3D NAND Flash Ready Neural Networks: Learnings and Afterthoughts

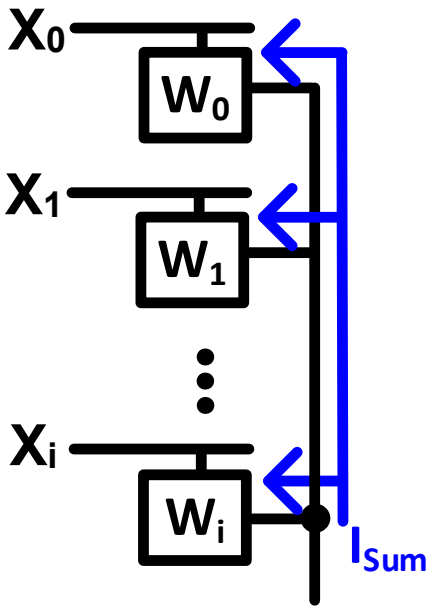
Chris Kim

University of Minnesota

chriskim@umn.edu, chriskim.umn.edu

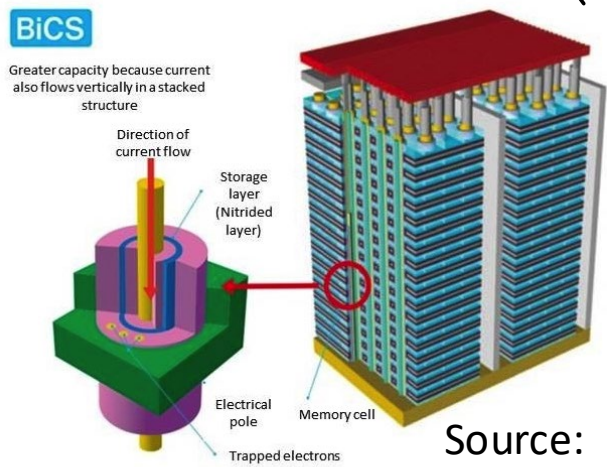
In-Flash-Memory Computing

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



Analog multiply and accumulate (MAC)

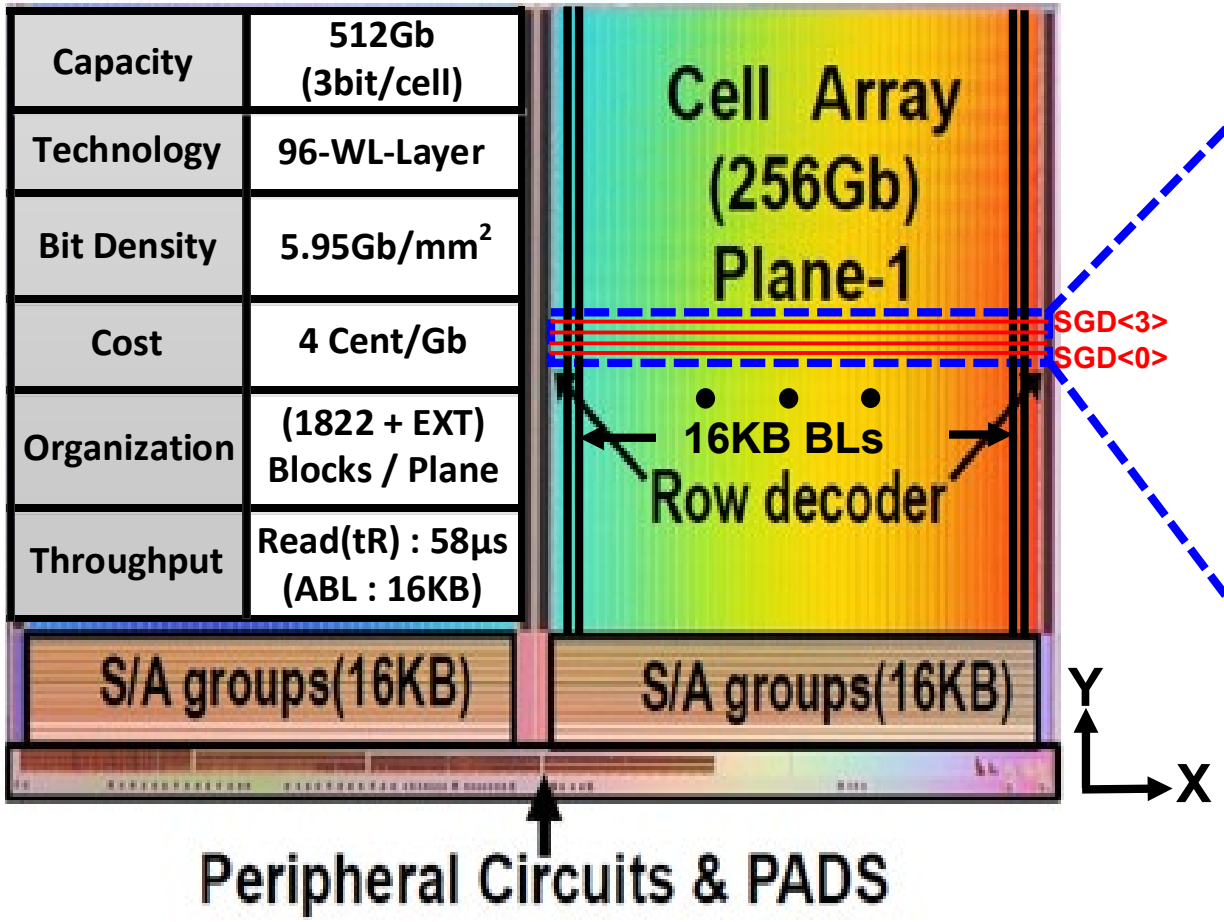
- **Ultra-high density, reduced data traffic, low cost, mature technology**
- **Low program/erase speed (but fast read speed), limited endurance cycles (but fine for neural network applications)**



