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]

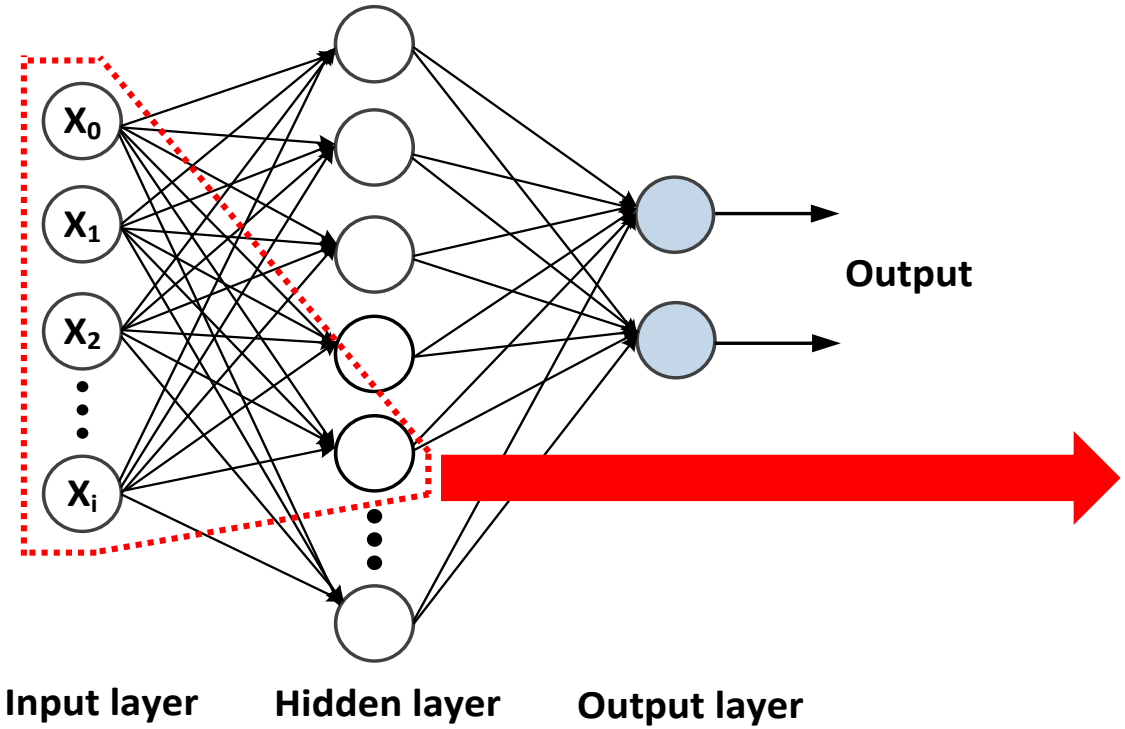
[1]Dept. of ECE, University of Minnesota
[2]Anaflash Inc.

Outline

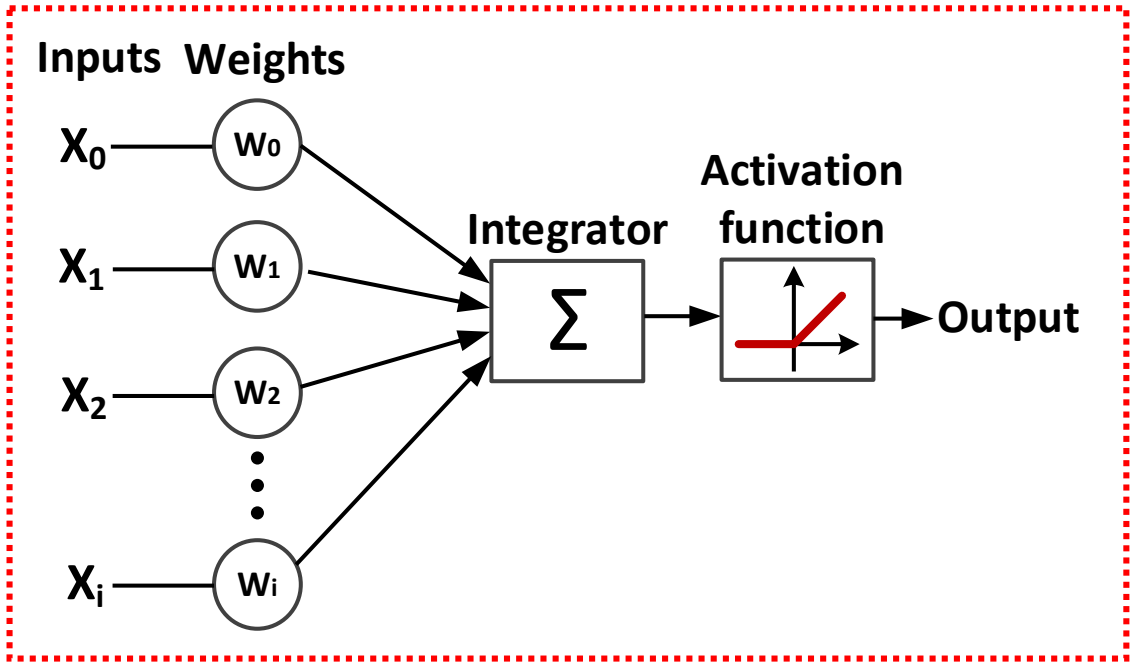
- Background
- Logic-Compatible eFlash based Synapse
- Neuromorphic Core Design
- 65nm Test Chip Results
- Conclusions

Artificial Neural Network (ANN)

Multi-layer perceptron (MLP)



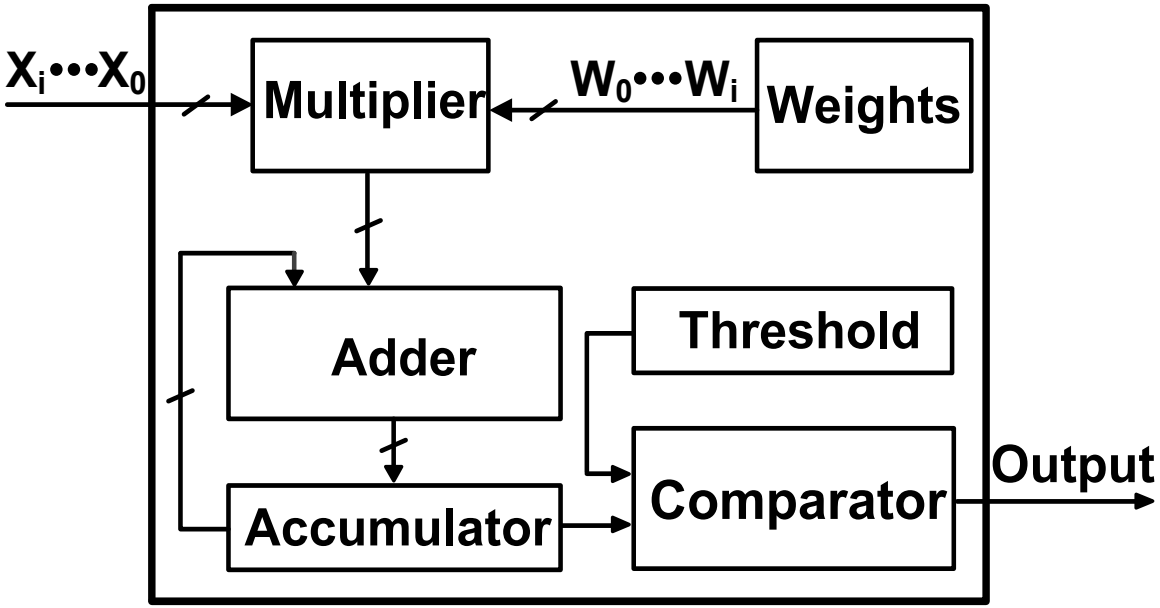
Unit perceptron



- MLP: input/hidden/output layers
- Unit perceptron: multiply-accumulate operation + activation function

