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]

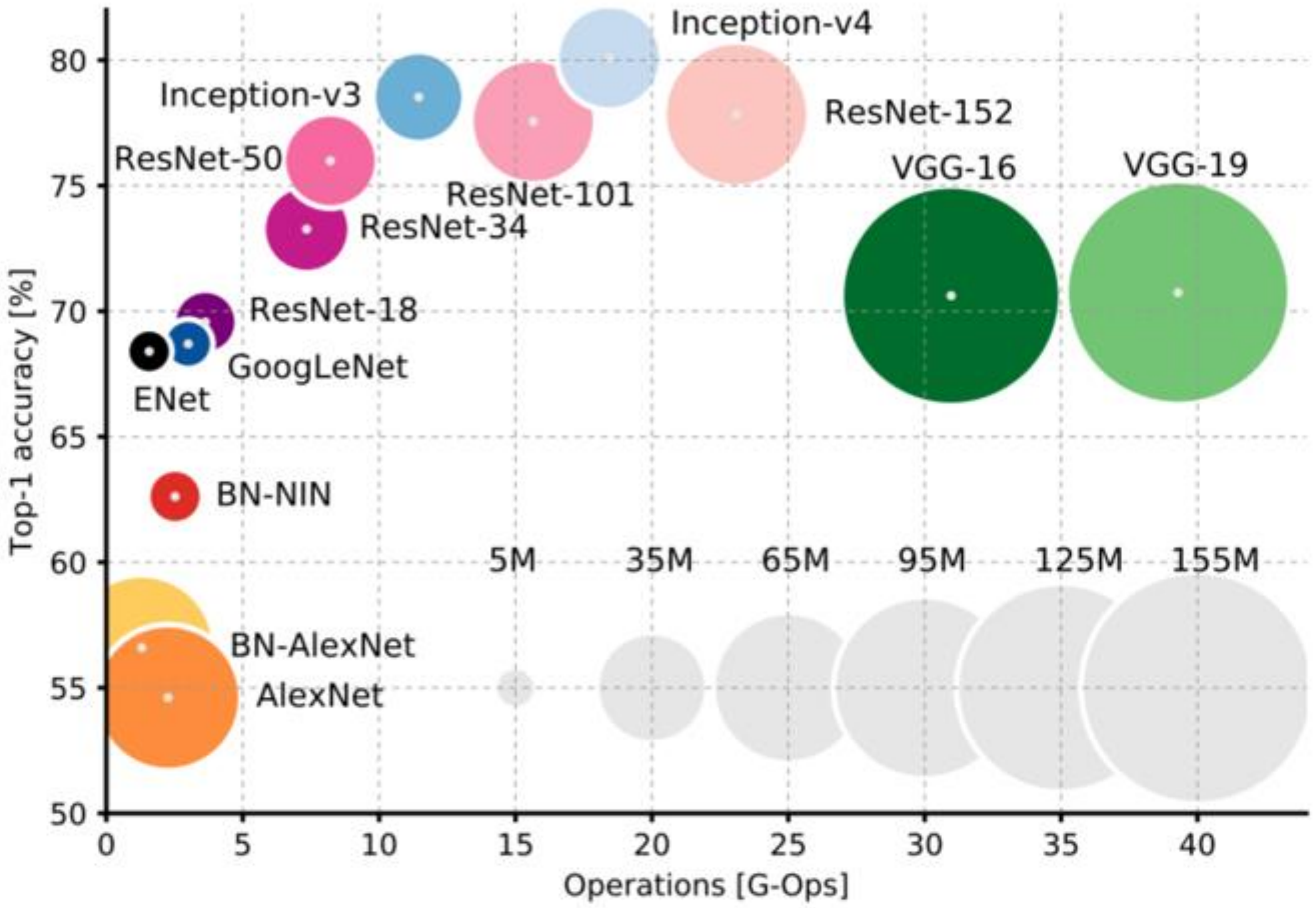
[1]Dept. of ECE, University of Minnesota

[2]Anaflash Inc.

Outline

- **3D NAND based Neural Network**
- **Prototyping in a Standard Logic Process**
  - **Architecture, synapse cell, memory array design**
- **65nm Test Chip Results**
  - **Programming results, MNIST demonstration, retention**
- **Conclusions**

Deep Neural Network Complexity

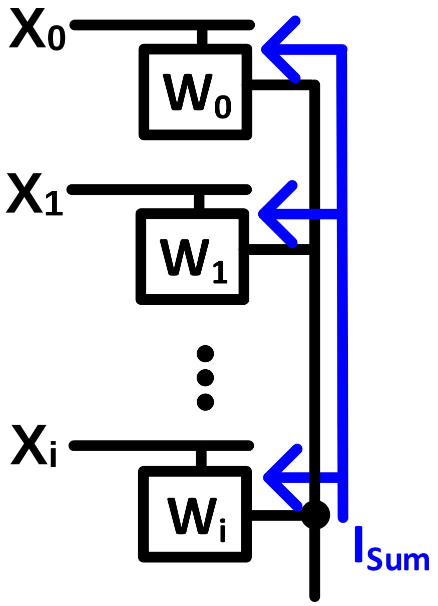


- **State-of-the-art deep neural networks: ~150 layers, ~150M parameters, ~100MB memory per image**

An Analysis of Deep Neural Network Models for Practical Applications, Arxiv, 2017

In-Memory Computing and Flash-based Design

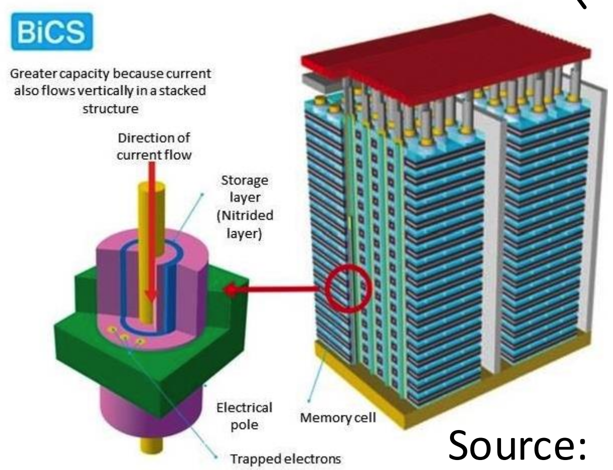
In-memory Computing (e.g. SRAM, RRAM)



Analog multiply and accumulate (MAC)

SRAM (e.g. 22MB) ⟺ Data ⟺ DRAM (e.g. 64GB) ⟺ Data ⟺ Flash SSD (e.g. 2TB)

- **Ultra-high density, reduced data traffic, low cost, mature technology**
- **Low program/erase speed (but fast read speed), limited endurance cycles (but fine for neural network applications)**