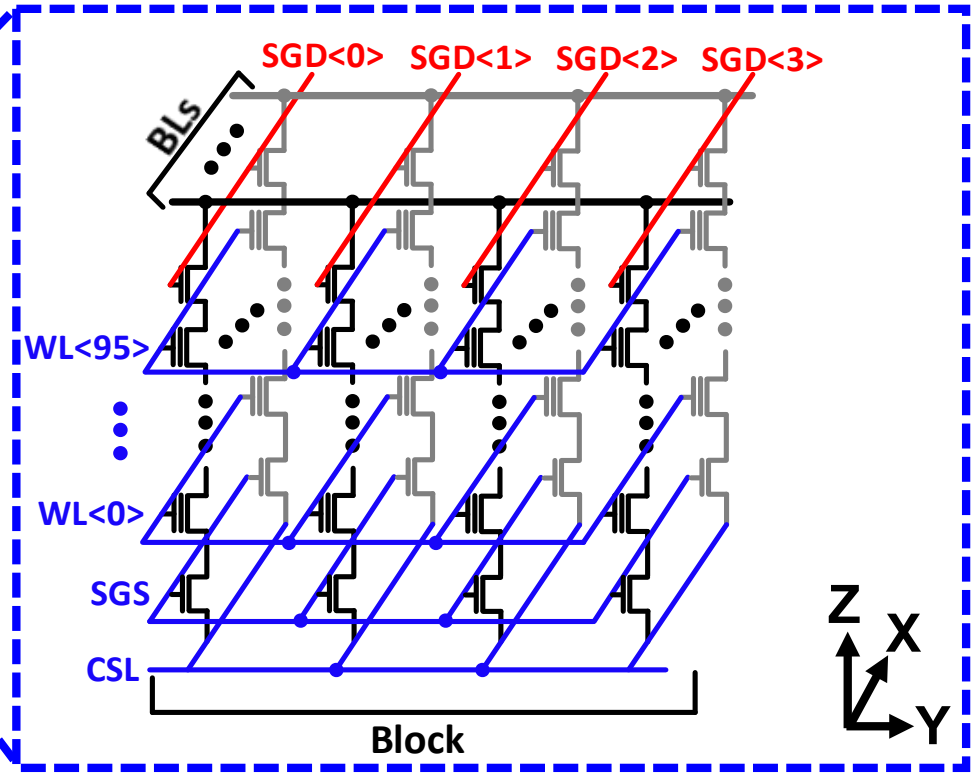
Source: Toshiba

3D NAND Cell and Array

3D NAND Flash Memory

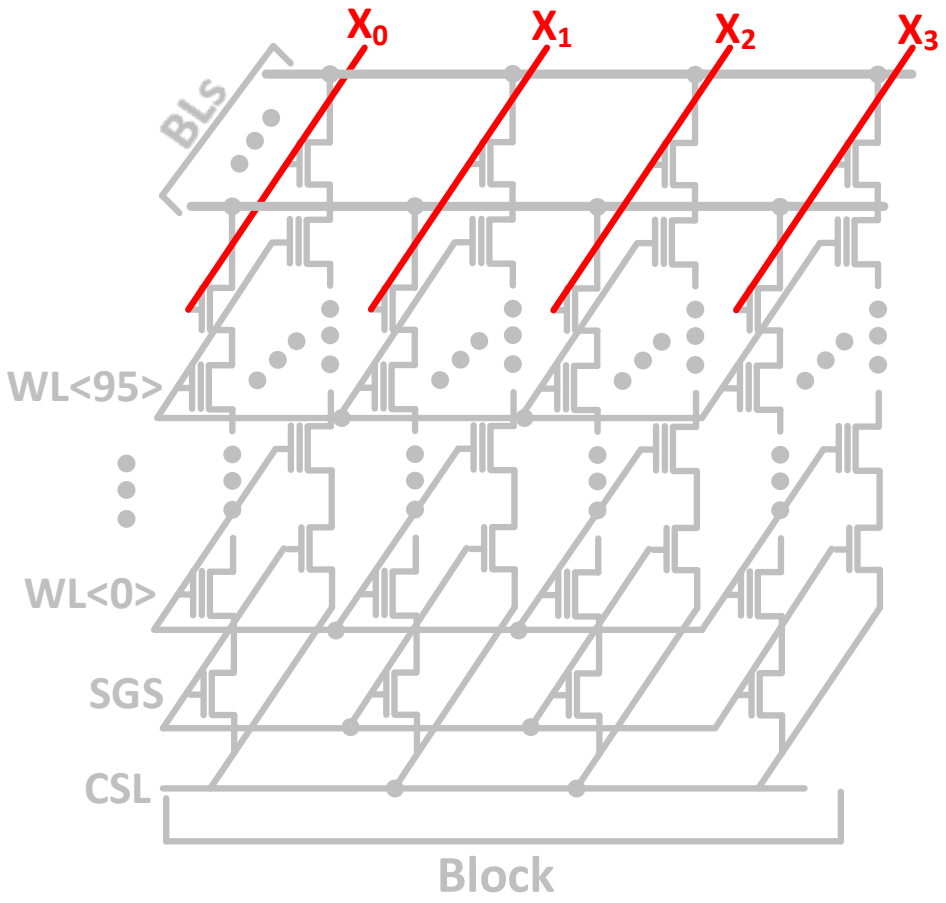


H. Maejima, Toshiba, ISSCC 2018



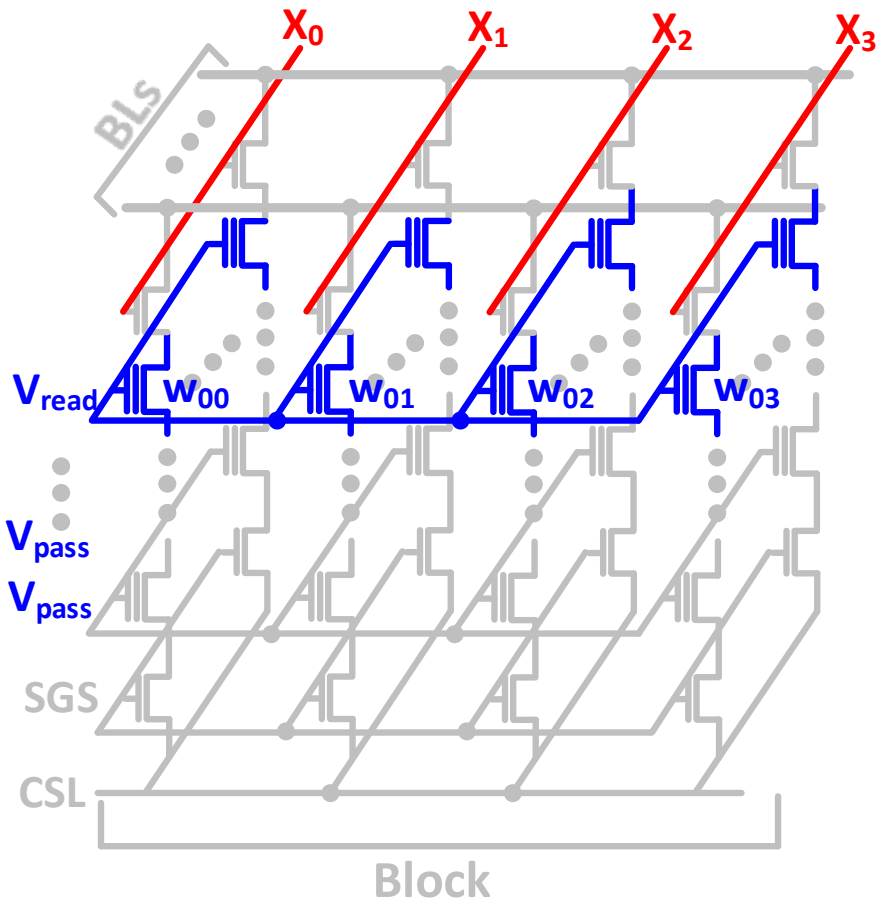
- (x,y,z) = (BL, SGD, WL)
  - BL shared across multiple blocks
  - WL is a shared plane (not line)
  - SGD can be individually controlled

