A 3D NAND Flash Ready 8-Bit Convolutional Neural Network Core Demonstrated in a Standard Logic Process

M. Kim[1], M. Liu[1], L. Everson[1], G. Park[1], Y. Jeon[2], S. Kim[2], S. Lee[2], S. Song[2], and C. H. Kim[1]
[1]Dept. of ECE, University of Minnesota, Minneapolis, MN, USA email: kimx4916@umn.edu
[2]ANAFLASH Inc., San Jose, CA, USA

Abstract – A convolutional neural network (CNN) core that can be readily mapped to a 3D NAND flash array was demonstrated in a standard 65nm CMOS process. Logic-compatible embedded flash memory cells were used for storing multi-level synaptic weights while a bit-serial architecture enables 8 bit multiply-and-accumulate operation. A novel back-pattern tolerant program-verify scheme reduces the cell current variation to less than 0.6µA. Positive and negative weights are stored in eFlash cells in adjacent bitlines, generating a differential output signal. Our eNAND based neural network core achieves a 98.5% handwritten digit recognition accuracy which is close to the software accuracy of 99.0% for the same precision. This work represents the first physical demonstration of an embedded NAND Flash based neuromorphic chip in a standard logic process.

I. INTRODUCTION

Deep neural networks (DNNs) contain multiple computation layers each performing a massive number of multiply-and-accumulate (MAC) operations between the input data and trained weights. Due to huge amount of data processing and computation need, the performance and energy-efficiency of DNN chips can be limited by available memory bandwidth and MAC engines. A promising approach to alleviate this issue is the compute-in-memory (CIM) paradigm where the computation occurs where the data is stored, with massively parallelized analog MAC engines. One embodiment of this approach is a resistive RAM (ReRAM) crossbar array where weights are stored in the form of transconductance of ReRAM cells while the MAC operation is performed in the analog domain. Despite their tremendous potential, past CIM demonstrations are mostly limited to small-scale networks with low precision operands (e.g. 1-4 bits) due to fabrication difficulties [1-5]. Many array level studies are based on software models extracted from individual device probing which cannot capture the intricate device and circuit level behaviors. To make CIM a practical reality for future DNNs with hundreds of convolutional layers, it is imperative to focus on a memory technology that is non-volatile, ultra-high density, low cost, and highly manufacturable. 3D NAND Flash technology is the leading candidate that satisfies all these requirements, however, almost no experimental works exist on 3D NAND Flash based CIM designs due to the proprietary nature of the technology.

In this work, we investigate the hardware and software design aspects of a 3D NAND Flash based CNN accelerator through a physical chip implementation. A demonstrator chip that mimics a 3D NAND Flash array was fabricated in a 65nm logic process. Logic-compatible single-poly eFlash memory was used for the synaptic device. The proposed design features multi-level non-volatile weight storage, single cycle current integration, bit-serial MAC operation, and multi bit output sensing. One of the highlights of this work is the back-pattern tolerant program-verify sequence which reduces the cell current variation to less than 0.6µA, allowing 28 individual cell currents to be summed up in a single cycle while delivering an MNIST classification accuracy of 98.5%.

II. NEURAL NETWORK CORE CIRCUIT DESIGN

Fig. 1 shows a bit column stacked (BiCS, [6]) 3D NAND array composed of 40 BLs, 4 SGDs, and 16 WLs, which was flattened for implementation in a standard logic process. Unlike NOR type designs where each memory cell requires a selecting device, transistors in a NAND string share the source/drain nodes and hence selecting devices are only required at the top and bottom of the stack. This results in an area reduction up to 20%. The bitline capacitance of NAND is reduced compared to NOR for the same capacity, enabling a faster bitline sensing operation. The logic-compatible 3T eFlash cell shown in Fig. 2 consists of two asymmetrically sized PMOS devices for efficient program and erase operation and an NMOS read device for the NAND string. Due to series resistance of the unselected devices, the I-V characteristics of the eNAND Flash cell vary depending on its location in the stack. Simulation results in Fig. 2 (right) show that such location dependent variation can be cancelled out by fine-tuning the floating gate (FG) voltage.

A pair of eNAND cells in adjacent bitlines are used to store a single weight 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. We included a repair block where three eNAND cells connected to each WL are reserved for redundancy and bias weights. Input data is simultaneously loaded onto the SGD lines which activate multiple memory cell currents. During this operation, the selected and unselected wordlines are held to VREAD and VPASS, respectively. The sum of the individual cell currents flows through each bitline. The bitline pair generates two

currents; i.e. positive weight and negative weight currents. Both currents are converted to the corresponding output voltages by the bitline regulator circuit (Fig. 4). Finally, a voltage controlled oscillator (VCO) circuit converts the output voltage to frequency which is measured using off-chip equipment for multi-bit sensing.

