A 68 Parallel Row Access Neuromorphic Core with 22K Multi-Level Synapses Based on Logic-Compatible Embedded Flash Memory Technology

M. Kim[1], J. Kim[1], G. Park[1], L. Everson[1], H. Kim[1], S. Song[1,2], S. Lee[2], and C. H. Kim[1]
[1]Dept. of ECE, University of Minnesota, Minneapolis, MN, USA email: chriskim@umn.edu
[2]Anaflash Inc, San Jose, CA, USA

***Abstract* –** A neuromorphic core utilizing logic-compatible embedded flash technology for storing multi-level synaptic weights is demonstrated in a 65nm standard CMOS process. A carefully-designed program-verify sequence along with a bitline voltage regulation scheme allows the individual cell currents to be programmed precisely. This makes it possible to enable a large number of rows in parallel without impacting the current summation accuracy. Furthermore, eflash based synapses are non-volatile and hence consumes zero standby power and supports instant on/off operation. Our design stores excitatory and inhibitory weights in adjacent bitlines whose voltage levels are regulated for accurate current programming and measurement. Output spikes are generated by comparing the excitatory and inhibitory bitline currents. Our logic-compatible eflash-based spiking neuromorphic core achieves a 91.8% handwritten digit recognition accuracy which is close to the accuracy of the software model with the same number of weight levels. The maximum throughput of the core is 1.28G pixels/s and the average power consumption of a single neuron circuit is 15.9μW.

I. Introduction

Deep neural networks contain multiple computation layers each performing a massive number of multiply-and-accumulate operations (i.e. $\sum x_i w_i$) between the input data and trained weights. The computation is typically performed by a digital processor while the data and weight are transferred back and forth between the DRAM and the on-chip buffer memory. To overcome the memory bottleneck, there is growing interest in so-called "compute-in-memory" architectures where the weights are stored in a dense memory while the multiply-and-accumulate function is performed in the analog domain. Here, the input data is typically loaded on to the memory wordlines, activating multiple cell currents at the same time. The individual cell currents are summed up and compared to a pre-defined threshold by the local "neuron" circuit. Ideally, memory cells used for compute-in-memory architectures should be non-volatile as this obviate the need for reloading the weights after a power down period. It is also highly desirable if the memory cell can support multi-level storage as this can enhance the accuracy of inference tasks.

Both volatile and non-volatile memories have been considered for synaptic weight storage including SRAM, magnetic tunnel junctions (MTJs) and resistive RAM (RRAM), flash, and phase-change memory (PCRAM) [1]-[6]. Each type of memory has its advantages and disadvantages. For instance, SRAM based synapses can be readily implemented in a standard CMOS process, but suffers from process variation which cannot be corrected after the chip has been fabricated. MTJs, RRAM, and PCRAM are non-volatile and dense. However, an MTJ can only store a 1 bit weight and the difference between the high resistance and low resistance states is only about 2X, rendering analog computing impractical. RRAM provides a wider resistance range, but the technology remains immature. Furthermore, robust multi-level programming has proven to be challenging for RRAM and PCRAM due to low controllability of the filament formation and heat diffusion [7]. Flash memory technology can easily store multiple levels by adjusting the number of electrons stored on a floating gate through row-by-row program-verify operation. However, conventional flash memory requires a specialized dual-poly or split-gate process which doesn't scale well below 40nm.

In this work, we demonstrate a logic-compatible eflash based spiking neuromorphic core in a 65nm standard CMOS process featuring multi-level non-volatile weight storage, and single cycle current integration and spike generation. The weights were tuned precisely using a carefully-designed program-verify sequence, allowing 68 individual cell currents to be summed up simultaneously, which to our knowledge, is the highest number ever reported.

II. Eflash-Based Neuromorphic Core Design

Fig. 1 shows a comparison between dual-poly and single-poly eflash cells. Dual-poly eflash cell stores charge on a floating gate fabricated between the control gate and channel. Single-poly eflash cell is implemented using back-to-back connected transistors and hence does not require any modification to the process. The detailed schematic of the 5T eflash cell used in this work is shown in Fig. 2 where two asymmetrically sized PMOS devices are used for high voltage program and erase operation while the NMOS read device is accessed through two additional NMOS switches [8][9]. Different cell currents can be programmed as shown in Fig. 2 (right).