Analog MAC in 3D NAND Array



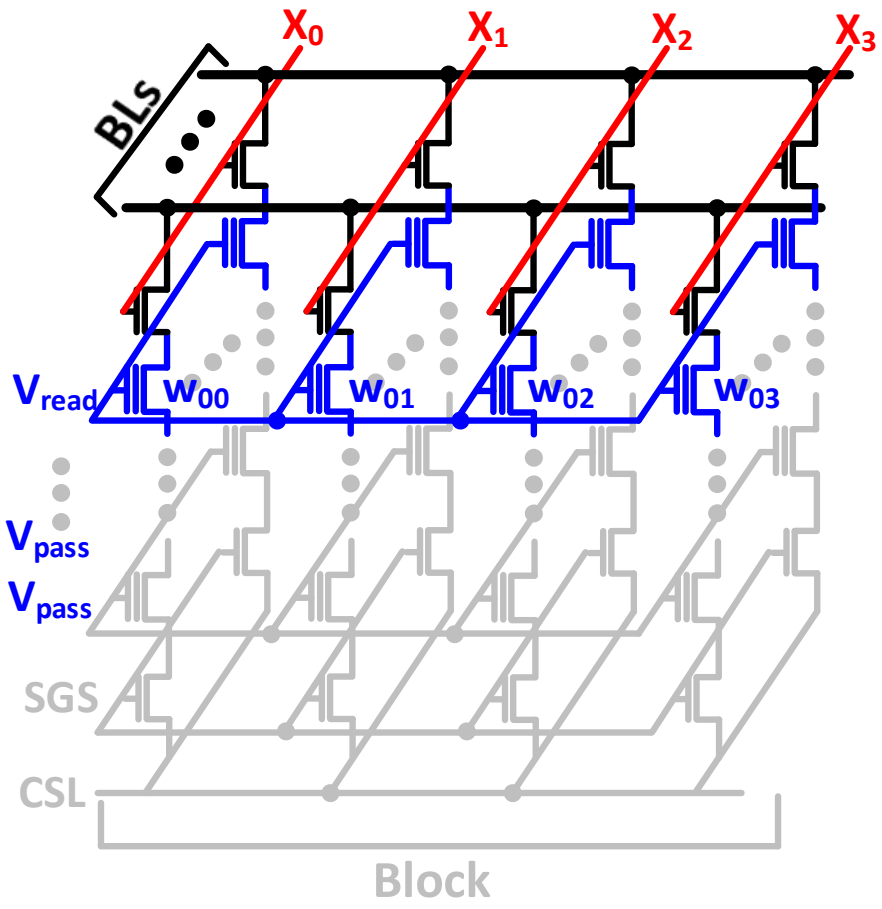
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - Σ (Accumulate) : Bitline currents of different blocks summed up
  - High resolution MAC can be realized using multiple weight cells, bit serial operation,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