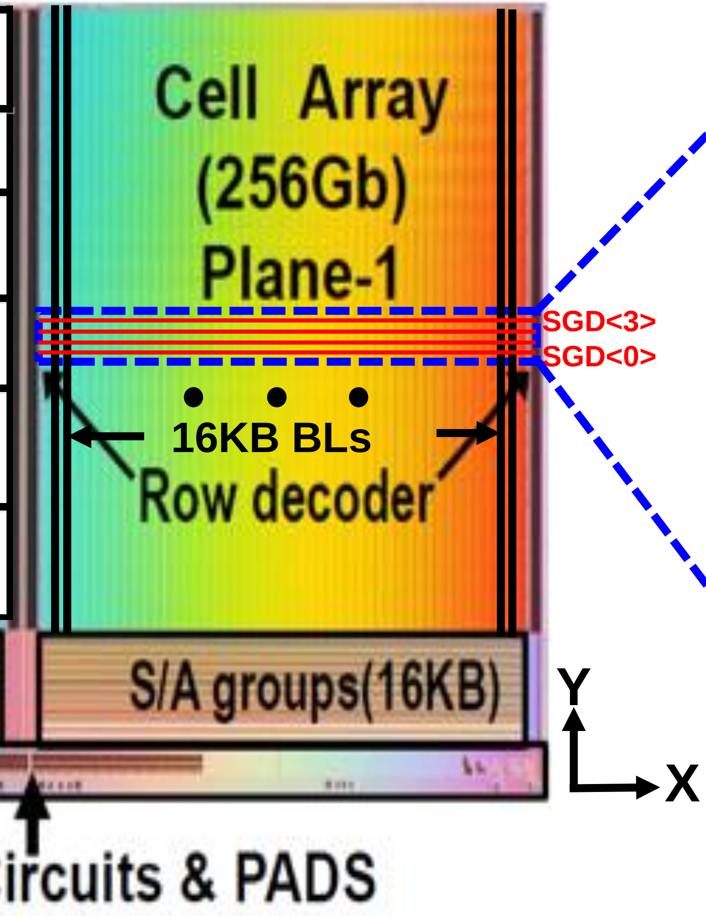


Source: Toshiba

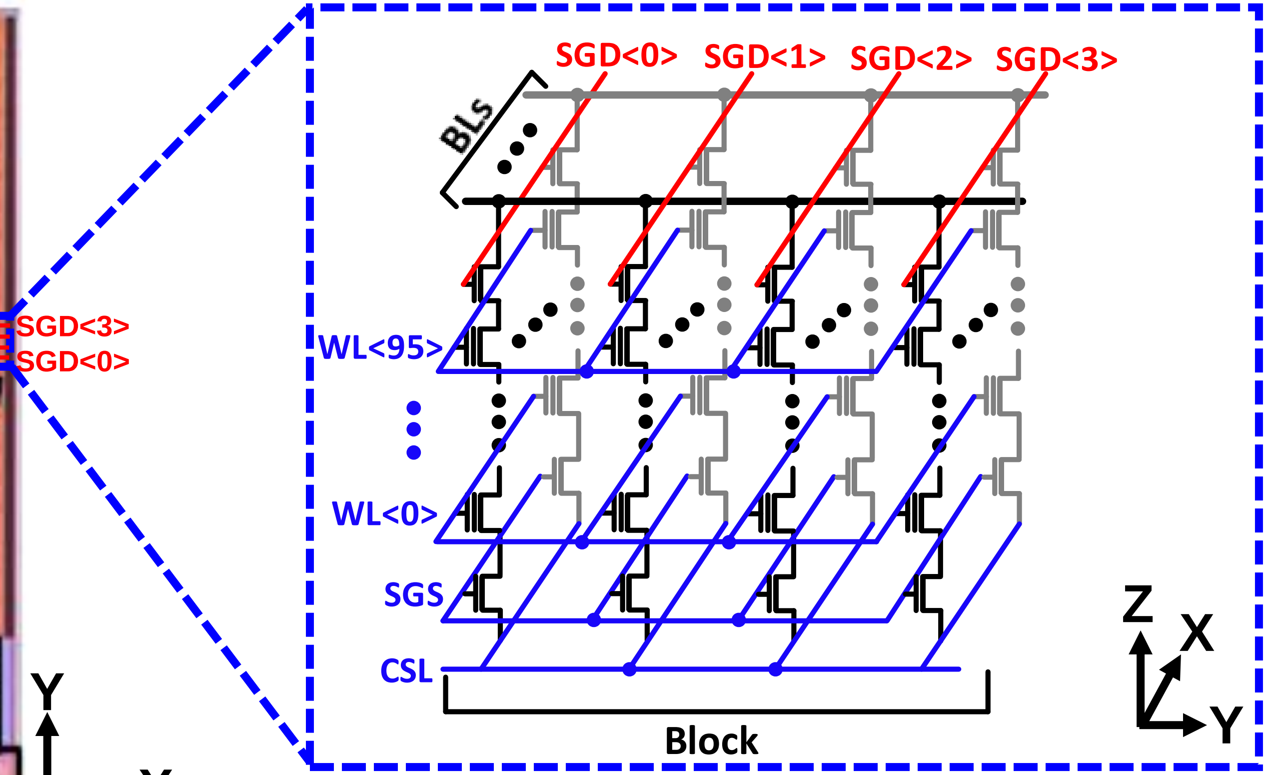
3D NAND Cell and Array

State-of-the-Art 3D NAND Flash Memory

Capacity	512Gb (3bit/cell)
Technology	96-WL-Layer
Bit Density	5.95Gb/mm^2
Cost	4 Cent/Gb
Organization	(1822 + EXT) Blocks / Plane
Throughput	Read(tR) : 58µs (ABL : 16KB)

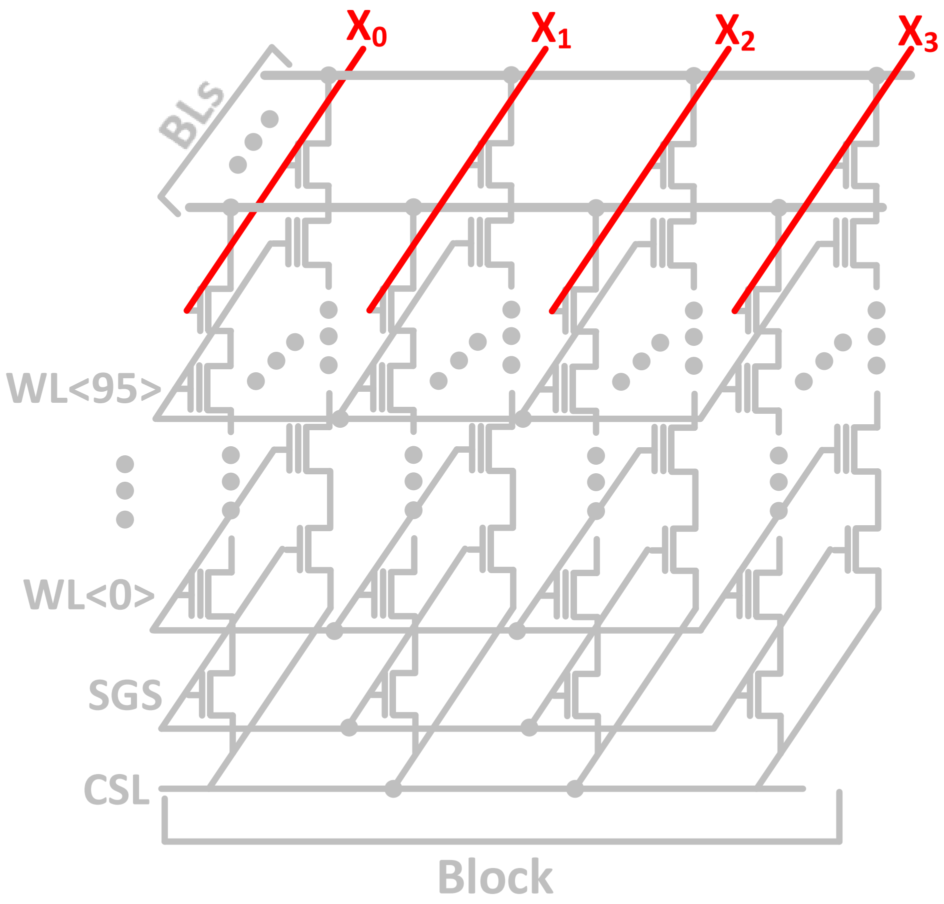


H. Maejima, Toshiba, ISSCC 2018



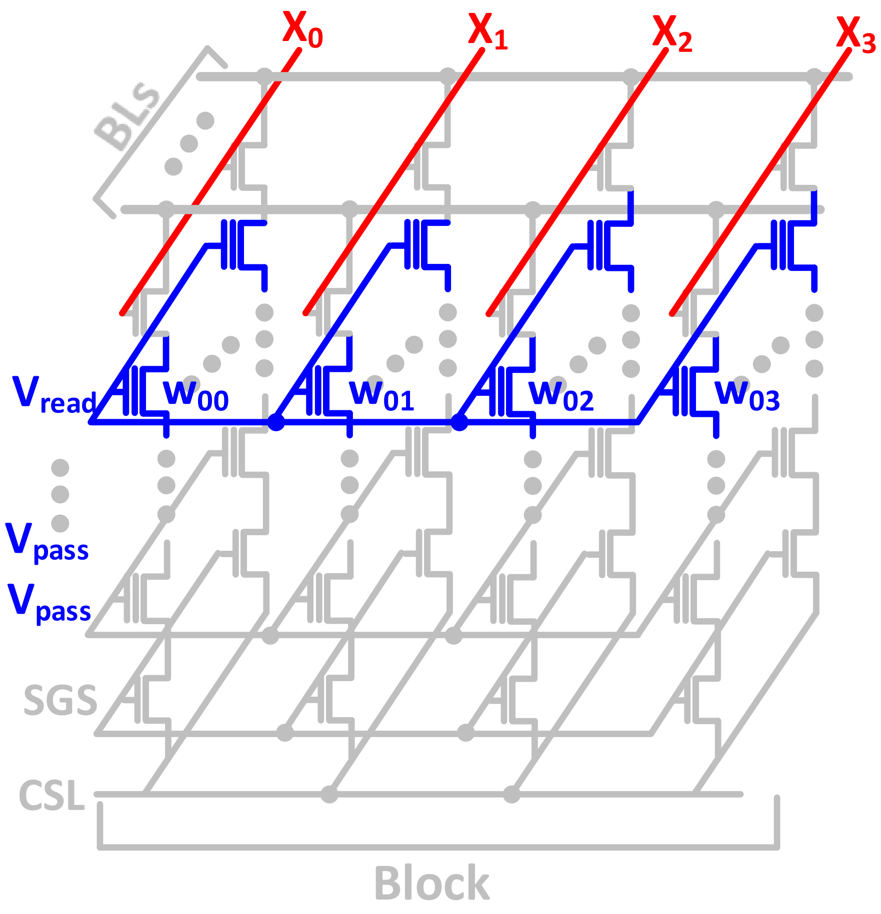
- (x,y,z) = (BL, SGD, WL)
  - BL shared across multiple blocks
  - WL is a shared plane (not line)
  - SGD can be individually controlled

