A 65nm Compute-In-Memory 7T SRAM Macro Supporting 4-bit Multiply and Accumulate Operation by Employing Charge Sharing


* Department of Electronics and Communication, Indian Institute of Technology, Roorkee, India
† Intel Technology India Pvt. Ltd., Bangalore, Karnataka, India
‡ NXP Semiconductors, Bangalore, Karnataka, India
§ Brookhaven National Laboratory, Upton, NY 11973 USA.
¶ Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA.

2022 IEEE International Symposium on Circuits and Systems
May 28- June 1, 2022 Hybrid Conference
Outline

- Motivation
- Contribution of this work
- Computing architectures
- Proposed 7T SRAM bit-cell
- Comparison of Read/Write access time
- Read stability analysis of 7T SRAM
- Reliability analysis of 7T SRAM
- Concept of MAC operation
- Proposed timing scheme for MAC
- CIM architecture
- Comparison to the state-of-the-art
- Conclusion
Motivation

- CNN workloads require parallel computing architectures.
- Conventional architectures have Von-Neumann bottleneck.
  - Limited on-chip memory bandwidth
  - High Latency
  - Less energy-efficiency
- The standard 6T SRAM bitcell has read disturb issue for CIM.
- Analog non-linearity disturbs the MAC operation.
- Single bit-precision of weights and inputs limits the accuracy.
Contribution of this work

• Improved throughput due to avoiding of read disturb issue using custom 7T SRAM bit-cell with decoupled read/write port.

• A novel timing scheme to apply the input pulses for MAC operation. Which improves the MAC operation linearity by minimizing reverse charging current.
Computing Architectures

Von-Neumann architecture

CIM architecture

Huge data transfer
Proposed 7T SRAM bitcell

(a) 7T SRAM bitcell

- The bit-cell area is $4.2 \, \mu m^2$. The bit-cell area of 7T is $1.3 \times$ higher than 6T and $1.1 \times$ lower than 8T.

(b) Bitcell Layout
Comparison of Read/Write access time

<table>
<thead>
<tr>
<th>Bit-cell</th>
<th>Write access time</th>
<th>Read Access time</th>
</tr>
</thead>
<tbody>
<tr>
<td>6T</td>
<td>21.58ps</td>
<td>36.19ps</td>
</tr>
<tr>
<td>7T*</td>
<td>22.62ps</td>
<td>44.49ps</td>
</tr>
<tr>
<td>8T</td>
<td>22.68ps</td>
<td>64.73ps</td>
</tr>
</tbody>
</table>

* Bit-cell proposed in this work

- The advantage of 7T bit-cell over the 6T are higher RSNM and read disturb free MAC operation with multiple row assertion.
- The 7T bit-cell is better than 8T in terms of less area and high speed of read operation.
Read Stability Analysis of 7T SRAM

(a) RSNM at worst case corner (FS) at $V_{RWL} = 0.6 \, V$.

(b) Distribution of read current of 7T bit-cell. The normalized standard deviation is 11.4%.

- The 7T bit-cell has $1.1 \times$ higher RSNM than 6T and $1.21 \times$ lower RSNM than 8T.
Reliability Analysis

(a) Standard 6T SRAM bit-cell supports only 4 rows activation simultaneously.

(b) Proposed 7T SRAM bit-cell supports 64 rows activation without read disturb.
### Concept of MAC operation

<table>
<thead>
<tr>
<th></th>
<th>Input = 0</th>
<th>Input = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight = 0 (QB = 1)</td>
<td>RWL</td>
<td>RWL</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\Delta V_{BL} \approx 0$</td>
<td>$\Delta V_{BL} \approx 0$</td>
</tr>
<tr>
<td>Weight = 1 (QB = 0)</td>
<td>RWL</td>
<td>RWL</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\Delta V_{BL} \approx 0$</td>
<td>$\Delta V_{BL} \approx 1 \times \Delta V$</td>
</tr>
</tbody>
</table>

(a) Bitwise multiplication concept for single 7T bit-cell.

(b) Multiple row activations performing bitwise multiply and accumulate on RBL.
Proposed Timing Scheme for MAC

(a) Input pulse generator circuit design using counter.

(b) Input timing schemes to apply input pulses into the SRAM array on RWL rows.

\[ V_{RBL}(t = t_1) = V_{DD} \]

\[ V_{RBL}(t = t_2) = V_{DD} - 2 \times \Delta V \]

\( V_{RBL(t_1)} > V_{RBL(t_2)} \)

\( I_{Rev2} > I_{Rev1} \)
CIM Architecture

(a) Proposed Compute-In-Memory Architecture

(b) Control Signals for a complete CIM operation and MAC operation linearity of proposed CIM macro
Comparison to the state-of-the-art

<table>
<thead>
<tr>
<th>Parameters</th>
<th>[1]</th>
<th>[2]</th>
<th>[3]</th>
<th>[4]</th>
<th>This work (Simulation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>65nm</td>
<td>55nm</td>
<td>65nm</td>
<td>40nm</td>
<td>65nm</td>
</tr>
<tr>
<td>Array Size</td>
<td>16 Kb</td>
<td>3.75 Kb</td>
<td>NA</td>
<td>32 Kb</td>
<td>4 Kb</td>
</tr>
<tr>
<td>Bitcell Type</td>
<td>10T</td>
<td>T8T</td>
<td>6T</td>
<td>8T</td>
<td>7T</td>
</tr>
<tr>
<td>Input Bits</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>2-4</td>
<td>4</td>
</tr>
<tr>
<td>Weight Bits</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>2-5</td>
<td>4</td>
</tr>
<tr>
<td>Output Bits</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>2-6</td>
<td>4</td>
</tr>
<tr>
<td>Model</td>
<td>CNN</td>
<td>CNN</td>
<td>k-NN</td>
<td>CNN</td>
<td>CNN</td>
</tr>
<tr>
<td>Supply Voltage (V)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.1 &amp; 0.9</td>
<td>1</td>
</tr>
<tr>
<td>Throughput (GOPS)</td>
<td>8 @6×1b</td>
<td>21.2 @4×5b</td>
<td>10.2 @8×8b</td>
<td>122 @2×2b</td>
<td>212.9* @4×4b</td>
</tr>
<tr>
<td>Energy Efficiency (TOPS/W)</td>
<td>40.3 @6×1b</td>
<td>18.4 @4×5b</td>
<td>1.94 @8×8b</td>
<td>17 @4×5b</td>
<td>28.9* @4×4b</td>
</tr>
<tr>
<td>Accuracy (CIFAR-10)</td>
<td>NA</td>
<td>90.42%</td>
<td>NA</td>
<td>88.5%</td>
<td>89.1%</td>
</tr>
</tbody>
</table>

*4b×4b MAC = 2OP

[1] A. Biswas et al. JSSC’19
[2] X. Si et al. JSSC’20
Conclusion

- 7T SRAM bitcell supports parallelism of MAC operation.
- CIM macro achieves a throughput of 212.9 GOPS.
- CIM macro achieves energy efficiency of 28.9 TOPS/W at 1V supply.
- The classification accuracy is 89.1% for CIFAR-10 dataset.
THANK YOU !!!

This work is supported in part by Semiconductor Research Corporation (SRC)