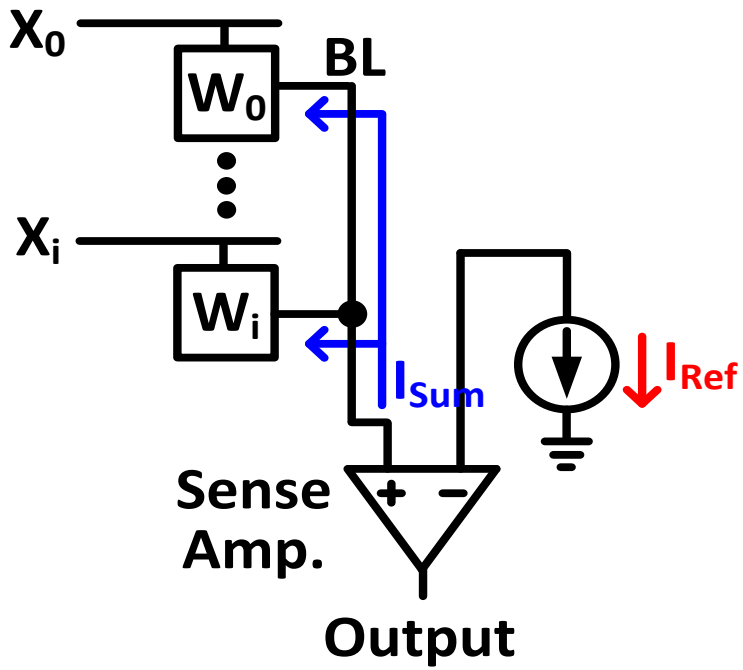
ANN Digital vs Analog Implementation

Digital implementation



- Pros : Digital CMOS
- Issues : Large area, large power consumption

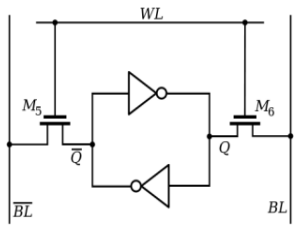
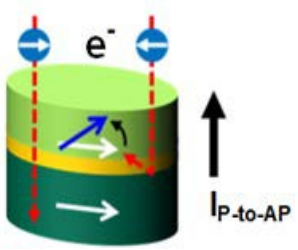
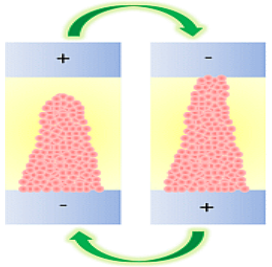
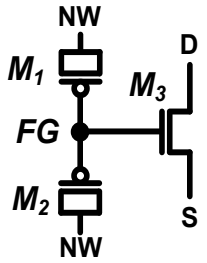
Analog Implementation



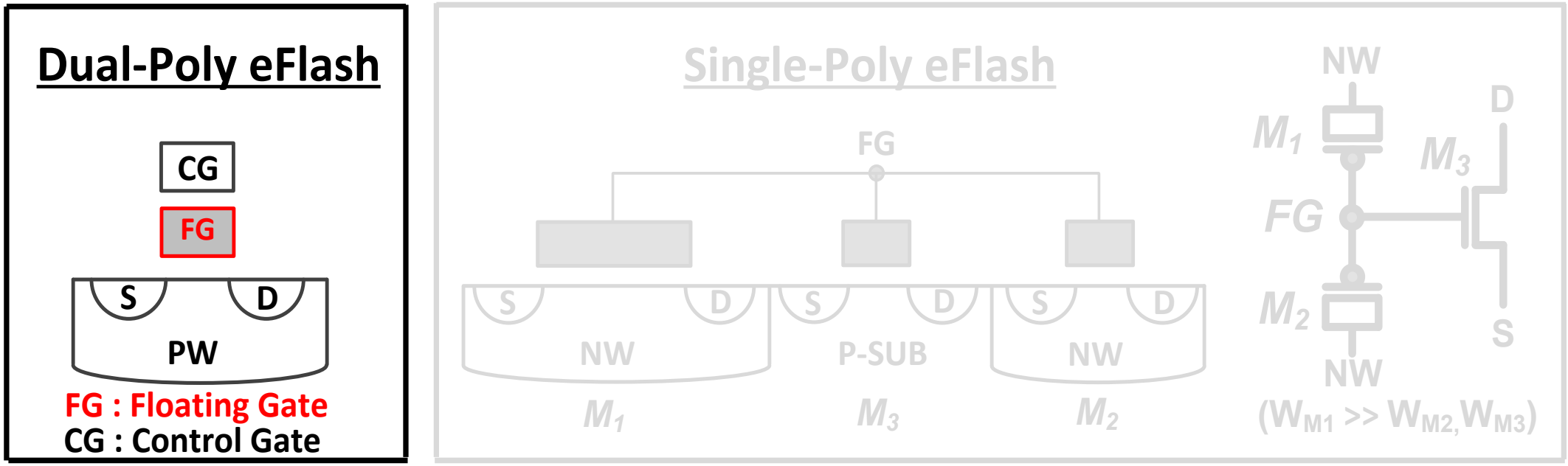
- Pros : Current summation replaces complex digital blocks
- Issues: Sensitive to PVT variation, requires good synaptic device

Memory Options for Synapse Circuit

This work

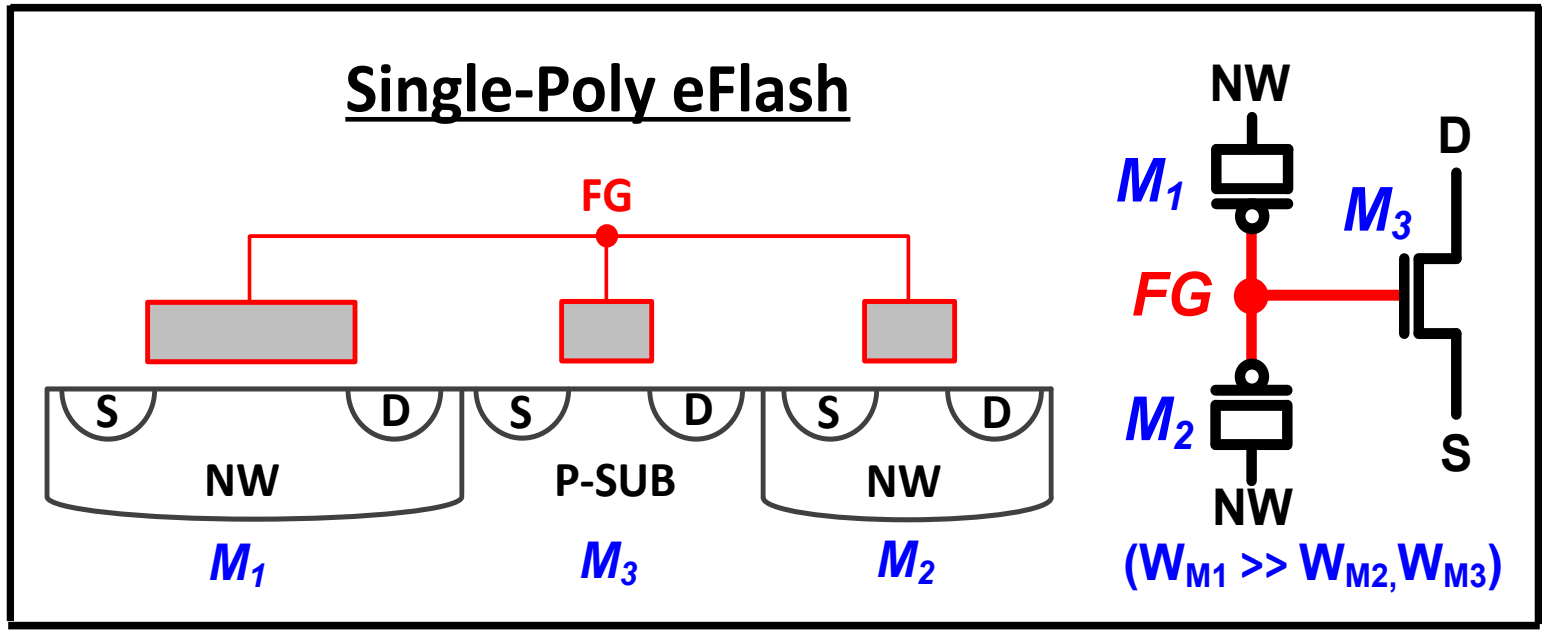
Device	SRAM	MRAM	RRAM	PCRAM	eFlash
Cell Configuration					
Nonvolatile?	No	Yes	Yes	Yes	Yes
Tunable?	No	No	No	Yes	Yes
Logic Compatible?	Yes	Not yet	No	No	Yes
Multi level Weights?	No	No	No	Yes	Yes
Area/bit	$150F^2$	$40F^2$	$60F^2$	$40F^2$	$\sim 500F^2$

Dual-poly vs. Single-poly eFlash



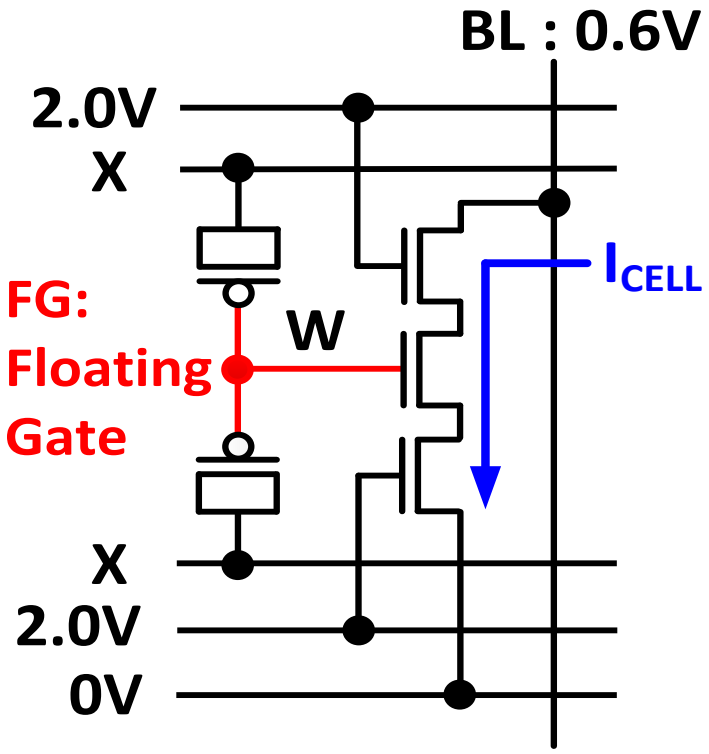
- **Needs additional masks to form floating gate (FG)**

Proposed Synapse: Logic Compatible eFlash

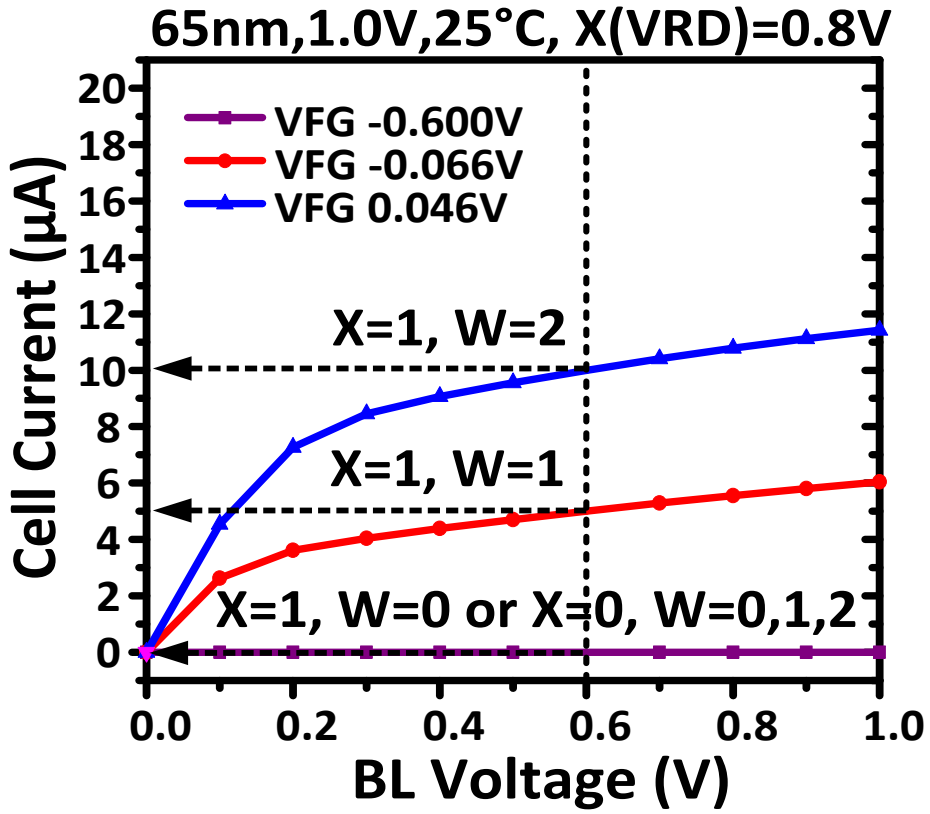


- Floating gate: Back-to-back connected gate
- Logic-compatible nonvolatile memory solution
- Program verify allows precise weight programming

Proposed Synapse: Logic Compatible eFlash

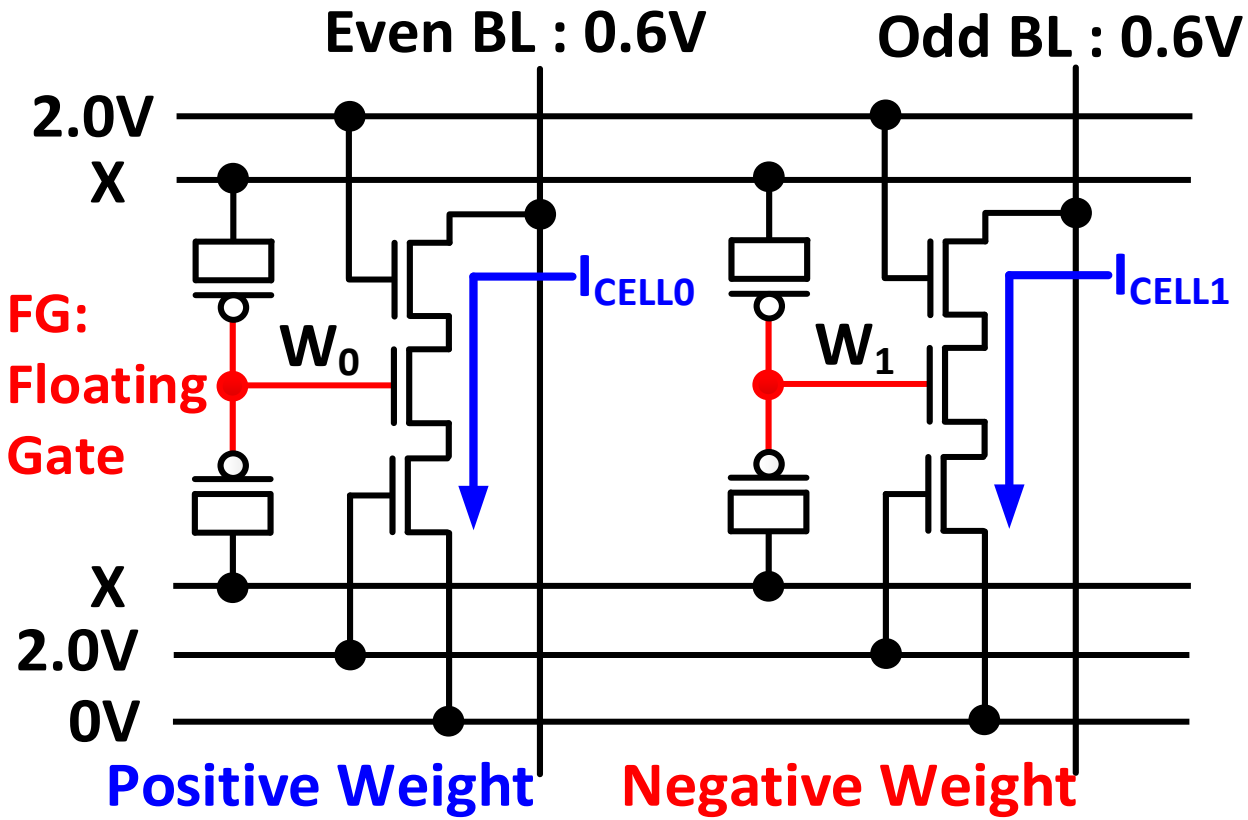


X	W	X·W	I_{CELL}
0	0	0	0μA
0	1	0	0μA
0	2	0	0μA
1	0	0	0μA
1	1	1	5μA
1	2	2	10μA



- Cell current proportional to X·W (=0μA, 5μA, or 10μA)
- BL voltage pinned at 0.6V during read operation

Proposed Synapse: Logic Compatible eFlash

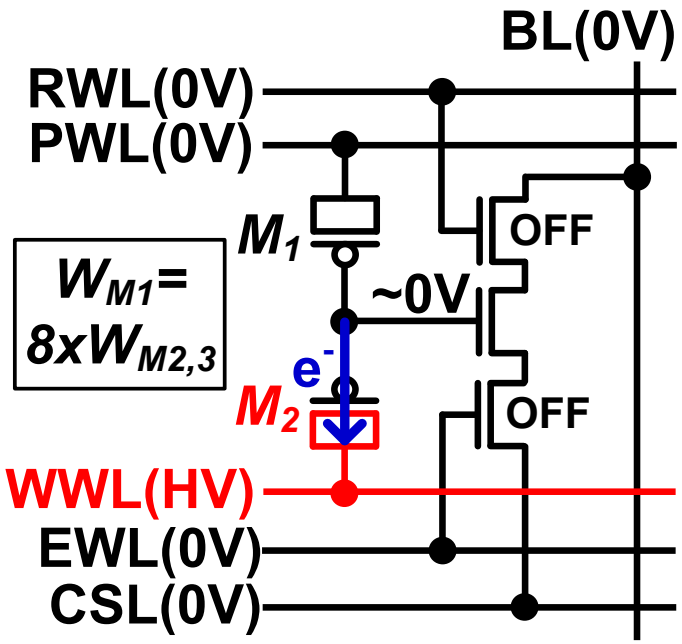


X	W_0	W_1	$X \cdot W_0$	$X \cdot W_1$	I_{CELL0}	I_{CELL1}	ΔI	
1	0	2	0	2	0μA	10μA	-10μA	Negative Weights
1	0	1	0	1	0μA	5μA	-5μA	
1	0	0	0	0	0μA	0μA	0μA	
1	1	0	1	0	5μA	0μA	5μA	Positive Weights
1	2	0	2	0	10μA	0μA	10μA	