Analog MAC in 3D NAND Array



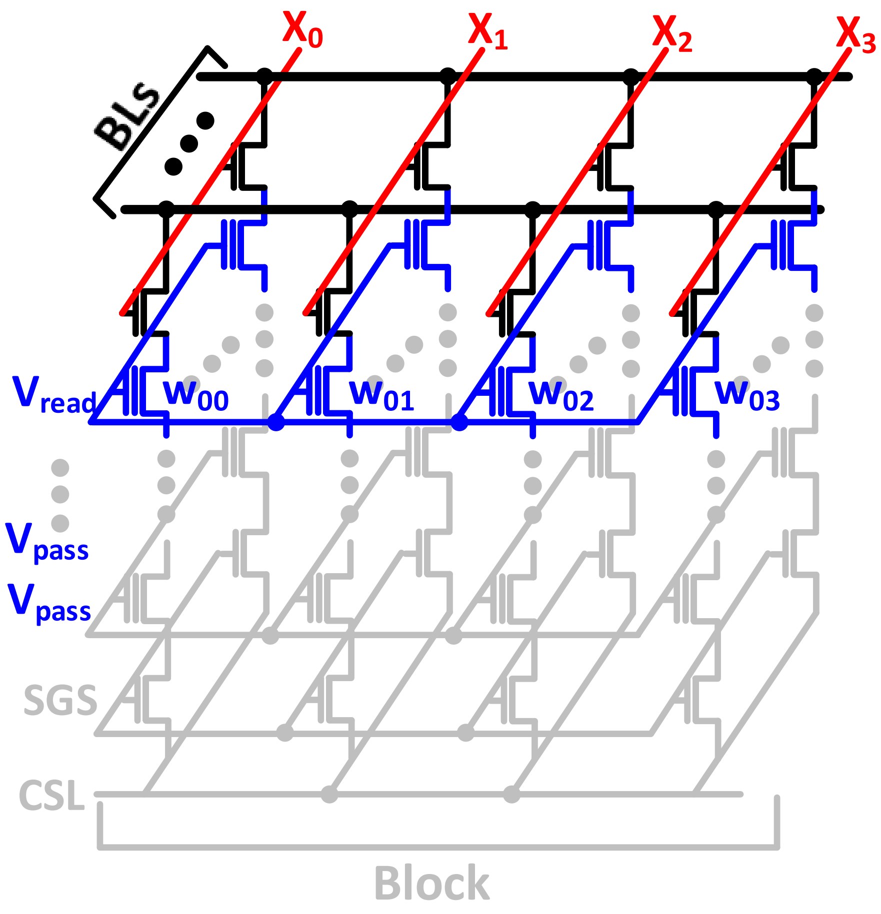
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - Σ (Accumulate) : Bitline currents of different blocks summed up
  - High resolution MAC can be realized using multiple weight cells, bit serial operation, and partial product post-processing

Analog MAC in 3D NAND Array



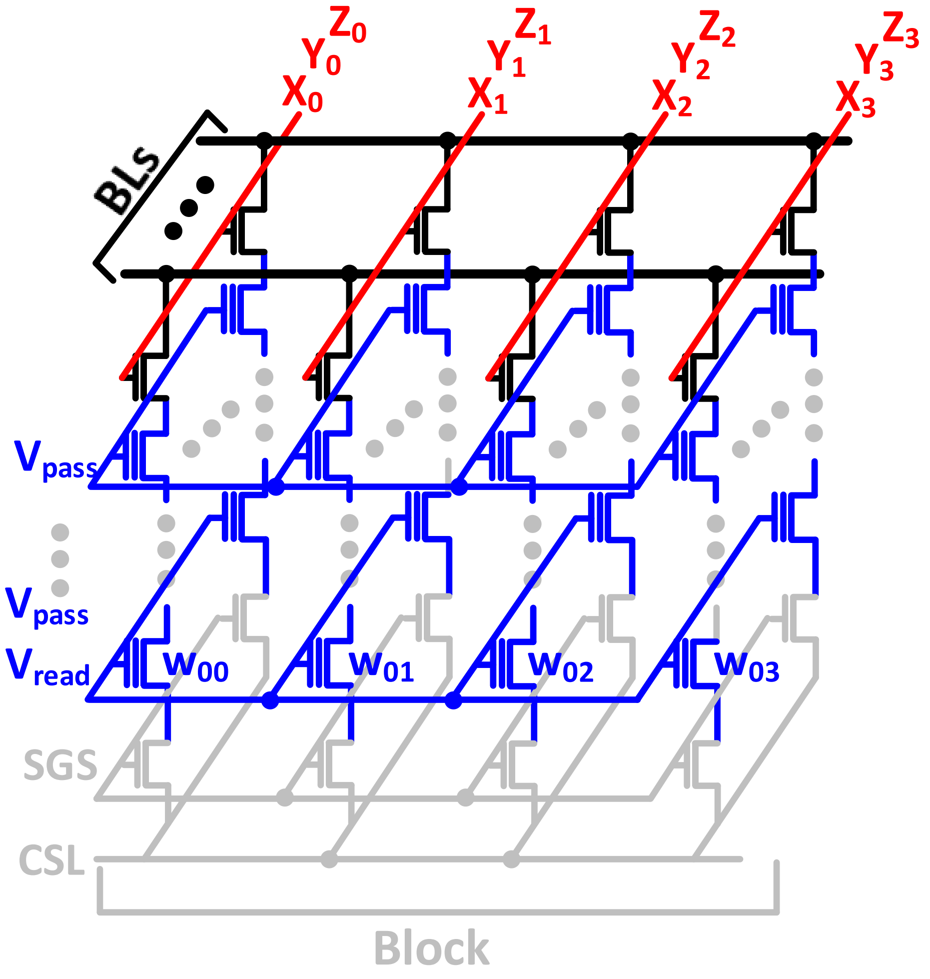
- $\Sigma\ X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - Σ (Accumulate) : Bitline currents of different blocks summed up
  - High resolution MAC can be realized using multiple weight cells, bit serial operation, and partial product post-processing

