Proposal of Analog In-Memory Computing with Magnified Tunnel Magnetoresistance Ratio and Universal STT-MRAM Cell

Yanan Guo
Supervisor: Associate Prof. Hao Cai
National ASIC System Engineering Research Center
Southeast University
Digital and analog realization of IMC

- In-memory computing (IMC) is an effectual solution for energy-efficient artificial intelligence applications.
- Compared with digital IMC, analog IMC shows great merit in terms of high on-chip bandwidth and computation-area efficiency.
The challenges of analog IMC based on STT-MRAM

- Limited TMR causes mismatch between the impact of external input data and internal stored data on the output.
- The analog computation behavior is more vulnerable to transistor nonlinear effects than its digital counterpart.
- Multiple bit-lines are required to enable simultaneously for cost amortization. In order to avoid inconsistency operations, processing units must fulfill the calculation requirements of different loads.

\[ TMR = \frac{R_{AP} - R_P}{R_P} \]

100%-200% regular tunnel magnetoresistance (TMR) ratio is difficult to fulfill the requirement of analog in-memory computing.
Proposed analog IMC based on universal STT-MRAM

- MTJ resistance is converted into the gate voltage of transistor N1 by the latch structure. The TMR is ultimately amplified to the m-TMR:

\[ m\text{-TMR} = \frac{R_{off} - R_{on}}{R_{on}} \]
Latching structure

Integrating current mirror with feedback (CMF)
Power consumption of latch structure

Simulated latch performance with different TMR ratio

<table>
<thead>
<tr>
<th>TMR (%)</th>
<th>$R_{ref}$ (kΩ)</th>
<th>$V_1$ (mV)</th>
<th>Power (fJ)</th>
<th>Yield (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>7.7</td>
<td>700</td>
<td>97.8</td>
<td>75.8</td>
</tr>
<tr>
<td>100</td>
<td>8.5</td>
<td>600</td>
<td>78.4</td>
<td>86.8</td>
</tr>
<tr>
<td>150</td>
<td>9</td>
<td>600</td>
<td>74.3</td>
<td>93.8</td>
</tr>
<tr>
<td>200</td>
<td>9.5</td>
<td>600</td>
<td>70.8</td>
<td>95.2</td>
</tr>
<tr>
<td>250</td>
<td>9.5</td>
<td>600</td>
<td>68.8</td>
<td>97.5</td>
</tr>
</tbody>
</table>

- 5000 Monte-Carlo runs analysis is performed for $R_{ref}$ and latching voltage $V_1$ under different TMR.
- Comprehensive analysis of the influence of $R_{ref}$ and $V_1$ on latching and selection of appropriate parameters can reduce power consumption and improve yield.
Simulation results

- The integral nonlinearity (INL) can be reduced by 57.6% by using a feedback structure.
- The proposed IMC uses latches to magnify the m-TMR by 7500× when TMR equals 200%.
- The digital output quantized by 4-bit SAR-ADC is close to the ideal output. The maximum INL is 1.25 LSB.
Analysis of overall energy efficiency

- Latching power consumes the main part, more than 30%.

- When 64 rows open, the proposed structure achieves 25.4 TOPS/W which is 3.05× more than conventional MRAM.

- The delay of in-MRAM computing cannot change as the amount of calculated data increases and decreases 83.8% of digital MRAM when 64 lines are turned on.
Conclusion

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>180nm</td>
<td>90nm</td>
<td>130nm</td>
<td>22nm</td>
<td>28nm</td>
</tr>
<tr>
<td>Bit-cell</td>
<td>1T1FGT</td>
<td>1PCM</td>
<td>2T2R</td>
<td>2T2M</td>
<td>1T-1M</td>
</tr>
<tr>
<td>Computing Bit</td>
<td>4bIN-4bW-8bOUT</td>
<td>8bIN-4bW-8bOUT</td>
<td>1bIN-TbW-1bOUT</td>
<td>N/A</td>
<td>2bIN-1bW-4bOUT</td>
</tr>
<tr>
<td>Energy Efficiency</td>
<td>37¹ TOPS/W</td>
<td>11.9 TOPS/W</td>
<td>78.4 TOPS/W</td>
<td>N/A</td>
<td>9.47-25.4 TOPS/W</td>
</tr>
<tr>
<td>Write Voltage ²</td>
<td>&gt; 7V</td>
<td>&gt; 2V</td>
<td>&gt; 3V</td>
<td>&lt; 1V</td>
<td>&lt; 1V</td>
</tr>
<tr>
<td>Process Deviation ²</td>
<td>σ &lt;15%</td>
<td>σ &lt;15%</td>
<td>σ &gt;15%</td>
<td>σ &lt;15%</td>
<td>σ &lt;15%</td>
</tr>
</tbody>
</table>

¹ Only MAC operation, other peripheral work was not included.
² A part of the comparison refers to [4].

- **Challenge:** Little superiority of energy efficiency
- **Advantage:** Low write voltage and low process deviation
- **Favorable scenario:** artificial intelligence applications on IoT edge devices, normally-off and instant-on storage and computing units

Thanks

Yanan Guo
E-mail: guoyn@seu.edu.cn
Supervisor: Associate Prof. Hao Cai
National ASIC System Engineering Research Center
Southeast University