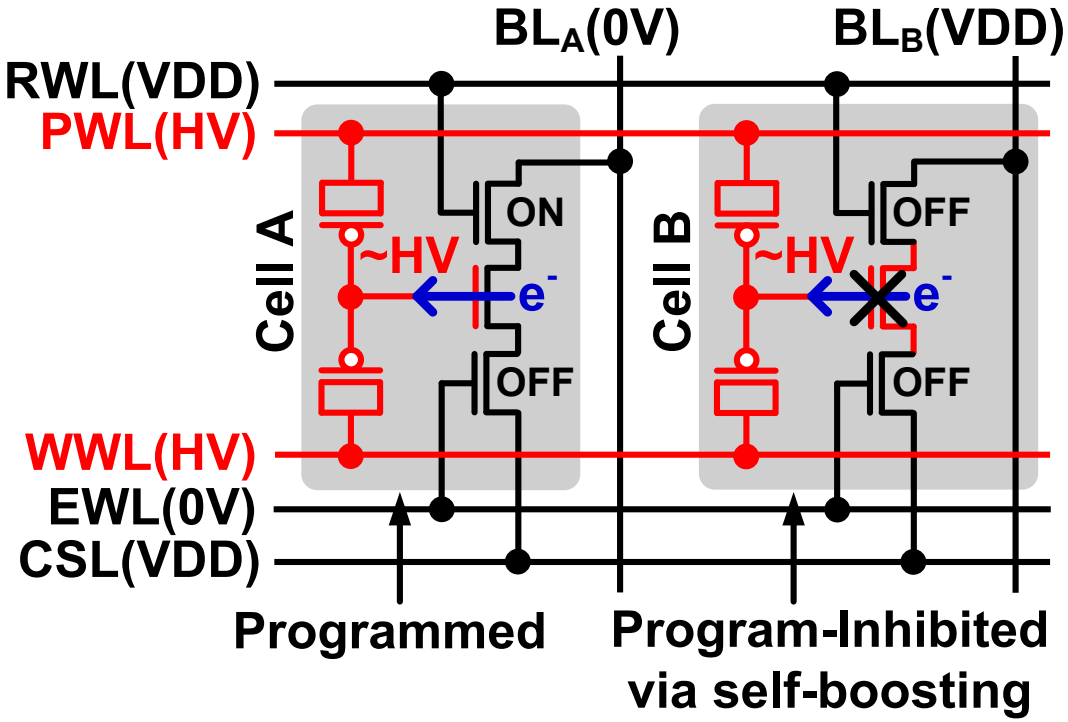
- Even and odd bitline pair realizes 5 level weights

Detailed Erase and Program Operations

Erase Operation



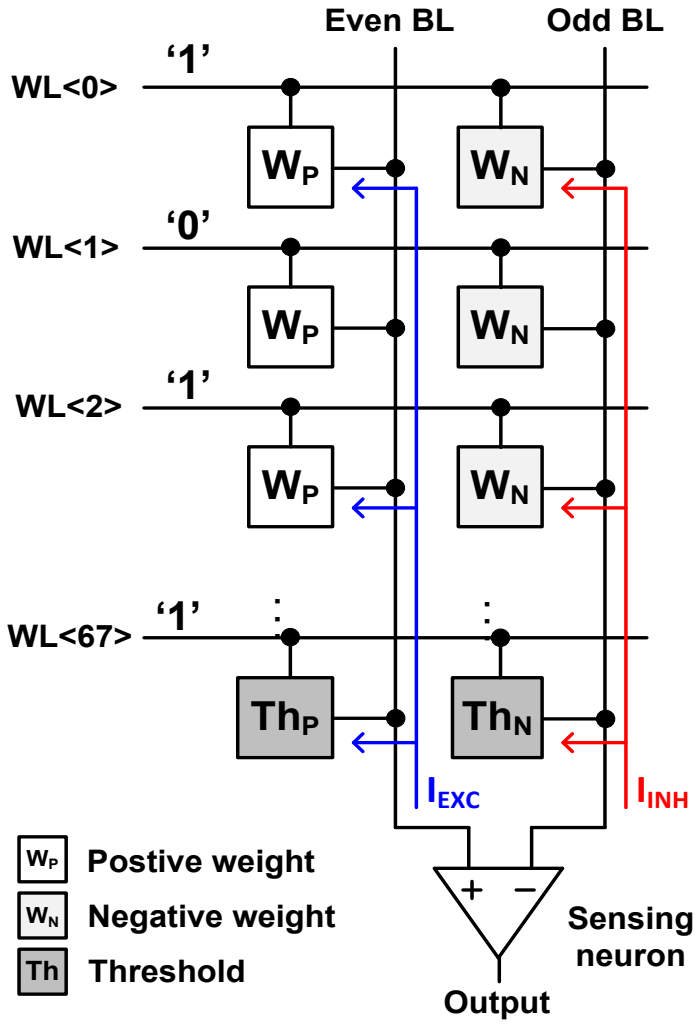
Program Operation



[8] S. Song, JSSC 2013 (UMN)

- FN tunneling utilized for erase and program
- Program inhibition of unselected cells via self-boosting