Analog MAC in 3D NAND Array



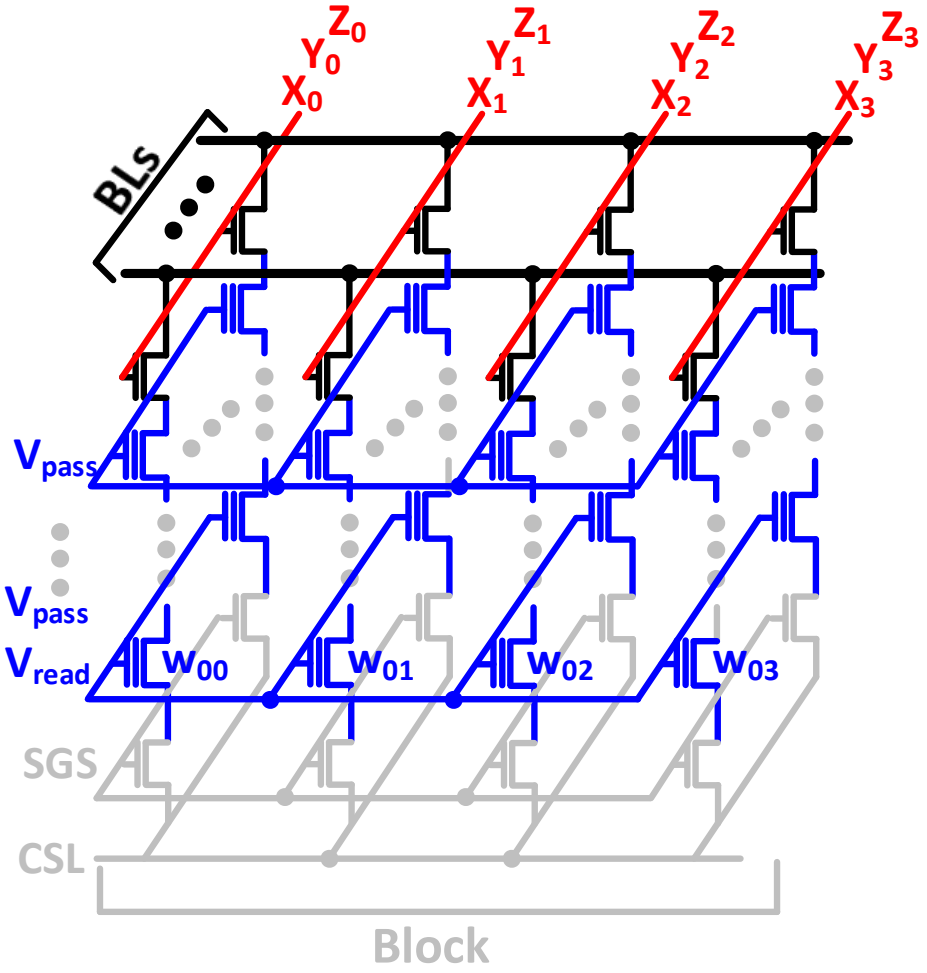
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - **W_i: Multi-level weight (MLC, TLC, QLC)**
  - Σ (Accumulate) : Bitline currents of different blocks summed up
  - High resolution MAC can be realized using multiple weight cells, bit serial operation,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

