An MRAM-based Deep In-Memory Architecture for Deep Neural Networks

Ameya D. Patil\textsuperscript{1}, Haocheng Hua\textsuperscript{1}, Sujan Gonugondla\textsuperscript{1}, Mingu Kang\textsuperscript{2}, and Naresh R. Shanbhag\textsuperscript{1}

\textsuperscript{1}: University of Illinois at Urbana-Champaign
\textsuperscript{2}: IBM T. J. Watson Research Center
Challenge of Bringing Intelligence to the Edge

Intelligence at the edge

von Neumann Bottleneck

Programmable

Memory ↔ Processor

Read energy

Processing energy

>10 : 1

[Horowitz-ISSCC-2014]

Need disruptive innovations encompassing circuits and architectures
Deep In-memory Architecture (DIMA)

- read multiple-bits per column per access
- massively parallel mixed-signal computing
- bandwidth efficient digital output

\[ \Delta V_{BL} \propto \text{multiple-bits per column} \]

PE: processing elements

[Kang-ICASSP-2014]
[Kang-ISCAS-2015]
[Kang-JSSC-2018]
[Gonugondla-ISSCC-2018]
Recent In-memory Computing Works

**SRAMs**

- **Our work**
  - [Kang-JSSC-2018]
  - [Jiang-VLSI-2018]
  - [Valavi-VLSI-2018]
  - [Gonugondla-ISSCC-2018]

- [Biswa-ISSSC-2018]
- [Khwa-ISSCC-2018]

**Floating-gate analog**

- [Lu-JSSC-2018]
- [Mahmoodi-DAC-2018]

**Multi-state RRAM/PCM**

- [Bayat-NatureComm-2018]
- [Gallo-NatureElec-2018]
In which Memory Should the Compute Take Place?

**Emerging non-volatile memories**
- At sweet-spot for in-memory computing: size, density, non-volatility
- MRAM: on the verge of commercialization

**SRAM**
- small capacity (~20MB)
- low density, volatile
- 10X access/MAC cost

**DRAM**
- volatile – requires expensive refresh
- 100X access/MAC cost

**Flash**
- large in size
- Data wrapped under file system

Conventional MRAM Array

- 1T-1R bitcell, storing 1-bit/cell
- Crossbar architecture: SL ⊥ BL
- Muxing ratio of L:1
- Constant current-based voltage sensing for read
- Typical values:
  - Read current $I_{\text{read}} = 40\mu A$
  - Mux ratio: L=8:1
  - Access time: $T_{\text{on}} = 3\text{ns}$

SL: source-line, BL: bitline, WL: word-line
MRAM-based Deep In-memory Architecture (MRAM-DIMA)

- Preserves conventional bitcell array (BCA) structure
- Executes multi-bit matrix-vector multiplication (MVM) as single operation

\[ \mathbf{a} = \mathbf{W} \mathbf{x} \]

- \( \mathbf{W} \): \( M \times N \) matrix stored in BCA
- \( \mathbf{x} \): \( N \times 1 \) digital input vector
- \( \mathbf{a} \): \( M \times 1 \) digital output vector
Word Row Block (WRB) Details

WRB computing one N-dimensional dot-product

BL analog voltages: input X

Modulated WL pulse widths

Current integrator (CI)

Integrating accumulated SL currents

Weight storage in column-major format
MRAM-DIMA Operation

(functional read phase)

\[ \Delta V_{x,i} \]

Input BL voltages

\[ \text{Binary weighted WL pulses} \]

WL0 is OFF

\[ \Delta V_0 \]

\[ \text{at the end of FR} \]

\[ \Delta V_{0,FR} = \frac{T_0 V_{lsb}}{C_o} \left[ \Delta G \sum_{j=1}^{N} w_{ij} x_j + (2^{B_w-1} - 1) \sum_{j=1}^{N} G_j x_j \right] \]
MRAM-DIMA Operation
(bias removal phase)

Change in BL voltages

Only WL0 is ON

\[ \Delta V_{x,i} \]

\[ \Delta V_{o,FR} \]

\[ \Delta V_{o,dp} \]

\[ \Delta V_{o} \]