Proposed Bitline Pair Implementation

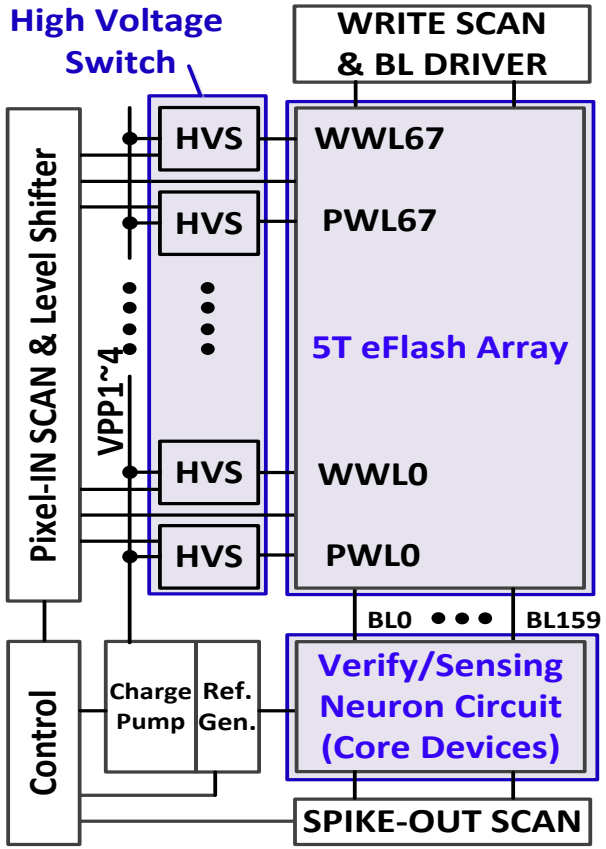


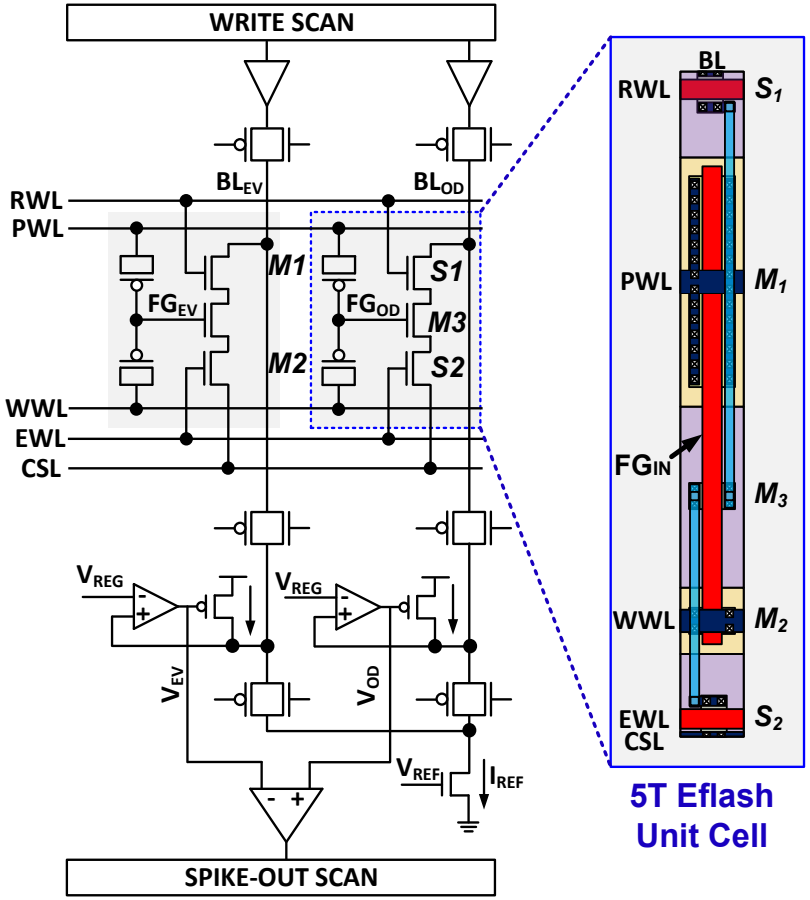
- **Spiking criteria**

$$\sum W_P - (-Th_P) > \sum W_N + (+Th_N)$$

- **Output generated in a single current summation and thresholding cycle**

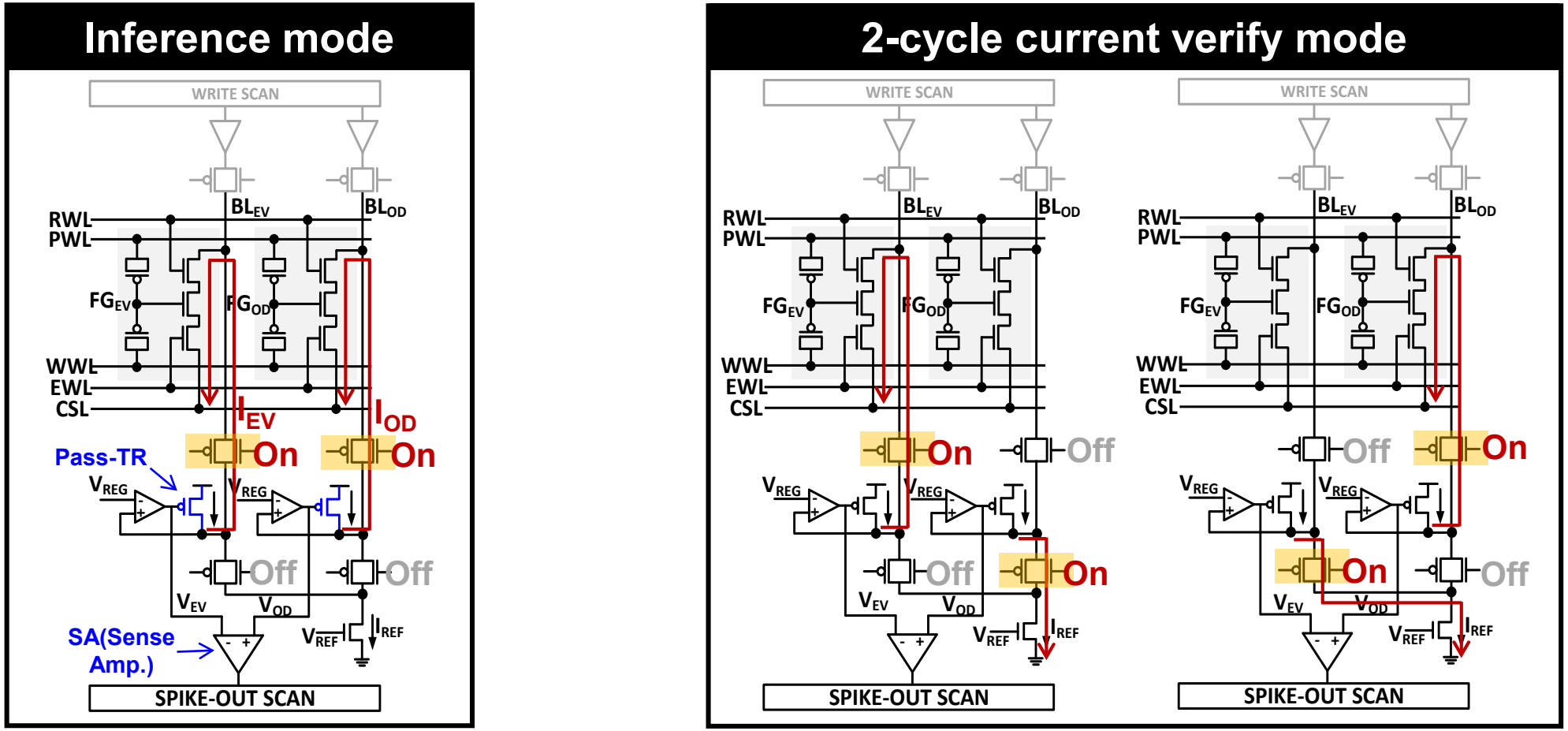
68 Row x 160 Column Core Architecture





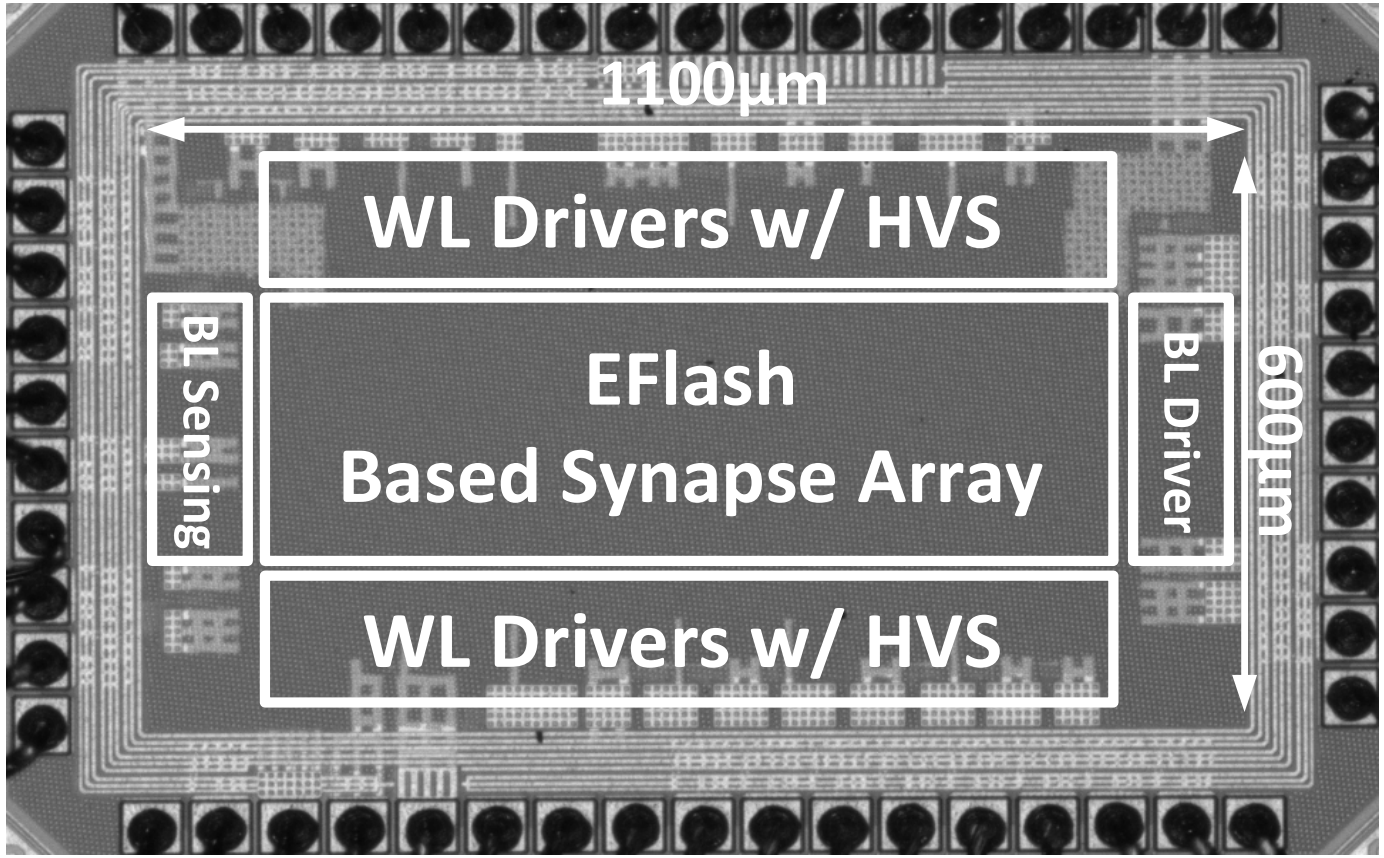
- 5T eFlash array, high voltage switches, BL sensing circuits
- Input data loaded on to 64 read wordlines, 4 rows for threshold

Inference and Verify Operations



- **Inference mode: Compares BL pair currents**
- **Verify mode: Compare BL currents with a common ref. current**

