MAGNETIC TUNNEL JUNCTION (MTJ) AS STOCHASTIC NEURONS AND SYNAPSES:
STOCHASTIC BINARY NEURAL NETWORKS
BAYESIAN INFERENCING
OPTIMIZATION PROBLEMS

Kaushik Roy & Abhronil Sengupta
Gopal Srinivasan, Chamika Liyanagendra, Akhilesh Jaiswal,
Parami Wijesinghe
Center for Brain-Inspired Computing (C-BRIC)

Center for Brain-Inspired Computing (C-BRIC), National Science Foundation, DARPA,
Vannevar Bush Fellowship, Office of Naval Research, Semiconductor Research Corporation, Intel
Barrier height of magnet (determined by the magnet volume) dictates data retention time.

As the magnet is scaled down to barrier heights < 5JT, it enters superparamagnetic telegraphic switching regime.

Appropriate choice of barrier height required depending on the application under consideration – memory or computing.
Stochastic Switching of MTJs

Sigmoid like Switching

- Write Pulse

- Write Current

MTJ with Underlying HM Layer

- Separate read (terminals 1&2 and) write paths (terminals 3&2)
- Constant resistance across the write path
- High spin Injection efficiency of the spin-orbit-coupling phenomenon
Earlier work on STT-based Stochastic Devices

**True Random Number Generator**
(Bias magnet to 50% probability)

**Stochastic Logic**
Roy Raghunathan DATE 2015

**Stochastic Synapse**
Querlioz TBioCAS 2015

**Stochastic Neurons & STDP-Based Stochastic Synapses**

**Comparative Analysis of High and Low Barrier Magnets for Stochastic NNs**
Roy PR Applied 2017

**Belief Networks**
Datta Scientific Reports 2016

**“p bits” – Scaled 1KT stochastic magnets for Spin Logic, Ising Computing**
Datta Scientific Reports 2017, PRX 2017, Magnetic Letters 2017

Magnets biased to a particular switching probability

Magnets behaving as “Stochastic Bits”
Stochastic Bits in Other Technologies

- Stochasticity in memristors can be exploited for learning (disadvantages: high power, reduced reliability) and neural inference (disadvantages: power hungry converters required for interfacing with crossbar arrays)

- CMOS implementations require hardware-expensive random number generator circuits
Stochastic Computing: From Devices to Circuits and Systems

Neuromorphic Computing

Stochastic spin device

Combinatorial optimization (Ising model)

Random Walk (SAT solver)

Probabilistic Inference (BN)

STOCHASTIC NEURAL NETWORKS:
STOCHASTIC BITS AS NEURONS & SYNAPSES
Biological & Artificial Neural Networks

Biological Neural Network

Artificial Neural Network

Algorithms, Computing Architecture, Neurons, Synapses.......  

- Number of Synapses about 10,000 times the number of neurons; need for memory compression  
- Architecture: Memory bottleneck, compute-in-memory can help
Artificial Neural Networks: Simple Model

Artificial NN

Input spikes

axon

time

Signal transmission

weighted summation

Thresholding function

Cross-bar array of programmable synapses

Stochastic Neurons
MTJ Mimics Biological Spiking Neuron

- The leaky fire and integrate can be approximated by an MTJ – the magnetization dynamics mimics the leaky fire and integrate operation.
- Barrier height (Eb ~25KT or higher)
Stochastic Neuron

**LLG Magnetization Dynamics**

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

\(\hat{m}\) : magnetization direction

\(I_s\) : input spin current

\(H_{eff}\) : effective field including thermal noise

\[H_{thermal} = \sqrt{\frac{\alpha}{1 + \alpha^2 \gamma \mu_0 M_s V \delta t} \frac{2k_B T}{G_{0,1}}}\]
Stochastic Neuron: Read/Write Circuits

Design concerns

- Low-Eb: Highly sensitive device -- “Read” current and noise can significantly bias the stochastic switching
- Low-Eb: “read” and “write” currents are active concurrently for the volatile magnet, “read” circuit needs to be highly optimized
Synaptic Behavior: Spike Timing Dependent Plasticity

Strength of the synapse should increase (decrease) as post and pre neurons appear to be temporally correlated (uncorrelated)

\[ w_i^{new} = w_i^{old} + \Delta w(t_i) \times w_{max} \]

\( i = 1, 2, 3 \)

