tinyML® Research Symposium

Enabling Ultra-low Power Machine Learning at the Edge

March 27, 2023

www.tinyML.org
Memory-Oriented Design-Space Exploration of Edge-AI Hardware for XR Applications

Vivek Parmar¹, Syed Shakib Sarwar², Ziyun Li², Hsien-Hsin S. Lee², Barbara De Salvo²†, and Manan Suri¹

manansuri@ee.iitd.ac.in

¹Indian Institute of Technology Delhi
²Meta Reality Labs Research

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023
Motivation & Scope

Demonstrate benefits of memory-centric computing utilizing advanced NVM technology for XR-EAI applications

- Exploit normally-off computing due to nature of workload
- Analyze memory & power budgets for hybrid architectures through DTCO
- Estimates/Projections at multiple nodes and type of NVM devices
- Relevant Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>HTC Vive Pro</th>
<th>Ideal VR</th>
<th>Microsoft HoloLens2</th>
<th>Ideal AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (MP)</td>
<td>4.6</td>
<td>200</td>
<td>4.4</td>
<td>200</td>
</tr>
<tr>
<td>Refresh rate (Hz)</td>
<td>90</td>
<td>90-144</td>
<td>120</td>
<td>90-144</td>
</tr>
<tr>
<td>Motion-to-photon latency (ms)</td>
<td>&lt;20</td>
<td>&lt;20</td>
<td>&lt;9</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Power (W)</td>
<td>N/A</td>
<td>1.2</td>
<td>&gt;7</td>
<td>0.1-0.2</td>
</tr>
</tbody>
</table>

1. **Eye Segmentation**  
   Dataset: OpenEDS 2019  
   - Network: Unet (backbones: MobileNetv2)  
   - Framework: Tensorflow

2. **Hand detection**  
   Dataset: FPHAB*  
   - Network: DetNet (MegaTrack)  
   - Framework: PyTorch

---

* Indian Institute of Technology Delhi obtained and used the FPHAB dataset

---

XR-EAI Workloads: Impact of Quantization

- Comparable performance between full-precision and quantized versions
- Weight distribution profile changes due to use of additional scaling factors specific to layers during quantization
Performance on CMOS-based Systolic Accelerators

Performance on CMOS-based Systolic Accelerators

- Technology scaling based on DeepScale [1] for: 22 nm, 28 nm
- 7nm estimates based on TPUv4 [2] scaling factors
- **Benefits of scaling diminishing at 7nm**

Proposed NVM-based Enhancements

(a) Operation Breakdown for XR-AI accelerator

(b) Memory Activity Breakdown

(c) AI Inference - Memory Operation Breakdown

(i) Traditional Memory Mapping (baseline CPU, Eyeriss, Simba)

(ii) Proposed P0 Mapping (NVM for weight matrix)

(iii) Proposed P1 Mapping (NVM in all buffers)

Two flavours explored

1. P0: MRAM for only weights

2. P1: MRAM everywhere except compute registers
Performance Analysis for Proposed NVM-enhanced variants

Direct Area saving in all variants

<table>
<thead>
<tr>
<th>Architecture</th>
<th>7 nm Area (mm²)</th>
<th>Area savings</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRAM-only</td>
<td>P0</td>
<td>P1</td>
<td>P0</td>
</tr>
<tr>
<td>Simba</td>
<td>2.89</td>
<td>2.41</td>
<td>1.88</td>
<td>16.56%</td>
</tr>
<tr>
<td>Eyeriss</td>
<td>2.56</td>
<td>2.11</td>
<td>1.67</td>
<td>17.52%</td>
</tr>
</tbody>
</table>

28nm P0 savings

- **DetNet**: ~50% with CPU, ~80% with Eyeriss, ~70% with Simba
- **EDSNet**: ~ 7% with CPU, ~70% with Eyeriss, ~1% with Simba

Energy saving evident in some variants (28nm-P0, all applications) w.r.t SRAM only variants
Energy Breakdown for Compute & Memory

At 7nm energy estimated for NVM-based variants (P0,P1) > “SRAM-only” variant

- 7nm MRAM type considered is write-optimized (ref-IMEC). However, the XR application is Read Dominant.

- Gains @ 7nm can be obtained with a read optimized MRAM.

- Mem Read E > Mem Write E in P0 (all cases) → Reduced write operations in weight memory – inference dominated workload (not true for SRAM though)
IPS (Inferences Per Second over op time) / Effective Latency & not actual Inference Latency

A more relevant performance metric for edge XR-AI as inference operations may:

Invoke AI for XR in Asymmetric/Infrequent manner after long/erratic intervals

Configure:
- Min. Hand Detection IPS ~ 10 (use)
- Min. Eye segmentation IPS ~ 0.1 (use only during initiation of gaze tracking or authentication)
IPS-Analysis

Below IPS cross-over point $\rightarrow$ energy-saving while using advanced NVM compared to baseline SRAM only variant.
IPS Analysis - Summary

**TABLE III**
IPS Analysis summary for proposed architectures using PE configuration v2 (64x64).

<table>
<thead>
<tr>
<th>XR-AI Workload</th>
<th>Architecture</th>
<th>Inference Latency (ms)</th>
<th>$P_{Mem} \text{ Savings @ } IPS_{min}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P0</td>
<td>P1</td>
</tr>
<tr>
<td>DetNet $IPS_{min}=10$</td>
<td>Simba</td>
<td>0.34</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Eyeriss</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>EDSNet $IPS_{min}=0.1$</td>
<td>Simba</td>
<td>48.57</td>
<td>60.72</td>
</tr>
<tr>
<td></td>
<td>Eyeriss</td>
<td>45.22</td>
<td>45.22</td>
</tr>
</tbody>
</table>

Clear power saving even with write optimized MRAM!
Conclusion

1. Detailed study on 2 XR-AI workloads (hand-detection and eye-segmentation).
2. Design exploration for mapping workloads on CPU and systolic accelerators (QKeras & Timeloop + Accelergy frameworks).
4. Memory-oriented DTCO based on the use of different types of the emerging MRAM devices.
   a) Memory-Energy Savings ≥ 24% observed for hand detection (at IPS = 10) and eye segmentation (at IPS=0.1) for Simba-like NVM accelerator variant.
   b) Substantial area reduction (≥30%) due to the high-density feature of MRAM technology.
Thank You

manansuri@ee.iitd.ac.in
Copyright Notice

This presentation in this publication was presented at the tinyML® Research Symposium (March 27, 2023). The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org