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

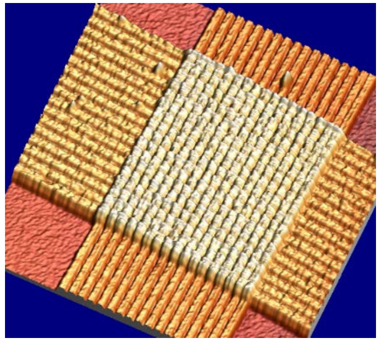
Miao Hu, Qiangfei Xia*, J. Joshua Yang*, R. Stanley Williams, and John Paul Strachan

Hewlett Packard Labs, Hewlett Packard Enterprise, Palo Alto CA
*UMass Amherst

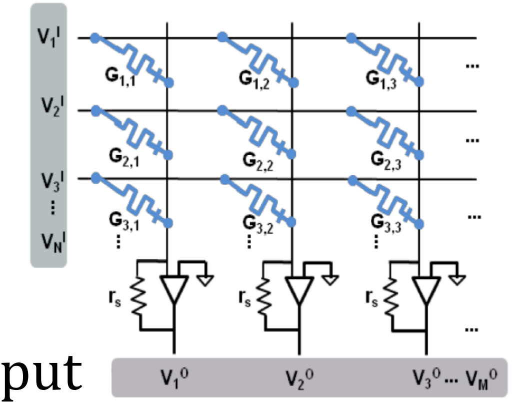
Hewlett Packard
Enterprise

Office of the Director of National Intelligence
IARPA
BE THE FUTURE

DPE: Memristor arrays for computing



Input Voltage vector



Output current

$$I_j^O = \sum_j G_{ij} \cdot V_i^I$$

Is this true in the real world?

- **Parallel multiply & add through Kirchoff and Ohm's law**

  1961, K. Steinbuch "*Die Lernmatrix*"– suggests using "ferromagnetic toroids"

- **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

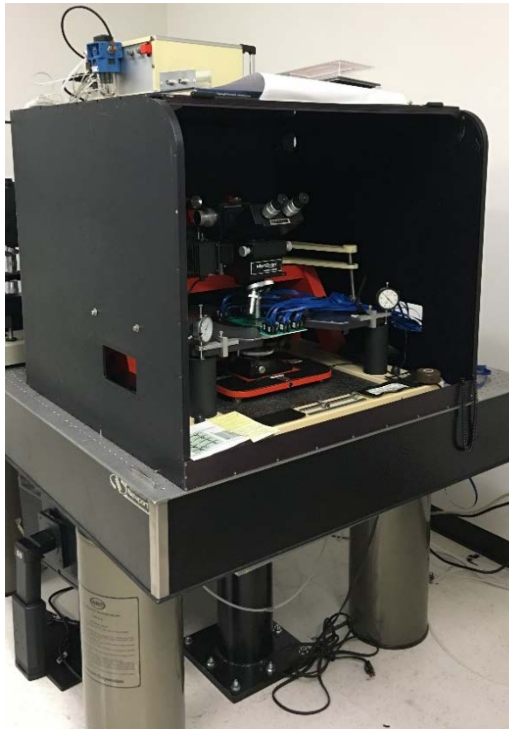
- **Many Teams have been working in this field:**

