Experimental demonstration of software-trained neural network inferencing in analog memristor crossbar arrays

Miao Hu, Qiangfei Xia*, J. Joshua Yang*, R. Stanley Williams, and John Paul Strachan

Hewlett Packard Labs, Hewlett Packard Enterprise, Palo Alto CA
*UMass Amherst
Parallel multiply & add through Kirchoff and Ohm’s law


Memristors as highly scalable, tunable analog resistors

High ON/OFF ratio (~10^{5}), supporting multiple levels

→ HPE differentiator vs competing accelerator designs

Advantages:

• Well suited for streaming workloads; Key advantage is in-memory processing; Many ways to scale up

• Many Teams have been working in this field:

  IBM, GeorgiaTech (Hasler), U Michigan (W.Lu), ASU (S. Yu), Duke (H.Li), and many others

But what actually does a crossbar do?
Dot-product Engine demonstrator

- Flexible peripheral circuit platform to study the behavior of actual memristor crossbars for in-memory computing.
Programming full memristor arrays

64x64 = 4096 memristors (TaOₓ)
~6 bits at each memristor (full range of accessible resistance)

Histogram of error around zero
MNIST Pattern recognition demonstration

Neural network
1 layer softmax
Neural network

Partition and program (100 \, \mu\text{s} to 700 \, \mu\text{s})

Programming error distribution

Pattern vector

Weight matrix

10 values, entry with maximum value is the prediction

Weight matrix

Target Conductance (\mu\text{s})

Columns

Programmed Conductance (\mu\text{s})

Occurrence

Programming Error in Device Conductance (\mu\text{s})

-500 0 500 1000 1500 2000 2500 3000

0 50 100 150 200 250 300 350 400 450

-500 0 50 100 150 200 250 300 350 400 450 500

-100 -80 -60 -40 -20 0 20 40 60 80 100 200

Hewlett Packard Enterprise
Computing accuracy of a 64x64 crossbar

- Crossbar parameters:
  - Wire per segment ≈ 1 ohm
  - Input/output resistance ≈ 1 ohm
  - Device resistance: 1.4k to 10k ohm (100 uS to 700 uS)

- Computing accuracy
  - 150k (2.5k * 60) data points.
  - Memristor is <4 bit for the given range
  - Output accuracy is ~4 bit.
  - Noise is nonlinear due to circuit parasitics.
Pattern “7” recognition

G\_IDEAL current = Vin \times G\_IDEAL; G\_PROG current = Vin \times G\_PROG

Output current per column

Simulation matches experiments

Significant Degradation Due to circuit parasitic

Pattern “7” current

G\_IDEAL current

G\_PROG current

Part 1 + Part 2 + Part 3 + Part 4 + Part 5 + Part 6 = Current
MNIST pattern recognition accuracy

- Using a software-trained weight matrix, a single 64x64 crossbar achieves 85% accuracy (90% is ideal) for MNIST with post processing
- Single-layer NN highly sensitive to even a few defects
- Next steps:
  - Better matrix to conductance mapping:
    - Implement the “conversion algorithm” taking non-idealities into account
  - Use Multi-layer NNs more resilient to defects: