Efficient Neural Computing Enabled by Magneto-Metallic Neurons and Synapses

Kaushik Roy
Abhronil Sengupta, Kartik Yogendra, Deliang Fan, Syed Sarwar, Priya Panda, Gopal Srinivasan, Jason Allred, Zubair Azim, A. Raghunathan
ECE, Purdue University

Presented By: Shreyas Sen, ECE, Purdue University
The Computational Efficiency Gap

IBM Watson playing Jeopardy, 2011

~200000 W

IBM Blue Gene supercomputer, equipped with 147456 CPUs and 144TB of memory, consumed 1.4MW of power to simulate 5 secs of brain activity of a cat at 83 times slower firing rates
**Neuromorphic Computing Technologies**

- **Approximate Neural Nets**, ISLPED ’14
- **Conditional Deep Learning**, DATE 2016
- ….

**Hardware Accelerators**

- **Spintronics-Enabled**
- **Approximate Computing, Semantic Decomposition, Conditional DLN**

**SW (Multicores/GPUs)**

1 uJ/neuron

- **Spin neuron**, IJCNN ’12, APL’15, TNANO, DAC, DRC, IEDM
- **Spintronic Deep Learning Engine**, ISLPED ’14
- **Spin synapse**, APL ’15
- ….
Device/Circuit/Algorithm Co-Design: Spin/ANN/SNN

**Top-Down**

Investigate brain-inspired computing models to provide algorithm-level matching to underlying device physics

Device-Circuit-Algorithm co-simulation framework used to generate behavioral models for system-level simulations of neuromorphic systems

**System Level Solution**

**Bottom-Up**

Investigate device physics to mimic “neuron/synapse” functionalities

Calibration of device models with experiments
**BUILDING PRIMITIVES: MEMORY, NEURONS, SYNAPSES**

Lateral Spin Valve (Local & Non-local)

Spin current

Domain Wall “transistor”

Spin current

Excitatory Input

Inhibitory Input

Preset

Synaptic current

\[ \Delta I_{in} = (I_{in}^+ - I_{in}^-) \]
Three terminal device structure provides decoupled “write” and “read” current paths
- Write current flowing through heavy metal programs domain wall position
- Read current is modulated by device conductance which varies linearly with domain wall position

**Universal device:** Suitable for **memory, neuron, synapse, interconnects**
Simple ANN: Activation

Artificial NN

transmitting neuron

Signal transmission

Summation of weighted inputs

Thresholding function

Spin Hall based Switching

Switch a magnet using spin current, read using TMR effect

DW-MTJ
Step and Analog ANN Neurons

- Neuron, acting as the computing element, provides an output current ($I_{OUT}$) which is a function of the input current ($I_{IN}$).
- Axon functionality is implemented by the CMOS transistor.
- Note: Stochastic nature of switching of MTJ can be used in Stochastic Neural nets.
## Benchmarking with CMOS Implementation

<table>
<thead>
<tr>
<th>Neurons</th>
<th>Power</th>
<th>Speed</th>
<th>Energy</th>
<th>Function</th>
<th>technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMOS Analog neuron 1 [1]</td>
<td>~12µW (assume 1V supply)</td>
<td>65ns</td>
<td>780fJ</td>
<td>Sigmoid</td>
<td>/</td>
</tr>
<tr>
<td>CMOS Analog neuron 2 [2]</td>
<td>15µW</td>
<td>/</td>
<td>/</td>
<td>Sigmoid</td>
<td>180nm</td>
</tr>
<tr>
<td>CMOS Analog neuron 3 [5]</td>
<td>70µW</td>
<td>10ns</td>
<td>700fJ</td>
<td>Step</td>
<td>45nm</td>
</tr>
<tr>
<td>Digital Neuron [3]</td>
<td>83.62µW</td>
<td>10ns</td>
<td>832.6fJ</td>
<td>5-bit tanh</td>
<td>45nm</td>
</tr>
<tr>
<td>Hard-Limiting Spin-Neuron</td>
<td>0.81µW</td>
<td>1ns</td>
<td>0.81fJ</td>
<td>Step</td>
<td>/</td>
</tr>
<tr>
<td>Soft-Limiting Spin-Neuron</td>
<td>1.25µW</td>
<td>3ns</td>
<td>3.75fJ</td>
<td>Rational/Hyperbolic</td>
<td>/</td>
</tr>
</tbody>
</table>