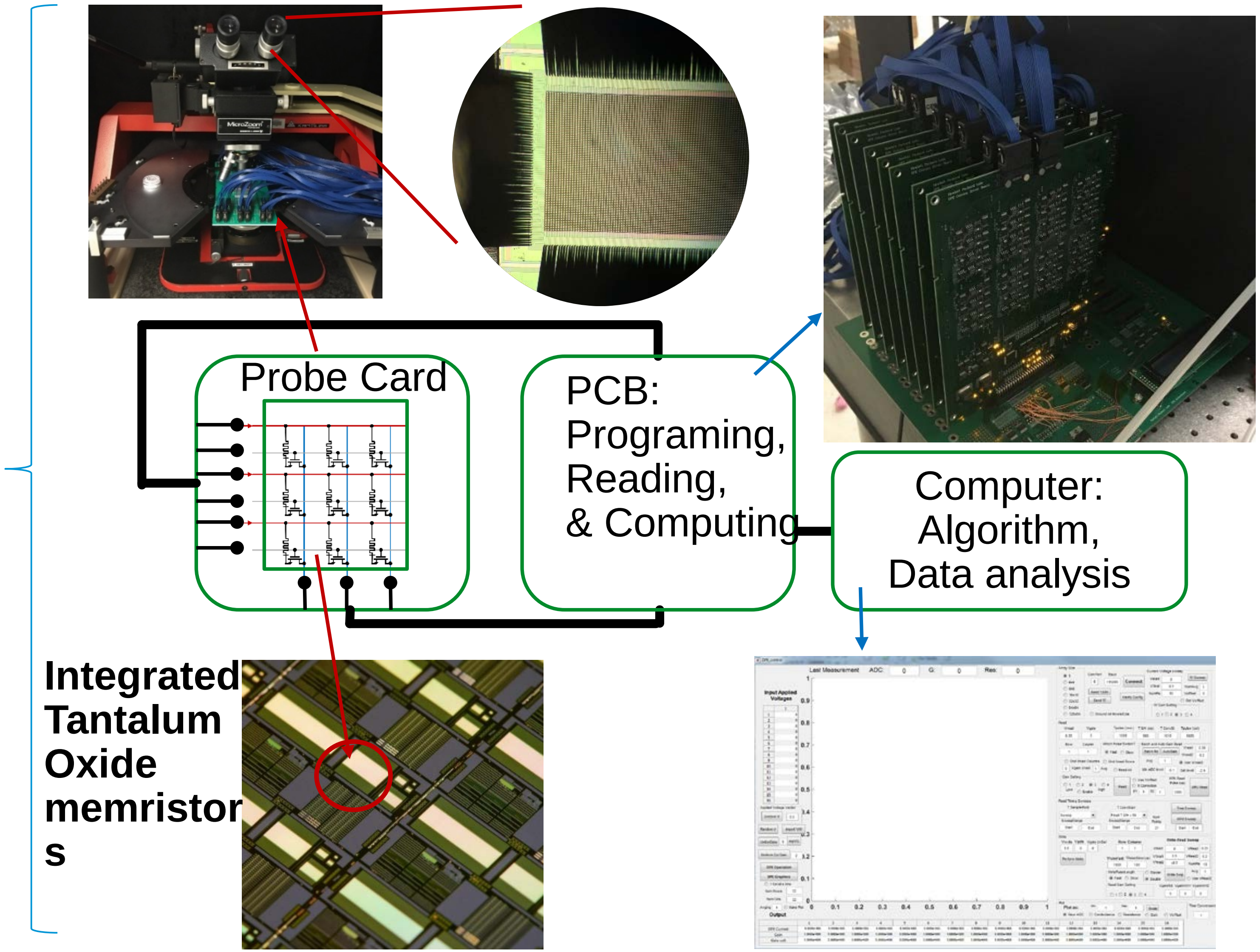
IBM, GeorgiaTech (Hasler), U Michigan (W.Lu), ASU (S. Yu), Duke (H.Li), and many others

- **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.