Analog MAC in 3D NAND Array



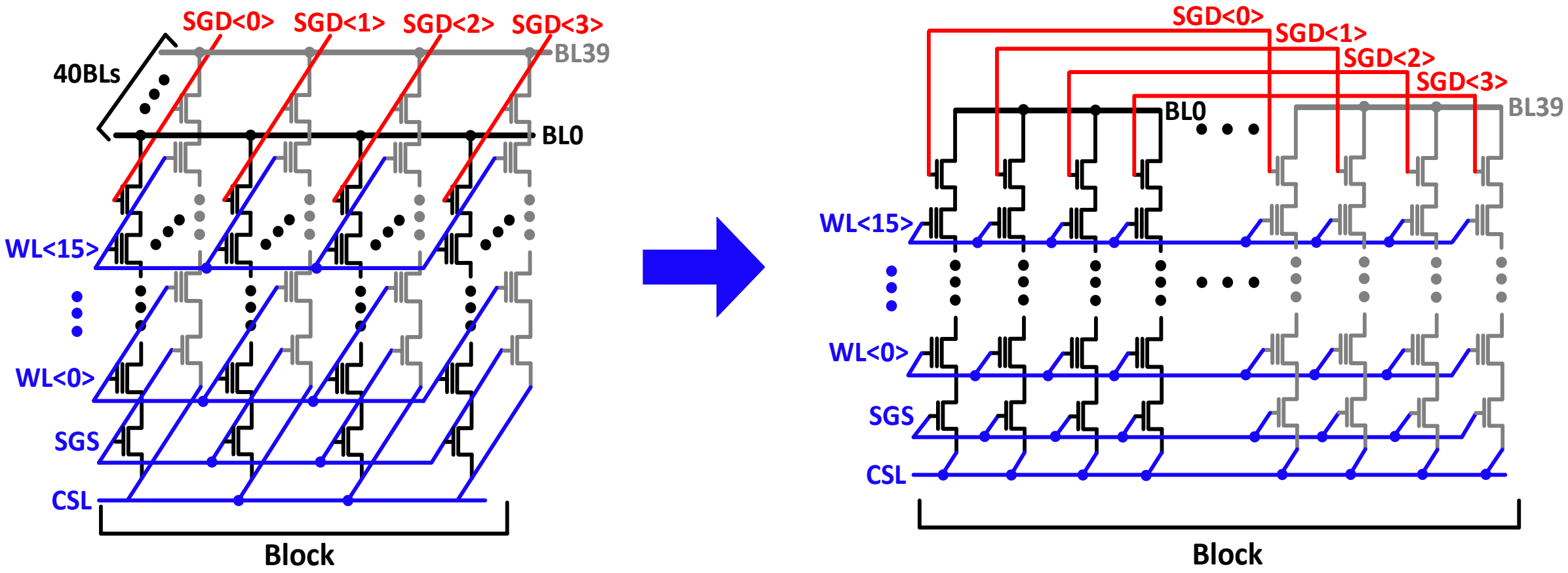
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - **Σ (Accumulate) : Bitline currents of different blocks summed up**
  - High resolution MAC can be realized using multiple weight cells, bit serial operation,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

Analog MAC in 3D NAND Array



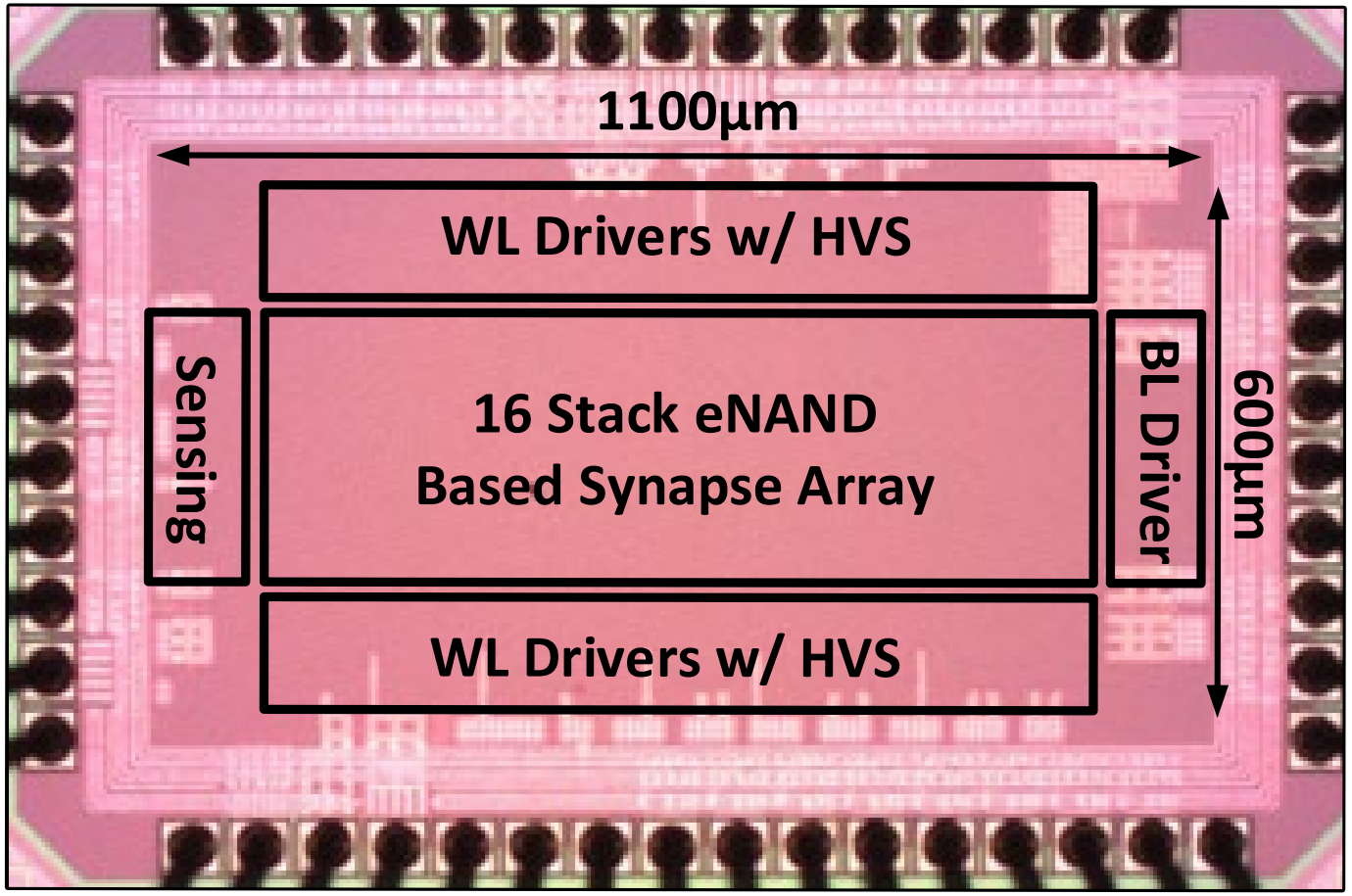
- $\Sigma X_i \times W_i$
  - X_i : Binary input applied to individual SGD lines (no analog voltages)
  - W_i: Multi-level weight (MLC, TLC, QLC)
  - Σ (Accumulate) : Bitline currents of different blocks summed up
  - **High resolution MAC can be realized using multiple weight cells, bit serial operation,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