Synaptic weights stored in two adjacent bitlines as illustrated in Fig. 3. If the weight is positive then the cell current of the left bitline is increased accordingly while the

cell current on the right bitline is programmed to <0.1μA, and vice versa. Four flash cells are reserved on each bitline for the spiking threshold. The input data is simultaneously loaded onto the wordline which activates multiple memory cell currents at the same time. The sum of the individual cell currents flows through each bitline. The bitline pair generates two currents: excitatory and inhibitory currents. The neuron circuit generates a spike depending on which bitline current is higher. Note that the weight multiplication, accumulation, and spike generation are all performed in a single cycle, which speeds up the overall computation.

The core architecture is shown in Fig. 4 (left). It contains high voltage switches (HVSs) for driving wordlines, neuron sensing circuits for current comparison and spike generation, and scan chains for data input/output. The HVS circuit must withstand a voltage as high as 10V during program and erase modes. We employed the multi-story latch based HVS circuit [8][9] because it is inherently immune to overstress issues and implementable using standard IO devices. The circuit diagram and layout of the unit 5T eflash cell are shown in Fig. 4 (right).

The weights are written to the array by first erasing the entire array and then adjusting the threshold voltage of each individual eflash cell. This is done by simultaneously programming each row through a selective program-verify operation. The cell currents were set to either 0, 5, or 10 μA while keeping the gate bias (VRD=0.8V) and drain bias (VBL=0.6V) fixed. This translates into five distinct weight levels (i.e. -10, -5, 0, 5, or 10 μA) using the bitline pair configuration described earlier in Fig. 3. Retention characteristics shown later in Fig. 15 confirm that a 5 μA margin between the different levels is sufficient to overcome charge loss issues. The wordline and bitline bias condition for erase, program, and program inhibition modes are denoted in Fig. 5. To obtain a precise cell current corresponding to the weight value, the bitline voltages were regulated to 0.6V during both verify and inference modes. Our neuron circuit shown in Fig. 6 employs a feedback loop to maintain a fixed 0.6V bitline voltage regardless of the amount of current flowing through the bitline. The bitline current is indirectly measured by reading out the feedback voltage driving the PMOS load. This makes it possible to compare the two bitline currents using a simple voltage sense amplifier circuit. The same neuron circuit was used for current-verify operation as shown in Fig. 6 (right). Here, the cell currents of the left and right bitlines were verified separately by activating one bitline at a time. The overall operation sequence of our neuromorphic core is shown in Fig. 7.

III. Experimental Results

We first measured the program and program inhibition characteristics of the 5T eflash cell. The average current of 100 cells was measured for different program voltages, program pulse widths, and program pulse counts. Fig. 8 confirms excellent program and program inhibition results. Based on the test results, we designed the program sequence in Fig. 9 for configuring the cell currents. To minimize cell disturbance, we first programmed the weight 0 cells (i.e. <0.1μA) while inhibiting the program of weight 1 and 2 cells. To ensure that the cell currents of all weight 0 cells are below 0.1μA, we use a high voltage (8.8V), long pulse width (20μs), and large pulse count (8). Next, the rest of the cells were programmed to an intermediate current level of about 15μA using a single 7.4V and 40μs pulse. Then, using smaller and shorter pulses (i.e. 7.1V, 5μs), we adjusted the weight 2 cells to 10μA, and finally the weight 1 cells to 5μA. Note that the unselected wordlines are not driven to a high voltage preventing the cells on those wordlines from being disturbed. Fig. 10 shows the cell currents for trained weights of the MNIST handwritten digit recognition algorithm. The variation for weight 0 cells is less than 0.1μA. Weight 1 and 2 cells also have a variation of only 0.8μA. The total number of program pulses applied to each wordline ranges from 25 to 32 as shown in Fig. 11. The average power consumption of a single neuron during inference mode is 15.9μW (Fig. 12). Fig. 13 provides further insight on how the cell current changes with more program pulses for weight 1 and weight 2 cells. It can be seen that that intrinsic cell current variation of 7μA is reduced to 0.8μA after the proposed program-verify sequence. This offers a significant advantage over SRAM or MRAM based implementations which do not have any post-silicon tuning capabilities.