Stochastic STDP in SHE-MTJ Synapse

- Stochastic STDP learning rule specifies the synaptic switching probability based on difference between pre- and post-spike times.
- Write current controls the switching probability of a SHE-MTJ.
- Switch the MTJ based on spike timing by passing the required write current through the heavy metal.
Decoupled Spike Transmission and Learning Current Paths

- When pre-neuron spikes, spike voltage $V_{\text{spike}}$ passes through MTJ and gets modulated by MTJ conductance
- PRE is programming signal that starts when pre-neuron spikes to implement the desired STDP
- POST is a short pulse of duration of few ns that gets activated when post-neuron spikes to implement STDP
Self-learning in Spiking Neural Networks

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.
- 36fJ energy consumption per synaptic event which is ~10-100x lower in comparison to emerging devices like PCM.
Results for Fully-connected SNN Composed of Stochastic Binary (one-bit) Synapses

- Synaptic representations of MNIST digits learned by two-layered fully-connected SNN consisting of 400 excitatory neurons.
S-STDTP: Hebbian Potentiation/ Anti-Hebbian Depression

- Stochastically switch the binary synapse with a constant probability based on spike timing.
  - Approximation of the biologically plausible STDP learning rule for simpler hardware implementation.
  - Distinct potentiation (Hebbian) and depression (Hebbian and Anti-Hebbian) timing windows.
  - Anti-Hebbian depression improves the efficiency of synaptic learning.
Stochastic STDP for Quantized (two-bit) Synapse

- Quantized synapse uses the entire timing window for learning.
- Synaptic weight and switching probability depend on spike timing.
- Incorporates Hebbian and Anti-Hebbian learning mechanisms.

\[ \Delta t = t_{\text{post}} - t_{\text{pre}} \]

<table>
<thead>
<tr>
<th>Synaptic logic levels</th>
<th>Synaptic weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘11’</td>
<td>1</td>
</tr>
<tr>
<td>‘10’</td>
<td>2/3</td>
</tr>
<tr>
<td>‘01’</td>
<td>1/3</td>
</tr>
<tr>
<td>‘00’</td>
<td>0</td>
</tr>
</tbody>
</table>
SNN with two-bit synapses provides 3.1% average improvement in the classification accuracy over one-bit synapses up to a network size of 800 excitatory neurons.

Both configurations offer comparable classification accuracy for larger networks (greater than 1600 excitatory neurons).
StOCNet: **Stochastic One-Bit Convolutional Spiking Neural Network**

- Deep StOCNet populated with biologically inspired Leaky Integrate & Fire (LIF) neurons.
- Convolutional layers are trained with (layer-wise) unsupervised stochastic STDP that enables weight kernels to extract characteristic input features.

- Shallow 36C5-2S-10FC SNN offers **95.46%** accuracy on the MNIST test set.
- Deeper 16C5-36C5-2S-500FC-10FC SNN yields **95.57%** accuracy.
All Spin Binary Stochastic Neural Network (weights learnt through backpropagation)

Train a Deep ANN with Binary Weights and Sigmoid Activation

Network Topology

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONV: 5x5 64 filters</td>
<td></td>
</tr>
<tr>
<td>Average Pooling [2x2]</td>
<td></td>
</tr>
<tr>
<td>CONV: 5x5 128 filters</td>
<td></td>
</tr>
<tr>
<td>Average Pooling [2x2]</td>
<td></td>
</tr>
<tr>
<td>CONV: 3x3 512 filters</td>
<td></td>
</tr>
<tr>
<td>Average Pooling [2x2]</td>
<td></td>
</tr>
<tr>
<td>Fully Connected 512/1024</td>
<td></td>
</tr>
<tr>
<td>Output layer 10/100</td>
<td></td>
</tr>
</tbody>
</table>

All Spin Binary Stochastic Neural Network
LOWER-BARRIER MAGNETS
Low Barrier MTJs (<5k_B T)

- As the magnets are further scaled, they will become thermally unstable – retention time is small
- Display random telegraphic switching at zero write current
- Average magnetization of the free layer can be modulated through the write current in the SHE layer

For a 1k_B T MTJ:
Synchronous Architecture (High $E_B$)

Input from previous layer

Digital Buffer

Memristors

1 time period
2ns

WRITE
READ
RESET
Asynchronous Networks

Input from previous layer

VS+

VS+

VS−

VS−

0
Resiliency to variations in the Supply Voltages

- Due to the higher sensitivity of telegraphic MTJs, asynchronous networks are sensitive to variations in the supply voltages.