Functional read (FR)
Bias removal (BR)

\[ \Delta V_{x,1} \]
\[ \Delta V_{x,2} \]

\[ \Delta V_{o} \text{ at the end (desired dot product)} \]

ECE ILLINOIS 10
Multi-bank MRAM-DIMA

Sub-array 0

Sub-array 1

Sub-array 2

$B_w$ $\frac{N}{3}$ shared SLs $\frac{N}{3}$ shared SLs $\frac{N}{3}$

$B_w$ $N$ $M$ WRBs $a_1$ $a_M$

Sub-array 0

Sub-array 1

Sub-array 2

FR FR BR BR

BR FR BR BR

BR BR FR

$t = 0$ $t = T_{FR}$ $t = 2T_{FR}$ $t = 3T_{FR}$
Simulation Methodology

- SPICE simulation in commercial 22nm CMOS-MRAM process
- Analytical model MTJ process variations
- Estimate system-level accuracy and energy-delay benefits via component-level behavioral & variations models

Transfer function of scalar multiplication $w \times x$ simulated in commercial 22nm CMOS-MRAM process
Impact of MTJ Process Variations

- scalar multiplication $w \times x$:

$$\Delta V_{o,m} = S + \eta = \frac{T_0 V_{lsb} \Delta G}{C_o} (wx) + \eta$$

$\eta$: spatial noise arising due to MTJ process variations, $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$

$$\sigma_\eta^2 = \left( \frac{T_0 V_{lsb}}{C_o} x \right)^2 \left( (2^{B_w-1} - 1)^2 G_0^2 + \sum_{b=1}^{B_w-1} 4^{b-1} G_b^2 \right) \left( \frac{\sigma}{\mu} \right)_{G-bc}^2$$

$\left( \frac{\sigma}{\mu} \right)_{G-bc}$: std. to mean ratio of bitcell conductance variation

Estimated via simulations in 22nm CMOS-MRAM process

Enables evaluation of the impact on system-level accuracy
Energy-Delay Benefits

**Energy Breakdown**

- BCA
- SA
- Compute
- ADC+DAC
- CI

- MRAM-Digital
- MRAM-DIMA

Total MVM energy ($\times 10^1 n$)

- 3.5
- 3
- 2.5
- 2
- 1.5
- 1
- 0.5
- 0

**Delay Breakdown**

- Conventional
- DIMA

- 70x lower delay for $M = 64$

- $L = 8$; $N = 576$, $M = 64$, $B_w = 5$, $B_a = 4$

- MRAM-DIMA delay is independent of $M$

- energy reduction dominated by col. mux. factor $L$
System-level Accuracy

**LeNet-300-100 on MNIST**

- Networks trained using DoReFa-net [Zhou-CoRR-2016] training methodology
- MRAM bitcell variations emulated in PyTorch
- MRAM-DIMA maintains classification accuracy up to 4x higher \( \frac{\sigma}{\mu} \) than its typical value

(statistics is obtained for 100 instances of MRAM arrays for each value of \( \frac{\sigma}{\mu} \))

**9-layer CNN on CIFAR-10**

- Typical \( \frac{\sigma}{\mu} \) accuracy drop ≤ 0.5%
- Ideal fixed-point accuracy

- Typical \( \frac{\sigma}{\mu} \) accuracy drop ≤ 1%
- Ideal fixed-point accuracy

- MRAM-DIMA maintains classification accuracy up to 4x higher \( \frac{\sigma}{\mu} \) than its typical value

(statistics is obtained for 100 instances of MRAM arrays for each value of \( \frac{\sigma}{\mu} \))
Conclusion and Future Work

- Proposed MRAM-based deep in-memory architecture (DIMA) achieves 4.5x and 70x reduction in energy and delay, respectively.

- Next steps:
  - Investigate layout and pitch-matching constraints for DIMA peripheral circuits
  - Explore architectural dataflow mappings to realize efficient DNN system implementations
Thank You!

This material is based on research sponsored by Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-18-2-7866. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) or the U.S. Government.