M-2-01 (Invited)

# Memory devices in Neuromorphic Computing Systems

Carlo Reita[1],

[1] CEA-LETI, 17 Rue des Martyrs, Grenoble 38054 Cedex, France
Phone: +33-4-3878 4248 E-mail: carlo.reita@cea.fr

**Abstract**

**This paper will review some of the approaches for neuromorphic computing using the results of the new resistive memories technologies. The interrelation between the characteristics of each technology and the architecture chosen to solve a specific problem is very deep and it is not yet clear if it will be possible to arrive at some form of standardization or if each application will require or afford to have a dedicated solution. Some results obtained in the context of the H2020 project NeuRAM3 will also be shown.**

## 1. Introduction

Research activities in the field of brain-inspired computing have gained a large momentum in recent years. The main reason is the attempt to go beyond the limitation of the conventional Von Neumann architecture that is increasingly affected by the limitation of the bandwidth and latency of the memory-logic communication. In neuromorphic architectures, the memory is distributed and can be co-localised with the logic, in particular it is what the new resistive memories technologies could provide. While most of the attention is being directed to implementation of Deep Learning algorithms in large computing system, the impact on device and circuit technology has been mixed. On one hand, advanced standard CMOS technology has been used to develop GPU and specific circuit accelerators without making use of any “bio-inspired” hardware. On the other hand, emerging resistive memory devices (RRAMs) are considered good candidates to emulate a biologically plausible synaptic behavior at nanometer scale, because of the fact that they offer the possibility to modulate their conductance by applying low biases, and can be easily integrated with CMOS-based neuron circuits in a back-end process during the making of the chip. This has opened the way for the realization of compact and energy-efficient computing architectures based on artificial neural networks (ANNs) – mainly using unsupervised learning rules such as the Spike Timing Dependent Plasticity - but that have been restricted mostly to the research community due to the insufficient maturity of the technology.

An intermediate, and probably faster to market application of these new memory technologies will be their application as a slow-nonvolatile cache/fast mass storage as intermediate memory level in conventional accelerators. This will allow a reduction of the fast DRAM and SRAM cache areas while still reducing latency to access the mass storage.

In the following we will restrain to the application of RRAMs to more bio-inspired circuits for local low power inference and possibly adaptive or unsupervised learning.

## 2. Simulation and benchmarking environment

A fundamental element to carry out efficiently studies of the interdependence of technology, design and architecture is the availability of a simulation framework able of handling models, simulations and/or real data for the different part of the system being studied. CEA has developed such a framework for Deep Neural Network (DNN) simulation and full DNN-based applications building called N2D2 (for 'Neural Network Design & Deployment'). N2N2 is now an open source software [1] and allows the exploration of different network topologies and also to define independently each structure of the network. A set of dedicated plug-ins links it to "classical" parallel compute cores (Intel, ARM), to GPUs and DSPs, to FPGA synthesis and to dedicated ASICs. It is aimed to be a simple and effective simulation and exploration environment, which is designed from the beginning to integrate variability and stochasticity in both synaptic and neural models. It is also not restricted to conventional DNN supervised learning rules. Embedding various synaptic device models and event-based simulation, it can be used to estimate the synaptic power consumption. In N2D2, the neurons are functionally modeled for computational efficiency and uses a mixed and flexible event processing model. This event processing engine differs from some spiking neural networks simulators that do not compute the exact timings of events but just model the events order. Knowing the exact timings is necessary for using nanodevices models and to give direct feedback for technological optimizations. The input nodes of a network can be mapped to various types of external stimuli. For example, Address Event Representation (AER) data can be loaded directly to form the input stimuli. N2D2 is fundamentally synapse-centric, which differs from other neural network simulators which are mainly neuron-centric and where the synapses are usually modeled as a single floating number parameter. In N2D2, the neurons and associated synaptic learning rules can accurately emulate programming pulses on the physical synaptic devices: functional and/or semi-physical modeling is also possible.

N2D2 was developed to allow fast and efficient design exploration of "classical" and spiking neuromorphic architectures by providing a framework that can mix high-level behavioral modeling with hardware constraints integration and different computational approaches. This simulation environment allows us to benchmark different network type and different implementations for a given problem and helps providing an optimum solution, even for various device implementations.