The overall chip architecture is shown in Fig. 4 (left) which contains high voltage wordline drivers [3], BL sensing circuits, VCO and frequency divider circuits, and scan chains. The circuit diagram and layout of the 16 stack eNAND string are shown in Fig. 4 (right). The weights are written to the array by first erasing the entire array and then selectively programming the threshold voltage of each individual eNAND cell. Each cell can store a 2 bit weight segment with target cell current levels of 0, 3, 6 and 9μA. Retention characteristics shown later in Fig. 14 confirm that a 3 μA margin between the different levels is sufficient to overcome charge loss issues. During inference mode, the selected and unselected WL gates are biased at VREAD=1.1V and VPASS=2.6V, respectively, while the drain bias is fixed at VBL=0.8V. Positive and negative weights are stored in a bitline pair which generates a differential bitline current of -9, -6, -3, 0, 3, 6 or 9μA. The bias condition of wordline, SGD, bitline voltages for different operating modes are shown in Fig. 5. To obtain a precise cell current, the bitline voltage was regulated to 0.8V during both verify and inference modes. The bitline current is indirectly measured by reading out the feedback voltage driving the PMOS load as shown Fig. 4 (right) [3]. This voltage is fed to a VCO circuit which generates a frequency output for multi bit inference. The timing sequence of a 5x5 convolution operation with 8 bit input and 8 bit weight is shown in Fig. 6. The bit serial inner-product computing scheme was adopted where the operand is serialized and asserted one bit at a time to the array [7]. This novel architecture enables variable precision MAC operation without any significant modification to the CIM hardware.

III. EXPERIMENTAL RESULTS

Measured data in Fig. 7 confirms excellent program and program inhibition characteristics for the flash cells in the 16 stack NAND string. The output frequency of the VCO was proportional to the cell current (Fig. 8), which is essential for accurate program verify and inference operations. One critical requirement for precise cell current readout is that the VPASS bias applied to the unselected wordlines must be high enough to minimize the series resistance of the unselected cells. However, a higher VPASS can lead to unwanted threshold voltage shift in the unselected cells. We chose a VPASS voltage of 2.6V which offered a good compromise between the two competing effects.

Another critical challenge we faced while programming the weights into the NAND string is the so-called back pattern dependency where the programmed cell current is affected by the program state of other cells in the array. In our testing, we saw the cell current increase by up to 3μA depending on whether the rest of the array is in erase mode (lowest Vt) or weight 0 mode (highest Vt). To overcome this issue, we devised a novel back-pattern tolerant program-verify scheme in Fig. 10 which ensures that the programmed cell current remains constant irrespective of the weights stored in the rest of the array. The operating sequence is as follows. In phase one, we programmed the weight 0 cells on a given wordline while inhibiting the weight 1, 2 and 3 cells. To ensure that the cell currents of all weight 0 cells are below 0.1μA, we apply high voltage (8.0V) and long duration (20μs) pulses until the target current is reached. Once this is completed, we roughly program the weight 1, 2 and 3 cell currents close to their intended targets using additional program pulses as shown in Fig. 10. The first pulse is applied to weight 1, 2, 3 cells. The second pulse is applied to weight 1 and 2 cells, and the third pulse is applied to weight 1 cells. The specific program voltage of each pulse was carefully chosen based on the program and program inhibition characteristics in Fig. 7. The same sequence was repeated for the rest of the wordlines until the entire array is programmed. In phase two, using smaller and shorter pulses (i.e. 7.0V, 10μs), we fine-tuned the weight 3 cells to 9μA, the weight 2 cells to 6μA and the weight 1 cells to 3μA. A VPASS voltage of 2.6V was used throughout the entire program-verify sequence. Fig. 11 shows how the cell current changes with increasing number of program pulses for weight 1, 2 and 3 cells. It can be seen that the cell current variation is reduced from 3μA to 0.6μA after the program-verify operation. This offers a significant advantage over SRAM or MRAM based neuromorphic implementations which do not have any post-silicon tuning capabilities. Fig. 12 shows the measured cell current and output frequency after programming the entire array using MNIST trained weights [8].