**Compared with analog/digital CMOS based neuron design, spin based neuron designs have the potential to achieve more than two orders lower energy consumption**

[4]: D. Coue, etc "A four-quadrant subthreshold mode multiplier for analog neural network applications", TNN, 1996
[5]: M. Sharad, etc, “Spin-neurons: A possible path to energy-efficient neuromorphic computers”, JAP, 2013
In-Memory Computing (Dot Product)

Artificial NN

nucleolus

transmitting neuron

In-portion:

\[
V_1 \cdot w_{i1} + V_2 \cdot w_{i2} + V_3 \cdot w_{i3} = \sum V_{i} \cdot w_{i}
\]

Programmable resistors/DWM

Output:

\[
nucleolus
\]

Signal transmission

Summation of weighted inputs

Thresholding function

Input Voltages

WRITE

READ

MTJ “Free Layer”

MTJ “Pinned Layer”

Domain Wall

Heavy Metal

“Pinned Layer”

GND

WRITE
All-Spin Artificial Neural Network

- All-spin ANN where spintronic devices directly mimic neuron and synapse functionalities and axon (CMOS transistor) transmits the neuron’s output to the next stage
- Ultra-low voltage (~100mV) operation of spintronic synaptic crossbar array made possible by magneto-metallic spin-neurons
- System level simulations for character recognition shows maximum energy consumption of 0.32fJ per neuron which is ~100x lower in comparison to analog and digital CMOS neurons (45nm technology)
Spiking Neural Networks (Self-Learning)
Spiking Neuron Membrane Potential

The leaky fire and integrate can be approximated by an MTJ – the magnetization dynamics mimics the leaky fire and integrate operation.

**LIF Equation:**

\[ C \frac{dV}{dt} = -\frac{V}{R} + \sum_j w_j I_{post,j} \]

**Input spikes to neuron**

**Input spin current (uA)**

**Neuron Membrane Potential**

**MTJ Spiking Neuron**

**LLGS Equation:**

\[ \frac{d\hat{m}}{dt} = -\gamma (\hat{m} \times H_{eff}) + \alpha (\hat{m} \times \frac{d\hat{m}}{dt}) + \frac{1}{qN_s} (\hat{m} \times I_s \times \hat{m}) \]

**In-plane magnetization**

**Time (ns)**

0 1 2 3 4 5

0 1 2 3 4 5

0.4 0.6 0.8
MTJ as a Spiking Neuron

- MTJ magnetization leaks and integrates input spikes (LLG equation) in presence of thermal noise.
- Associated “write” and “read” energy consumption is ~ 1fJ and ~1.6fJ per time-step which is much lower than state-of-the-art CMOS spiking neuron designs (267pJ [1] and 41.3pJ [2] per spike).
Spiking Neurons

LLGS Based Spiking Neuron

Input Spikes

MTJ conductance

LLG Equation Mimicking Spiking Neurons

DW-MTJ base IF Neurons

Input Spikes

Membrane Potential

Output Spikes

DW Integrating Property Mimicking IF Neuron
Arrangement of DW-MTJ Synapses in Array for STDP Learning

Spike-Timing Dependent Plasticity

- Spintronic synapse in spiking neural networks exhibits spike timing dependent plasticity observed in biological synapses
- Programming current flowing through heavy metal varies in a similar nature as STDP curve
- Decoupled spike transmission and programming current paths assist online learning
- 48fJ energy consumption per synaptic event which is ~10-100x lower in comparison to SRAM based synapses /emerging devices like PCM
## Comparison with Other Synapses