Fig. 14 shows the overall work flow for demonstrating the handwritten digit recognition application on our test chip. During training phase, weights were trained based on 60,000 handwritten digit images from the MNIST dataset [10] and downloaded to the test chip. During inference phase, the neuromorphic core generates a spike signal based on the 16x16 pixel data and the programmed weights. 10,000 MNIST test images were processed to calculate the prediction accuracy. The accuracy measured from the test chip was 91.8% (Fig. 14) which is close to the software accuracy of 93.8% for the same number of distinct weight levels (i.e. 5). The small discrepancy can be attributed to noise effects and sense amplifier offset. Retention characteristics were measured after baking the chip for 16 hours at 150°C. Measured results in Fig. 15 confirm that the margin between the different current levels is not compromised, suggesting that storing more than 5 levels is also possible. Comparison with previous SRAM, RRAM, and NOR flash based neuromorphic core designs underscores the promising features of our logic-compatible eflash-based design (Fig. 16). The die photo and chip feature summary are given in Fig. 17.

References

[1] D. Kuzum, R. Jeyasingh, H.S. Wong, IEDM, pp.693-696, Dec. 2011. [2] W. Chen, W. Lin, L. Lai, et al., IEDM, pp. 657-660, Dec. 2017. [3] X. Guo, F. Merrikh Bayat, M. Bavandpour, et al., pp. 6.5.1-6.5.4, IEDM, Dec. 2017. [4] W. Khwa, J. Chen, J. Li, et al., ISSCC, pp. 496-497, Feb. 2018. [5] S. Gonugondia, M. Kang, N. Shanbhag, ISSCC, pp. 490-491, Feb. 2018. [6] W. Chen, K. Li, W. Lin, et al., pp. 494-495, ISSCC 2018. [7] D. Nminibapiel, D. Veksler, P. Shrestha, et al., pp. 736-739, IEEE EDL, June 2017. [8] S. Song, K. Chun, C. Kim, pp. 1302-1314, JSSCC, May 2013. [9] S. Song, K. Chun, C. Kim, pp. 1-4, CICC, 2013. [10] MNIST dataset, http://yann.lecun.com/exdb/mnist/index.html

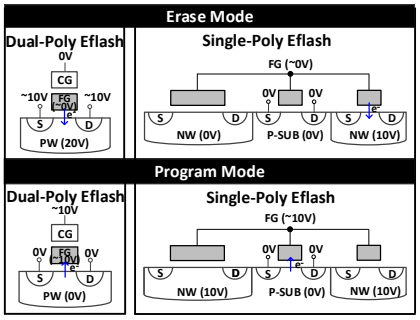


Fig. 1. Comparison between dual-poly eflash (left) and single-poly eflash (right).

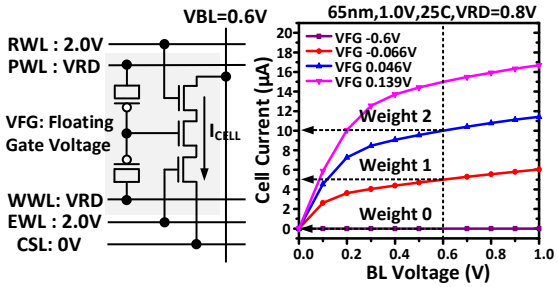


Fig. 2. Output characteristic of proposed 5T eflash cell for different floating gate (FG) node voltages. Multi-level weights can be stored precisely through program-verify.

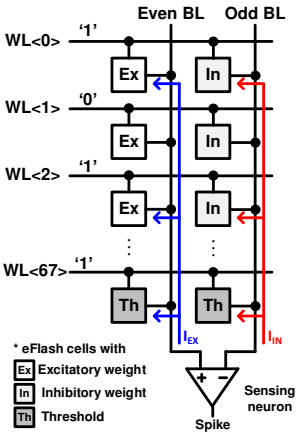


Fig. 3. Excitatory and inhibitory weight values are stored in two adjacent bitlines. Currents are summed up and compared for spike generation.

