A 65nm 6T-MRAM Mix-Precision Near-Memory-Compute Macro Using Input-Sparsity Aware and Toggle Rate Aware Learning Schemes

組別：A517　組員：吳承翰　指導教授：張孟凡

Abstract

To meet the energy and performance demands of edge AI, this project explores a 65nm MRAM-based near-memory computing architecture. By applying toggle-aware quantization (TAWT), we reduce bit-level switching during weight access, achieving energy savings with minimal accuracy loss. Circuit-level optimizations, including improved sensing timing and a DVME-VSA-XOR design for mixed-precision computation, further enhance efficiency.

Method

This work presents a 65nm STT-MRAM-based nvNMC architecture supporting hybrid 8-bit and 1-bit precision for edge AI. The 4-kbit macro (8×576) enables both MAC and XOR operations through shared bitlines, using DVME-VSA and DVME-VSA-XOR circuits.

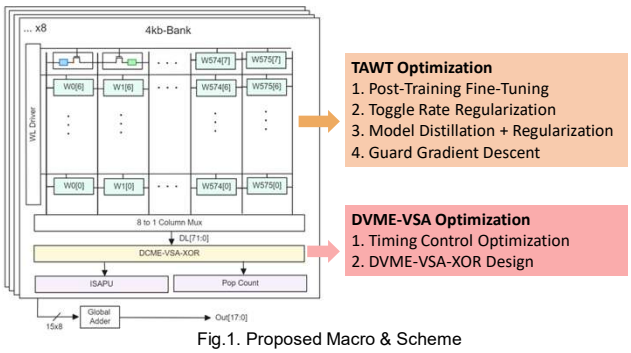

Fig.1. Proposed Macro & Scheme

On the **algorithm side**, we apply Toggle-Aware Weight Training (TAWT) with multiple strategies:

- **Post-training fine-tuning** reduces toggle rate by adjusting weights within a narrow range.
- **Toggle rate regularization** integrates bit-switching penalties into the loss function.
- **Knowledge distillation + regularization** balances accuracy and toggle efficiency.
- **Guarded gradient descent** selectively applies weight updates that benefit energy or accuracy.

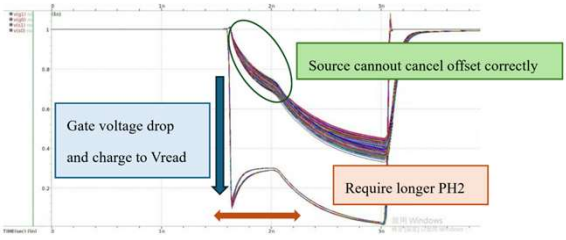

Fig. 2. Source & Gate Readout Waveform Before Optimization

On the **circuit level**, we introduce two key improvements:
1. Timing optimization of the DVME-VSA: By advancing the BLPRE signal by 0.3 ns, we ensure a smoother discharge path and better offset margin, improving sensing accuracy and reducing latency. This can resolve the problem shown in Fig. 2.
2. DVME-VSA-XOR design: As shown in Fig. 3, a novel sensing amplifier performs XOR directly during readout.

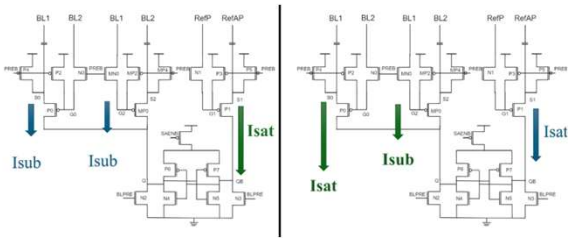

Fig. 3. Schematic of Proposed DVME-VSA
(Left : when input are different ; Right :when input are the same)

Result & Conclusion

As shown in **Fig. 4**, the proposed *Guarded Gradient Descent* reduces toggle rate by **31%** while cutting accuracy drop by **50%**, achieving high efficiency without sacrificing model precision.

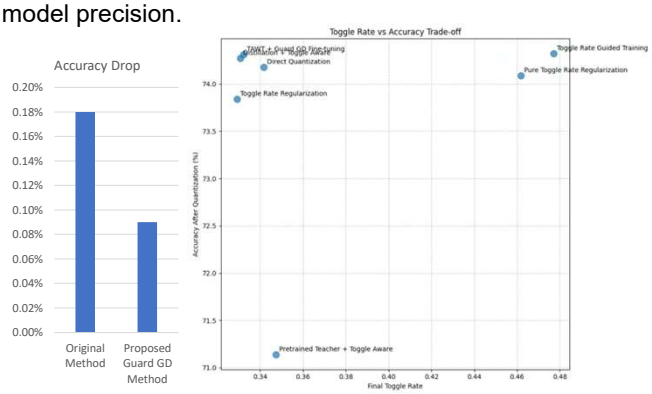

Fig. 4. Comparison Between Different Toggle Rate Reduction Method

Fig. 5 shows that with optimized DVME-VSA timing, the source discharges smoothly without recharging, reducing read latency. Under identical sensing windows, accuracy improves from **99.8%** to **100%**.

Moreover, our DVME-VSA-XOR achieves an average **28.4% power reduction** versus the traditional 6T XOR implementation.

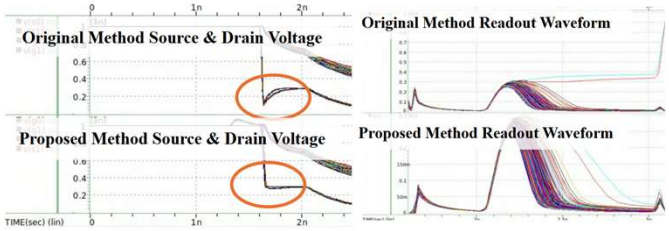

Fig. 5. DVME-VSA Simulation Waveform Comparison

This work proposes a 65nm STT-MRAM-based near-memory compute macro supporting both 8-bit and 1-bit precision to meet the energy and flexibility demands of edge AI. On the algorithm side, toggle-aware training strategies—especially Guarded Gradient Descent—effectively reduce toggle rate by 31% with only minimal accuracy degradation (less than 0.3%), cutting the accuracy drop in half compared to traditional methods. Combined with a 28.4% power reduction in XOR operations and improved readout accuracy from 99.8% to 100%, these results validate the macro's effectiveness for low-power AI inference at the edge.