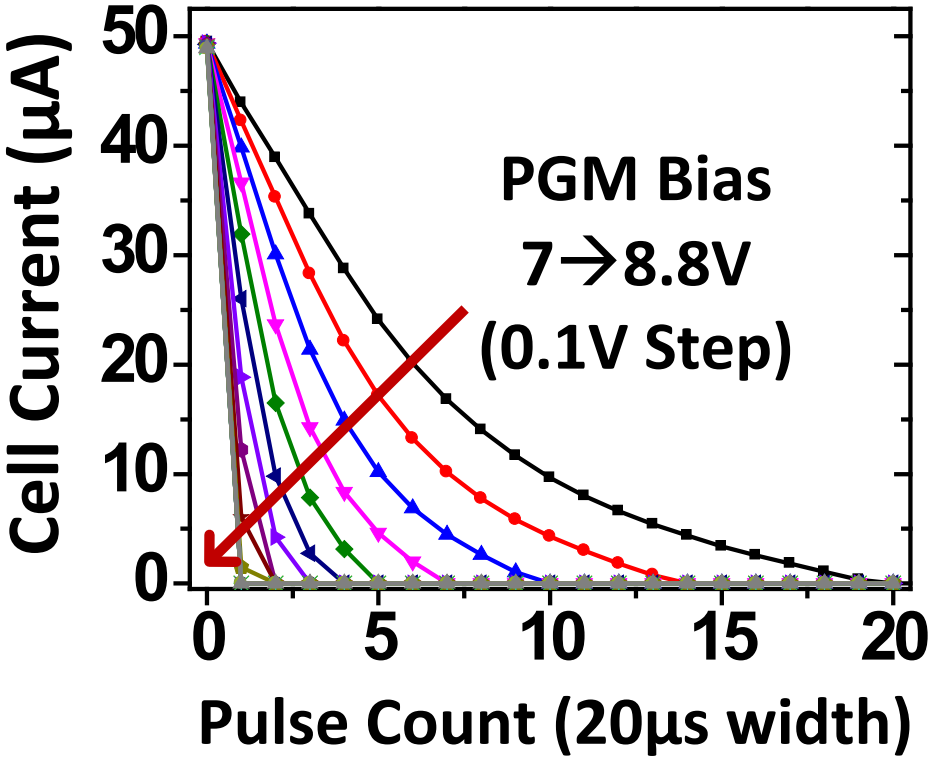
65nm Die Photo and Feature Summary



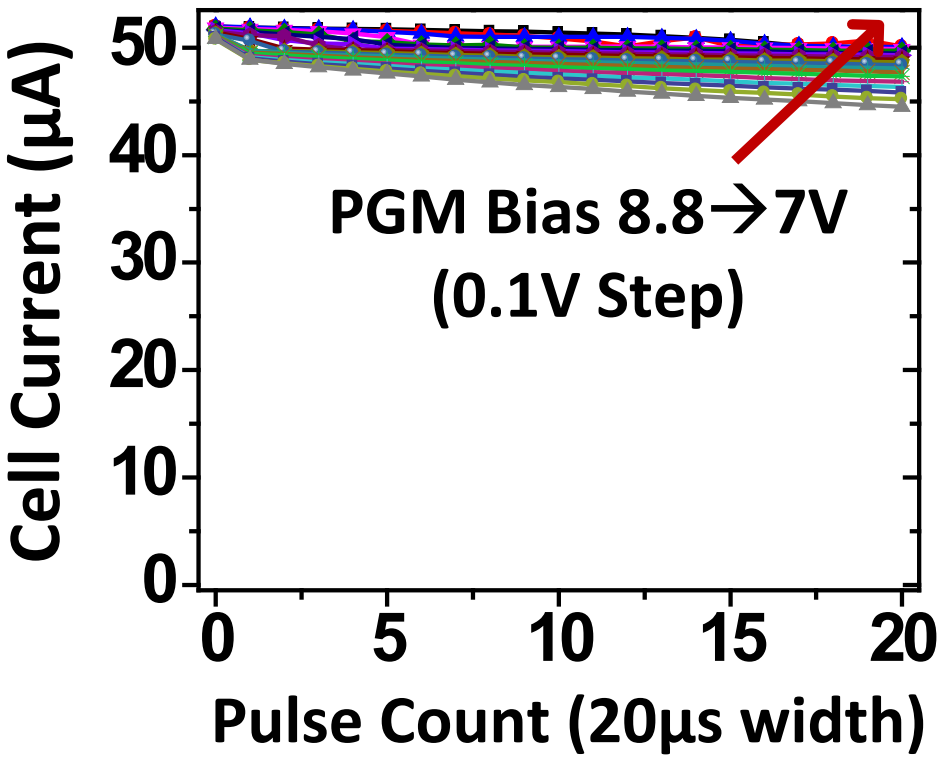
Technology	65nm CMOS
Circuit Area	1100 X 600 μm^2
VDD (Core, IO)	1.0V / 2.5V
# of Neurons	320
# of Synapses	22K (=68x320)
Throughput	1.28G pixels/s per core (tREAD : 50ns)
Power	15.9µW (per neuron)