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

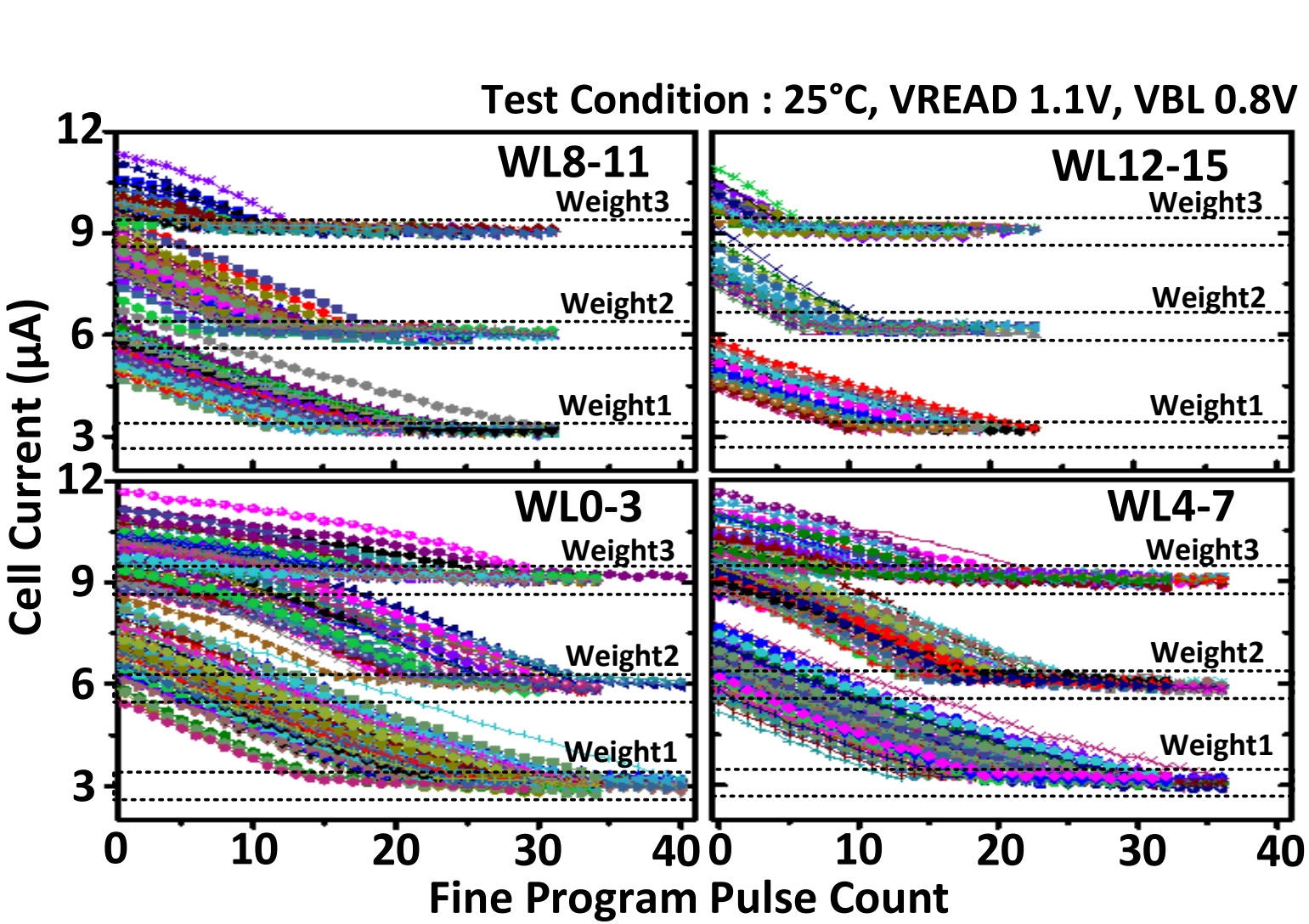
65nm Die Photo and Feature Summary

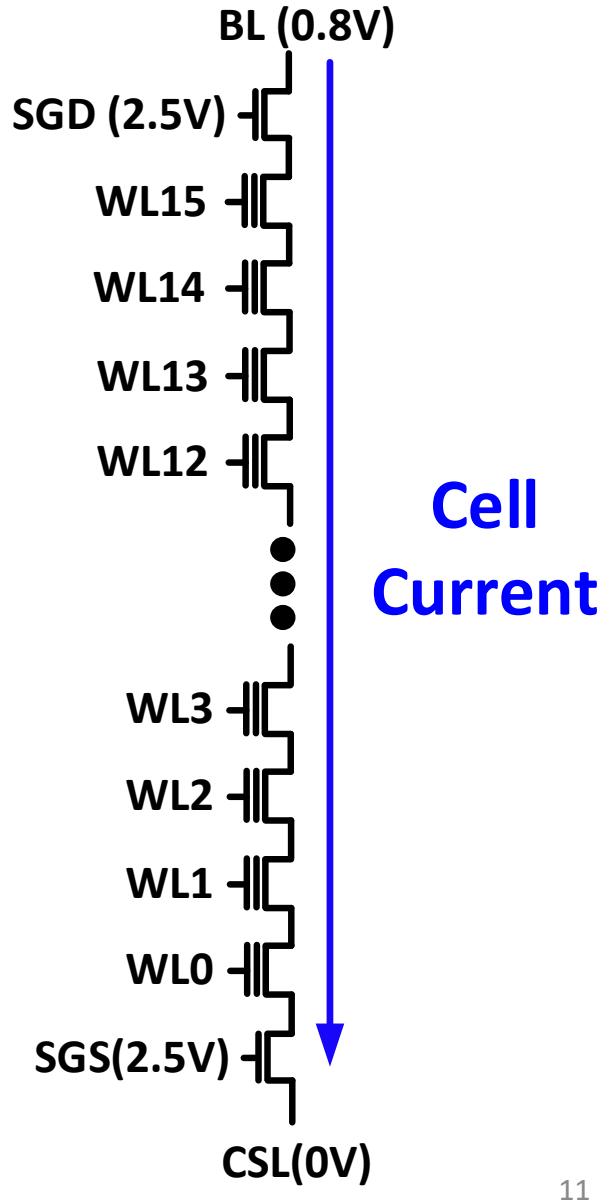


Technology	65nm Logic
Core Size	1100 X 600 µm^2
VDD (Core / IO)	1.2V / 2.5V
# of 8bit Weights	2,560
# of Synapses	20K (=64x320)
Throughput w/o VCO	0.5G pixels/s per core (tREAD : 50ns)
Power	4.95 µW (per bitline)