Analog MAC in 3D NAND Array



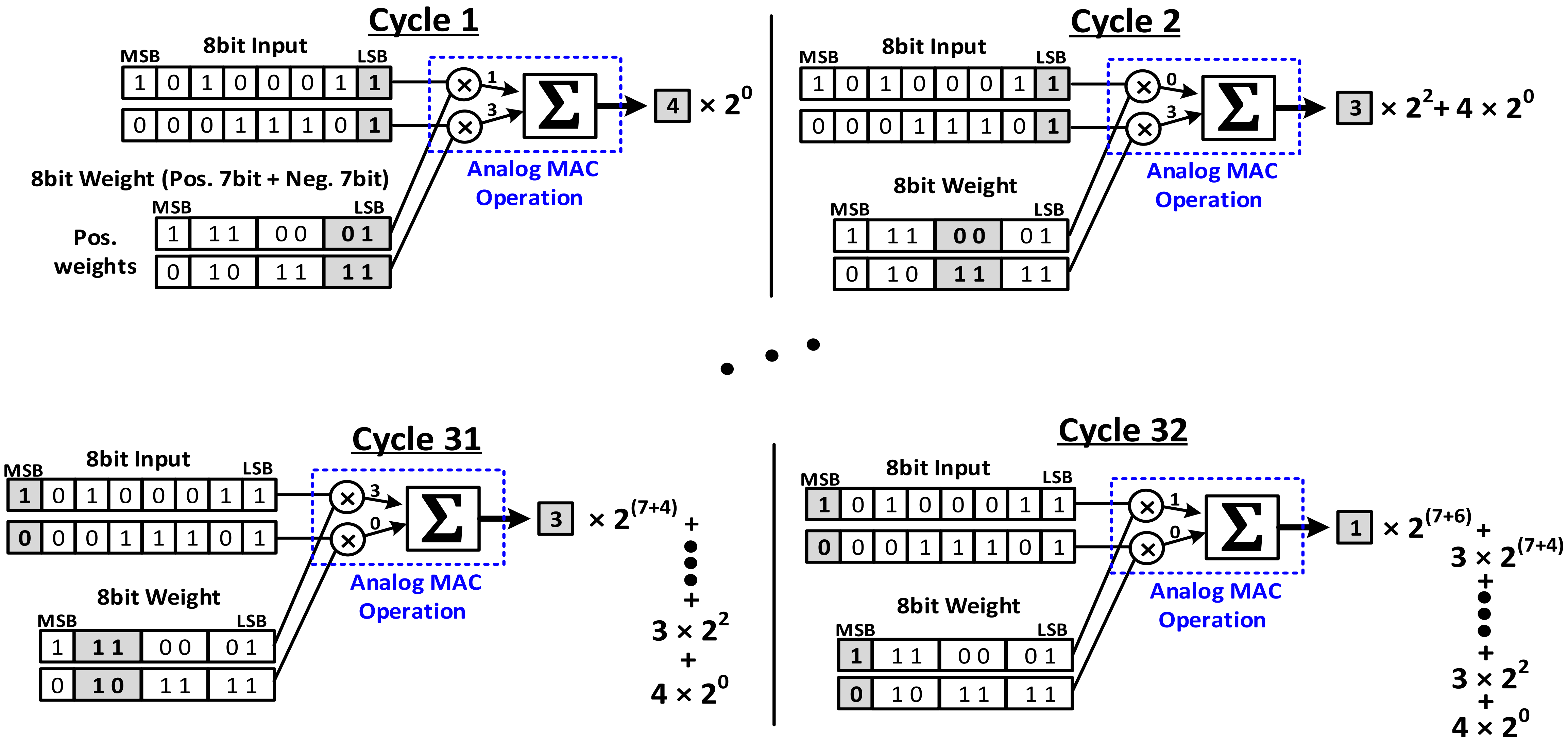
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - **Σ (Accumulate) : Bitline currents of different blocks summed up**
  - High resolution MAC can be realized using multiple weight cells, bit serial operation, and partial product post-processing

Analog MAC in 3D NAND Array



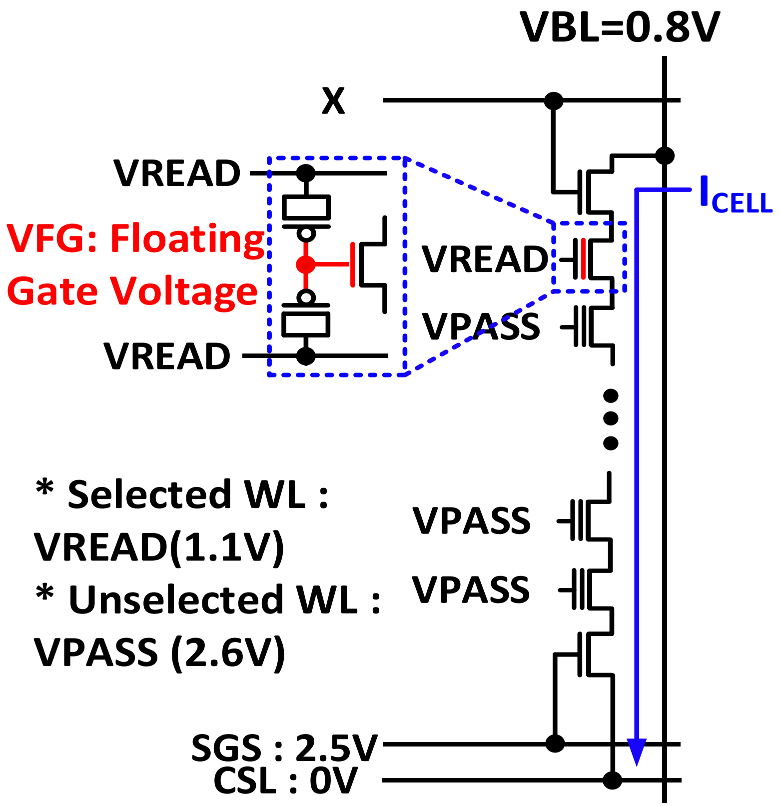
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - Σ (Accumulate) : Bitline currents of different blocks summed up
  - **High resolution MAC can be realized using multiple weight cells, bit serial operation, and partial product post-processing**