Program and Program Inhibition Characteristics

Avg. current of 100 cells, 25°C, VRD=0.8V, VBL=0.6V

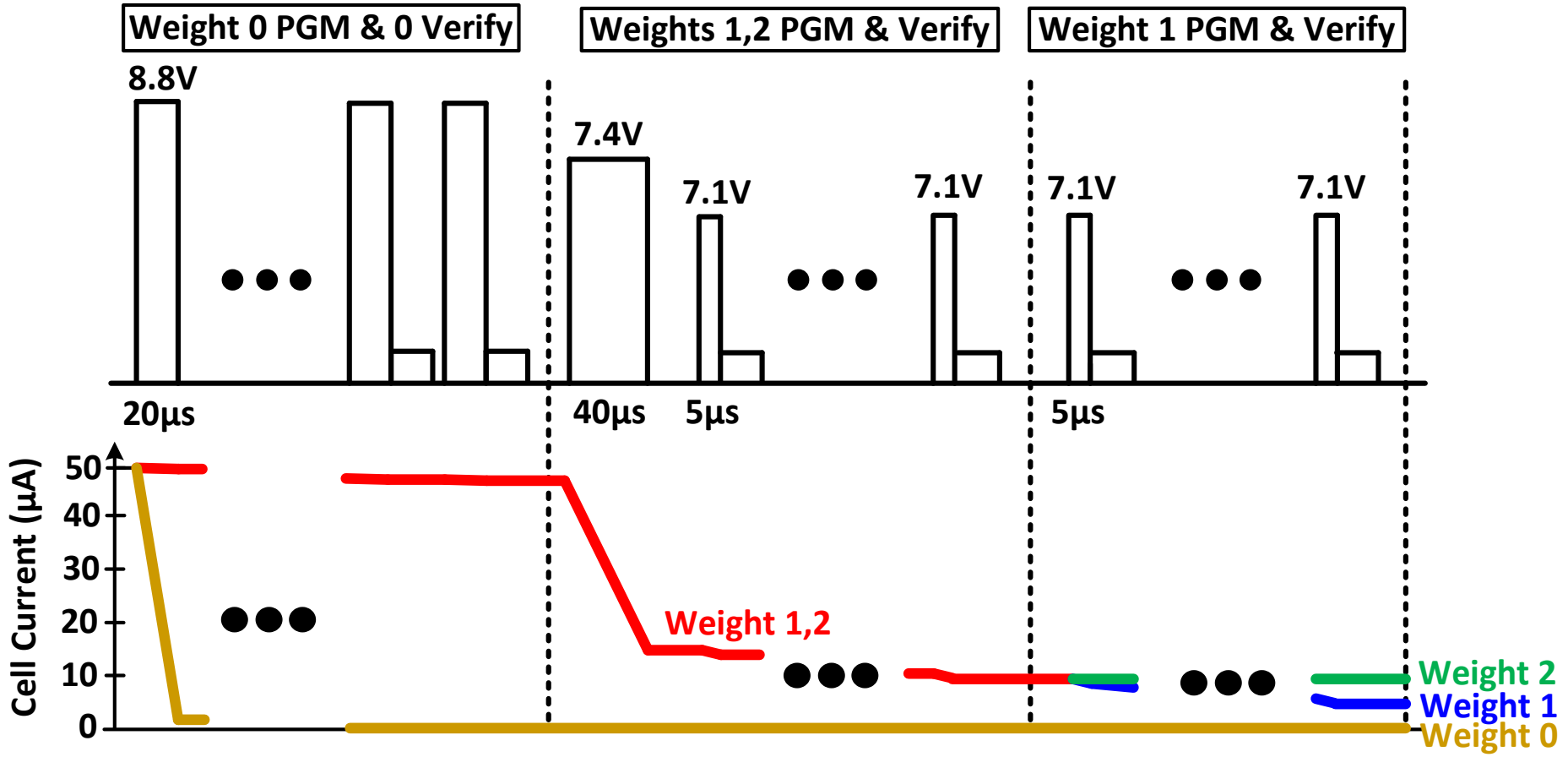


Program mode



Program inhibition mode

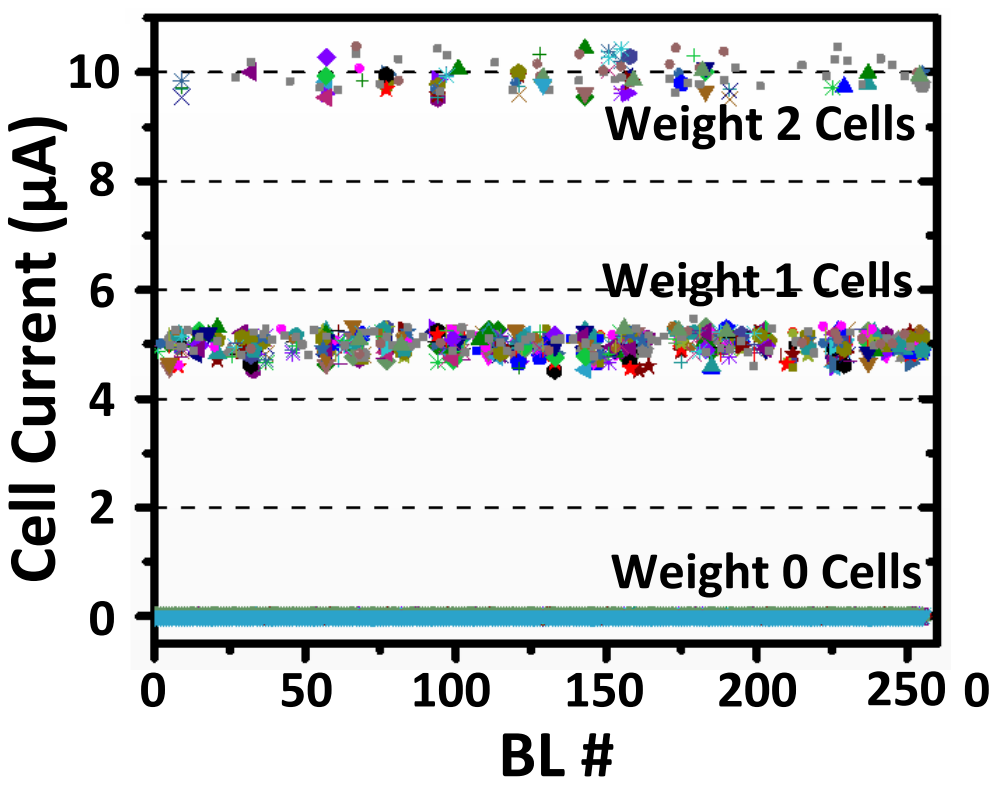
Multi-level Program Sequence

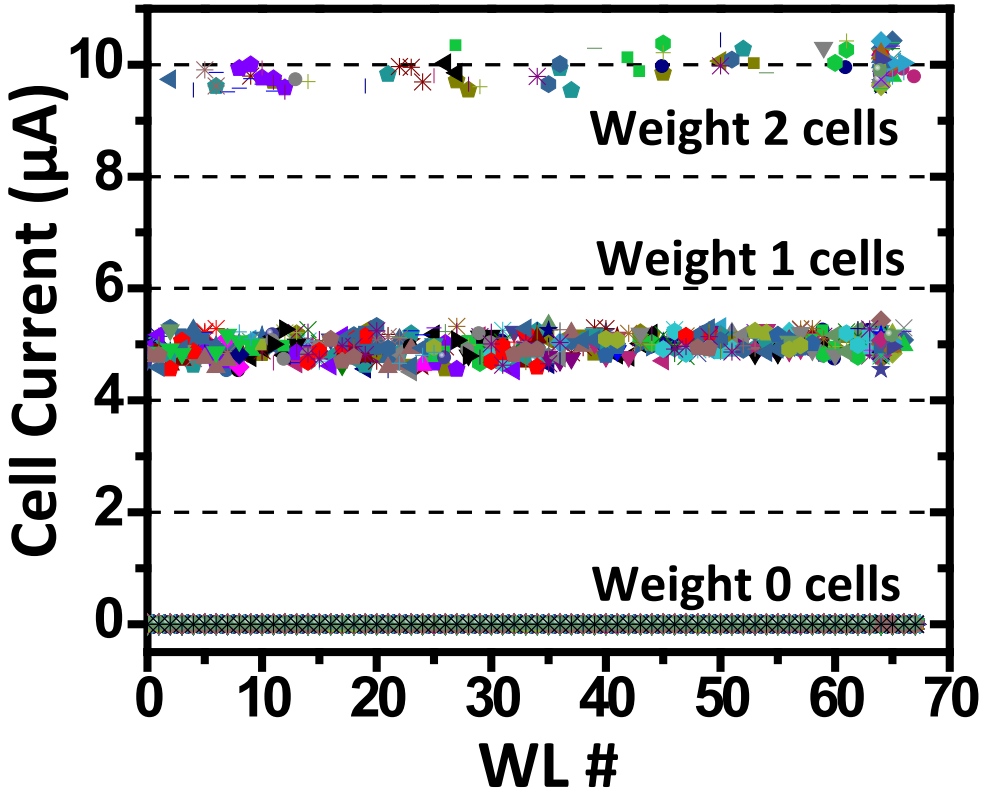


- **Program bias: 8.8V → 7.4V → 7.1V**
- **Target cell current: 0µA, 5µA , and 10µA**

Programmed Cell Current Variation

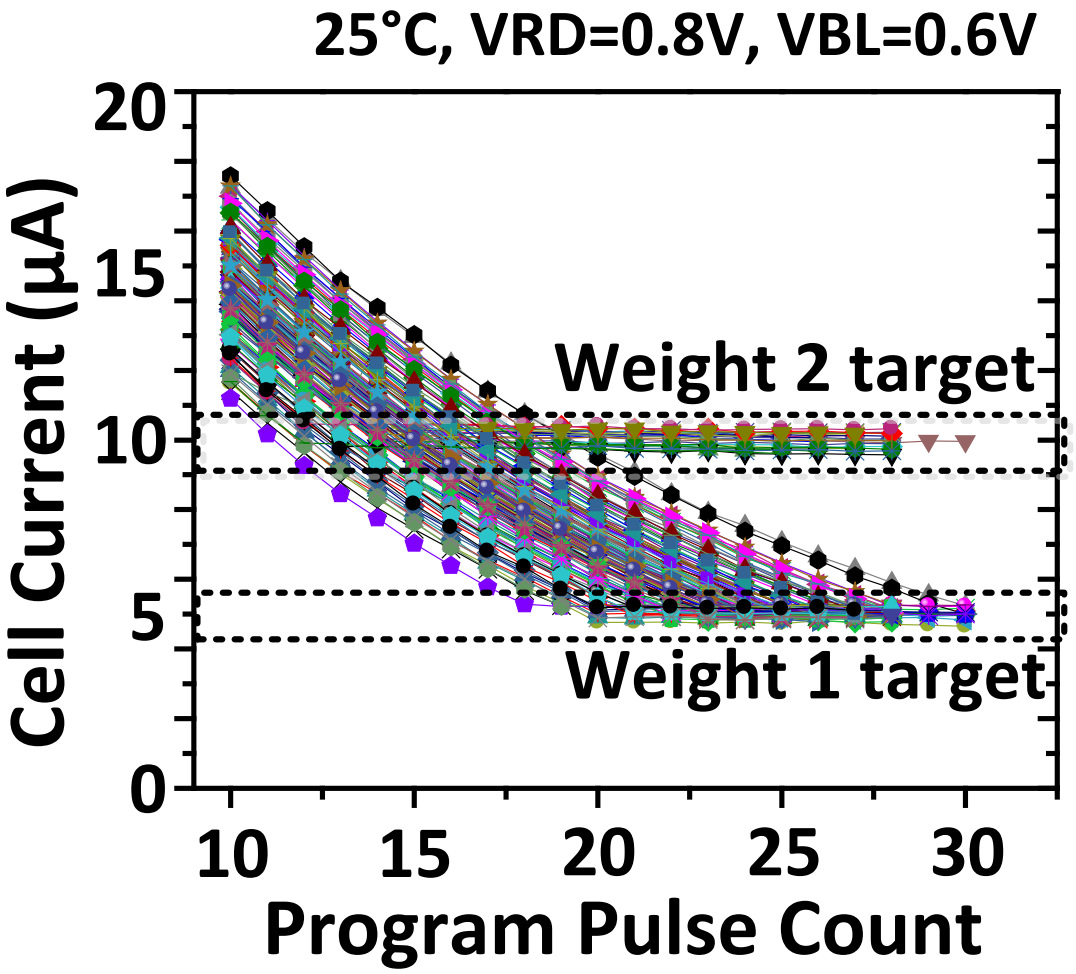
Weight 0,1,2 Cell Current Variation, 25°C, VRD=0.8V, VBL=0.6V





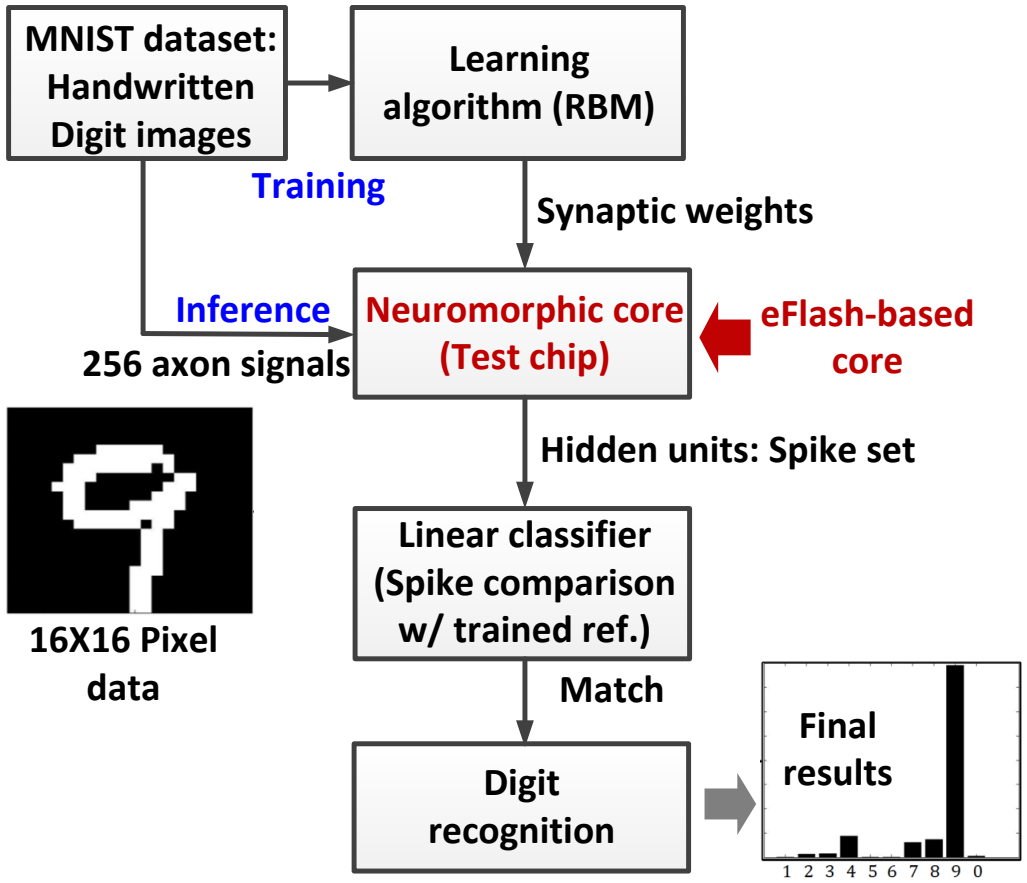
- Cell current variation less than 0.8µA after program-verify operations
- Data shows that a higher number of levels is possible