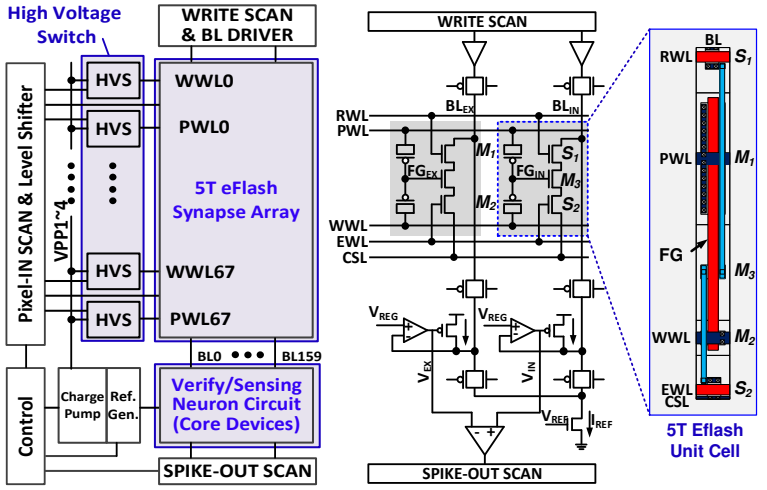


Fig. 4. (a) Overall neuromorphic core with high voltage switch, eflash array, and neuron sensing circuit. (b) Single column pair and 5T unit cell layout.

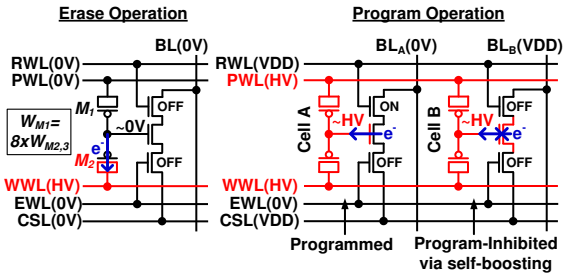


Fig. 5. Bias conditions of the proposed 5T eflash cell for erase and program operations [8].

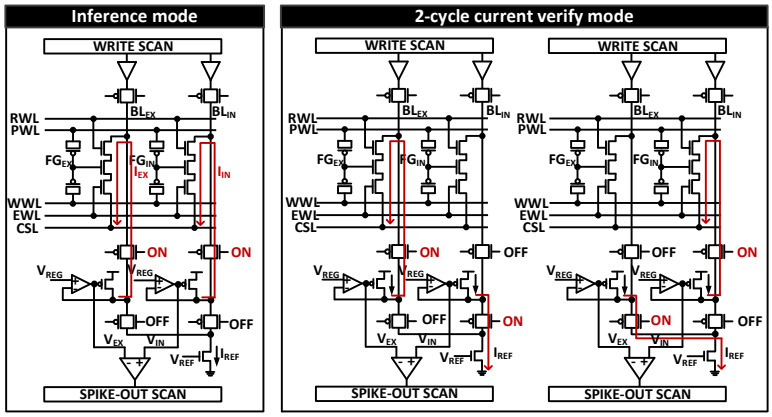


Fig. 6. Neuron circuit with regulated bitline voltage; (left) inference mode for spike generation and (right) weight programming mode with program-verify operation.

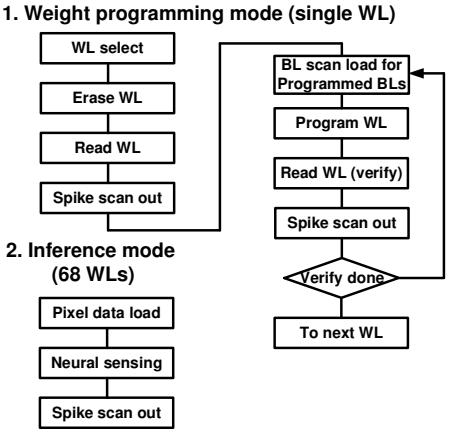


Fig. 7. Overall neuromorphic core operation sequence: Weight programming mode (upper) and inference mode (lower).