Probe Card

PCB: Programing, Reading, & Computing

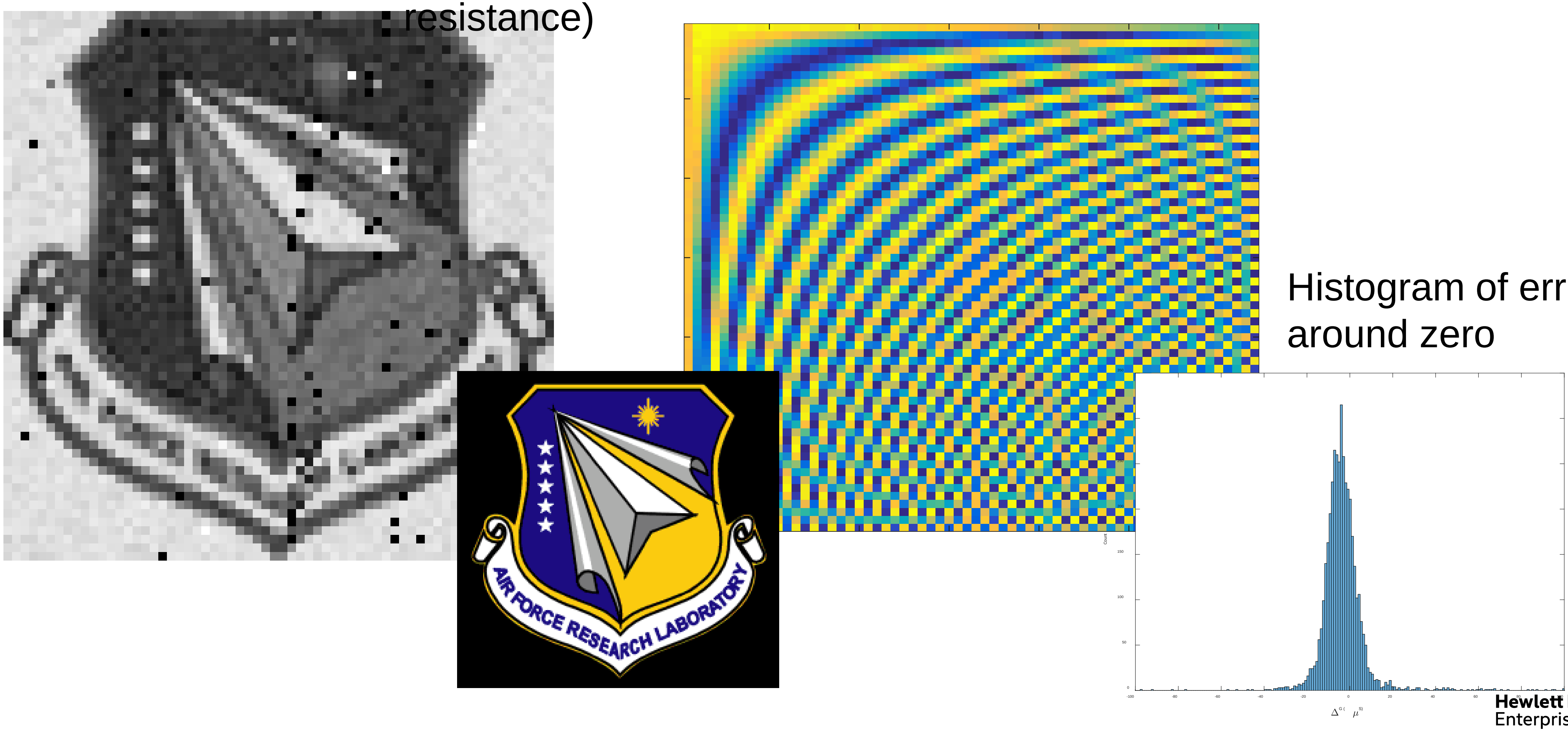
Computer: Algorithm, Data analysis

Integrated Tantalum Oxide memristors

Programming full memristor arrays

64x64 = 4096 memristors (TaO_x)

~6 bits at each memristor (full range of accessible resistance)

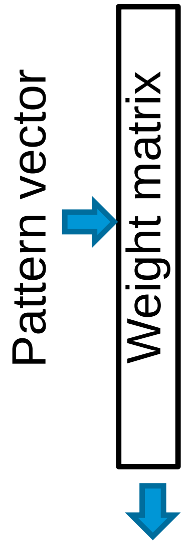


Histogram of error around zero

MNIST Pattern recognition demonstration

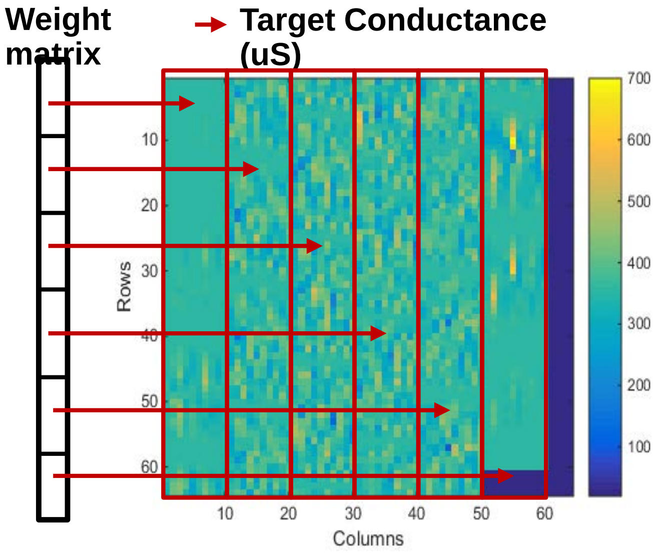
Neural network

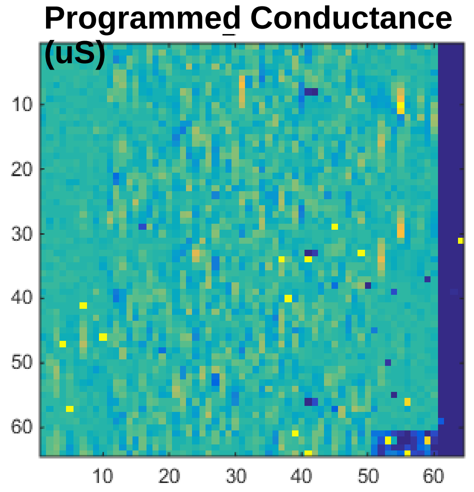
1 layer softmax
Neural network



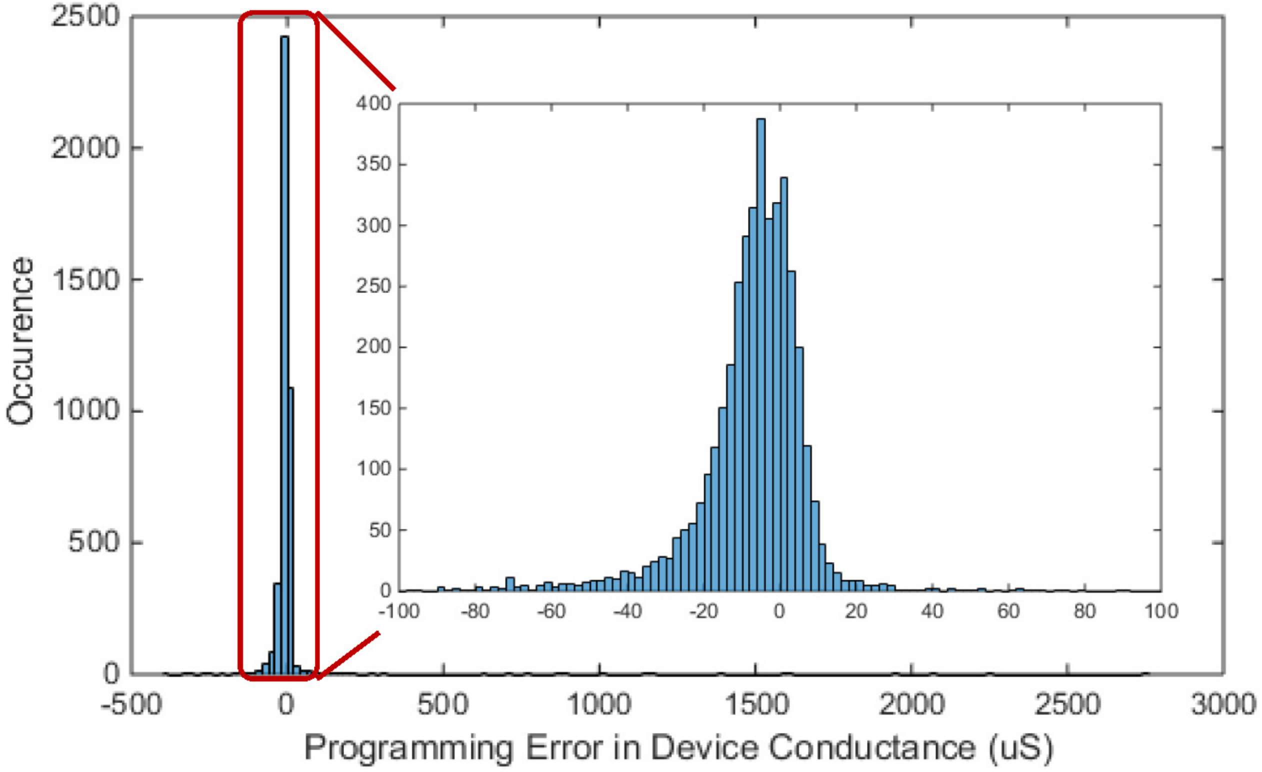
10 values, entry with maximum value is the prediction

Partition and program (100 uS to 700 uS)



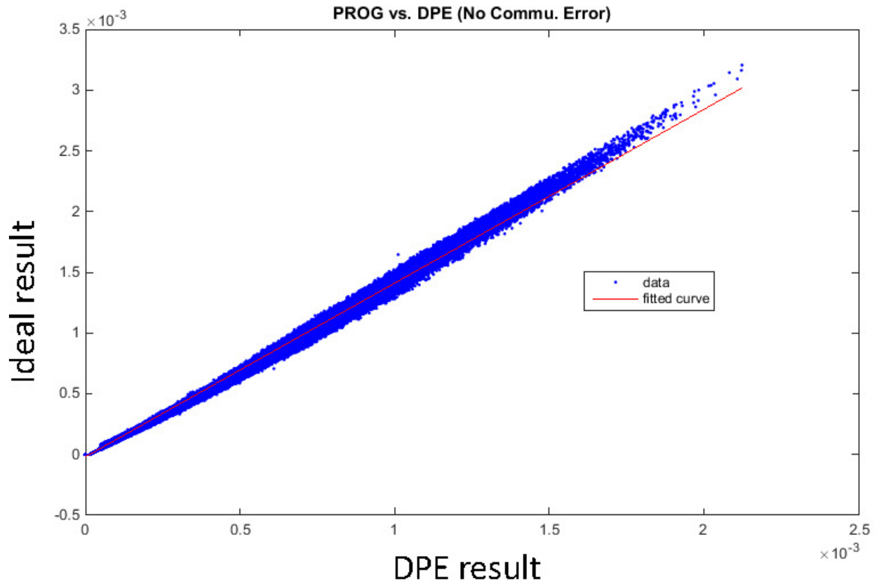


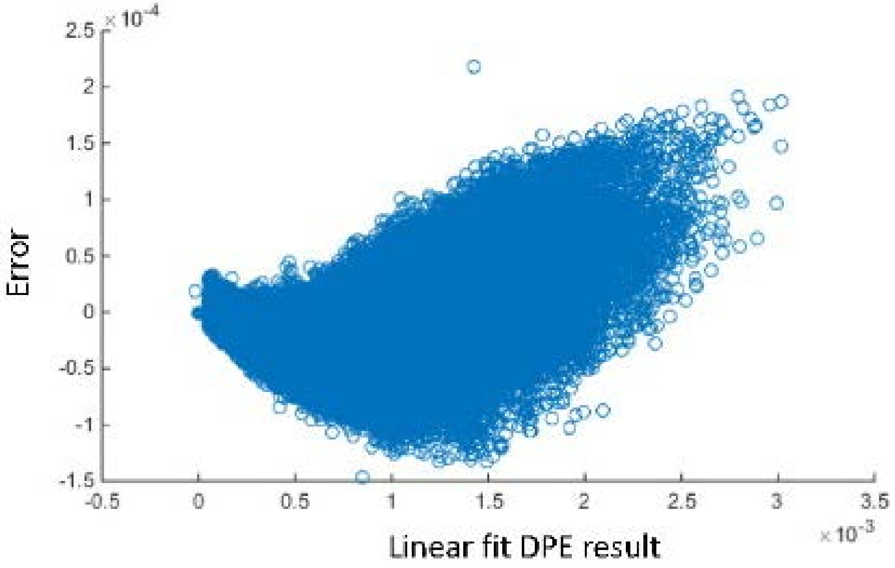
Programming error distribution



Computing accuracy of a 64x64 crossbar

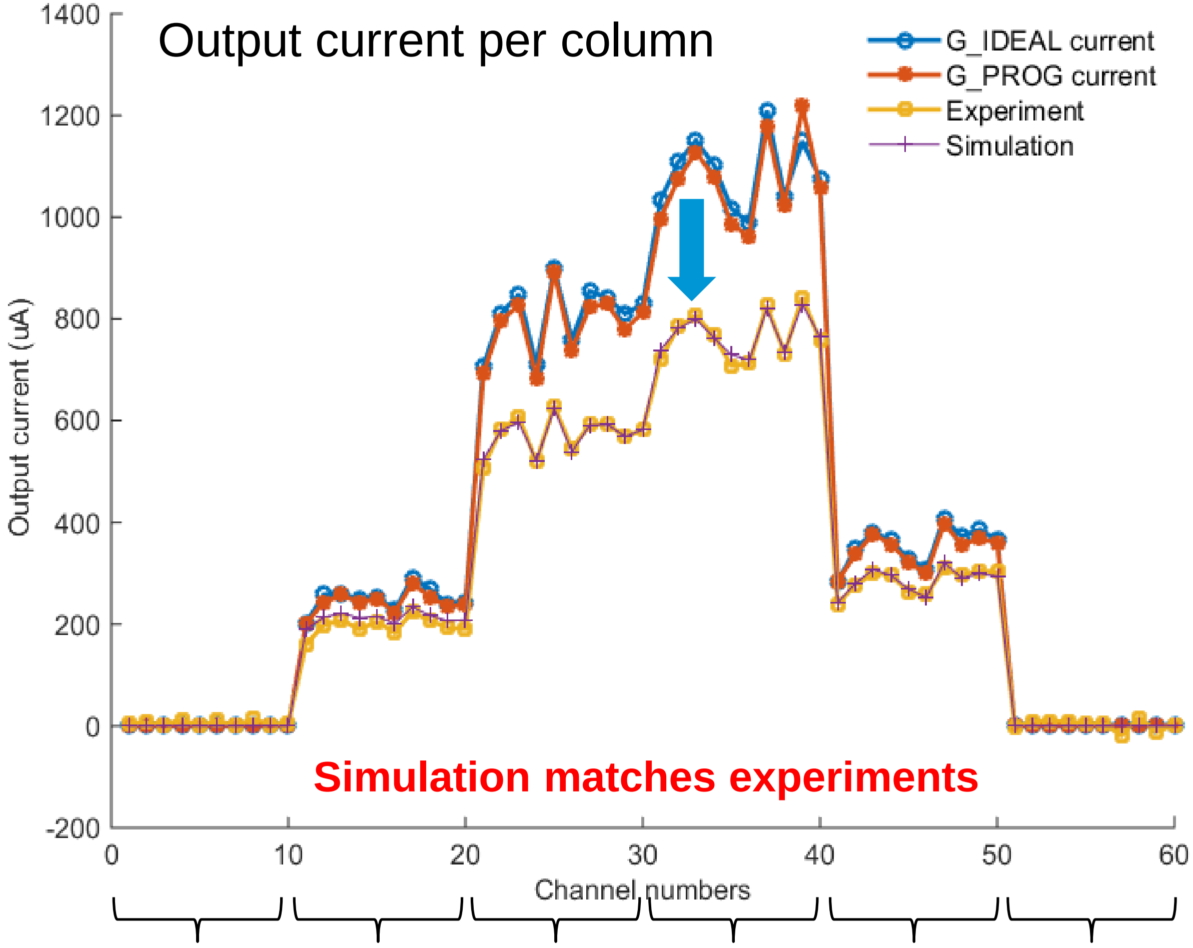
- **Crossbar parameters:**
- Wire per segment ≈ **1 ohm**
- Input/output resistance ≈ **1ohm**
- Device resistance: **1.4k to 10k ohm (100 uS to 700 uS)**
- **Computing accuracy**
- 150k (2.5k * 60) data points.
- Memristor is <4 bit for the given range
- Output accuracy is ~4 bit.
- **Noise is nonlinear** due to circuit parasitics.





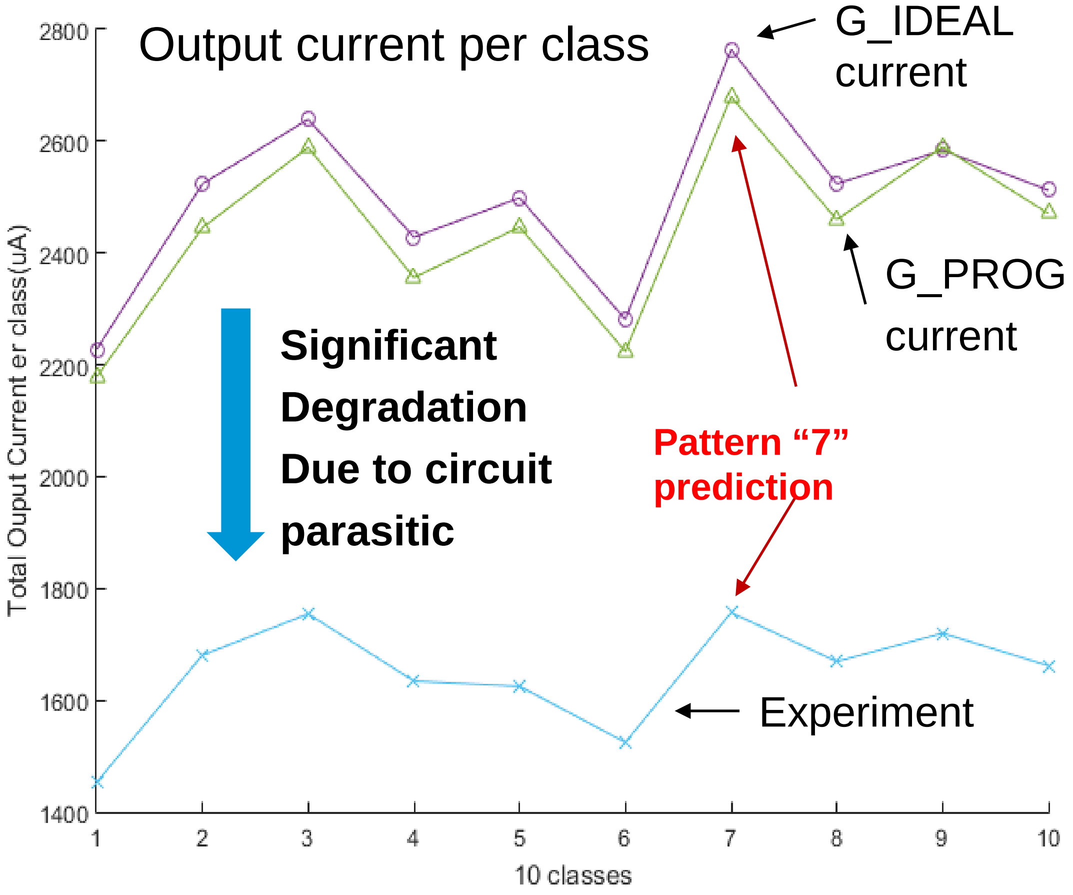
Pattern “7” recognition

G_IDEAL current = Vin * G_IDEAL; G_PROG current = Vin * G_PROG



Output current per column

Simulation matches experiments



Output current per class

G_IDEAL current

G_PROG current

Significant Degradation Due to circuit parasitic

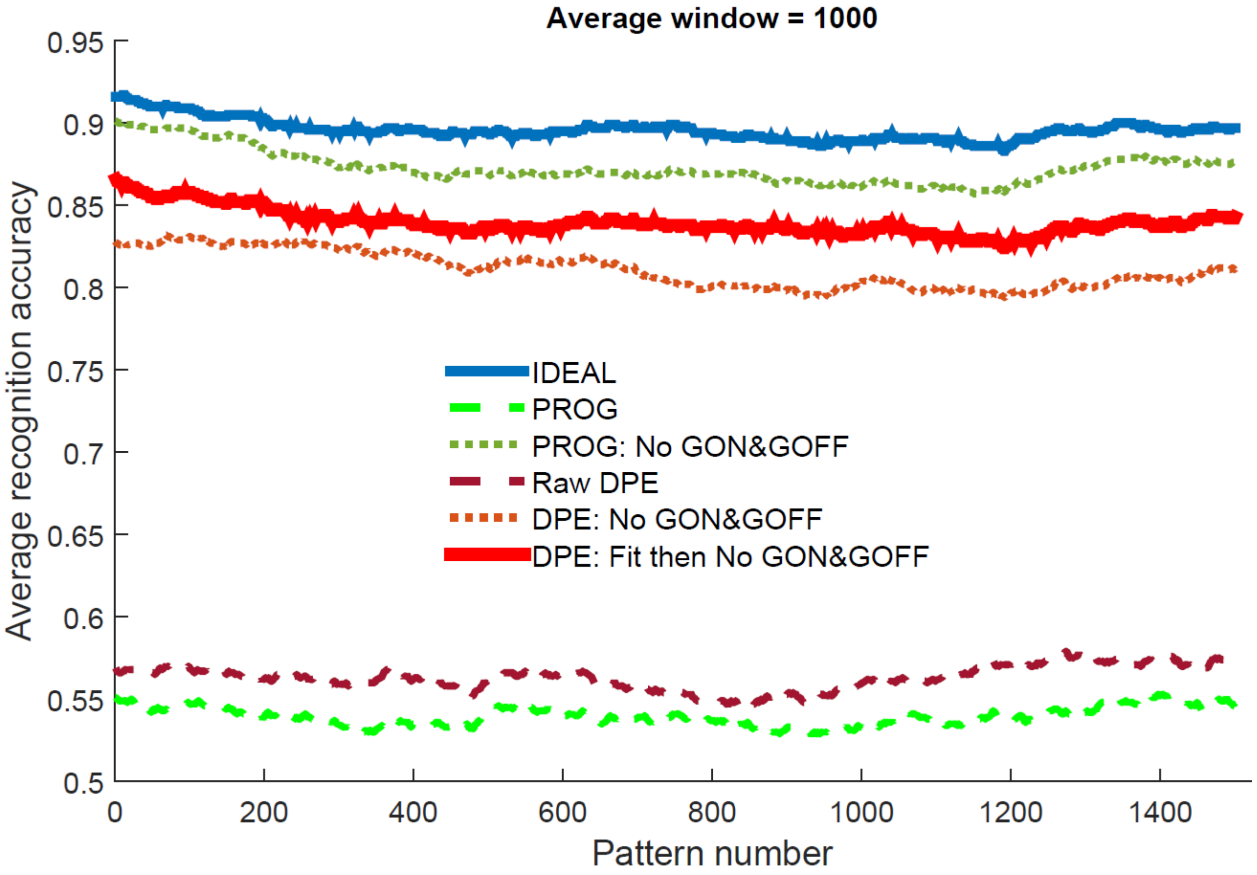
Pattern “7” prediction

Experiment

Part1 + Part2 + Part3 + Part4 + Part5 + Part6 = Current per class

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:



Hewlett Packard Enterprise