
Sangyeob Kim¹, Sangjin Kim¹, Soyeon Um¹, Soyeon Kim¹, Kwantae Kim², and Hoi-Jun Yoo¹

¹ School of Electrical Engineering, KAIST
² Institute of Neuroinformatics, University of Zurich and ETH Zurich
Computing-in-Memory (CIM) Accelerator

- Multi WLs Driving ➔ Low Energy Efficiency by ADC (<100 TOPS/W)

Pros
- MEM Access Reduction
- Analog Accumulation

Cons
- 1 WL ➔ Multi Cells Active
- 1 Col. ➔ Multi WLs Active
- ADC/DAC ➔ Large Power ➔ Large Area

NPU Architecture

CIM Architecture

Digitization

MAC Array

Input Memory

Weight Memory

Bottleneck

Output Memory

Digital Input Memory

Digital Output Memory

ADC

DAC

Digital Input Memory

Digital Output Memory

Pros

Cons

\[ \sum X_i \times W_{ij} \]
Limitation of Previous CIMs

1. High Precision ADC is Required for Digital Output Activations
2. Low Energy Efficiency due to Low Sparsity in Real Conditions

<Power Breakdown (VLSI 21)>

<Low Performance @ Real Application>

![Energy Efficiency (TOPS/W)](image)

- Peak Efficiency
- @ CIFAR-10
- @ ImageNet
Neuromorphic CIM Processor

- ADC and DAC are Not Necessary ➔ Power/Area Reduction
- Event-driven operation ➔ Input sparsity, but low weight sparsity

Pros
- High Input Sparsity
- No ADC/DAC

Cons
- 1 WL ➔ Multi Cells
- 1 Col. ➔ Multi WLs
- Limited range of $V_{BL}$

ADC and DAC are Not Necessary ➔ Power/Area Reduction
Event-driven operation ➔ Input sparsity, but low weight sparsity

Pros
High Input Sparsity
No ADC/DAC

Cons
1 WL ➔ Multi Cells
1 Col. ➔ Multi WLs
Limited range of $V_{BL}$
**Limited $V_{BL}$ Range of Neuromorphic CIM**

- **$V_{BL}$ Resolution and Range** are Limited by the Number of WL ($N_{WL}$)
  - Range of $V_{BL}$: $0 \sim N_{WL} \times \Delta V_{BL}$, Resolution of $V_{BL}$: $V_{BL,MAX} / N_{WL}$

### Diagram Details

- **Step 1**: $\Delta V_{BL}$
  - $W \times \frac{1}{C_{BL}} \int_{0}^{T_{WL}} I(t) \, dt$

- **Step 2**: $V_{BL}$
  - $\sum_{i=0}^{N} W \times \frac{1}{C_{BL}} \int_{0}^{T_{WL}} I(t) \, dt$

- **Output Memory**: $N_{WL}$ Cells Accum

- **Output Code**

- **Ideal Case**: $V_{MAX}$
  - Resolution = $V_{BL,MAX} / N_{WL}$

- **Error**: $V_{BL,MAX}$
  - $N$ Increase $\Rightarrow$ Error Increase

---

HOTCHIPS 2022

High Accuracy of Neuromorphic CIM

- Spiking-Neural-Network (SNN) Conversion from trained CNN
  - After CNN training, transfer weight to SNN ⇒ Highly Accurate SNN \([1,2]\)

\[ V(t+1) = V(t) + \sum S_i(t+1) \& W_i \]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Acc.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cifar-10</td>
<td>CNN</td>
<td>90.80</td>
</tr>
<tr>
<td>Cifar-100</td>
<td>SNN</td>
<td>90.05([1])</td>
</tr>
<tr>
<td>ImageNet</td>
<td>CNN</td>
<td>70.08</td>
</tr>
<tr>
<td>ImageNet</td>
<td>SNN</td>
<td>69.00([2])</td>
</tr>
</tbody>
</table>


Overall Architecture

- 16 Macros with 16 Banks ➔ Each Bank has 64x10 8T cell array
- Cap-based Adder, Voltage Folding Logic, Comparator array, PGA

### Overall Architecture

- **16 Macros with 16 Banks ➔ Each Bank has 64x10 8T cell array**
- **Cap-based Adder, Voltage Folding Logic, Comparator array, PGA**
Overall Architecture

1. MSB Word Skipping w/ '-1' Flag
2. Early Stopping w/ Sub-WL driver
3. Mixed-mode Firing w/ Voltage Folding
Characteristics of Weight Stored in CIM

- **High Negative Sign Extended Bits** (‘1111xxxx’) Ratio of MSB Part
  - Weight has gaussian distribution ➔ Most MSB has negative sign extended bits
  - 45% of total computation power is consumed by processing negative sign

![Graph showing weight distribution and sign extended bits ratio](image)

- **Weight Distribution**
  - 8'b 1111xxxx
  - 8'b 0000xxxx

- **Sign Extended Bits Ratio**
  - 95% > 4b MSB
  - 4b MSB 4b LSB

---

HOTCHIPS 2022

**MSB Word Skipping (MWS) with ‘-1’ Flag**

- **MSB BL Activity Reduction**
  - MSB ‘-1’ Flag BL enables BL voltage not to switch.
  - 4bit Negative sign extend bits (‘1111’) ➔ 5bit ‘-1’ flag + zeros (‘10000’)

### Diagram

- **Enable**
- **Disable**
- **Main CWL**
- **Sub-CWL 1**
- **Sub-CWL 2**
- **RWL**
- **Bit Conversion**
- **Peripheral Logic**
- **No Bit Conversion**

#### 8bit Weight (11111010)

- **1** 1 1 1 1 1 0 1 0
MSB Word Skipping (MWS) with ‘-1’ Flag

- **MSB BL Activity Reduction**
  - Two ‘-1’ Flag BLs enable BL voltage not to switch.
  - 2bit Negative sign extend bits (‘11’) $\rightarrow$ 3bit ‘-1’ flag + zeros (‘100’)

![Diagram of MSB Word Skipping (MWS) with ‘-1’ Flag]

- **MSB BL Activity Reduction**
  - Two ‘-1’ Flag BLs enable BL voltage not to switch.
  - 2bit Negative sign extend bits (‘11’) $\rightarrow$ 3bit ‘-1’ flag + zeros (‘100’)

![Diagram of MSB Word Skipping (MWS) with ‘-1’ Flag]
**MSB Word Skipping (MWS) with ‘-1’ Flag**

- **Performance of MSB Word Skipping**
  - 38% power consumption reduction @ 8b weight mode case
  - 25% power consumption reduction @ 4b weight mode case

**<Power Consumption>**

<table>
<thead>
<tr>
<th>Power Consumption (mW)</th>
<th>8b Mode</th>
<th>4b Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o MWS</td>
<td>241</td>
<td>169</td>
</tr>
<tr>
<td>w/ MWS</td>
<td>112</td>
<td>84</td>
</tr>
</tbody>
</table>

- **38%** power reduction
- **25%** power reduction

**<Measured Waveforms for MWS>**

- **V_{BL} Switching Reduction**
- **Time (ns)** 0 5 10 15 20 25 30 35 40
- **Voltage (V)** 0.0 0.1 0.2 0.3 0.4 0.5
- **V_{BL} w/o MWS**
- **V_{BL} w/ MWS**
Motivation of Early Stopping (ES)

- **Power Reduction by Eliminating Redundant Operation**
  - Small membrane voltage ($V_{MEM}$) neuron → No output spike
  - If $V_{MEM} < V_{ES}$ (Hyper Param) @ $T_{ES}$ → Early stopping

**<Firing Neuron>**

- \[ \sum S_{Pre} \cdot W_{Pos} > \sum S_{Pre} \cdot W_{Neg} \]

**<Non-Firing Neuron w/ ES>**

- Early Stopping → Power Reduction
  - Very Small!
  - Stop Next Operations

<table>
<thead>
<tr>
<th>Pre 0</th>
<th>Pre 1</th>
<th>Pre 2</th>
<th>Neuron</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre 0</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Voltage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{TH}$</td>
</tr>
<tr>
<td>$V_{MEM}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Voltage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{ES}$</td>
</tr>
<tr>
<td>$V_{MEM}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next</td>
</tr>
</tbody>
</table>

No Spike!

Stop Next Operations
(1) $V_{\text{POS}}$ Generation

- Connecting Bridge Cap. and Accumulating 2’s Complement Weight
  - Only BL voltage of data bit lines $\rightarrow$ pos. adder (Flag & Sign bit not transferred)
  - Positive adder generates **positive part of ($\sum W&S$ - Threshold)**

![Diagram showing the process of $V_{\text{POS}}$ generation with connecting bridge cap and accumulating 2's complement weight.](Image)
(2) $V_{\text{NEG}}$ Generation

- Connecting Bridge Cap. and Accumulating 2’s Complement Weight
  - BL voltage of ‘-1’ flag bits and sign bit are transferred to neg. adder
  - Negative adder generates negative part of $(\sum W&S - \text{Threshold})$
(3) Stopping before $T_{ES}$

- **1-bit Analog Comparator Generates Output Spike**
  - $V_{POS} > V_{NEG}$ in comparator $\Rightarrow$ output spike firing
  - Spike counter stores the number of output spikes for output memory

$V_{POS} > V_{NEG} \Rightarrow Output Spike$

![Diagram showing the firing logic and comparator](image)

<Measured Waveform of Neuron Operation>

- $V_{POS}$
- $V_{NEG}$
- $+V_{TH}$
- $+V_{NEG}$
- $+V_{POS}$
- $V_{POS} > V_{NEG}$

MEM Potential > $V_{TH}$

Fire
(4) Stopping @ $T_{ES}$

- Predicting Non-Firing Neuron and Stopping Neuron Operation
  - At $T_{ES}$, compare $V_{DIFF}$ and early stop voltage ($V_{ES}$)
  - If $V_{DIFF} < V_{ES}$, processing is stopped to reduce power

**Diagram:**

- $V_{DIFF} < V_{ES} \Rightarrow$ Sub-WL EN Low
- $T_{ES} (ES_{EN} \text{ High})$
- $V_{ES}$
- $V_{DIFF} \Rightarrow$ Early Stop
- $V_{DIFF} < V_{ES}$
- $ES_{EN} @ T_{ES}$
- $<ES \text{ Measured waveforms}>$

**Diagram Components:**

- Cell
- Gating
- Low
- Pos/Neg Adder
- Control Logic
- Spike Counter
- $+V_{POS}$
- $+V_{NEG}$
- $V_{POS}$
- $V_{NEG}$
- $GND$
- $V_{DIFF}$ ($V_{POS}$ $-$ $V_{NEG}$)
- $ES_{EN} @ T_{ES}$
- $GND (V_{POS} - V_{NEG})$
- $Firing \ Logic \ Gating$
- $Low \ Gating$
- $+V_{POS}$
- $+V_{NEG}$
Performance of Early Stopping

- Reducing Power Consumption by Early Stopping (@ CIFAR-10)
  - 50~70% of neurons in each layer are early terminated by prediction
  - 37.6% power consumption is reduced by early Stopping

![Early Stop Ratio (%) – Layer (@ ResNet-12)](image)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Early Stopping Ratio</th>
<th>Accuracy (%)</th>
<th>Power (mw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30%</td>
<td>92%</td>
<td>168.8</td>
</tr>
<tr>
<td>2</td>
<td>40%</td>
<td>90%</td>
<td>105.4</td>
</tr>
<tr>
<td>3</td>
<td>50%</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>60%</td>
<td>86%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>70%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>80%</td>
<td>82%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>90%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>95%</td>
<td>78%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>99%</td>
<td>76%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>100%</td>
<td>&lt;1% loss</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>100%</td>
<td>&lt;1% loss</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy (%) Power (mw)

Base Base ES ES
Motivation of Voltage Folding

- Virtually Increasing $V_{\text{LSB}}$ of Membrane Voltage ($V_{\text{MEM}}$)
  - By Voltage Folding $V_{\text{MEM}} \rightarrow$ Folding Count + Residue Voltage
  - Amplifying the residue voltage $\rightarrow$ Increasing the range and $V_{\text{LSB}}$ of $V_{\text{MEM}}$

Motivation of Voltage Folding

- Virtually Increasing $V_{\text{LSB}}$ of Membrane Voltage ($V_{\text{MEM}}$)
  - By Voltage Folding $V_{\text{MEM}} \rightarrow$ Folding Count + Residue Voltage
  - Amplifying the residue voltage $\rightarrow$ Increasing the range and $V_{\text{LSB}}$ of $V_{\text{MEM}}$
### Voltage Folding

- **Virtually Increasing Range of Membrane Voltage ($V_{MEM}$)**
  - $V_{IN}$ is folded @ $(V_1+V_2)/2$, $(V_2+V_3)/2$, $(V_3+V_4)/2$ …
  - Generating two voltages ($V_X$ and $V_Y$) with a phase difference of 180 degrees

![Voltage Folding Circuit](image)

- **Voltage Folding Circuit**
  - $V_{DD}$
  - $V_{IN}$
  - $V_X$
  - $V_Y$
  - $Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, Q_7, Q_8$

- **Graphs**
  - $V_Y$ vs $V_{IN}$
  - $V_X$ vs $V_{IN}$
  - $(V_1+V_2)/2$, $(V_2+V_3)/2$, $(V_3+V_4)/2$
Voltage Folding

- **Virtually Increasing Range of Membrane Voltage (V\text{MEM})**
  - Positive slope voltage selection between $V_X$ and $V_Y$ in Folding Circuit
  > increasing range of $V_{IN}$ continuously

![Voltage Folding Circuit](image)

Range: $x_1$, $x_2$, $x_3$

Residue: Positive Slope

$V_X = \frac{(V_1+V_2)/2}{(V_2+V_3)/2}$

$V_Y = \frac{(V_3+V_4)/2}{(V_3+V_4)/2}$

$Q_1$, $Q_2$, $Q_3$, $Q_4$, $Q_5$, $Q_6$, $Q_7$, $Q_8$
Increasing the Range of Voltage and Generating Folding Count

- $V_{MEM} \rightarrow \text{Analog}$ Folded output voltage ($V_{SEL}$) + \text{Digital} folding count
- High Virtual Range $\rightarrow$ Multi-Macro Aggregation w/o High Precision ADC

**Measured Waveforms of Voltage Folding**

- $V_{X}$ Selected
- $V_{Y}$ Selected
- $V_{POS,SEL}$ w/o folding
- $V_{POS,SEL}$ w/ folding

**High Reconfigurability**
Chip Summary

- Chip Photograph and Performance

### Chip Summary Table

<table>
<thead>
<tr>
<th>Technology</th>
<th>28nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die Area</td>
<td>3228μm x 900μm</td>
</tr>
<tr>
<td>Single Macro Area</td>
<td>0.048</td>
</tr>
<tr>
<td>Total CIM Storage</td>
<td>32 KB</td>
</tr>
<tr>
<td>Digital SRAM</td>
<td>32 KB</td>
</tr>
<tr>
<td>Supply</td>
<td>1.1V</td>
</tr>
<tr>
<td>Frequency</td>
<td>200MHz</td>
</tr>
<tr>
<td>Macro Power (mW)</td>
<td>15.8&lt;sup&gt;1&lt;/sup&gt; – 36.2&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>System Power (mW)</td>
<td>105.4&lt;sup&gt;1&lt;/sup&gt; – 241.4&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>System Energy Efficiency (TOPS/W)</td>
<td></td>
</tr>
<tr>
<td>l=4b, W=1b: 310.4&lt;sup&gt;3&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>l=4b, W=4b: 124.2&lt;sup&gt;3&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>l=4b, W=8b: 62.1&lt;sup&gt;3&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>

1) w/ MWS & ES  
2) w/o MSB & ES  
3) CIFAR-10 with ResNet-18
Conclusion

- **Neuro-CIM: An Energy-Efficient Neuromorphic CIM+SNN Processor**
  - CIM: Reducing memory access and multiple WL driving
  - SNN: Generating input sparsity and eliminating high-precision ADC

- **For Energy-Efficient Neuromorphic CIM Processing**
  - **MSB Word Skipping** ➔ Reducing 25~38% power consumption
  - **Early Stopping** ➔ Reducing 37% power consumption
  - **Mixed-mode Neuron Firing** ➔ Increasing the voltage range x3

A 310.4 TOPS/W 1034.6 TOPS/W Bank
Neuromorphic CIM Processor
for Energy Efficient Neural Network Processing
Thank You!

- Questions? Feel Free to Contact Me!
  - E-mail: sangyeob.kim@kaist.ac.kr
  - LinkedIn: https://www.linkedin.com/in/sangyeob-kim-871818179/
  - Zoom Meeting: https://us05web.zoom.us/j/3753663353?pwd=dzNlMFh3M0pleWJkM0dVZmxLNVVoUT09

- Acknowledgement
  - This work was supported by the Samsung Electronics.
    (IO201207-07799-01)