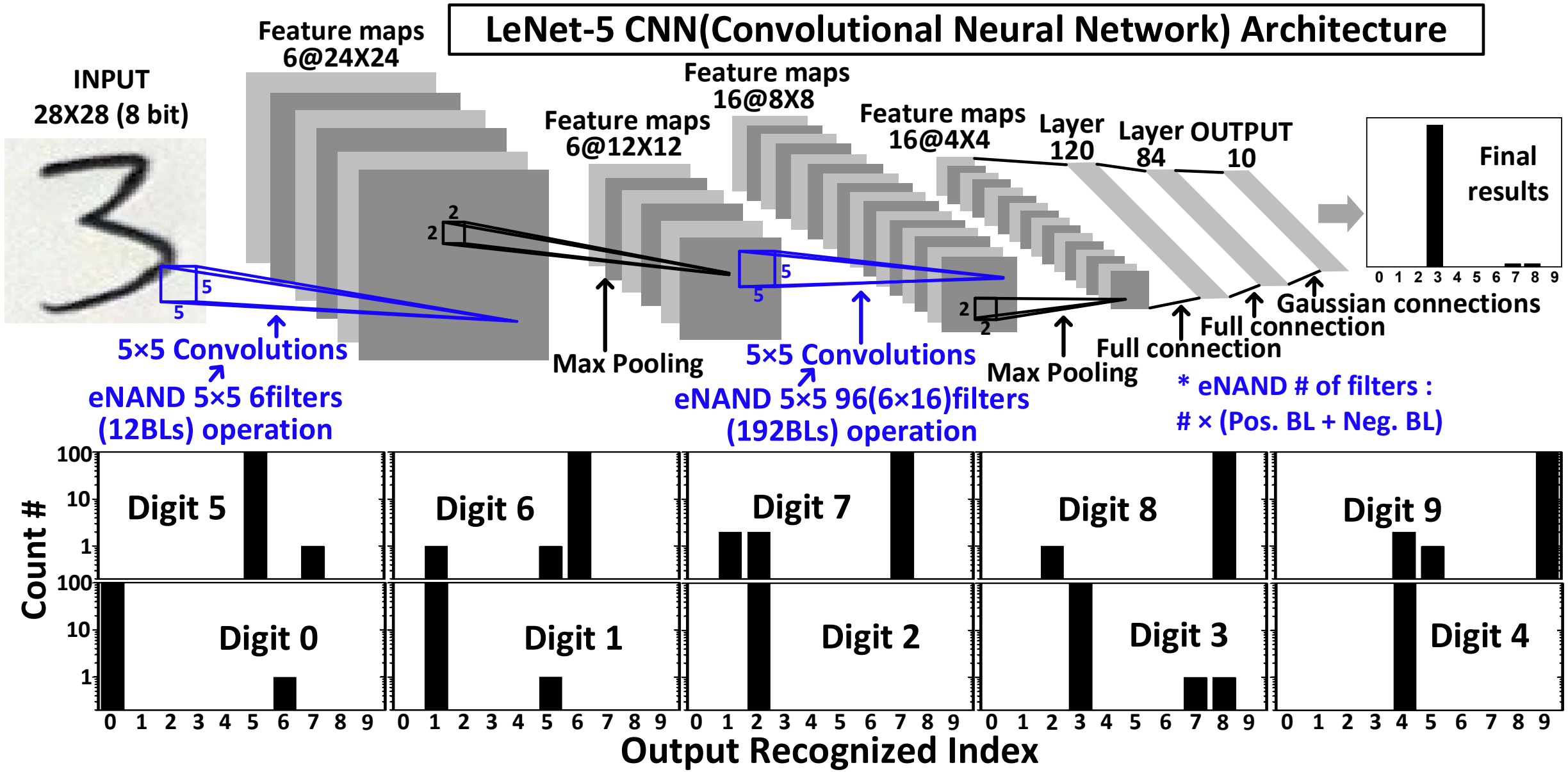
M. Kim, et al., IEDM 2018

NAND String Programming Results



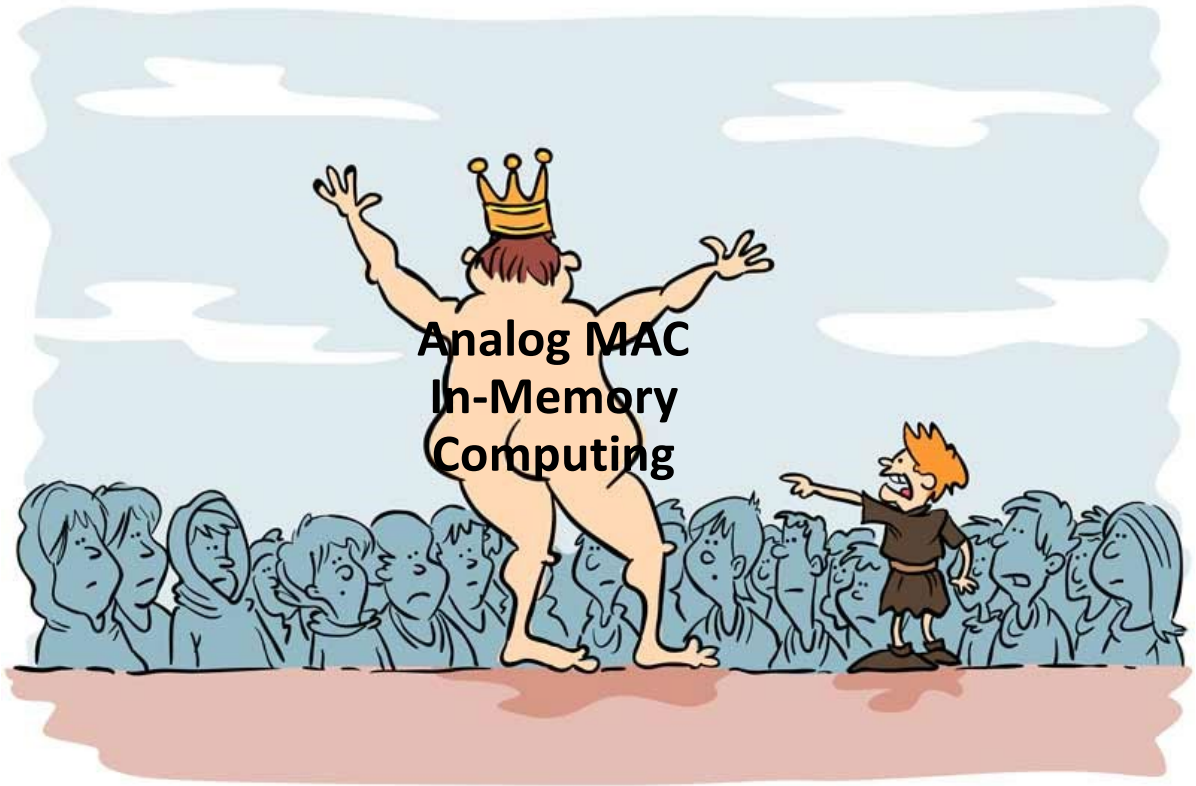


LeNet-5 CNN(Convolutional Neural Network) Architecture



- 98.5% recognition accuracy (test chip data)

Two Sides of the *In-Memory Computing* Coin



https://www.bitglass.com/blog/naked-emperors

From a research appeal, publication, and attention-grabbing standpoint (~2015)

From a GPU/FPGA/ASIC replacement standpoint