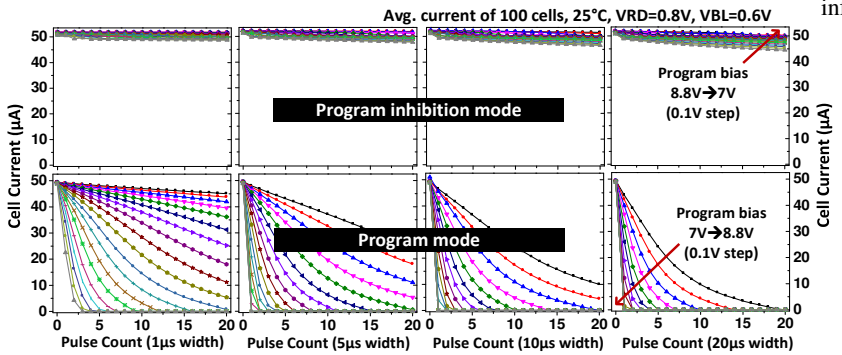


Fig. 8. Cell current versus number of program pulses for program inhibited cells (upper row) and programmed cells (low row). Results are shown for different pulse widths (i.e. 1μs, 5μs, 10μs, 20μs) and 0.1V program bias increments from 7.0V to 8.8V. Test chip data shows reliable programming and minimal program disturbance.

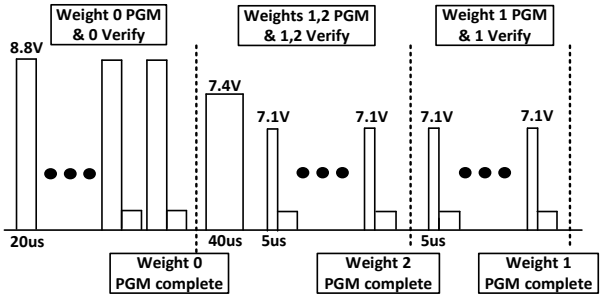


Fig. 9. Pulse sequence for programming weights 0, 1, and 2 into the eflash neuromorphic core. Multi-level weights can be programmed precisely owing to the carefully-design program-verify sequence.

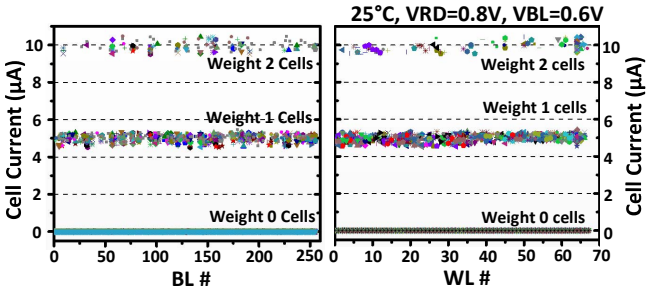


Fig. 10. Individual cell currents in bitline and wordline direction for MNIST trained weights. The exceptionally tight current distribution suggests that storing 3 or more levels of weights is possible.

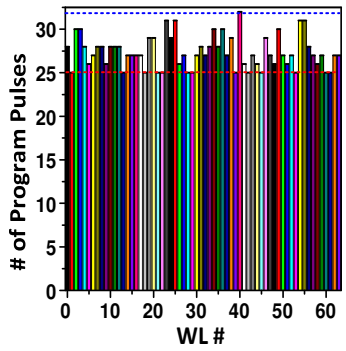


Fig. 11. The number of program pulses applied to each wordline for MNIST handwritten digit recognition.

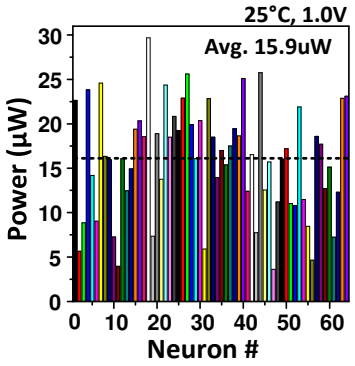


Fig. 12. Power consumption of each neuron during inference mode at VDD=1.0V.

