High-Performance Stochastic Memristive Networks for Neurocomputing and Neurooptimization

Dmitri Strukov
UC Santa Barbara

PRiME
October 2020 (virtual)
Vector-by-matrix multiplication (dot-products with the same input vector) is the most common operation
Noise in Biological and Artificial Neural Networks

Molecular-level operations in the brain, e.g. neurotransmitter release in synaptic clefts and voltage gating of ion channels, are stochastic.

Example: fluctuations in K channel

--- channel closed --------- channel open -----

fluctuations due to thermal noise

10 ms

voltage applied

Stochastic (binary) neuron

$$x_i = \sum_{j=1}^{N} w_{ij} y_j$$

$$p(y=1) = \frac{1}{1 + e^{-x/T}}$$

need efficient hw for dot-product and stochastic transfer function

Stochastic neural networks:
- (Restricted) Boltzmann machines
- Stochastic Hopfield networks
- Deep believe networks
- Bayesian networks
- ...
Focus of This Talk

Stochastic vector-by-matrix multiplication is the most common operation

\[
p(x_i = 1) = \frac{1}{1 - e^{-\sum_{j} w_{ij} y_j}}
\]
Radical Improvement with Analog Computing

**Vector-by-Matrix-Multiplication (VMM):**

basic neuromorphic operation…

![VMM Diagram](image)

\[ x_i = \sum_{j=1}^{N} w_{ij} y_j \]

**Analog VMM:**

…using the Ohm & Kirchhoff laws

![Analog VMM Diagram](image)

\[ I_\Sigma = \sum_{j=1}^{N} G_j V_j \]

\[ U_\Sigma \approx 0 \]

**Features:**

- physical-level (very compact) and in-memory computation → fast and very energy-efficient
- proposed by Widrow in 1960s, popularized by Mead and his students (CalTech) in the 1980s
- no dense adjustable-conductance crosspoint devices - until recently
Tunable Non-Volatile Memory Device Options

- STTRAM
- PCRAM
- 2D FeRAM
- CBRAM
- ReRAM
- 2D NOR
- 3D NAND
- CBRAM
- 2D RRAM
- 3D FeRAM?
- 3D RRAM

- Active “1T1R”
- Passive “0T1R”

Most important specs: Density, retention, analog switching!

Cell density

few 100's $F^2$ for current, potentially down to 25 $F^2$

$F = \text{feature size}$

40 $F^2$ now, <20 $F^2$ with FinFET

Active “1T”

1-10 TB/in²

would allow to fit extra-large models on chip!
Long-Term Option: (3D) Passive Metal-Oxide Memristors

- 64 × 64 passive crossbar circuit


Typical I-V characteristics

Details:
- Al₂O₃/TiO₂ₓ active bilayer by reactive sputtering
- CMOS-compatible CMP/dry etching process and TiN/Al electrodes for higher conductance
- ~250 nm wide lines, passive (0T1R) integration (e.g. >250x/10,000x better memristor / memory cell density compared to 1T1R work at comparable complexity and yield
- The largest functional analog-grade passive memristor crossbar circuit supported by proper statistics

H. Kim et al. arXiv 2019
Most Important Metric: Yield and Switching Threshold Variations in 64×64 Xbar

- Raw data (voltage ramp) …

- Switching threshold spatial map:
  - set
  - reset

- ... and processed statistics

- Switching threshold is defined as voltage at which current changes by > 10% when applying voltage ramp
- Dark blue dots: ~1% devices that cannot be switched

H. Kim et al. 2019 arXiv
Conductance Tuning in 64×64 Memristor Crossbar

- Desired pattern
- Actual pattern

- Color encoding: 256 levels from white (10 µS) to black (100 µS) @ 0.2V
- < 5% / < 3% absolute / relative tuning error using automated algorithm, with reserves for improvement
Near-Term Option: Floating-Gate Devices

- **NOR eFlash device & chip**

- **Vector-by-Matrix Multiplier Circuit**

- **2-layer MLP classification results**

  (10,000 MNIST test patterns)

**Summary:**
- 28x28 B/W input, 10-class output, >100,000 NOR flash synapses, 64 hidden layer CMOS neurons, 180-nm process with eFlash
- 94.65% experimental fidelity (96.5% theoretical)
- < 1-µs latency, < 20 nJ energy per pattern (reserves for improvement for both with better neuron design)
- Much better in speed and energy efficiency over digital circuits at comparable MNIST fidelity (10^6 better energy-delay than IBM TrueNorth)
- Reproducible, temperature insensitive, no change in performance after 7 months shelf-time, without any cell retuning
- More recent work using 55-nm ESF3 NOR-flash technology (CICC’17, IEDM’18’19), scalable to 28 nm
New Result #1: Stochastic Analog Vector-by-Matrix Multiplier

**Basic Idea:**
add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator

$$I_{\Sigma} = I_{n,ext} + \sum_{i=1}^{N} G_i V_i + I_{n,i}$$

$$V = \begin{cases} V_{ON}, & I_{\Sigma} \geq 0 \\ 0, & I_{\Sigma} < 0 \end{cases}$$

**Two Implementation Options:**

0T1R memristor cell (works for 1T1R as well)

Floating gate transistor
New Result #1: Stochastic Analog Vector-by-Matrix Multiplier

**Basic Idea:**
add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator

$$V = \begin{cases} V_{ON} , I_\Sigma \geq 0 \\ 0 , I_\Sigma < 0 \end{cases}$$

**Experimental Demo:**
using 20×20 passive array with externally-injected noise from readout circuitry

**Features:**
- Sigmoid slope (i.e. SNR or compute temperature $T$) controlled dynamically by the applied voltage $V_{ON}$
- Some smearing of output probabilities due to input-dependent noise and device imperfections

M.R. Mahmoodi et al. Nature Communications, 2019
New Result #1: Stochastic Analog Vector-by-Matrix Multiplier

**Basic Idea:**
add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator

$$V = \begin{cases} V_{ON}, & I_{\Sigma} \geq 0 \\ 0, & I_{\Sigma} < 0 \end{cases}$$

$$I_{\Sigma} = I_{n,ext} + \sum_{i=1}^{N} G_i V_i + I_{n,i}$$

**Experimental Demo:**
using 180nm embedded ESF1 NOR-flash memory technology

**Features:**
- Sigmoid slope (i.e. SNR or compute temperature $T$) controlled dynamically by the applied gate voltage

M.R. Mahmoodi et al. Nature Communications, 2019
New Result #2: Restricted Boltzmann Machine Demo

- 10-input 8-hidden neuron RBM network

![Diagram of RBM network]

- Experiment (solid) vs. simulation (dash-dot)

![Plot with energy and counts]

Details:
- Hardware injected noise with software-emulated neuron functionality
- Random weights (from -32 µS to +32 µS) mapped to 10×16 portion of memristor xbar
- Neuron input currents sampled at 1 MHz bandwidth after applying Random Input → Visible → Hidden → Visible → Hidden → …

M.R. Mahmoodi et al. Nature Communications, 2019
Solving Optimization Problems with Hopfield Neural Network

- **Combinatorial optimization problems**

<table>
<thead>
<tr>
<th>Application</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistics / package delivery</td>
<td>Traveling salesman</td>
</tr>
<tr>
<td>Power grid</td>
<td>Maximum flow</td>
</tr>
<tr>
<td>Design automation</td>
<td>Vertex cover</td>
</tr>
<tr>
<td>Molecular dynamic simulations</td>
<td>Graph partitioning</td>
</tr>
</tbody>
</table>

- **Solving TSP with Hopfield neural network**

Traveling Salesman Problem: NP hard → use heuristics, e.g.
- single route = specific neuron outputs
- finding optimal solution = minimizing “energy” function of neuron outputs
- dynamics of the recurrent network with proper weights minimizes energy function over time

- **Example of continuous time / binary neuron Hopfield network**

∑ = sum amp & comparator
Earlier Work: (Deterministic) Hopfield Network Experimental Demonstration with Discrete Memristors

- Hopfield network for A-to-D conversion
  - Implemented with memristors
  - Experimental results

- Major features:
  - 4-bit ADC implemented as a Hopfield network
  - The first demo for the memristor-based Hopfield neural network
  - CMOS discrete IC neurons
  - Discrete packaged memristors
  - Fine-tuning to cope with offsets and variations

- Chips with single-device memristors
- TL074CN opamp

---

- Analog Input Voltage (V)
  - V0
  - V1
  - V2
  - V3

- Analog output Vr (V)
- Digital output (V)

- Digital error

---

- X. Guo et al., Frontiers in Neuroscience 9, art. 488, Dec. 2015

---

- 16
Local Minima in Hopfield Network

Local minima present problems!

Color background:
- Baseline Hopfield neural network

$\sum = $ sum amp & comparator
Simulated Annealing with Generalized Hopfield Network (Boltzmann Machine)

Local minima present problems!

**Solution:** employ probabilistic neurons (stochastic VMMs) to implement simulated annealing

**Color background:**
- Baseline Hopfield neural network
- Stochastic annealing

\[ \Sigma = \text{sum amp & comparator} \]
\[ \times = \text{scaling} \]
Emerging (Custom) Hardware for Combinatorial Optimization

Nanomagnets / P-bits

- Exponentially measured ground states for the network consisting of up to 3 coupled magnetic devices with fixed coupling;
  - Limited (near neighbor, fixed) coupling and/or …
  - Integer (up to 945) factorization with 8 p-bits
  - … high CMOS overhead

Borders et al. IEDM 2016

CMOS

- Experimental results for solving maximum-cut problem with 2x30K-spin Ising network 40-nm 23.65-mm² SRAM-based chips
  - Not in-memory (bulky, slower, power hungry)
  - Binary weights