## 3. Hardware Platforms for bio-inspired computing

From a technological perspective, RRAMs are a good candidate for neuromorphic applications because of CMOS compatibility, high scalability, strong endurance, and good retention characteristics. However, defining practical implementation strategies and useful applications of large-scale co-integrated hybrid neuromorphic systems (CMOS neurons with resistive memory synapses) remains a difficult challenge

The possibility of using resistive memory devices as synapses in bio-inspired hardware has received growing interest. Resistive RAM (RRAM) devices like Phase Change Memories (PCM), Conductive Bridge RAM (CBRAM) and Oxide RAM (OxRAM) in particular, have been proposed to emulate biologically inspired features of synaptic functionality. Among the different types of emulated synaptic features, spike-timing-dependent plasticity (STDP) has gained a lot of significance recently and.an in-depth analysis of RRAM implementations is given in this same conference [2].

The major impact of an effective plasticity mechanism at circuit level is the possibility of using it for on-chip learning of some sort. This feature is not critical today, where the usage and impact of neural networks comes mostly from DNN that use large indexed database and backpropagation algorithms for the learning phase, but it will become increasingly relevant to apply machine learning concepts to system where is difficult or non-economical to use an indexed database and where adaptation on the fly is critical.. For this reason, the recent successes of Deep Learning should and digital accelerators should not be seen as a reason to slow down research in alternative approaches.

An important market today for neuromorphic circuits comes from the local ("edge") execution of the inference function. A large number of applications cannot afford, for latency, bandwidth, privacy reasons, to send data to large computers offsite (the "cloud") in order to exploit the data analysis capability of large networks. For this reason, interest exists for chips that can implement already trained network and can perform only the inference phase. Today, depending on the complexity of the task, we observe FPGA implementations for highly customized applications, pure software implementations running on MCUs or CPUs or dedicated neuromorphic cores/accelerators with highly parallel architectures similar to GPUs. All these approaches can also benefit from the availability of local nonvolatile memory that could lead to more compact FPGAs, more optimized memory hierarchy for MCU/CPUs and other implementations of neuromorphic cores.

In particular, asynchronous spiking analog architectures can easily exploit crossbar arrays of RRAM for much denser implementations than the multiplier-accumulator (MAC) blocks of formal neurons and would allow even more parallelism than in digital approaches. This, coupled with computing in the time domain and the elimination of clocking, will bring compact low power systems into applications that today cannot afford the high-power consumption associated with the digital implementations. While very promising, this approach is still not largely accepted by the industry which points to the difficulty of designing, verifying, characterizing and certifying analog asynchronous designs and by the difficulty of scaling analog solutions.

The use of the N2D2 modeling tools has showed that for such systems the tolerance to device to device variations is quite high and can reach 15% without serious degradation. Furthermore, it has been shown that high success rate for these systems is obtained with a much lower precision for the network parameter than that required by the backpropagation learning phase [3, 4]. These advantages should contribute to a future major acceptance.  On the scaling issue, promising work is being carried out in the frame of the NeuRAM3 project where a 28nm FDSOI implementation of an asynchronous analog architecture has been designed and is currently being fabricated (albeit without RRAM in the first phase) which shows promising energy efficiency [5]

One of the limitations of neuromorphic systems comes from the interconnection challenges and, when using RRAM, the size of the crossbar arrays. In order to overcome these limitations, a combination of 3D techniques and memory architectures can greatly improve the system level solution. In particular, the use of dedicated versions of monolithic 3D integration, with RRAM planes intercalated among analog neuron planes, can produce much more compact and less power hungry system. 3DTSV and 3D by Cu-Cu bonding are also promising candidate to have compact neuromorphic systems comprising the various elements in a highly-integrated architecture.

## 3. Conclusions

A review of the impact RRAM can have in the bio-inspired computing systems has been conducted and some promising results and concept discussed.

## Acknowledgements

The author wish to thanks his colleagues Elisa Vianello, Alexandre Valentian, Olivier Bichler, Marc Duranton for the contributions to this paper as this work is an overview of the projects carried out at CEA Tech in recent years and by multiple groups. The support of the French programs ARCH2NEU, NEMESIS, and by the H2020 European Program NeuRAM3 (Grant Agreement 687299) are also acknowledged.

## References

[1] https://github.com/CEA-LIST/N2D2

[2] S. La Barbera *et al* , *Extended Abstracts of the 2017 International Conference on Solid State Devices and Materials* (2017)

[3] D.Querlioz *et al., Proceedings of the IEEE, Vol 103, No 8, August 2015,* 2015

[4] J. Binas *et al* , *arXiv 1606.07786.* 2016.

[5] N. Qiao, G. Indiveri, *Biomedical Circuits and Systems Conference, (BioCAS 2016),* 2016