<table>
<thead>
<tr>
<th>Device</th>
<th>Reference</th>
<th>Dimension</th>
<th>Prog. Energy</th>
<th>Prog. Time</th>
<th>Terminals</th>
<th>Prog. Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeSbTe memristor</td>
<td>D. Modha ACM JETCAS, 2013 (IBM)</td>
<td>40nm mushroom and 10nm pore</td>
<td>Average 2.74 pJ/event</td>
<td>~60ns</td>
<td>2</td>
<td>Programmed by Joule heating (Phase change)</td>
</tr>
<tr>
<td>GeSbTe memristor</td>
<td>H.-S. P. Wong Nano Letters, 2012 (Stanford)</td>
<td>75nm electrode diameter</td>
<td>50pJ (reset) 0.675pJ (set)</td>
<td>10ns</td>
<td>2</td>
<td>Programmed by Joule heating (Phase change)</td>
</tr>
<tr>
<td>Ag-Si memristor</td>
<td>Wei Lu Nano Letters, 2010 (U Michigan)</td>
<td>100nmx100nm</td>
<td>Threshold voltage~2.2V</td>
<td>~300µs</td>
<td>2</td>
<td>Movement of Ag ions</td>
</tr>
<tr>
<td>FeFET</td>
<td>Y. Nishitani JJAP, 2013 (Panasonic, Japan)</td>
<td>Channel Length-3µm</td>
<td>Maximum gate voltage – 4V</td>
<td>10µs</td>
<td>3</td>
<td>Gate voltage modulation of ferroelectric polarization</td>
</tr>
<tr>
<td>Floating gate transistor</td>
<td>P. Hasler IEEE TBIOCAS, 2011 (GaTech)</td>
<td>1.8µm/0.6µm (0.35µm CMOS technology)</td>
<td>Vdd - 4.2V Tunneling Voltage – 15V</td>
<td>100µs (injection) 2ms (tunneling)</td>
<td>3</td>
<td>Injection and tunneling currents</td>
</tr>
<tr>
<td>SRAM synapse</td>
<td>B. Rajendran IEEE TED, 2013 (IIT Bombay)</td>
<td>0.3µm² (10nm CMOS technology)</td>
<td>Average 328fJ for 4-bit synapse</td>
<td>-</td>
<td>-</td>
<td>Digital counter based circuits</td>
</tr>
<tr>
<td>Spintronic synapse</td>
<td>NRL Purdue</td>
<td>340nmx20nm</td>
<td>Maximum 48fJ /event</td>
<td>1ns</td>
<td>3</td>
<td>Spin-orbit torque</td>
</tr>
</tbody>
</table>
MTJ Enabled All-Spin Spiking Neural Network

Probabilistic Spiking Neuron
• A pre-neuronal spike modulated by synapse to generate current that controls the post-neuronal spiking probability.
• Exploit stochastic switching behavior of MTJ in presence of thermal noise.
MTJ Enabled All-Spin Spiking Neural Network

Stochastic Binary Synapse

- Synaptic strength proportional to temporal correlation between pre- and post-spike trains.
- **Stochastic STDP** – Synaptic learning embedded in the switching probability of binary synapses.
MTJ Enabled All-Spin Spiking Neural Network

Stochastic SNN Hardware Implementation

- Crossbar arrangement of the spin neurons and synapses for energy efficiency.
  - Average neuronal energy of 1fJ and 1.6fJ per timestep for write and read operations, and 4.5fJ for reset.
  - Average synaptic programming energy of 70fJ per training epoch.

Classification accuracy of 73% for MNIST digit recognition.
Summary

- Spintronics do show promise for low-power non-Boolean/brain-inspired computing
  - Need for new leaning techniques suitable for emerging devices
  - Materials research, new physics, new devices, simulation models
- An exciting path ahead…