Takemoto, et al. ISSCC 2018

Josephson Junction

- Experimentally measured ground state of random spin glass problems based on 108-qubit D-Wave One system (with evidence of quantum annealing)
  - Low temperature operation
  - Many issues unsolved


Photonics

- Experimental results for solving max-cut problems with up to 2,000 nodes with Ising network based on degenerate optical parametric oscillators
  - Slow due to high overhead of the electronic feedback used for updating spatial light modulator

Adjustable Energy Function Annealing

\[ \text{Energy} = E_{\text{original}} + \exp(-\text{time})E_{\text{addon}} \]

Another solution inspired by quantum annealers: Dynamically adjustable energy function

Color background:
- Baseline Hopfield neural network
- Adjustable energy function / weight annealing

\[ \sum = \text{sum amp} \& \text{comparator} \]
\[ \times = \text{scaling} \]

M.R. Mahmoodi et al. Nature Communications, 2019
Yet Another Approach: Chaotic Annealing

Color background:
- Baseline Hopfield neural network
- Chaotic annealing

$\sum = \text{sum amp & comparator}$
$x = \text{scaling}$

M.R. Mahmoodi et al. Nature Communications, 2019
New Result #3: Flexible-Annealing Mixed-Signal Generalized Hopfield Networks for Combinatorial Optimization

Color background:
- Baseline Hopfield neural network
- Stochastic annealing
- Adjustable energy function / weight annealing
- Chaotic annealing

\[ \sum = \text{sum amp & comparator} \]
\[ \times = \text{scaling} \]
New Result #3: Combinatorial Optimization Demo with FG

- Weighted graph partitioning problem...
  (finding two mutually exclusive, set of nodes with maximally balanced node weights and minimized edge weights between two sets)

... and experimental results using 10×20 180-nm NOR flash memory array

---

M.R. Mahmoodi et al. Nature Comm, 2019
New Result #3: Combinatorial Optimization Demo with Passive $64 \times 64$ Metal-Oxide Memristive Crosbar Circuits

- 5-node maximum-weighted clique problem
- 12-node maximum-weighted vertex cover problem

M.R. Mahmoodi et al., Proc. IEDM'19
New Result #3: Combinatorial Optimization Demo with Passive 64×64 Metal-Oxide Memristive Crosbar Circuits

- 10-node maximum-weight independent set problem
- 6-node maximum-weight graph partitioning problem

M.R. Mahmoodi et al., Proc. IEDM’19
Summary

- In-memory analog computing based on emerging analog grade memory devices to enable very energy-efficient, compact, and fast analog VMMs
  - Near term: Metal oxide memristors (the most dense though least mature)
  - Long term: Embedded NOR floating gate memories (available at foundries now)

- Intrinsic noise of memory devices to implement stochastic transfer function or stochastic vector-by-matrix multiplication

- Experimental demonstration of Boltzmann machines based on small-scale stochastic VMMs circuits with applications in deep belief networks and combinatorial optimization

- Major memristor challenges: poor yield, device uniformity, high cell currents

<table>
<thead>
<tr>
<th>Performance estimates &amp; comparison to competition*</th>
<th>Conventional</th>
<th>Emerging technology</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CPU</td>
<td>GPU</td>
<td>D-Wave</td>
</tr>
<tr>
<td>Time to solution (µs)</td>
<td>220</td>
<td>10</td>
<td>10^{10}</td>
</tr>
<tr>
<td>Energy to solution (µJ)</td>
<td>4000</td>
<td>2500</td>
<td>250×10^{12}</td>
</tr>
</tbody>
</table>

* benchmarked on noisy mean-field algorithm, adapted from ArXiv:1903.11194
Relevant References

- **Stochastic neurocomputing and neuro-optimization demos**
  - M.R. Mahmoodi et al., "An analog neuro-optimizer with adaptable annealing based on 64×64 0T1R crossbar circuit", *Proc. IEDM’19*
  - M.R. Mahmoodi et al., "Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization", *Nature Communications* 10, art. 5113, 2019

- **Passive metal-oxide memristors**
  - H. Kim et al. arXiv 2019
  - F. Merrikh Bayat et al., "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits", *Nature Communications* 9, art. 2331, 2018
  - G.C. Adam et al., “3-D memristor crossbars for analog and neuromorphic computing applications”, *IEEE TED* 64 (1), pp. 312-318, 2017
  - M. Prezioso et al., "Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al2O3/TiO2-x/Pt memristors", *Proc. IEDM’15*, pp. 17.4.1 – 17.4.4

- **NOR flash VMM-level experimental demos**
  - X. Guo et al. “Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology”, *Proc. IEDM’17*, pp. 6.5.1-6.5.4
Questions?!
strukov@ece.ucsb.edu

Paper co-authors:
UC Santa Barbara: Zahra Fahimi, Hyungjin Kim, Hussein Nili, M. Reza Mahmoodi
Linkoping U., Norrkoping, Sweden: Leo Sedov and Val Polishchuk


Sponsors (past and present):