SNNs with analog neurons and RRAM synapses

A. Valentian, F. Rummens, E. Vianello, T. Mesquida, O. Bichler, C. Reita, F. Heitzmann

alexandre.valentian@cea.fr, CEATECH, Grenoble/Paris, France

Summary

Data-centric workloads exhibited by Deep Neural Network (DNN) applications call for circuit architectures where data movement is reduced to a minimum.

This has motivated architectures in which memories are spatially located near the computing elements. These memories must be very dense, preferably non-volatile and inserted into the computational dataflow, therefore RRAMs are excellent candidates to this purpose.

Spiking Neural Networks (SNN), also called 3rd generation of NNs, are promising to further reduce the computational power. Spikes are unary, discrete events, that take place at point in time, rather than continuous values.

Combining analog spiking neurons with RRAM synapses enables a natural in-memory computing, without the need for ADCs or DACs.

Motivation

Edge computing devices, although making progress compared to cloud-based solutions, are still far from the energy efficiency of biology.

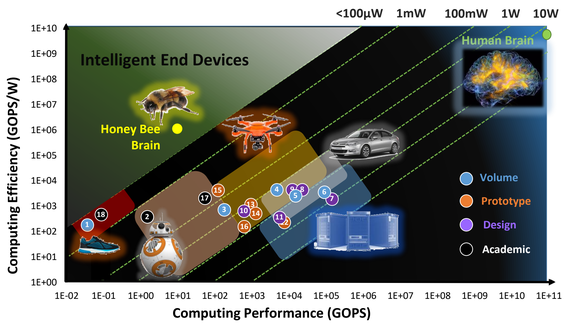


Figure 1: Comparison of the energy efficiencies of various EdgeAI implementations with biology

Brain-inspired solutions might just be the Key.

Human brain	Brain inspired
✓ Massively parallel – 10^{11} neurons and 10^{15} synapses	✓ High density storage, close to neurons
✓ Doing processing using memory elements	✓ Computational storage
✓ Analog computation – Neuron soma = synaptic current integrator	✓ Analog neuron
✓ Digital communication – Spikes = unary events, very robust to noise	✓ Spike coding

Demonstration

We made a proof-of-concept circuit combining:

- Spike coding
- RRAM synapses
  - Weighted input thanks to Ohm's law
- Analog neurons
  - Inputs summation thanks to Kirchhoff's law

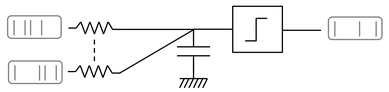


Figure 2: Simplified representation of an analog neuron and its resistive synapses

Application

When you experiment with a new implementation, such as this proof-of-concept circuit, you always consider a simple application. So we considered the MNIST database, for handwritten digits classification, which is the "Hello World" of neural networks.

We have chosen a frequency coding of the inputs, or rate code, to be equivalent to classical coding. This choice was guided by the learning strategy. The grey level of pixels is thus frequency encoded: the brighter the pixel, the higher the frequency of spikes.

Figure 3: Frequency coding of pixel intensity

Learning strategy

Bio-inspired unsupervised learning rules, such as the Spike Timing Dependent Plasticity one, give poorer results than the Gradient Descent algorithm. Decision was made to do offline learning in the classical coding domain, and then to transcode into spikes.

For this, we have used an in-house learning framework, called N2D2. It has the advantage of being able to support both classical coding and spike coding.

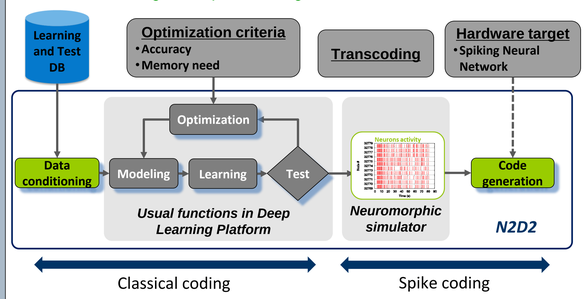
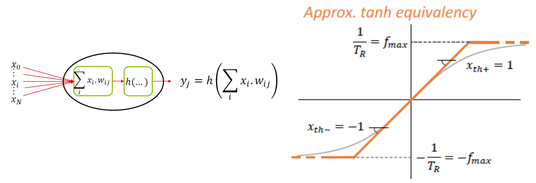