Fig. 13 shows the LeNet-5 CNN demonstration flow for the MNIST handwritten digit recognition application. Weights were trained based on 60,000 handwritten digit images from the MNIST dataset and were preloaded to the test chip. During inference mode, the neural network core generates a frequency output based on 28x28 pixel 8 bit grayscale images and the preloaded 8 bit weights. Due to the limited test time, this paper presents classification results of 1,000 randomly chosen MNIST images (full test results will be available soon). The classification accuracy measured from the test chip was 98.5% (Fig. 13) which is close to the software accuracy of 99.0% for the same weight and data precision. The small discrepancy can be attributed to noise and variation effects in a real chip. Charge loss was minimal when the chip was baked at 150°C for 16 hours. Comparison with previous NOR type neuromorphic core designs in various memory technologies is shown in Fig. 15. The die photo and chip feature summary are given in Fig. 16.

REFERENCES

[1] X. Guo, F. Merrikh Bayat, M. Bavandpour, et al., pp. 6.5.1-6.5.4, IEDM, Dec. 2017. [2] W. Chen, K. Li, W. Lin, et al., ISSCC, pp. 494-495, Feb. 2018. [3] M. Kim, J. Kim, G. Park, et al., IEDM, pp. 15.4.1-15.4.4, Dec. 2018. [4] C. Xue, W. Chen, J. Liu, et al., ISSCC, pp. 388-389, Feb. 2019. [5] X. Si, J. Chen, Y. Tu, ISSCC, pp. 396-397, Feb. 2019. [6] H. Maejima, K. Kanda, S. Fujimura, et al., ISSCC, pp. 336-337, Feb. 2018. [7] P. Judd, J. Albercio, A. Moshovos, ICAL, Vol. 16, pp.80-83, 2017. [8] Y. LeCun, L. Bottou, Y. Bengio, et al., PROC, pp.1-36, Nov. 1988.

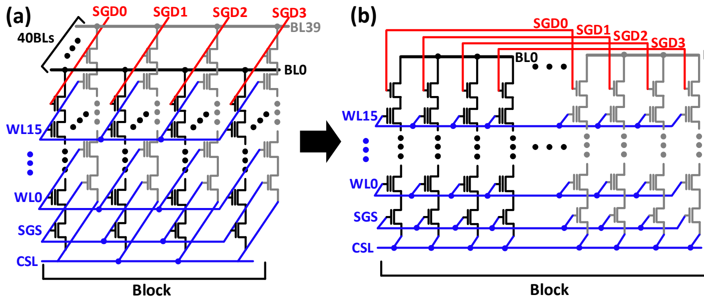


Fig. 1. (a) 3D NAND BiCS architecture [6] with 40 BLs x 4 SGDs x 16 WLs. (b) Flattened 3D NAND architecture for demonstrating an eNAND flash based deep neural network accelerator in a standard logic process.

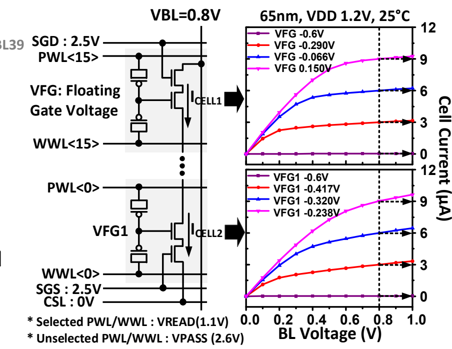


Fig. 2. Simulated I-V characteristics of top and bottom eFlash cells of a 16 stack NAND string. Series resistance varies depending on the location in the stack which can be compensated by incremental programming.

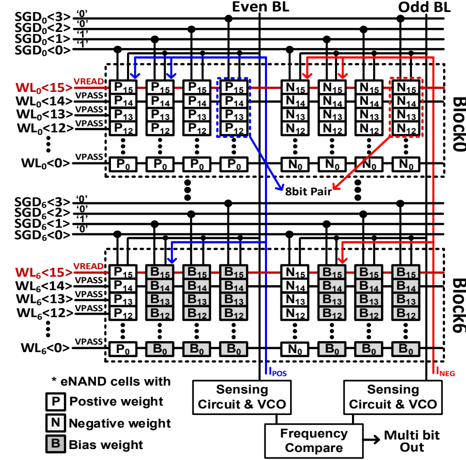


Fig. 3. Positive, negative and bias weight values are stored in two adjacent bitlines. 28 individual NAND string currents are summed up and converted to frequencies for multi-bit sensing.

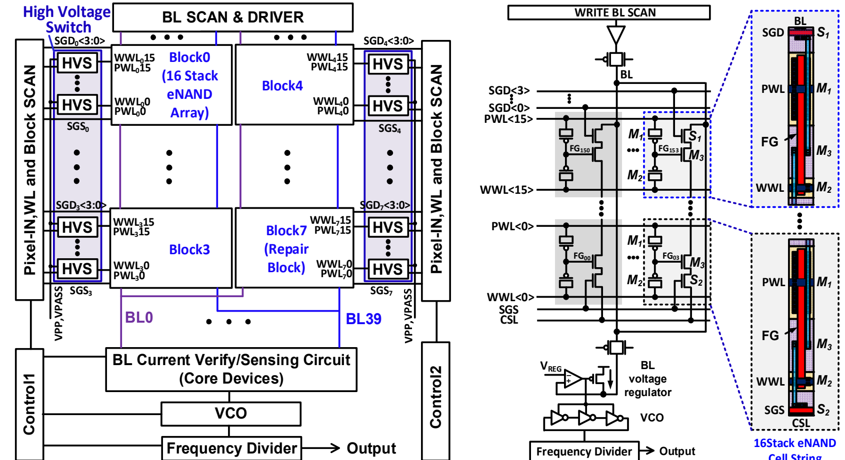


Fig. 4. (Left) Overall diagram of CNN core with high voltage wordline drivers, a 16 stack eNAND Flash array, and BL sensing circuit. (Right) BL pair and readout circuit along with the layout of a 16 stack eNAND string.

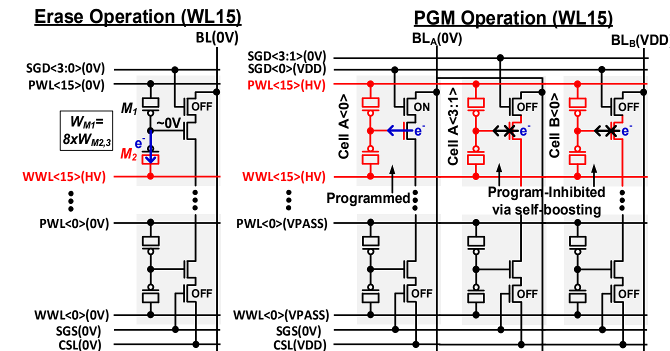


Fig. 5. Bias conditions of logic-compatible eNAND Flash cell for erase and program operations.

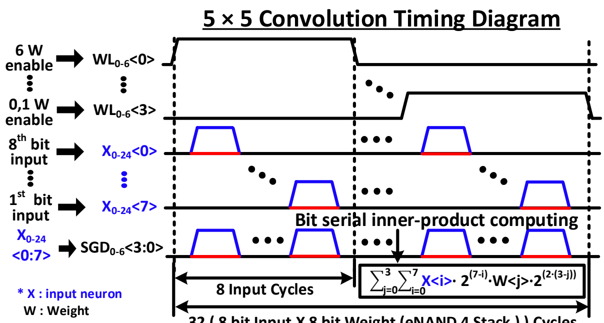


Fig. 6. Timing diagram of 5 × 5 convolution operation with 8 bit data and 8 bit weights. 4 cells are used to store a single 8 bit weight. Bit serial operation produces a multi-bit inner product result.

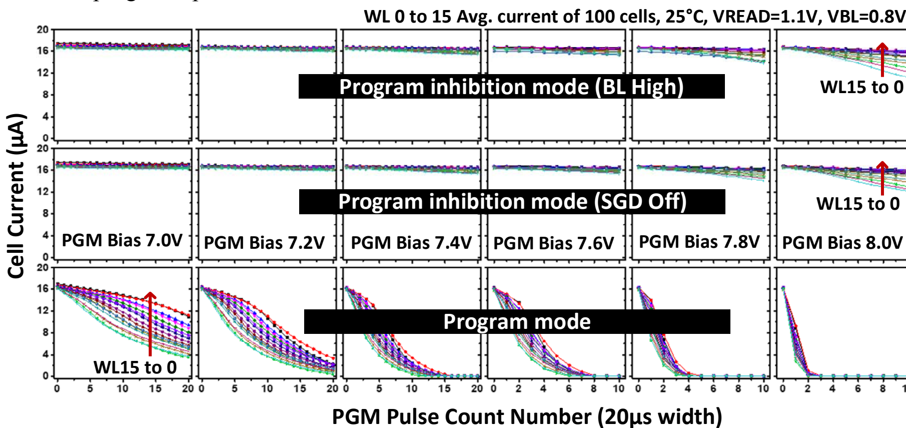


Fig. 7. Cell current versus the number of program pulses for two program inhibition modes (BL high and SGD off) and program mode (bottom row). The average current of 100 cells is shown for each wordline and 0.2V program bias increments from 7.0V to 8.0V for a constant pulse width of 20μs. Test chip data shows reliable programming with minimal program disturbance.

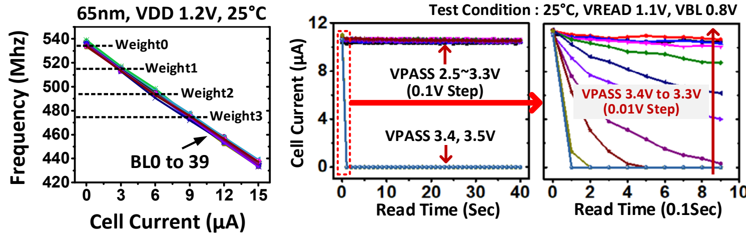


Fig. 8. VCO frequency versus cell current for BL0 to 39, and different weights.

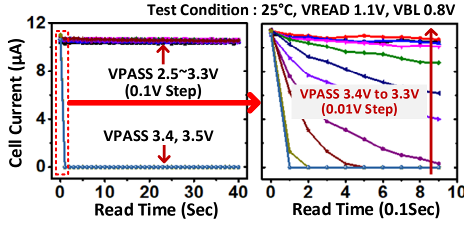


Fig. 9. VPASS disturbance characteristics of eNAND cells during read operation. Cell current remains constant for VPASS<3.3V.

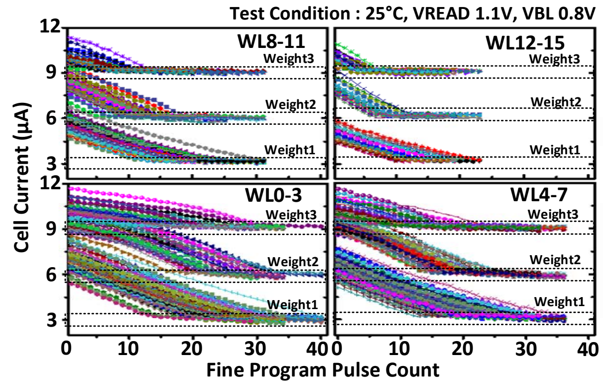


Fig. 11. Cell current versus program pulse count. The program-verify operation ensures precise weight storage.

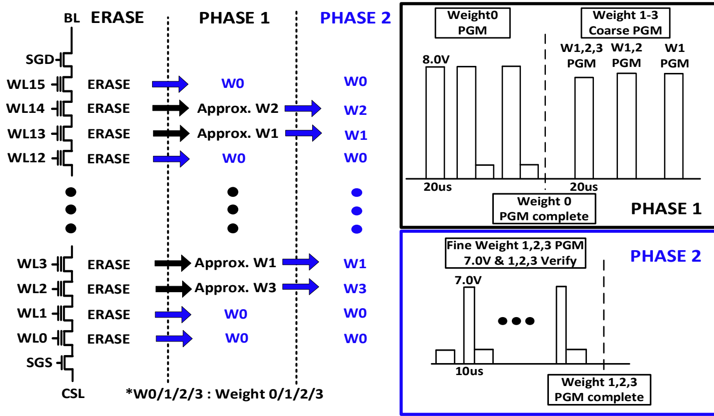


Fig. 10. Pulse sequence for programming weights 0, 1, 2 and 3 into the 16 stack eNAND array. 2 bit weight segments can be programmed into each cell using the proposed back-pattern tolerant program-verify scheme.

