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

M. Kim¹, M. Liu¹, L. Everson¹, G. Park¹, Y. Jeon²,
S. Kim², S. Lee², S. Song² and C. H. Kim¹

¹Dept. of ECE, University of Minnesota
²Anaflash 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

- **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**

Analog multiply and accumulate (MAC)
State-of-the-Art 3D NAND Flash Memory

- **Capacity**: 512Gb (3bit/cell)
- **Technology**: 96-WL-Layer
- **Bit Density**: 5.95Gb/mm²
- **Cost**: 4 Cent/Gb
- **Organization**: (1822 + EXT) Blocks / Plane
- **Throughput Read(tR)**: 58µs (ABL : 16KB)

- **Peripheral Circuits & PADS**

- **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

\[ \sum X_i \times W_i \]

- \( X_i \): Binary input applied to individual SGD lines (no analog voltages)
- \( W_i \): Multi-level weight (MLC, TLC, QLC)
- \( \sum \) (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)
  - $\Sigma$ (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)
- \( \Sigma \) (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

\[ \sum X_i \times W_i \]

- \( X_i \): Binary input applied to individual SGD lines (no analog voltages)
- \( W_i \): Multi-level weight (MLC, TLC, QLC)
- \( \sum \) (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

**Cycle 1**

8bit Input:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

8bit Weight (Pos. 7bit + Neg. 7bit):

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Analog MAC Operation:

\[ \sum (4 \times 2^0) \]

**Cycle 2**

8bit Input:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

8bit Weight:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Analog MAC Operation:

\[ \sum (3 \times 2^2 + 4 \times 2^0) \]

**Cycle 31**

8bit Input:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

8bit Weight:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Analog MAC Operation:

\[ \sum (3 \times 2^{(7+4)} + 3 \times 2^2 + 4 \times 2^0) \]

**Cycle 32**

8bit Input:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

8bit Weight:

<table>
<thead>
<tr>
<th>MSB</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSB</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Analog MAC Operation:

\[ \sum (1 \times 2^{(7+6)} + 3 \times 2^{(7+4)} + 3 \times 2^2 + 4 \times 2^0) \]

P. Judd, ICAL 2017
Logic Compatible 3T eFlash Based NAND String

• Cell current proportional to \(X \cdot W\) (=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

<table>
<thead>
<tr>
<th>X</th>
<th>W₀</th>
<th>W₁</th>
<th>X·W₀</th>
<th>X·W₁</th>
<th>I_{CELL0}</th>
<th>I_{CELL1}</th>
<th>ΔI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0µA</td>
<td>9µA</td>
<td>-9µA</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0µA</td>
<td>6µA</td>
<td>-6µA</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0µA</td>
<td>3µA</td>
<td>-3µA</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0µA</td>
<td>0µA</td>
<td>0µA</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3µA</td>
<td>0µA</td>
<td>3µA</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6µA</td>
<td>0µA</td>
<td>6µA</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>9µA</td>
<td>0µA</td>
<td>9µA</td>
</tr>
</tbody>
</table>

Even BL=0.8V  Odd BL=0.8V

* Selected WL : VREAD (1.1V)
* Unselected WL : VPASS (2.6V)
Erase and Program Operations in NAND String

**Erase Operation (WL15)**

- SGD<3:0>(0V)
- PWL<15>(0V)
- WWL<15>(HV)
- WL<1>(0V)
- WL<0>(0V)
- SGS(0V)
- CSL(0V)

**PGM Operation (WL15)**

- SGD<3:1>(0V)
- SGD<0>(VDD)
- PWL<15>(HV)

**Programmed**

- Cell A<0>

**Programmed - Inhibited via self-boosting**

- Cell A<3:1>

- Cell B<0>

- WWL<15>(HV)

- WL<1>(VPASS)
- WL<0>(VPASS)
- SGS(0V)
- CSL(VDD)

**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 μm²
- **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

• 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>65nm</td>
<td>55nm</td>
<td>55nm</td>
<td>65nm</td>
<td>65nm</td>
<td>180nm</td>
</tr>
<tr>
<td><strong>Voltage</strong></td>
<td>1.2V</td>
<td>1.0V</td>
<td>1.0V</td>
<td>1.0V</td>
<td>1.0V</td>
<td>2.7V</td>
</tr>
<tr>
<td><strong>Cell Type</strong></td>
<td>NAND</td>
<td>NOR</td>
<td>NOR</td>
<td>NOR</td>
<td>NOR</td>
<td>NOR</td>
</tr>
<tr>
<td><strong>Non volatile?</strong></td>
<td>YES (eFlash)</td>
<td>YES (ReRAM)</td>
<td>NO (SRAM)</td>
<td>YES (eFlash)</td>
<td>YES (ReRAM)</td>
<td>YES (eFlash)</td>
</tr>
<tr>
<td><strong>Logic Compatible?</strong></td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td><strong>Program-verify?</strong></td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td><strong>Weight Resolution</strong></td>
<td>8 Bits</td>
<td>3 Bits</td>
<td>5 Bits</td>
<td>2.3 Bits</td>
<td>3 Bits</td>
<td>2 Bits</td>
</tr>
<tr>
<td><strong>Input Resolution</strong></td>
<td>8 Bits</td>
<td>2 Bits</td>
<td>2 Bits</td>
<td>1 Bit</td>
<td>3 Bits</td>
<td>1 Bit</td>
</tr>
<tr>
<td><strong># of Currents Summed Up</strong></td>
<td>28 Cells</td>
<td>8 Cells</td>
<td>32 Cells</td>
<td>68 Cells</td>
<td>14 Cells</td>
<td>4 Cells</td>
</tr>
<tr>
<td><strong>Neural Net Architecture</strong></td>
<td>CNN</td>
<td>CNN</td>
<td>CNN</td>
<td>MLP</td>
<td>CNN</td>
<td>MLP</td>
</tr>
</tbody>
</table>

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