Average classification accuracy (for 50 Monte Carlo simulations) with variations in the supply voltage up to 25mv.
Dimension Variations

- Even small dimension variations can shift $E_B$

Average classification accuracy (for 50 Monte Carlo simulations) with variations in the EB up to $0.4k_B T$
Accuracy and Speed

• The network was tested with 10,000 images from the MNIST data set
• Synchronous networks are faster than the asynchronous networks
  – Synchronous 96% accuracy reached under (20ns)
  – Asynchronous 96% accuracy reach at 80ns and 250ns (for 1k\textsubscript{B}T & 2 k\textsubscript{B}T)
• Maximum Accuracies
  – 98.1% and 97.6% for 10 k\textsubscript{B}T and 20 k\textsubscript{B}T networks
  – 97.5% and 97.2% for 1 k\textsubscript{B}T and 2 k\textsubscript{B}T networks
STOCHASTIC BITS FOR COMBINATORIAL OPTIMIZATION
Stochastic Computing: From Devices to Circuits and Systems

Ising Computing Model (Restricted Boltzman Machine): Combinatorial Optimization (fonts)

- **Ising computing model:**
  - Efficient computing model based on behavior of magnetic spins and coupling between them

1) **Ising spin model**

2) **System energy (H) vs spin state**

- **Solution of the problem**
  - Spin state at lowest system energy (H)

- **Main Issues:**
  1. Coupling with neighbors → ‘Majority vote’
  2. Sub-optimal solutions → overcome w/ ‘Annealing’

![Diagram of Ising spin model and system energy vs spin state]
Majority vote function with stochastic MTJ

- Efficient majority vote function can be implemented with 1) decoupled write path & 2) dependency of $P_{SW}$ on $I_q$

- Single spin model with peripherals for majority vote function

- Corresponding $P_{SW}$ vs $I_q$

- Coupling with neighbors controls amount of charge current through current sources and switches

- High(Low) $P_{SW}$ with many(less) voters ➔ Mimics majority vote functionality
Application: Graph Coloring

- Graph coloring: **Coloring n-vertices with k-colors** such that no two adjacent vertices have the same color

**Details of the problem:**
- Total number of spin needed: $n \times k$ (i.e. 4 vertices & 3 color $\rightarrow$ 12 spins)
- Interconnect map can be generated from ‘penalty Hamiltonian’\(^1\)

1) A. Lucas, vol. 2, article 5, frontiers in physics, 2014

- Problem setting & solutions from proposed Ising solver

<table>
<thead>
<tr>
<th>3 color problems</th>
<th>4 Vertices / 4 Edges</th>
<th>5 Vertices / 6 Edges</th>
<th>6 Vertices / 12 Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connectivity</td>
<td><img src="image" alt="4 Vertices / 4 Edges" /></td>
<td><img src="image" alt="5 Vertices / 6 Edges" /></td>
<td><img src="image" alt="6 Vertices / 12 Edges" /></td>
</tr>
<tr>
<td># of spins</td>
<td>12</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Avg. Iterations</td>
<td>115</td>
<td>435</td>
<td>4,002</td>
</tr>
<tr>
<td>Solutions from Ising solver</td>
<td><strong>12 types</strong></td>
<td><strong>12 types</strong></td>
<td><strong>6 types</strong></td>
</tr>
</tbody>
</table>
Simulation Framework

Spin-based Neuromorphic Architectures
Goal: Large-scale spin-based neuromorphic architectures

Simulation Framework
Neuromorphic Architecture Simulator
Neuron/Synapse Behavioral Models

Spintronic Neural Circuits
Goal: Ultra-low-power spintronic circuits for neural networks

SPICE models (spin-neurons)
Parameter variations

Spin Devices as Neurons and Synapses
Goal: Spin devices with neural and synaptic characteristics

Spin-diffusion transport
NEGF-LLG

Convolutional Networks
Spiking Networks
Conclusions

- Other than memory, STT devices also show promise for a class of computing models such as “brain-inspired computing”, stochastic computing, combinatorial optimization,..

- Exploit intrinsic stochasticity of STT devices (as natural annealer) in the presence of thermal noise to realize neuronal and synaptic dynamics, stochastic logic, and computing models for combinatorial optimization.

- STT-devices as “neurons” and “synapses” for both ANN, SNN show the possibility of large improvement in energy (and synaptic memory compression) compared to standard implementation.

- Computing paradigms like CiM can be efficient for vector operations and neuromorphic computing.