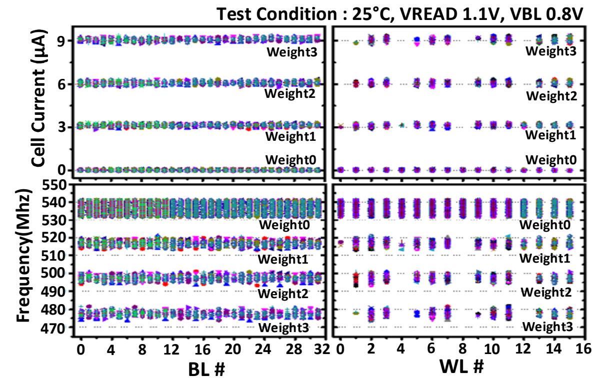


Fig. 12. Individual cell currents and frequency measurements in bitline and wordline directions for MNIST trained weights

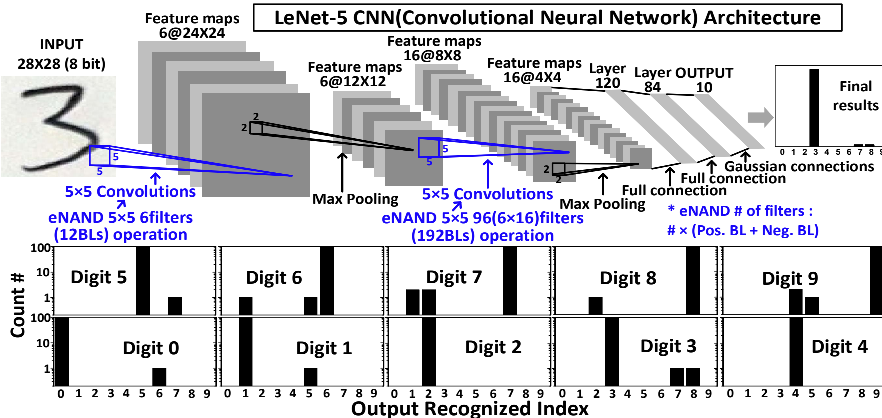


Fig. 13. (Upper) LeNet-5 Convolutional Neural Network (CNN) flow [8] using the proposed neural network core. (Low) Hand written digit recognition results measured from the test chip for 1,000 8 bit grayscale MNIST test images.

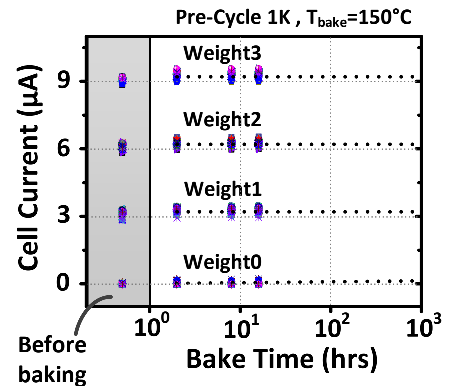
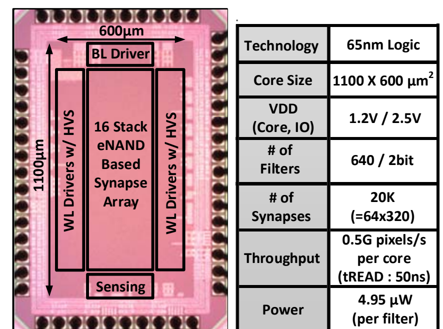


Fig. 14. Retention characteristics of weight 0, 1, 2 and 3 cell currents confirm minimal charge loss. Baking temperature was 150°C.

	This work	ISSCC'19 [4]	ISSCC'19 [5]	IEDM'18 [3]	ISSCC'18 [2]	IEDM'17 [1]
Technology	65nm	55nm	55nm	65nm	65nm	180nm
Voltage	1.2V	1.0V	1.0V	1.0V	1.0V	2.7V
Cell Type	NAND	NOR	NOR	NOR	NOR	NOR
Non volatile?	YES (eFlash)	YES (ReRAM)	NO (SRAM)	YES (eFlash)	YES (ReRAM)	YES (eFlash)
Logic Compatible?	YES	NO	YES	YES	NO	NO
Program-verify?	YES	NO	NO	YES	NO	YES
Weight Resolution	8 Bits	3 Bits	5 Bits	2.3 Bits	3 Bits	2 Bits
Input Resolution	8 Bits	2 Bits	2 Bits	1 Bit	3 Bits	1 Bit
# of Currents Summed Up	28 Cells	8 Cells	32 Cells	68 Cells	14 Cells	4 Cells
Neural Net Architecture	CNN	CNN	CNN	MLP	CNN	MLP

Fig. 15. Comparison with prior works.



Technology	65nm Logic
Core Size	1100 X 600 μm^2
VDD (Core, IO)	1.2V / 2.5V
# of Filters	640 / 2bit
# of Synapses	20K (=64x320)
Throughput	0.5G pixels/s per core (tREAD : 50ns)
Power	4.95 μW (per filter)

Fig. 16. Die microphotograph and test chip feature summary.