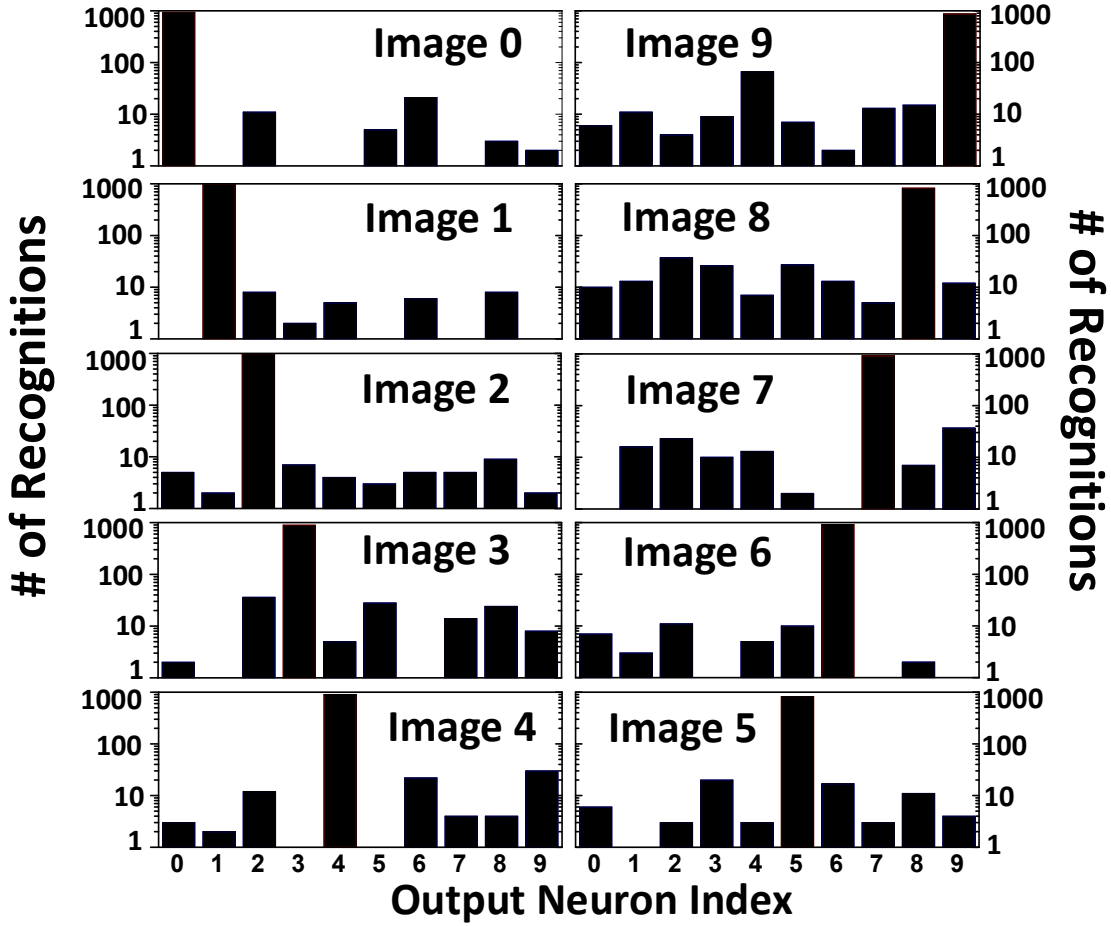
Number of Program Pulses



- Initial cell current variation is 8μA
- Weight 2 and weight 1 cells completely programmed after 20 and 30 pulses, respectively.

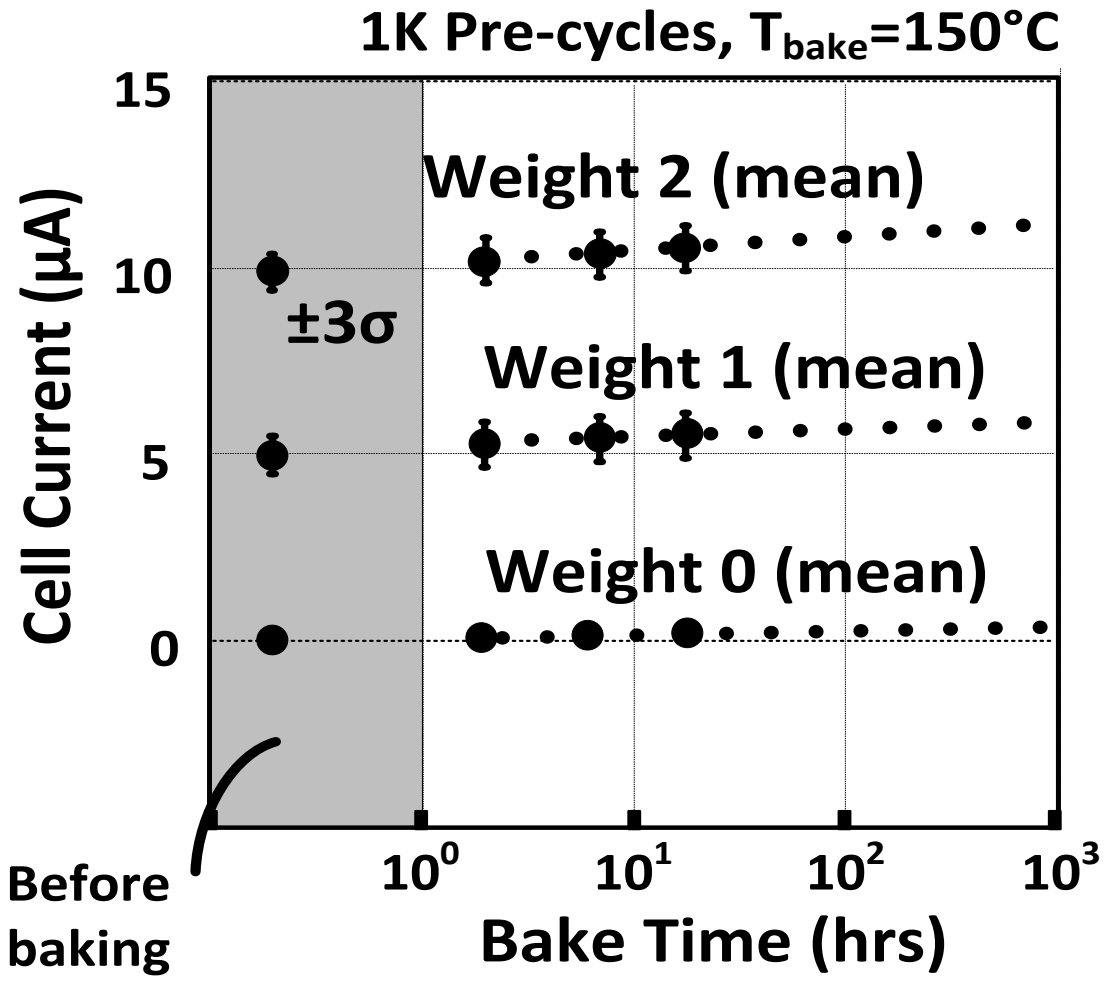
MNIST Digit Recognition Accuracy Results





- Recognition accuracy is 91.8% for 10K MNIST test images, which is close to the software model's 93.8% accuracy

Retention Test Results



- **Excellent retention characteristics**
  - **Reprogramming is not necessary**
  - **A higher number of levels is possible**

Comparison Table

	This work	ISSCC'18 [6]	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

[2] W. Chen, et al., IEDM, 2017.
[3] X.Guo, et al., IEDM, 2017.
[4] W.Khwa, et al., ISSCC, 2018.
[5] S. Gonugondia, et al., ISSCC, 2018.
[6] W.Chen, et al., ISSCC, 2018.

Conclusions

- **A logic-compatible 5T eFlash based neuromorphic core demonstrated in 65nm CMOS**
- **Key features**
  - **Non-volatile weight storage in a logic CMOS process**
  - **Precise multi-level weights enabled by program verify**
  - **68 row parallel access**
- **Test chip results show 91.8% digit recognition accuracy**