addition of NN engine with the former architecture. NN engine is a generic engine that can be used to implement multi-layer NN to predict the test image uploaded by the GUI based software (Fig. 6(b)). For our application, we implement a 2-layer NN. Two internal block RAMs (BRAM) of the FPGA are used to interact with the NN engine. We term the BRAMs as image memory and weight memory (Fig. 7). Image memory contains the scaled input RGB image. Weight memory contains the trained weights and biases of each layer of fully connected NN read from the NVM chip. Following operations are performed while prediction of the test image (i) Initially, the trained weights and the biases are uploaded in the emerging NVM chips. (ii) The pixel values of the input RGB image are uploaded in the image memory. (iii) NN engine is started for image prediction. The NN engine on initiation performs inference by reading the trained weights and biases and loads it to the FPGA BRAM (weight memory) for faster execution. Finally, it starts calculating the results sequentially of each hidden neuron followed by an output neuron to distinguish desired images from undesired images.

C. Experimental Results

Extensive experiments are performed to calculate weights’ write latency and current consumption of the NVM technologies while performing NN applications. Comparative study of the NVM technologies with the state-of-art Flash technology and SRAM is also performed. The DE algorithm used for the application provides approximately 72.25 % accuracy on the data set containing 209 images. During the initialization phase, number of weight/bias byte write operations to NVM chip is 5391 and number of image byte write operations on FPGA BRAM is 768. Similar number of read operations are performed after initiating the NN engine for test image prediction. Table I illustrates the study of the write latency and the electric current consumption using the proposed setup for different memory technologies. It can be observed that off-the-shelf emerging NVM chips are proficient candidates for replacement of Flash and SRAM technologies.

V. Conclusion

The paper presents a unified FPGA based test platform for characterizing different off-the-shelf emerging NVM technologies. Detailed electrical characterization and benchmarking study for multiple NVM chips using the test setup is performed. Current consumption due to different data patterns and aging effect is analyzed on emerging NVM technologies such as MRAM, FeRAM, ReRAM, and CBRAM. Moreover, nature of error and the distribution of error are analyzed at byte and page level granularity. Finally, the proposed test platform is utilized for NN image classification application. A comparative study of the emerging NVM chips with state-of-art Flash and SRAM technology is performed. Obtained results show that off-the-shelf emerging NVM chips are suitable candidates for future memory applications.

ACKNOWLEDGMENT

This work was supported in part by MHRD, SERB-CRG/2018/001901, and CYRAN.