Bit Serial Operation for 8bit x 8bit MAC

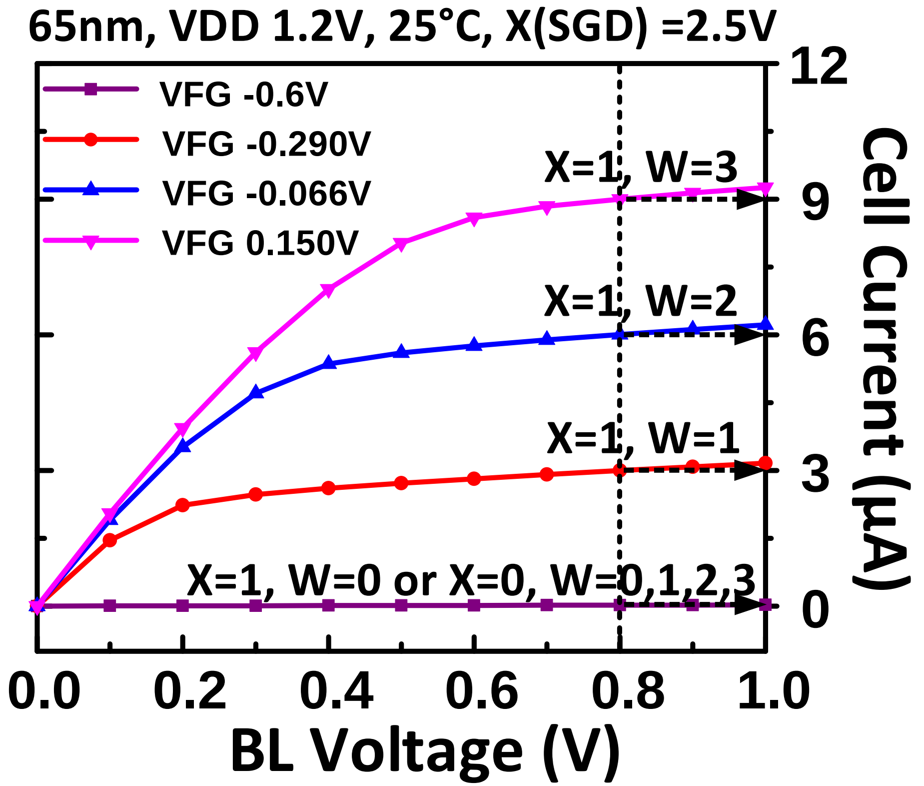


P. Judd, ICAL 2017

Logic Compatible 3T eFlash Based NAND String

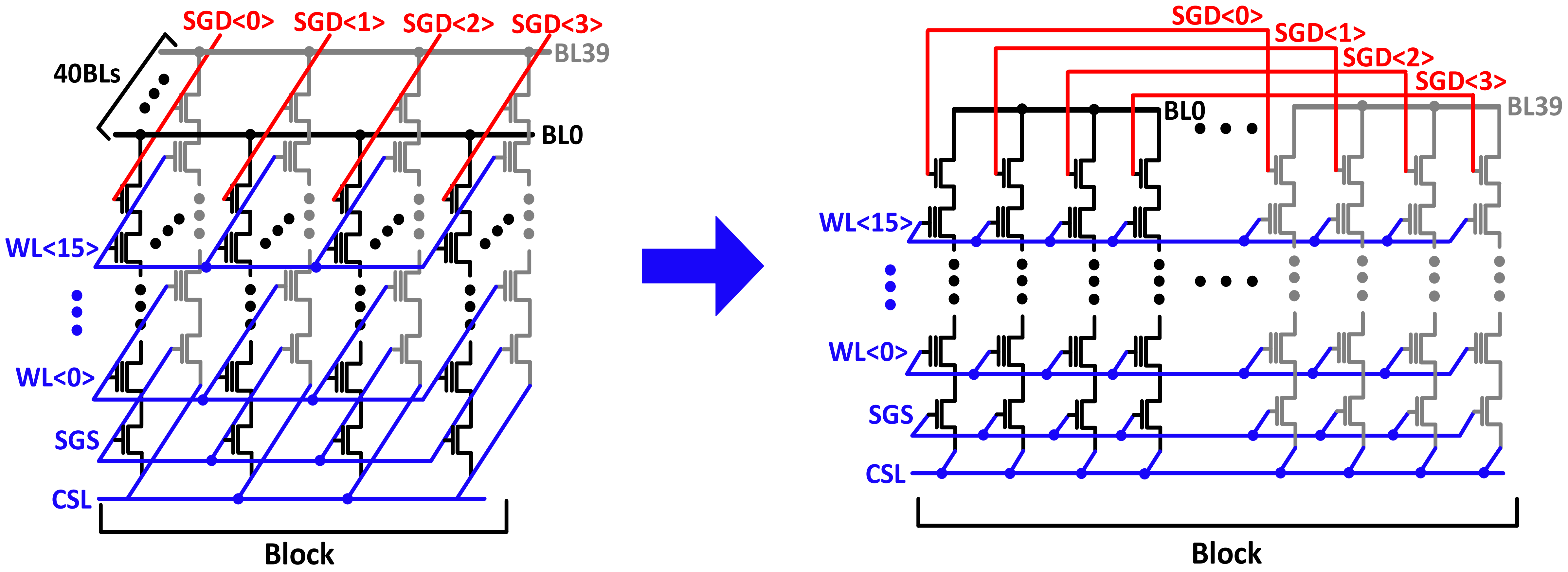


X	W	X·W	I_{CELL}
0	0	0	0μA
0	1	0	0μA
0	2	0	0μA
0	3	0	0μA
1	0	0	0μA
1	1	1	3μA
1	2	2	6μA
1	3	3	9μA



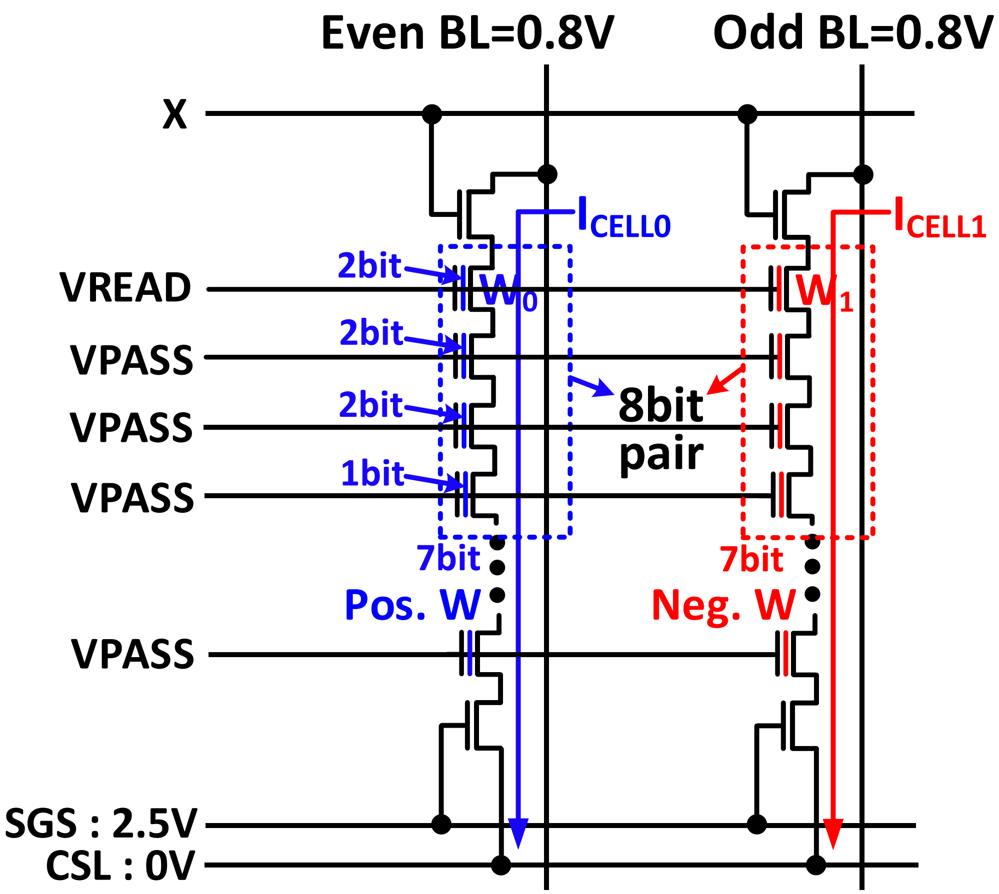
- Cell current proportional to X·W (=0μA, 3μA, 6μA, or 9μA)
- BL voltage pinned at 0.8V during read and verify operation

Prototype Design in a Standard Logic Process



- **Flatten to 2D while preserving 3D NAND array architecture**
- **Unit block size: 4 SGD x 16 WL x 40 BL**

Positive and Negative Weight Storage



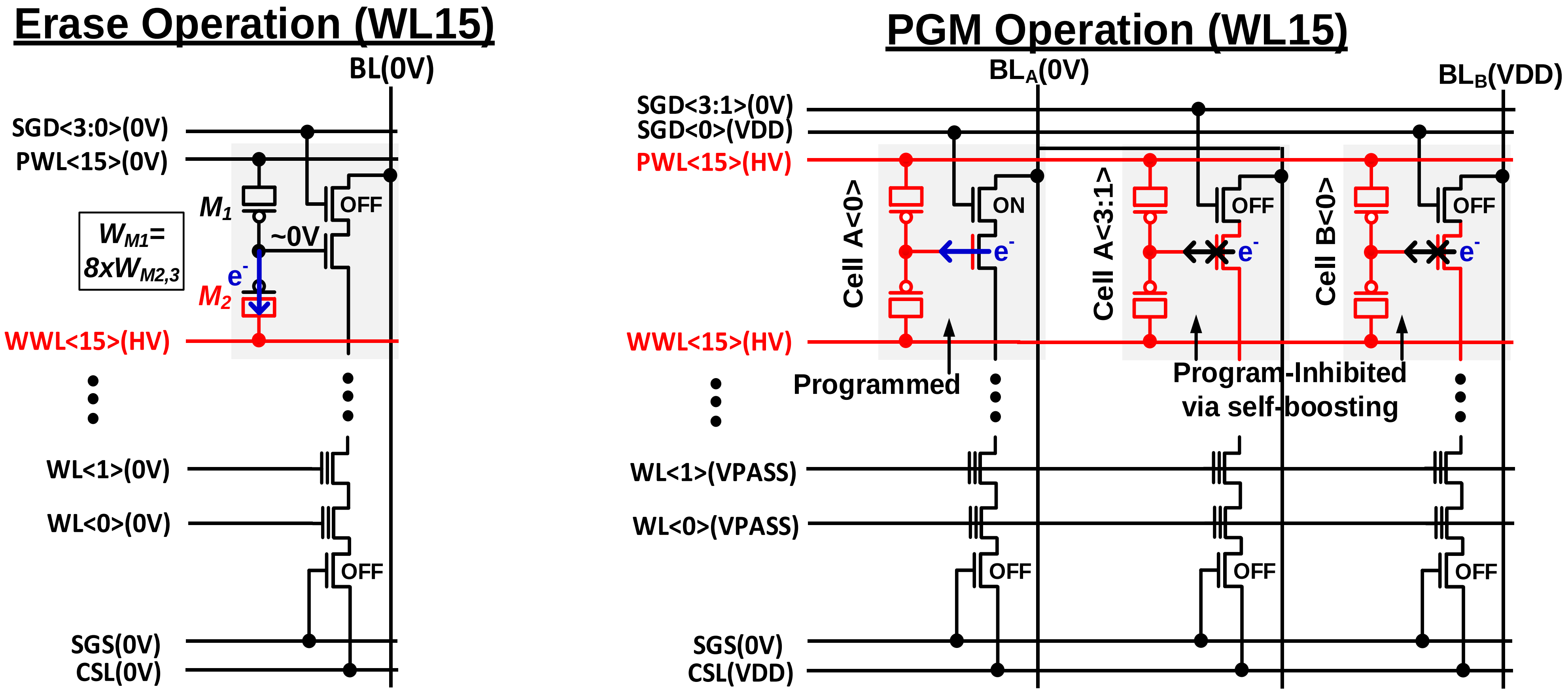
* Selected WL : VREAD(1.1V)
* Unselected WL : VPASS (2.6V)