Figure 4: N2D2 learning framework

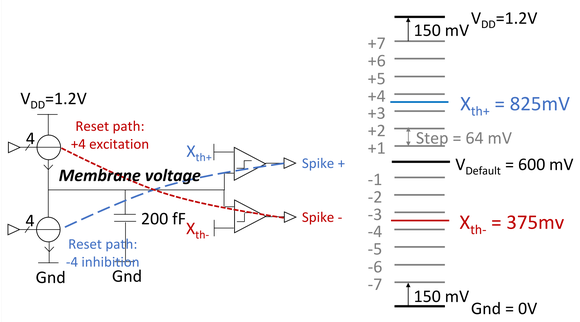
Neuron model and implementation

For being able to learn in the classical domain and transcode into the spike one, we need to have a mathematical equivalence between the two neuron models. This is ensured thanks to a specially-developed 'Integrate and Fire' model, which abstracts the TANH function.



(a) (b)

Figure 5: (a) classical domain neuron model; (b) equivalent Integrate and Fire model



(a) (b)

Figure 6: (a) Neuron schematic, with reset paths for ensuring model equivalence; (b) Voltage levels in membrane

Synapse implementation

Synaptic quantization experiments done on the learning framework have shown that integer values, ranging from -4 to +4, are a good compromise between classification accuracy and RRAM number.

Since RRAMs are used in binary mode (LRS and HRS states), multiple cells are put in parallel for enabling various weights. Synapses are arranged in a matrix, for sharing Word Line, Source Line and Bit Line drivers.

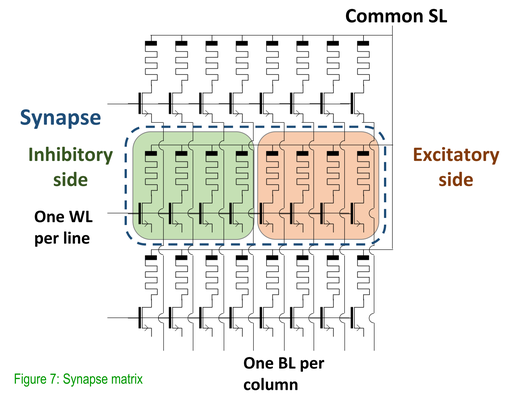


Figure 7: Synapse matrix

Fabrication and characterization results

The RRAMs are fabricated between Metal4 and Metal5, on Bulk CMOS 130nm base wafers.

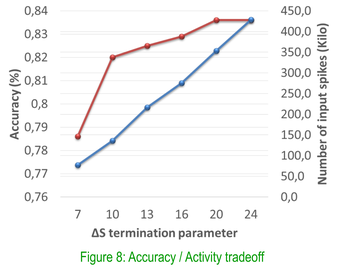


Figure 8: Accuracy / Activity tradeoff

Classification accuracy is measured on the 10K test images at 84% (compared to 88% in simulation). The energy gain is equal to 5X compared to classical coding. On average, 136 spikes are needed to make a decision, leading to an energy dissipation of 24,5nJ per inference.

Discussion

The energy and density figures obtained can be further improved thanks to technology scaling and multiple-level RRAM cells. Figure 9 shows the area scaling moving from 130nm to 28nm. The energy per synaptic event is divided by 10X. Figure 10 shows that analog multiple-level cells can be implemented, increasing synaptic density by 4X.

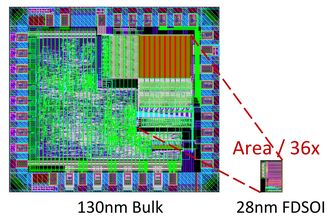


Figure 9: Area scaling

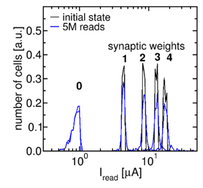


Figure 10: Multiple valued cell

Conclusion

We have demonstrated of a fully-functional Spiking Neural Network, combining analog neurons and RRAM synapses. It Exhibits a 5x energy gain compared to classical coding. Moving to the 28nm node leads to a 10x energy reduction and a 30x density gain. Synaptic density can be further improved by a factor of 4x by using RRAMs as a Multiple Level Cells.

Work is now ongoing to combine SNNs with Piezoelectric Micromachined Ultrasound Transducer (PMUT) sensors, exploiting a temporal coding of the inputs. Application is hand gesture recognition.