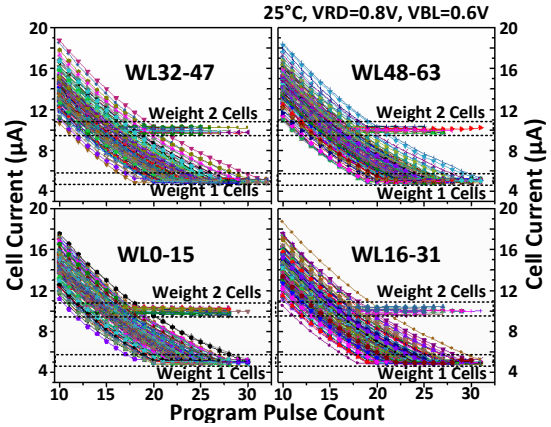
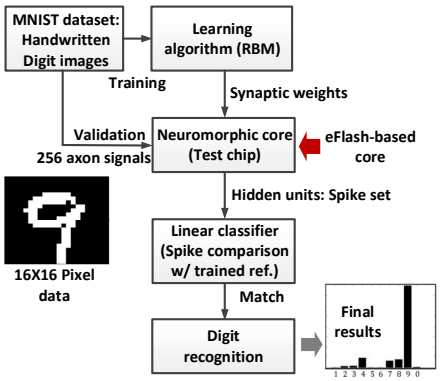


Fig. 13. Cell current versus program pulse count when applying the program-verify sequence.



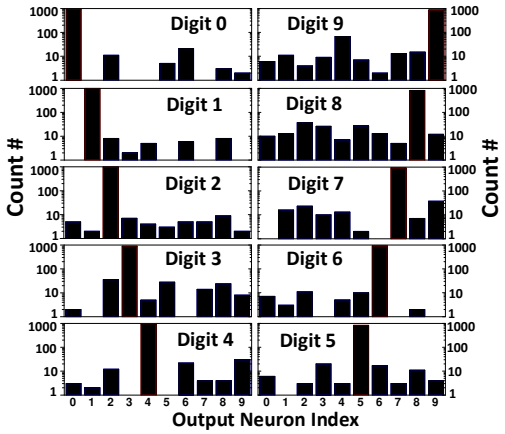


Fig. 14. (Left) Demonstration flow of hand-written digit recognition algorithm using the proposed neuromorphic core. (Right) Histograms of neuron output for 10,000 MNIST test images measured from the eflash-based neuromorphic core chip.

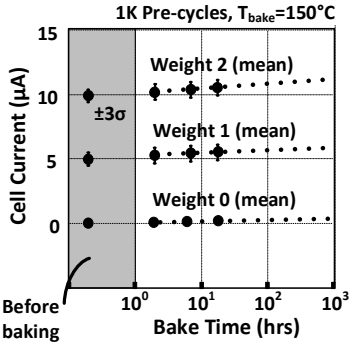
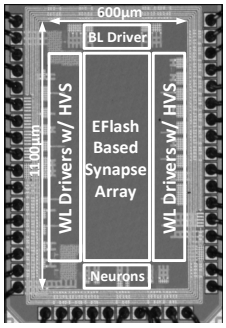


Fig. 15. Retention characteristics of weight 0, 1, and 2 cell currents confirm that the margin between the different states remains constant. Baking temperature was 150°C.

	This work	ISSCC'18 [7]	ISSCC'18 [4]	ISSCC'18 [5]	IEDM'17 [2]	IEDM'17 [3]
Application	Handwritten digit recognition	Handwritten digit recognition	Handwritten digit recognition	Machine learning classifier	Computing in memory	Handwritten digit recognition
Technology	65nm	65nm	65nm	65nm	150nm	180nm
Voltage	1.0V	1.0V	1.0V	1.0V	1.8V	2.7V
Non volatile?	YES (Eflash)	YES (ReRAM)	NO (SRAM)	NO (SRAM)	YES (ReRAM)	YES (Eflash)
Logic Compatible?	YES	NO	YES	YES	NO	NO
Program-verify?	YES	NO	NO	NO	NO	YES
Weight Resolution	2.3 Bits (5 levels)	3 Bits	1 Bit	1 Bit	2 Bits	N/A
# of Currents Summed Up	68 Cells	14 Cells	30 Cells	4 Cells	2 Cells	N/A

Fig. 16. Comparison with prior art.



Technology	65nm Logic
Core Size	1100 X 600 μm²
VDD (Core, IO)	1.0V / 2.5V
# of Neurons	320
# of Synapses	22K (=68x320)
Throughput	1.28G pixels/s per core (tREAD : 50ns)
Power	15.9μW (per neuron)

Fig. 17. Die microphotograph and test chip feature summary.