X	W_0	W_1	$X \cdot W_0$	$X \cdot W_1$	I_{CELL0}	I_{CELL1}	ΔI
1	0	3	0	3	0μA	9μA	-9μA
1	0	2	0	2	0μA	6μA	-6μA
1	0	1	0	1	0μA	3μA	-3μA
1	0	0	0	0	0μA	0μA	0μA
1	1	0	1	0	3μA	0μA	3μA
1	2	0	2	0	6μA	0μA	6μA
1	3	0	3	0	9μA	0μA	9μA

Negative 2bit Weights

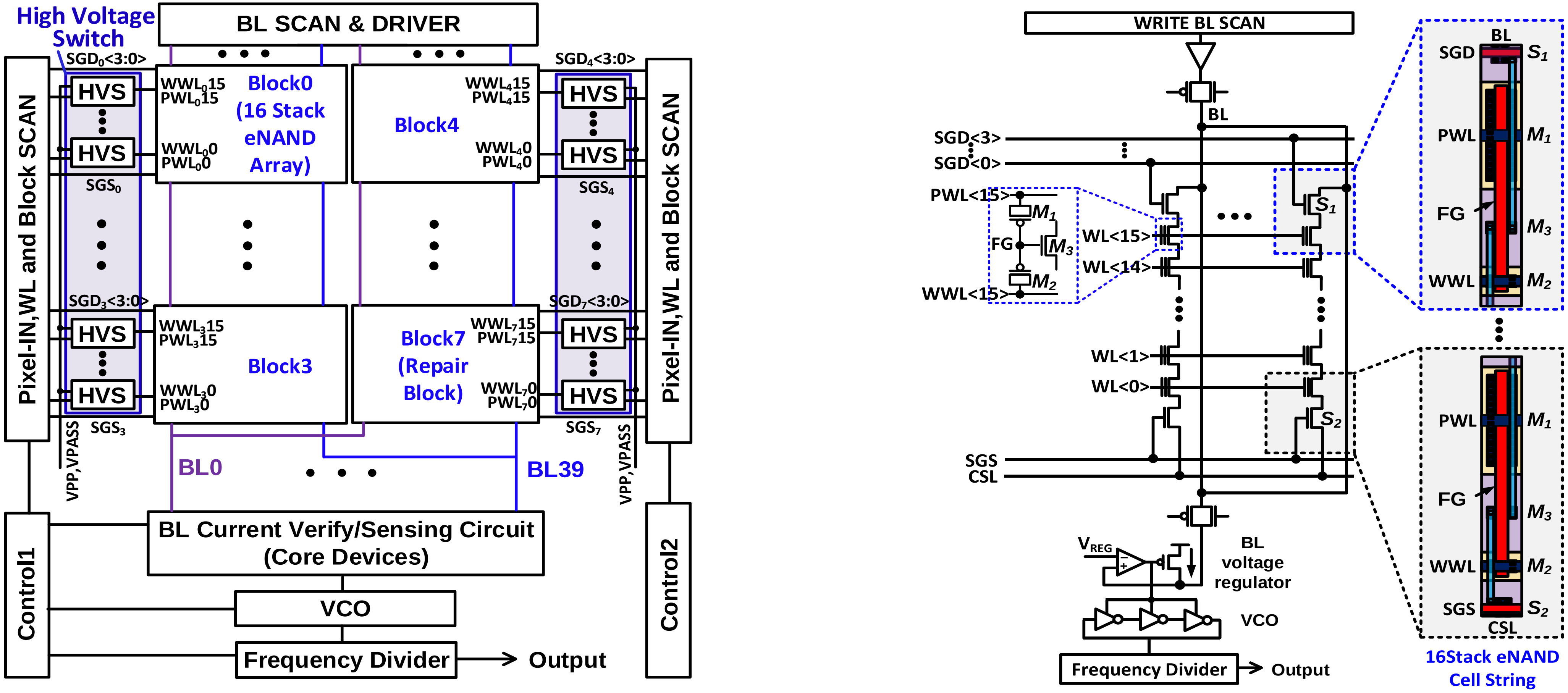
Positive 2bit Weights

Erase and Program Operations in NAND String



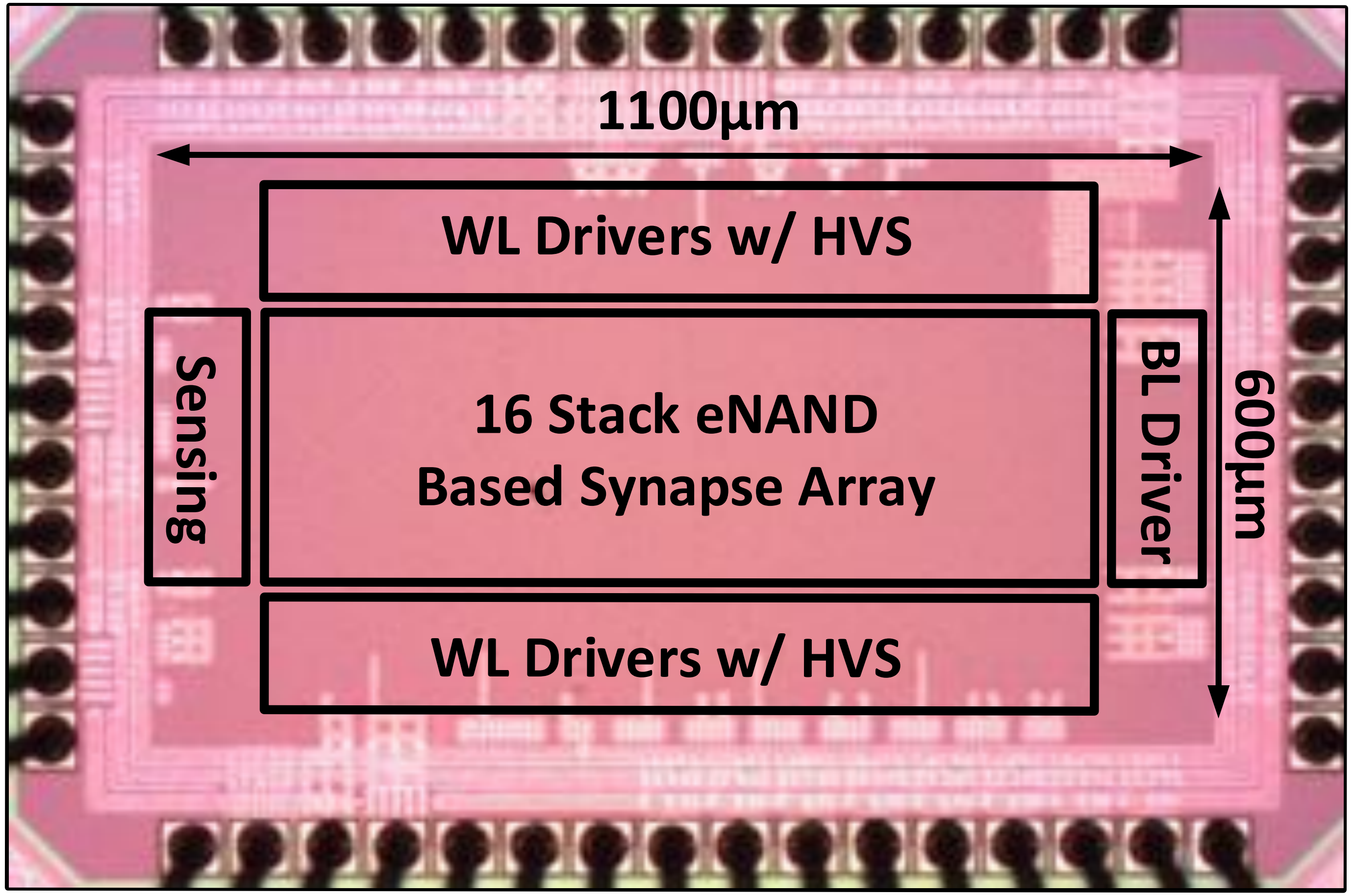
- FN tunneling utilized for erase and program
- Program inhibition of unselected cells via self-boosting

64 Row x 320 Column Core Architecture



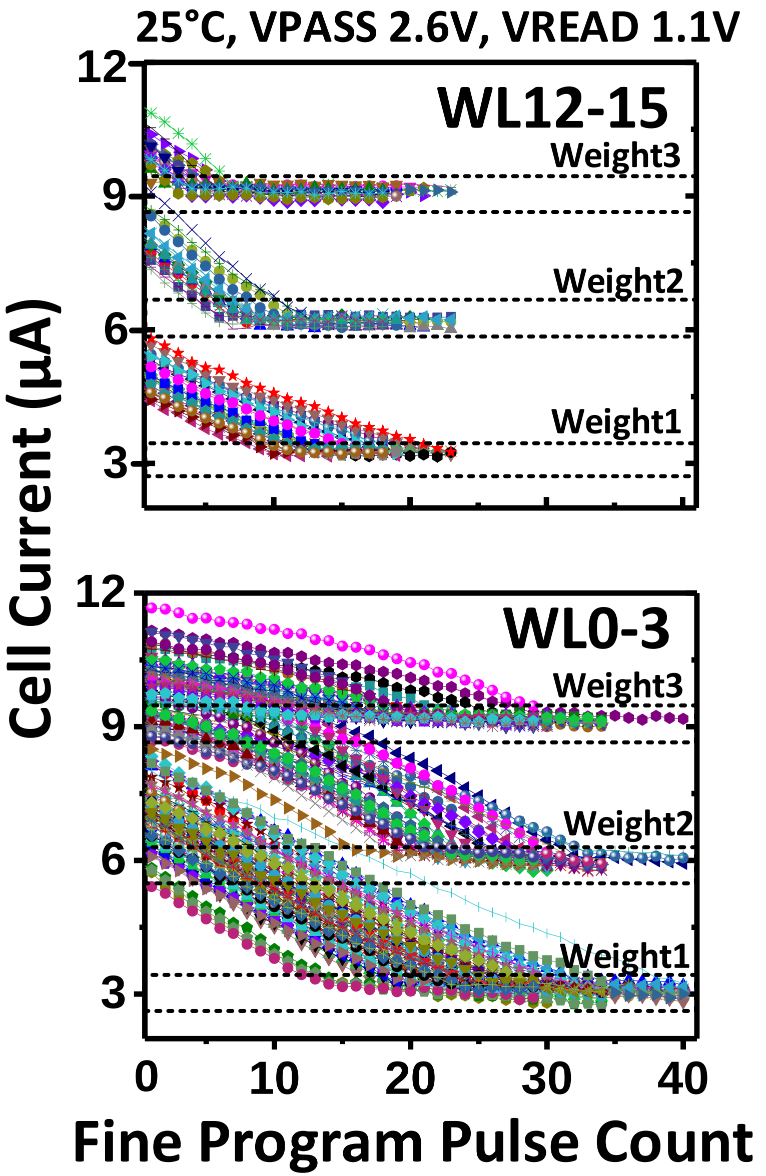
- 16 stack eNAND array, high voltage switches, BL sensing circuits
- Input data loaded on to 25 SGD lines, 3 SGD lines for bias

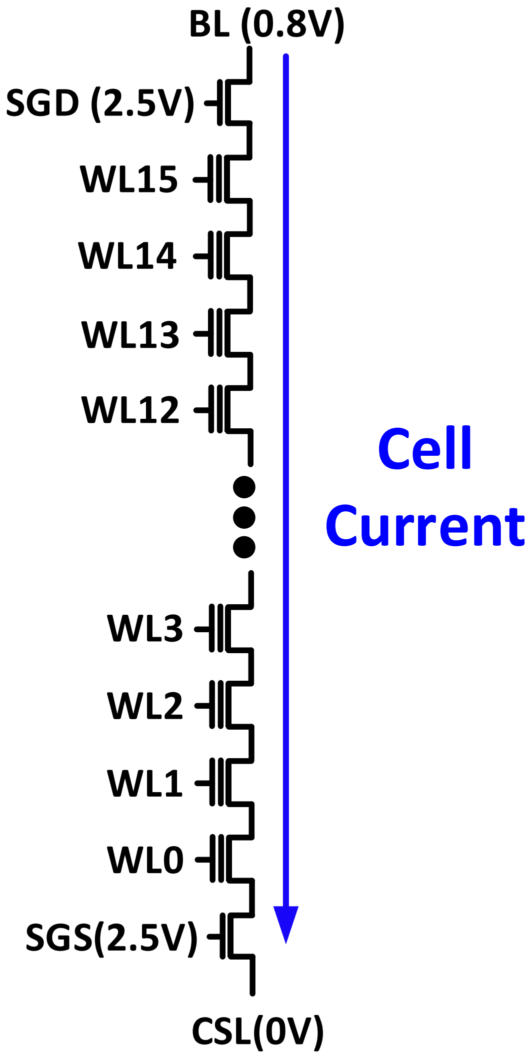
65nm Die Photo and Feature Summary



Technology	65nm Logic
Core Size	1100 X 600 μm^2
VDD (Core / IO)	1.2V / 2.5V
# of 8bit Weights	2,560
# of Synapses	20K (=64x320)
Throughput w/o VCO	0.5G pixels/s per core (tREAD : 50ns)
Power	4.95 µW (per bitline)

Program Characteristics vs. NAND String Location

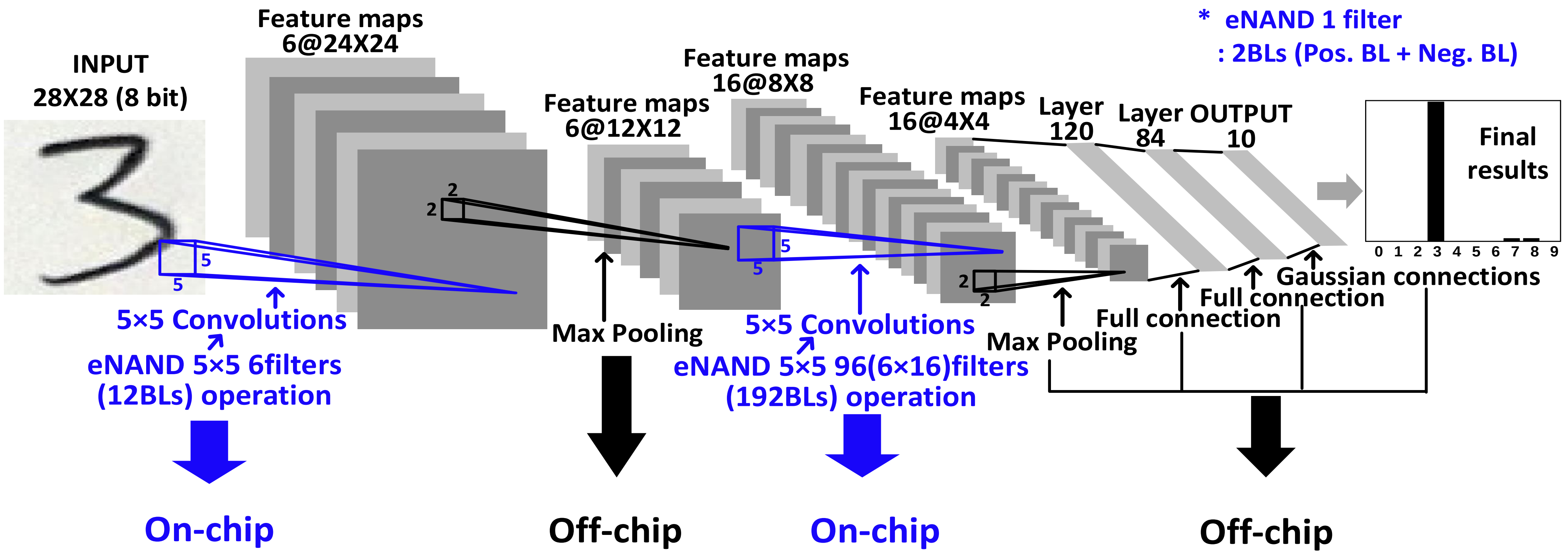




*Selected WL : VREAD,
Unselected WL : VPASS

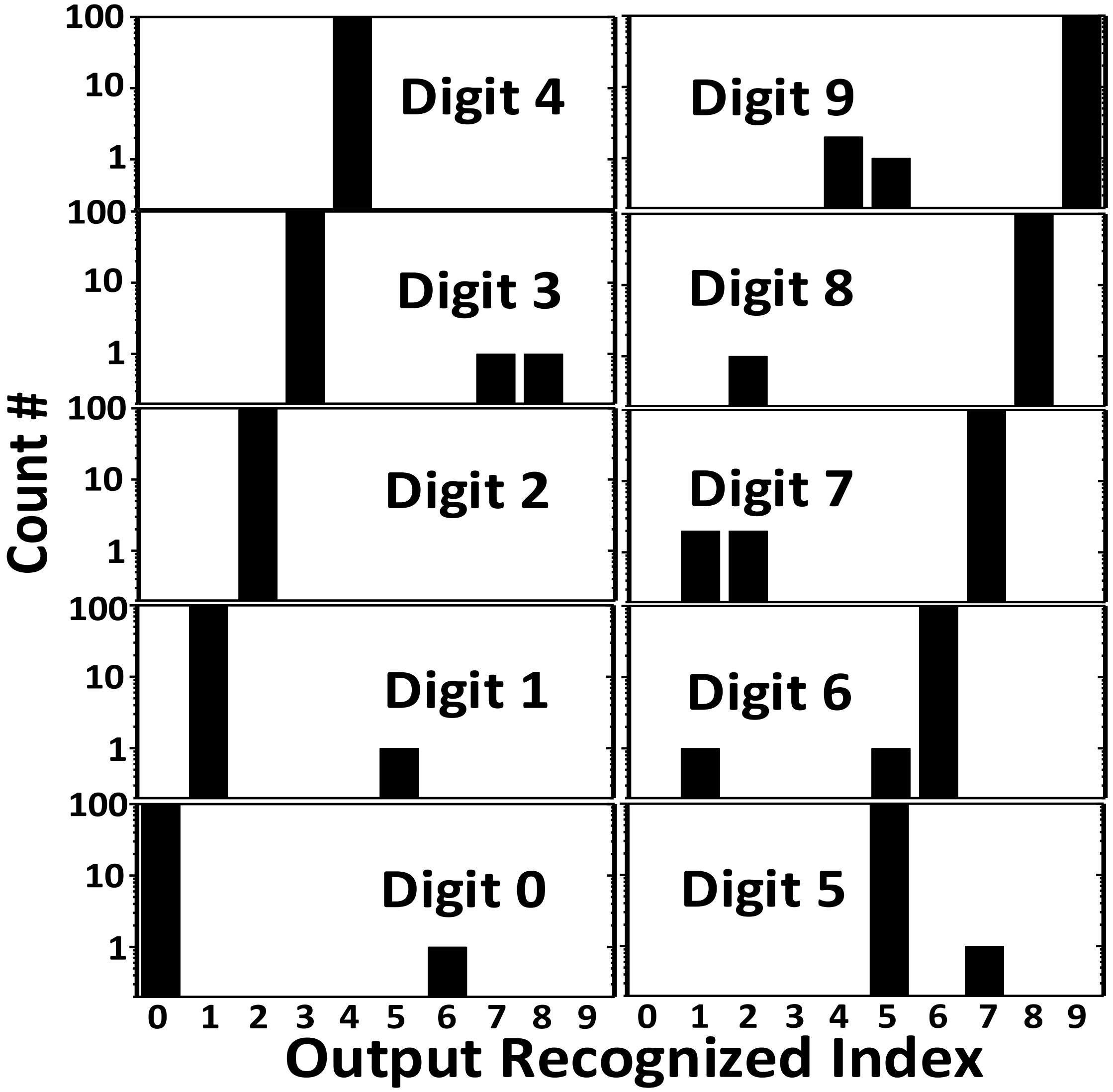
- Different Vgs and Vds depending on the stack location
- Requires different Vth for the same cell current
- Longer programming time for cells closer to the bottom
- Cell current variation less than 0.6μA after program-verify operations

LeNet-5 CNN Demonstration



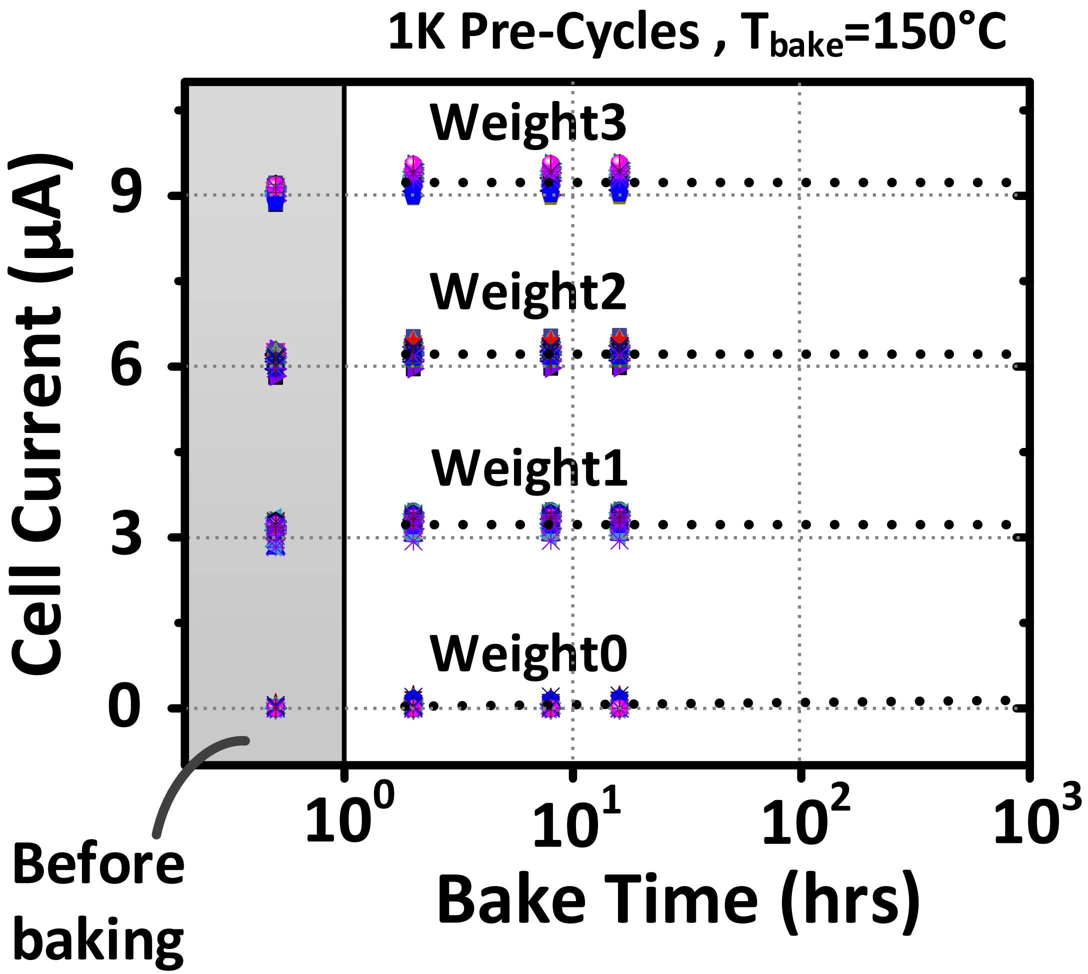
- 5 x 5 8bit convolution performed on chip
- ADC, pooling, bit serial operation performed off chip

MNIST Digit Recognition Accuracy Results



- Recognition accuracy is 98.5% which is close to the software model's 99.0% accuracy

Retention Test Results



- **1K erase and program cycles before baking test**
- **Excellent retention characteristics → additional weight levels possible (e.g. 3 bit per cell)**

Comparison Table

	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

[1] X. Guo, et al., IEDM, 2017.
[2] W. Chen, et al., ISSCC, 2018.
[3] M. Kim, et al., IEDM, 2018.
[4] C. Xue, et al., ISSCC, 2019.
[5] X. Si, et al., ISSCC, 2019.

Conclusions

- **3D NAND Flash ready 8bit x 8bit convolutional neural network core demonstrated in a 65nm standard logic process**
  - **16 stack NAND string**
  - **2 bit per cell weight storage**
  - **Bit serial operation**
  - **Back-pattern tolerant program verify (details in paper)**
- **LeNet5 2-layer CNN test results show 98.5% digit recognition accuracy**